McGill University

Doctoral Thesis

Radiomics: Enabling Factors Towards Precision Medicine

Author:
Martin Carrier-Vallières

Supervisor:
Dr. Issam El Naqa

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

in the

Medical Physics Unit
Department of Physics
McGill University, Montreal
June 2017

© Martin Carrier-Vallières 2017

This thesis is dedicated to all patients who have fought, are fighting, or will fight cancer. I am with you, and every research action I take is focused on how to better help you.

"N'est-ce pas dans le rêve
cependant que naissent la
plupart des projets qui en
valent la peine?"

– René Lévesque

Abstract

Most tumours do not represent a homogeneous entity, but rather are composed of multiple clonal sub-populations of cancer cells forming complex dynamical systems that can exhibit rapid evolution as a result of different therapy perturbations. Tumours exhibiting higher heterogeneous characteristics are typically associated with higher risk of resistance to treatment, progression, metastasis or recurrence, which can lead to patients' death. The quantification of intratumoural heterogeneity for decoding tumour phenotypes is thus an active area of research in oncology. Being acquired at multiple time points of treatment management for almost every patient with cancer, medical images would in fact carry an immense source of potential data for decoding tumour phenotypes. This hypothesis is at the core of the new emerging field of "Radiomics", a field which refers to the characterization of tumour phenotypes via the extraction of high-dimensional mineable data from all types of medical images, and whose subsequent analysis aims at supporting clinical decision-making. More specifically, texture analysis – a sub-branch of radiomics – is one the most promising methods for the characterization of intratumoural heterogeneity, as it involves the quantitative description of the spatial distribution of different gray-levels within a given region of interest. Ultimately, improved characterization of tumour aggressiveness and prediction of tumour outcomes (e.g., metastases, local recurrences, etc.) at diagnosis via quantitative imaging biomarkers would allow physicians to better personalize treatments for each patient, and hopefully, save more lives.

In this thesis, the major aim is to develop radiomic-based models for the accurate prediction of tumour outcomes via advanced machine learning. We first showed that the optimization of how texture features are extracted from medical images (different isotropic voxel sizes, image quantization schemes, etc.) is fundamental for best tumour outcome prediction. We then integrated the texture optimization process into a robust multivariable modeling methodology developed for the construction of radiomic-based prediction

models. This multivariable modeling methodology employs logistic regression to linearly combine radiomic features. Using this methodology, we were able to develop a model that can predict the development of lung metastases in soft-tissue sarcomas with high accuracy. This model combines texture features extracted from functional FDG-PET and anatomical MRI pre-treatment images. Following this initial work, we demonstrated how the predictive properties of imaging textures composing such prediction models could be further enhanced by optimizing the way images are acquired. The proof of concept for the enhancement of the prediction of lung metastases in soft-tissue sarcomas was carried out using computerized simulations of FDG-PET and MR image acquisitions with tumour and clinical scanner models, by varying different physical parameters employed during image acquisitions. Next, in another study, we developed a strategy for personalizing treatments for soft-tissue sarcoma patients identified at diagnosis to be at higher risks of developing lung metastases (using radiomic-based prediction models); specifically, we verified the feasibility of applying double nested radiation dose boosting to the hypermetabolic and hypoxic soft-tissue sarcoma sub-regions to counteract the progression of more aggressive parts of tumours. For the purpose of radiation treatment planning, contours defining the hypermetabolic and hypoxic tumour sub-regions were obtained from FDG-PET and low-perfusion DCE-MRI functional images. Finally, in our last study, we developed a methodology allowing to integrate radiomic imaging data with clinical prognostic factors into comprehensive prediction models using a random forest algorithm. We tested our methodology in head-and-neck cancer to better assess the risk of locoregional recurrences and distant metastases, this time using functional FDG-PET and anatomical CT pre-treatment images in conjunction to clinical data. The clinically-integrated radiomic models that we developed possess high prognostic power, leading to patient stratification into two sub-groups for the risk assessment of locoregional recurrences (low, high) in head-and-neck cancer, and into three groups for distant metastases (low, medium, high).

Overall, in this thesis, we demonstrated that radiomics analysis is an enabling method towards precision medicine. The different radiomic techniques and models developed in this work could have a major impact on the design of new clinical trials aiming at a better personalization of cancer treatments. One can envision different treatment regimens being delivered

to patients based on different radiomic-based risk assessments of specific tumour outcomes.

Abrégé

La plupart des tumeurs ne représentent pas une entité homogène mais sont plutôt composées de multiples sous-populations clonales de cellules cancéreuses formant des systèmes dynamiques complexes qui peuvent présenter une évolution rapide en raison de différentes perturbations thérapeutiques. Les tumeurs présentant des caractéristiques hétérogènes plus élevées sont typiquement associées à un risque plus élevé de résistance au traitement, à la progression, aux métastases ou à la récidive, ce qui peut entraîner la mort des patients. La quantification de l'hétérogénéité intratumorale pour le décodage des phénotypes tumoraux est donc un domaine actif de recherche en oncologie. Étant acquises à divers moments lors de la gestion du traitement pour presque tous les patients atteints du cancer, les images médicales seraient en fait une source immense de données potentielles pour le décodage des phénotypes tumoraux. Cette hypothèse est au coeur du nouveau domaine émergent de la "Radiomique", domaine qui se réfère à la caractérisation des phénotypes tumoraux grâce à l'extraction de données à grande dimension à partir de tous les types d'imagerie médicale et dont l'analyse subséquente vise à soutenir la prise de décision clinique. Plus précisément, l'analyse texturale – une sous-branche de la radiomique – est l'une des méthodes les plus prometteuses pour la caractérisation de l'hétérogénéité intratumorale, car elle implique la description quantitative de la répartition spatiale des différents niveaux de gris dans une région d'intérêt donnée. Ultimement, une caractérisation améliorée de l'agressivité et du devenir des tumeurs (par exemple, métastases, récidives locales, etc.) au moment du diagnostic à l'aide de biomarqueurs d'imagerie quantitatifs permettraient aux médecins de mieux personnaliser les traitements pour chaque patient et ainsi, espérons-le, sauver plus de vies.

Dans cette thèse, l'objectif principal est de développer des modèles basés sur la radiomique pour la prédiction du devenir des tumeurs grâce à l'apprentissage machine avancé. Nous avons d'abord démontré que l'optimisation de la façon dont les caractéristiques texturales sont extraites des images médicales est fondamentale pour une meilleure prédiction du devenir des tumeurs.

Nous avons ensuite intégré le processus d'optimisation des textures dans une méthodologie de modélisation multivariable robuste développée pour la construction de modèles de prédiction basés sur la radiomique. Cette méthodologie de modélisation multivariable utilise la régression logistique afin de combiner linéairement les données radiomiques. À l'aide de cette méthodologie, nous avons pu développer un modèle permettant de prédire avec une grande précision le développement des métastases pulmonaires pour les patients atteints de sarcomes des tissus mous. Ce modèle combine les valeurs de caractéristiques texturales extraites des images FDG-TEP fonctionnelles et IRM anatomiques acquises avant le traitement des tumeurs. À la suite de ce travail initial, nous avons démontré comment les propriétés prédictives des textures d'imagerie composant de tels modèles de prédiction pourraient être améliorées en optimisant la façon dont les images sont acquises. Une preuve de concept pour l'amélioration de la prédiction des métastases pulmonaires dans les sarcomes des tissus mous a été réalisée à l'aide de simulations informatisées d'acquisitions d'images FDG-PET et IRM, en variant les différents paramètres physiques utilisés pendant les acquisitions d'images. Ensuite, dans une autre étude, nous avons développé une stratégie permettant de mieux personnaliser les traitements pour les patients atteints de sarcomes des tissus mous identifiés comme étant à plus haut risque de développer des métastases pulmonaires (à l'aide de modèles de prédiction basés sur la radiomique); plus précisément, nous avons vérifié la faisabilité d'appliquer une double augmentation de la dose de radiation aux sous-régions tumorales hypermétaboliques et hypoxiques des sarcomes des tissus mous afin de contrecarrer la progression des parties plus agressives des tumeurs. Aux fins de la planification du traitement par radiations, les contours définissant les sous-régions tumorales hypermétaboliques et hypoxiques ont été obtenus à partir d'images fonctionnelles FDG-TEP et IRM. Enfin, dans notre dernière étude, nous avons développé une méthodologie permettant d'intégrer des données d'imagerie radiomique avec des facteurs pronostiques cliniques dans des modèles de prédiction, en utilisant cette fois un algorithme informatique dit à "forêt aléatoire". Nous avons testé notre méthodologie dans le cancer de la tête et du cou afin de mieux évaluer le risque de récidives locorégionales et de métastases distantes, cette fois en utilisant des images fonctionnelles FDG-TEP et des images anatomiques CT acquises avant le traitement en conjonction avec les données cliniques. Les modèles radiomiques cliniquement intégrés que nous avons développés possèdent une forte puissance de prédiction, conduisant à une stratification des patients en deux sous-groupes pour l'évaluation du risque de récidive locorégionale (faible, élevé) dans le cancer de la tête et du cou, et en trois sous-groupes pour les métastases distantes (faible, moyen, élevé).

Dans l'ensemble, nous avons demontré dans cette thèse que l'analyse radiomique est une méthode propice à la médecine de précision. Les différentes techniques et modèles radiomiques développés dans ce travail pourraient avoir un impact majeur sur la conception de nouveaux essais cliniques visant à une meilleure personnalisation des traitements contre le cancer. On peut imaginer que dans un avenir rapproché, des régimes de traitement différents seront administrés aux patients en fonction de différentes évaluations du risque d'un devenir tumoral spécifique par l'analyse radiomique.

Acknowledgements

First, I would like to sincerely thank my supervisor Issam El Naqa for his great help and guidance throughout my PhD research. Thank you for introducing me to the world of radiomics, your research vision will always be an inspiration. Let me also appreciate the stellar work of Jan Seuntjens as the Director of the Medical Physics Unit (MPU) of McGill University. The dedication he has for the success of his students is simply enormous. He is definitely the leader of MPU and truly takes great care of his students, at any time of the day (or night).

J'aimerais aussi sincèrement remercier Sébastien Laberge pour son implication dans mes recherches. Ta grande volonté de réussir m'a toujours poussé à en donner plus. Merci aussi à Ives R. Levesque pour ses précieux conseils et toujours passionnantes discussions. Puissent-ils perdurer encore longtemps. De plus, comment oublier tous les conseils et l'aide informatique reçus de la part de Marc-André Renaud? Un vrai gourou avec un grand sourire!

Thanks also to Sonia R. Skamene, François Deblois, Vincent Hubert-Tremblay, Luc Ouellet, Isabelle Gauthier, Jean-François Carrier, Chantal Boudreau and Karim Zerouali for their help in retrieving imaging data. Many thanks to Alex Zwanenburg for providing useful figures and for leading the IBSI. My sincere thanks go also to Dr. Carolyn Freeman, an expert radiation oncologist that fed me with plenty of medical advice since the start of my research. I am grateful for all the time she took with me for patient contouring.

I also would like to thank the Natural Sciences and Engineering Research Council (NSERC) and the Faculty of Medicine of McGill University for the doctoral fellowship awards.

Merci aussi au Québec, terre d'opportunité et pays où il fait le mieux vivre au monde.

Merci aussi infiniment à Paul et Cécile, qui m'ont offert l'hospitalité dans une période charnière de mon doctorat. Votre présence et tous vos petits gestes me rendant la vie plus facile m'auront procuré un profond bonheur tout au long de mon séjour.

Mes sincères remerciements reviennent aussi à Suzanne, Joannie et Simon. Votre support inconditionnel fût tout simplement indipensable lors de cette grande épopée. Milles fois merci encore.

Finalement, merci à l'infini et du plus profond de mon coeur à ma douce Roxanne. Ta présence apaisante à tout moment et ta façon de rendre heureux les gens autour de toi seront pour toujours la clé de ma réussite. Tu as enduré pendant tellement longtemps mes constantes périodes de stress et mes heures de travail impossibles, mais c'est le début d'un temps nouveau, je te l'assure!

Table of Contents

Abstract i

Abrégé iv

Acknowledgements vii

Table of Contents xii

List of Figures xiii

List of Tables xvi

List of Abbreviations xviii

Preface and Contributions xx

Chapter 1: Introduction 1

1.1 The era of precision medicine 1
1.2 Decoding tumour phenotype via quantitative imaging 6
1.2.1 Quantification of intratumoural heterogenity 7
1.2.2 Definition and hypothesis of radiomics 11
1.2.3 Major objective of radiomics 11
1.2.4 Overview of the role of radiomics in oncology 13
1.3 Paradigm of radiation dose painting 17
1.3.1 Radiation therapy 18
1.3.2 The Biological Target Volume (BTV) 20
1.4 Thesis hypotheses, objectives and organization 24
1.5 References 27

Chapter 2: Background on radiomics modeling 35

2.1 Medical imaging in cancer management 35
2.1.1 Positron Emission Tomography 36
2.1.2 Computed Tomography 39
2.1.3 Magnetic Resonance Imaging 41
2.2 Image processing 51
2.2.1 Spatial filtering and image fusion 51
2.2.2 Tumour segmentation 56
2.2.3 Image interpolation 57
2.2.4 Image quantization 59
2.2.5 Texture extraction parameters 63
2.3 Texture analysis 63

2.3.1 Gray-Level Co-occurence Matrix . . . 66
2.3.2 Gray-Level Run-Length Matrix . . . 67
2.3.3 Gray-Level Size Zone Matrix . . . 68
2.3.4 Neighborhood Gray-Tone Difference Matrix . . . 70
2.4 Machine learning . . . 71
2.4.1 Mathematical notations and concepts . . . 72
2.4.2 Logistic regression . . . 73
2.4.3 Bootstrapping . . . 74
2.4.4 Random forests . . . 75
2.4.5 Performance measures . . . 78
2.4.6 Estimation of prediction performance . . . 80
2.5 References . . . 84

Chapter 3: Development of a radiomic model for the early prediction of lung metastases 89
3.1 Foreword . . . 90
3.2 Abstract . . . 90
3.3 Introduction . . . 91
3.4 Materials and Methods . . . 94
3.4.1 Data . . . 95
3.4.2 Tumour volume definition . . . 97
3.4.3 Imaging data pre-processing . . . 97
3.4.4 FDG-PET/MRI volume fusion . . . 98
3.4.5 Feature extraction . . . 99
3.4.6 Univariate analysis . . . 103
3.4.7 Multivariable analysis . . . 105
3.5 Results . . . 109
3.5.1 Univariate analysis . . . 109
3.5.2 Multivariable analysis . . . 113
3.6 Discussion . . . 116
3.7 Conclusion . . . 119
3.8 Online resources . . . 120
3.9 Acknowledgments . . . 120
3.10 Supplementary Material . . . 120
3.10.1 Definition of texture features . . . 120
3.10.2 Texture features with highest predictive value . . . 121
3.10.3 Degrees of freedom on texture extraction parameters . 121
3.10.4 Uncertainty analysis . . . 123
3.10.5 Permutation tests . . . 126
3.11 References . . . 126

Chapter 4: A strategy for treatment personalization 130
4.1 Foreword . . . 130
4.2 Abstract . . . 131
4.3 Introduction . . . 132
4.4 Materials and Methods . . . 133
4.4.1 Patients . . . 134
4.4.2 Standard-of-care radiotherapy planning and treatment 134

4.4.3 Study design . . . 134
4.4.4 Image analysis . . . 136
4.4.5 Dose painting feasibility study . . . 137
4.5 Results . . . 137
4.5.1 Prediction of lung metastases development . . . 138
4.5.2 General imaging findings . . . 138
4.5.3 Tumour sub-region analysis . . . 139
4.5.4 Dose painting . . . 141
4.6 Discussion . . . 143
4.7 Acknowledgments . . . 145
4.8 Supplementary Material . . . 146
4.8.1 Supplementary results . . . 146
4.8.2 Image acquisition protocols . . . 146
4.8.3 Image registration . . . 150
4.9 References . . . 150

Chapter 5: Enhancement of radiomic-based prediction models via the optimization of imaging acquisition protocols 153
5.1 Foreword . . . 154
5.2 Abstract . . . 154
5.3 Introduction . . . 155
5.4 Methods . . . 160
5.4.1 Imaging dataset . . . 160
5.4.2 Construction of a radiomic prediction model . . . 161
5.4.3 Simulations of PET imaging acquisitions . . . 163
5.4.4 Simulations of MR imaging acquisitions . . . 168
5.4.5 STAMP: a software tool for texture optimization . . . 172
5.4.6 Optimization of a texture-based model . . . 174
5.5 Results . . . 175
5.5.1 Texture variations with acquisition parameters . . . 175
5.5.2 Evaluation on a texture-based model . . . 181
5.6 Discussion . . . 185
5.7 Conclusion . . . 189
5.8 Acknowledgments . . . 189
5.9 Supplementary Material . . . 190
5.9.1 Description of 3D radiomic features . . . 190
5.9.2 Construction of radiomic models . . . 190
5.9.3 Software integration for texture analysis of PET and MR simulated images . . . 191
5.9.4 Comparison of clinical and simulated textures . . . 194
5.10 References . . . 195

Chapter 6: Integration of radiomic-based prediction models with clinical prognostic factors 200
6.1 Foreword . . . 201
6.2 Abstract . . . 201
6.3 Introduction . . . 202
6.4 Results . . . 204

6.4.1 Summary of presentation of results 204
6.4.2 Association of variables with tumour outcomes 204
6.4.3 Construction of prediction models 207
6.4.4 Performance of prediction models 208
6.4.5 Comparison with other prognostic factors 209
6.4.6 Risk assessment of tumour outcomes 213
6.5 Discussion 214
6.6 Methods 220
6.6.1 Data sets availability 220
6.6.2 Sample size and division of cohorts 222
6.6.3 Extraction of radiomic features 222
6.6.4 Construction of radiomic models 224
6.6.5 Combination of radiomic and clinical variables 225
6.6.6 Imbalance-adjustment strategy 225
6.6.7 Random forest training 226
6.6.8 Calculation of performance metrics 227
6.6.9 Code and models availability 228
6.7 Acknowledgments 228
6.8 Supplementary Material 229
6.8.1 Supplementary results 229
6.8.2 Supplementary methods 235
6.9 References 252

Chapter 7: Conclusions 257
7.1 Summary and novelty of work 257
7.2 The fundamental necessity of standardization, transparency and online sharing in radiomics 264
7.3 Other contributions 265
7.4 References 266

Appendix A: List of 3D radiomic features 268
A.1 Morphological (shape) features 268
A.2 Histogram-based (intensity) features 268
A.3 GLCM features 270
A.4 GLRLM features 271
A.5 GLSZM features 274
A.6 NGTDM features 276

List of Figures

Figure 1.1: Personalized cancer care continuum 4
Figure 1.2: Cycle of evidence in rapid-learning health care 5
Figure 1.3: Molecular, functional and anatomical imaging of cancer processes 7
Figure 1.4: Example of intratumoural heterogeneity at the anatomical scale 8
Figure 1.5: Image texture example 10
Figure 1.6: Principle of treatment response prediction via extraction of imaging biomarkers 10
Figure 1.7: Major objective of radiomics 12
Figure 1.8: Overview of the role of radiomics in oncology 14
Figure 1.9: Graphical representation of the target contours in radiation therapy 19
Figure 1.10: The concept of biological target volume 21

Figure 2.1: Physical principles of PET imaging 37
Figure 2.2: Physical principles of CT imaging 40
Figure 2.3: Alignment of nuclear spins in the presence of an external field 42
Figure 2.4: Principles of magnetization excitation and signal acquisition 43
Figure 2.5: Principles of MR image contrast 44
Figure 2.6: Schematic representation of a typical MRI spin-echo sequence 45
Figure 2.7: Schematic illustration of a DW-MRI sequence 48
Figure 2.8: Principles of DCE-MRI 50
Figure 2.9: Example of a wavelet basis function 53
Figure 2.10: Principles of 2D discrete wavelet decomposition . . . 54
Figure 2.11: Principles of wavelet image fusion 55
Figure 2.12: Schematic illustration of the image segmentation process 57
Figure 2.13: Schematic illustration of the image interpolation process 58
Figure 2.14: Schematic illustration of the image ROI extraction process 60
Figure 2.15: Schematic illustration of the image quantization process 61
Figure 2.16: Impact of different texture extraction parameters on processed images 64
Figure 2.17: Full 3D approach to calculate texture matrices 66
Figure 2.18: Sigmoidal curve of logistic regression models 74
Figure 2.19: Example classification using a decision-tree 76
Figure 2.20: Schematic representation of a random forest prediction 78

Figure 2.21: Concept of the AUC metric 81

Figure 3.1: FDG-PET and MRI diagnostic images of two patients with soft-tissue sarcomas of the extremities 93
Figure 3.2: Example of soft-tissue sarcoma tumour volume definition performed on the T_2-weighted fat-saturated scan of a patient of the *LungMets* group 97
Figure 3.3: Example fusion of PET and MR images 99
Figure 3.4: Workflow of extraction of texture features 104
Figure 3.5: Workflow of multivariable analysis. 110
Figure 3.6: Estimation of prediction performance of multivariable models constructed from FDG-PET scans, SEPARATE scans, and FUSED scans using optimal degrees of freedom on texture extraction parameters, for model orders of 1 to 10 114
Figure 3.7: Probability of developing lung metastases as a function of the response of the multivariable model proposed in this work, for all patients of the cohort 116
Figure 3.8: Estimation of prediction performance as a function of model order for multivariable models constructed from the FDG-PET, SEPARATE, and FUSED scans for three types of experiment 121
Figure 3.9: Distributions of performance metrics estimates of multivariable models of order 4 found in 1000 permutation tests using shuffled outcome vectors 126

Figure 4.1: Timeline of the study protocol 135
Figure 4.2: Response of the FDG-PET/MRI texture-based model developed in a previous work on a different retrospective cohort of patients when directly applied on the patient cohort of this study 139
Figure 4.3: Changes in relevant prognostic metrics of FDG-PET, FMISO-PET, DW-MRI and DCE-MRI intratumoural data at 3 time points during radiotherapy (PRE-RT, MID-RT, POST-RT) for patients with available images 140
Figure 4.4: Analysis of high-intensity tumour sub-regions over time 142
Figure 4.5: Dose painting example 143
Figure 4.6: General observations from MRI 148
Figure 4.7: Evolution of FDG-PET nd FMISO-PET intratumoural uptake of the central slice of the tumour of four example patients over the 3 time points during radiotherapy management (PRE-RT, MID-RT, POST-RT) 149

Figure 5.1: Imaging acquisition parameters varied in this study . 159
Figure 5.2: Construction of a radiomic model from pre-treatment clinical images for the prediction of lung metastases in soft-tissue sarcomas 164

Figure 5.3: Example of the creation of PET and MRI digital tumour models used for image acquisition simulations . 166
Figure 5.4: Workflow of this study 176
Figure 5.5: Example PET and MR simulated images acquired using different parameters 177
Figure 5.6: Texture variations with different acquisition parameters and resulting effects on associations with lung metastases in soft-tissue sarcomas 180
Figure 5.7: Optimization of a texture-based prediction model with respect to PET and MR simulated image acquisition parameters using bootstrapping experiments 183
Figure 5.8: Texture-based prediction model response enhancement using an optimal set of PET and MR image acquisition parameters 184
Figure 5.9: Radiomic feature selection 192
Figure 5.10: Example image of the PET simulation GUI of STAMP 192
Figure 5.11: Example image of the MRI simulation GUI of STAMP 193
Figure 5.12: Example image of the image analysis GUI of STAMP . 193
Figure 5.13: Comparison between textures extracted from clinical and simulated images 194

Figure 6.1: From radiomics analysis to treatment personalization 205
Figure 6.2: Models construction strategy and analysis workflow . 206
Figure 6.3: Prediction performance of selected models 210
Figure 6.4: Risk assessment of tumour outcomes 215
Figure 6.5: Summary of presentation of results 230
Figure 6.6: Choice of complexity of radiomic models 231
Figure 6.7: Prediction performance of selected models – complete results 232
Figure 6.8: Imbalance-adjustment strategy 237
Figure 6.9: Workflow of construction of radiomic models 238
Figure 6.10: Random forest training 239

List of Tables

2.1 GLCM computation example . . . 67
2.2 GLRLM computation example . . . 69
2.3 GLSZM computation example . . . 69
2.4 NGTDM computation example . . . 71
2.5 Classification of testing instances i: TP, FP, TN and FN . . . 79

3.1 Acronyms/definitions used in this study . . . 94
3.2 Characteristics of STS patient cohort . . . 95
3.3 Texture features used in this study . . . 102
3.4 Baseline texture extraction parameters . . . 108
3.5 Spearman's rank correlation (r_s) between non-texture features and lung metastases development in STSs . . . 111
3.6 Spearman's rank correlation between texture features and lung metastases development in STSs using *baseline* | *optimal* texture extraction parameters . . . 112
3.7 p-value of the Wilcoxon rank sum tests asserting the significance of the effects of texture extraction parameters on the correlation of texture features with lung metastases development in STSs . . . 113
3.8 Summary of permutation tests . . . 117
3.9 Texture features with the highest Spearman's rank correlation (r_s) with lung metastases in STS cancer and their associated extraction parameters, for the five types of scans investigated in this study . . . 121
3.10 Texture extraction parameters allowed to vary in the experiments yielding optimal prediction performance estimation . . 122
3.11 Volumes, texture features and texture features differences of the four features forming the multivariable model proposed in this work in the case where the tumour region was outlined with the contour excluding surrounding edema (*Mass*), and the case where the tumour was outlined with the contour including the surrounding edema (*Edema*) . . . 124
3.12 Response of the multivariable model proposed in this work on the entire patient cohort, and its associated 95 % confidence interval (CI) bounds. . . . 125

4.1 Dose-volume parameters for the different dose painting boost levels averaged over 14 patients . . . 141
4.2 Clinical characteristics of patients . . . 147

6.1 Comparison of prediction/prognostic performance of models constructed in this work with other variable combinations . . 212
6.2 Univariate analysis of radiomics variables 231
6.3 Univariate analysis of clinical variables 231
6.4 Performance of random forest classifiers constructed using *radiomic* variables only 235
6.5 Best predictive/prognostic and balanced random forest models found in this work 236
6.6 Characteristics of H&N1 cohort – HGJ 240
6.7 Characteristics of H&N2 cohort – CHUS 242
6.8 Characteristics of H&N3 cohort – HMR 245
6.9 Characteristics of H&N4 cohort – CHUM 247

List of Abbreviations

FDG	**F**luor**D**eoxy**G**lucose
PET	**P**ositron **E**mission **T**omography
MRI	**M**agnetic **R**esonance **I**maging
FMISO	**F**luoro**MISO**nidazole
DW-MRI	**D**iffusion-**W**eigthed **M**agnetic **R**esonance **I**maging
DCE-MRI	**D**ynamic **C**onstrast-**E**nhanced **M**agnetic **R**esonance **I**maging
CT	**C**omputed **T**omography
T2FS	**T2**-weighted **F**at-**S**aturated
STIR	**S**hort **T**au **I**nversion **R**ecovery
TCIA	**T**he **C**ancer **I**maging **A**rchive
ROI	**R**egion **O**f **I**nterest
GLCM	**G**ray-**L**evel **C**o-occurrence **M**atrix
GLRLM	**G**ray-**L**evel **R**un-**L**ength **M**atrix
GLSZM	**G**ray-**L**evel **S**ize **Z**one **M**atrix
NGTDM	**N**eighborhood **G**ray-**T**one **D**ifference **M**atrix
PVE	**P**artial **V**olume **E**ffect
DNA	**D**eoxyribo**N**ucleic **A**cid
GTV	**G**ross **T**umour **V**olume
CTV	**C**linical **T**arget **V**olume
PTV	**P**lanning **T**arget **V**olume
BTV	**B**iological **T**arget **V**olume
STS	**S**oft-**T**issue **S**arcoma
LOR	**L**ine **O**f **R**esponse
ML-EM	**M**aximum-**L**ikelihood **E**xpectation **M**aximization
OSEM	**O**rdered **S**ubsets **E**xpectation **M**aximization
HU	**H**ounsfield **U**nits
RF	**R**adio**f**requency Pulse
TR	**T**ime of **R**epetition
TE	**T**ime of **E**cho
ADC	**A**pparent **D**iffusion **C**oefficient
SUV	**S**tandard **U**ptake **V**alue
DWT	**D**iscrete **W**avelet **T**ransform
IDWT	**I**nverse **D**iscrete **W**avelet **T**ransform
ROC	**R**eceiver **O**perating **C**haracteristic (curve)
AUC	**A**rea **U**nder the ROC **C**urve
MIC	**M**aximal **I**nformation **C**oefficient
PIC	**P**otential **I**nformation **C**oefficient
PSF	**P**oint **S**pread **F**unction
FOV	**F**ield **O**f **V**iew
STAMP	**S**imulator for **T**exture **A**nalysis in **M**RI and **P**ET

GUI	**G**raphical **U**ser **I**nterface
H&N	**H**ead **&** **N**eck
LR	**L**ocoregional **R**ecurrence
DM	**D**istant **M**etastases
OS	**O**verall **S**urvival
CI	**C**oncordance-**I**ndex

Preface and Contributions

The work performed in this thesis was initiated by my supervisor Issam El Naqa. I believe he is the first scientist that developed in 2009 a complete methodology to predict tumour outcomes using textural analysis of pre-treatment medical images. The methods he developed in his initial breakthrough study provided strong foundations for me to develop more methods on the subject under his supervision, eventually leading to this thesis.

This thesis consists of four manuscripts: one published and three submitted. To the best of the authors' knowledge, prior to our work, no study in the literature had: I) developed a complete methodology for the construction of radiomic-based prediction models that includes texture optimization, and developed a radiomic-based prediction model for lung metastases in soft-tissue sarcomas; II) evaluated the feasibility of double nested dose boosting to hypermetabolic and hypoxic tumour sub-regions in soft-tissue sarcomas; III) demonstrated the possibility of enhancing texture-based prediction models via the optimization of image acquisition protocols; and IV) developed radiomic-based prediction models allowing for patient risk stratification into more than two sub-groups, and developed radiomic-based prediction models for locoregional recurrences and distant metastases in head-and-neck cancer.

I served as a first author for all the manuscripts integrated to this thesis. However, none of these studies would had been possible without the contribution of various co-authors and collaborators, as detailed below.

1. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60,** 5471-5496 (2015).

   Carolyn R. Freeman and Sonia R. Skamene provided the patient list with complete clinical information. Issam El Naqa and I designed the project. I performed imaging data collection, processing, ROI contouring (verified by Carolyn R. Freeman) and analysis, as well as manuscript writing. Issam El Naqa provided some programming code, expert knowledge and supervision throughout the project. All co-authors edited the manuscript.

2. Vallières, M., Serban, M., Benzyane, I., Ahmed, Z., Xing, S., El Naqa, I., Levesque, I. R., Seuntjens, J. & Freeman, C. R. The role of FDG-PET, FMISO-PET, DW-MRI and DCE-MRI in the management of soft-tissue sarcomas of the extremities with pre-operative radiotherapy and surgery: a feasibility study. *Radiother. Oncol.* [submitted on March 16, 2017]

   Carolyn R. Freeman, Lara Hathout, Issam El Naqa and I initiated the project design and the conduction of the prospective study protocol. Asha K. Jeyaseelan and Seema Ambereen conducted patient recrutement and study management. Ives R. Levesque and I attended MRI scans for quality assurance. I collected all imaging data from PACS and performed ROI contouring (verified by Carolyn R. Freeman). Ives R. Levesque and Zaki Ahmed processed the DCE-MRI data. Ives R. Levesque and Shu Xing processed the DW-MRI data. I processed the FDG-PET and FMISO-PET data. Ibtissam Benzyane and I performed various image analyses on all data, including the definition of the tumour sub-contours for dose painting. Monica Serban performed the radiotherapy planning that included dose painting. Jan Seuntjens, Carolyn R. Freeman, Ives R. Levesque and Issam El Naqa provided expert knowledge and supervision throughout the whole project. I wrote the manuscript, and all co-authors edited the manuscript.

3. Vallières, M., Laberge S., Diamant, A. & El Naqa, I. Enhancement of multimodality texture-based prediction models via optimization of PET and MR image acquisition protocols: a proof of concept. *Phys. Med. Biol.* [submitted on May 15, 2017]

   Issam El Naqa and I designed the project. Sébastien Laberge and I performed PET and MRI simulations, reconstructions and STAMP software development. André Diamant developed the methods and code for FDG-PET tumour modeling. I performed all other image processing and analyses, as well as manuscript writing. Issam El Naqa provided expert knowledge and supervision throughout the project. All co-authors edited the manuscript.

4. Vallières, M., Kay-Rivest, E., Jean Perrin, L., Liem, X., Furstoss, C., Aerts, H. J. W. L., Khaouam, N., Nguyen-Tan, P. F., Wang, C.-S., Sultanem, K., Seuntjens, J. & El Naqa, I. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci. Rep.* [submitted on March 16, 2017]

I designed and carried out the project from start to end. Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang and Khalil Sultanem provided patient databases and patient clinical information from their respective institutions. I made intial contact with all the different cancer institutions, wrote and brought to term sharing agreements and IRB approvals, collected the data at all institutions and performed all subsequent image processing and analyses, including manuscript writing. Hugo J. W. L. Aerts provided expert knowledge about the radiomic signature. Issam El Naqa provided supervision throughout the project. All co-authors edited the manuscript.

Chapter 1

Introduction

1.1 The era of precision medicine 1
1.2 Decoding tumour phenotype via quantitative imaging 6
1.2.1 Quantification of intratumoural heterogenity 7
1.2.2 Definition and hypothesis of radiomics 11
1.2.3 Major objective of radiomics 11
1.2.4 Overview of the role of radiomics in oncology 13
1.3 Paradigm of radiation dose painting 17
1.3.1 Radiation therapy 18
1.3.2 The Biological Target Volume (BTV) 20
1.4 Thesis hypotheses, objectives and organization 24
1.5 References 27

1.1 The era of precision medicine

In its simplest definition, "cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can result in death" [1]. One in seven deaths worldwide is caused by cancer. It is the second leading cause of death in high-income countries (being next to cardiovascular diseases) and the third leading cause of death in low- and middle-income countries (being next to cardiovascular and infectious/parasitic diseases) [2]. Cancer is caused by external as well as internal factors, and these factors may act together or in sequence. External factors include tobacco, infectious organisms and unhealthy diet, wherehas internal factors include inherited genetic mutations, hormones and immune conditions. Cancerous mass (i.e., tumour) development is accompanied with hallmark deviations from normal cellular functions such as sustained proliferative signaling, evasion for growth suppressors, activation of invasion and metastasis, replicative immortality, induction of angiogenesis, resistance to cell death, deregulation of cellular energetics and evasion of immune destruction [3].

Cancer diagnosis involves a careful clinical and pathological assessment of the tumour histology and phenotype (i.e., of the ensemble of observable and behavioral characteristics of a tumour), and is the first step to cancer management. Once cancer diagnosis is confirmed, usually via histological examination of biopsies, staging is performed to determine the extent and spread of cancer. Staging is an essential procedure in determining treatment choice and in providing a first assessment of prognosis (i.e., course of the disease and patients' chance of survival). "TNM staging" is the most widely used system in cancer management, and it classifies the extent of the primary tumour (T), the absence of presence of regional lymph node invasion (N) and the absence or presence of distant metastases (M) [4].

The prevalent modalities of cancer treatment presently can be divided into four main categories, and these may be used alone or in combination at different time points in the course of cancer management: I) *Surgery*, which refers to the removal of the tumour and affected surrounding tissues during a surgical operation. In many cases, stage and location of tumours may prevent surgery; II) *Chemotherapy*, which refers to the administration of one or more cytotoxic drugs to destroy or inhibit the growth and division of malignant cells. The high sensitivity but low degree of specificity of chemotherapeutic agents often results in good therapeutic response at the expense of high toxicity levels in the body of patients; III) *Radiotherapy*, which refers to the use of ionizing radiation to kill malignant cells via the damaging of the DNA. More than 50 % of cancer patients receive radiotherapy over the course of treatment management; and IV) *Hormone therapy*, which refers to the administration of synthetic hormones to block the body's natural hormones. This treatment option is at the moment limited to a few types of hormone receptor-positive tumours (e.g., androgen and estrogen suppression therapy for prostate and breast cancers, respectively).

Traditionally, tumour site and stage have been used to define patient populations, and medical diagnostics and treatments have been mainly focused on the general principles that work for the majority of patients. Although all cancer types involve uncontrolled cell division, each cancer patient is in reality unique owing to the heterogeneity at inter- and intra-tumour levels. Tumours of the same histopathological type may have different cancer driver mutations and/or proteomic profiles, which would lead to diverse treatment responses and prognosis. In the past decade, the prevention, diagnosis and treatment of cancer (i.e., oncology) has become increasingly understood at the cellular and molecular levels with the advent of next generation sequencing, as many cancer sub-types are now characterized as functions of biomarkers (i.e., measurable indicators of a particular disease state or some other physiological state of an organism) and tumour genetic mutations [5–7]. Worldwide initiatives such as The Cancer Genome Atlas (TCGA) [8] and the International Cancer Genome Consortium (ICGC) [9] have pioneered the characterization of the genome variant landscape of cancers [10]. Simplified frameworks of the cancer hallmark network are also being developed to facilitate the modeling of genome sequencing data for the prediction of cancer clonal evolution and associated clinical phenotypes [11]. In fact, the recent progress in "omics" technology has created unprecedented opportunities for characterizing the biological processes correlated with clinical phenotypes of tumours, notably in terms of understanding the structure of the genome ("genomics"), of DNA methylation landscapes ("epigenomics"), of gene expression ("transcriptomics") and of protein expression ("proteomics") [12]. Overall, the integration of all levels of biological function probing of cancer has now been recently coined with the term "panomics".

With the advent of "Big Data" analytics, we can now envision that panomics will leverage our capacity to make accurate clinical and tumour phenotypic predictions. Although the proper integration of panomics data into comprehensive models will remain a statistically and computationally challenging task [13, 14], the hope of revolutionizing how we improve health and treat disease, with the goal to "deliver the right treatment at the right time, every time, to the right person" – a concept known as "Precision Medicine" – has been embraced by political leaders and many scientists [15, 16], including the author of these lines. Overall, panomics tumour phenotype profiling offers the promise of guiding more personalized cancer treatments – here specifically defined as "Precision Oncology". The more we know about tumour phenotypes at the moment of diagnosis, the better we could potentially tailor, monitor, and adapt treatments to each individual patient. In this context, it has also been demonstrated that observational epidemiology studies could also provide advancements in the applications of precision oncology [17]. Figure 1.1 hereby depicts in more details the personalized cancer care continuum.

Given the immense dimensionality of the cancer problem, precision oncology relies heavily on the power of big data results to harness its full potential. The conventional path for the translation of new anticancer strategies

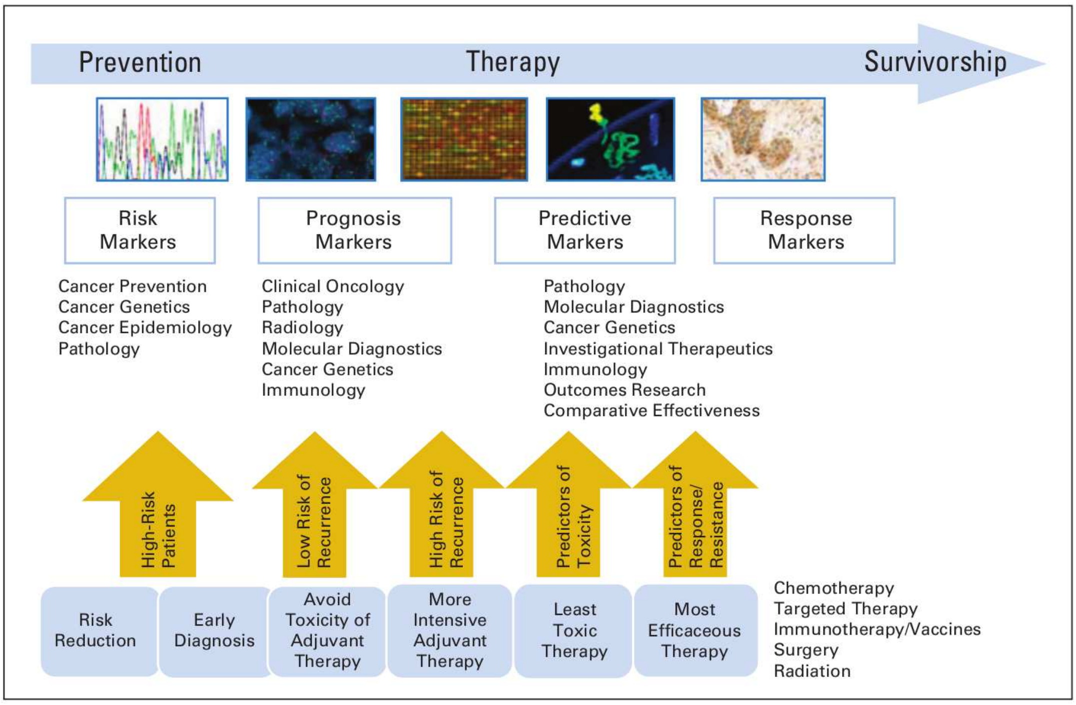


Figure 1.1: Personalized cancer care continuum. Reprinted with permission from [18]. © 2013 American Society of Clinical Oncology. All rights reserved.

from the lab to the clinic has been via clinical trials, but these may sometimes extend over years and would provide new data at a rather slow pace. To accelerate knowledge acquisition and reduce delays in getting promising treatments into the clinic, the Institute of Medicine (IOM) proposed a framework designed to use the data routinely acquired in the clinic. This framework is meant to drive scientific discovery at a faster pace and is called "Rapid-Learning Health Care" [19, 20]. Figure 1.2 presents the cycle of evidence in rapid-learning health care, which essentially consists of four phases that are continously iterated: I) the *data* phase, in which panomics data on past patients is collected; II) the *knowledge* phase, in which new knowledge is acquired from the panomics data; III) the *application* phase, in which the acquired knowledge is applied into clinical practice; and IV) the *evaluation* phase, in which the efficacy of the acquired knowledge applied on patients (i.e., improvements in tumour outcomes) is evaluated and after which the cycle starts again. Of important note, external knowledge coming, for example, from clinical trials is used to optimize every phase. The rapid-learning paradigm is thus really a complementary approach to evidence-based medicine, yielding different insights from the less controlled settings of routine clinical practice. Hence, rapid learning for precision oncology essentially consists of re-using routine clinical data in order to accelerate knowledge acquisition to form models that can predict cancer treatment outcomes, with the hypothesis being that the results of the treatment outcomes obtained in the past could be used to predict future ones [21]. Already, several researchers have proposed and piloted a " Global Cumulative Treatment Analysis" rapid-learning framework, in which treatment choice is continously revised based on updated performance statistics [22]. Overall, the increasing use of routinely acquired and retrospectively analyzed clinical data could be very effective for generating new hypotheses in the form of anticancer strategies that would be in-line with the precision oncology paradigm. At the same time, for best complementarity with evidence-based traditional oncology research and the sake of cancer patients, meaningful improvements in clinical outcomes obtained via rapid-learning healthcare ought also to be tested in rigorous randomized clinial trials [23].

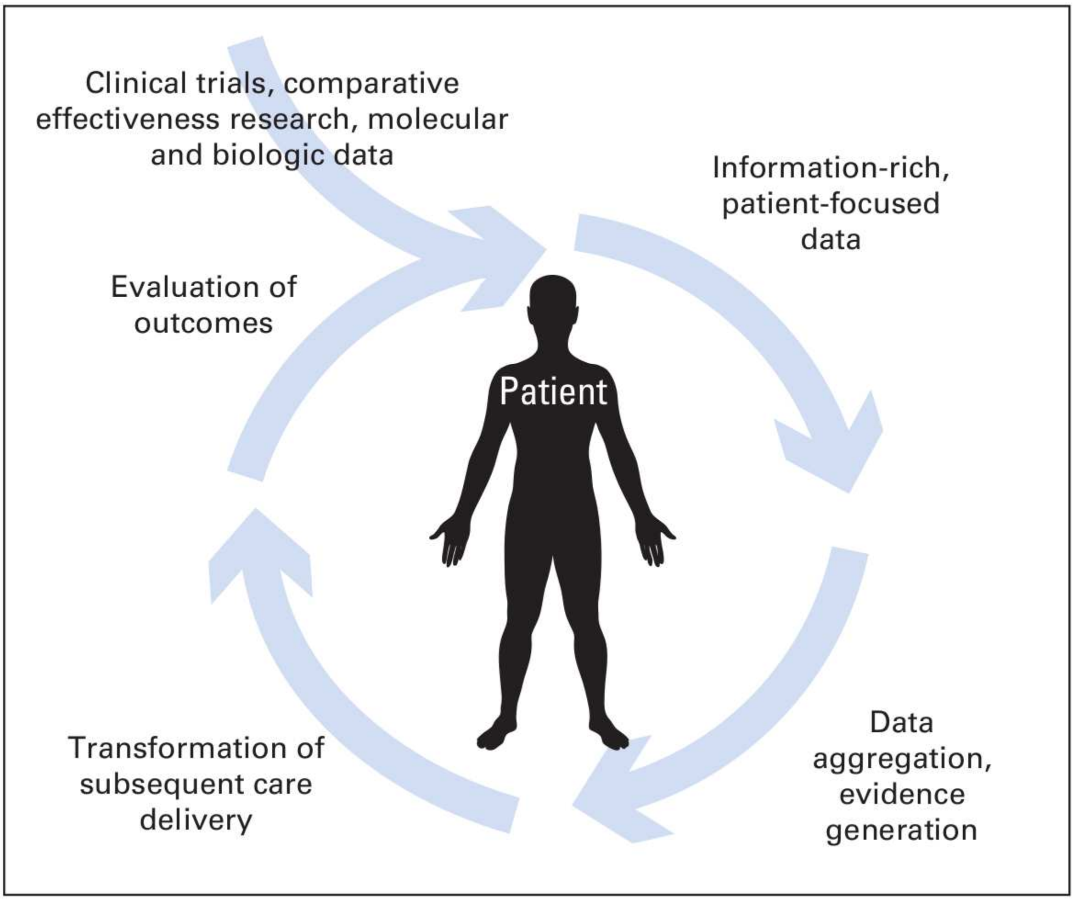

Clinical trials, comparative effectiveness research, molecular and biologic data
Information-rich, patient-focused data
Evaluation of outcomes
Patient
Transformation of subsequent care delivery
Data aggregation, evidence generation

Figure 1.2: Cycle of evidence in rapid-learning health care. In a patient-centered system of rapid-learning health care, patient-level data are aggregated to achieve population-based change, and results are applied to care of individual patients to achieve meaningful patient-level practice change. Reprinted with permission from [20]. © 2010 American Society of Clinical Oncology. All rights reserved.

1.2 Decoding tumour phenotype via quantitative imaging

Medical imaging of the anatomy and/or physio-pathology is acquired for almost every patient with cancer. From tumour diagnosis to tumour staging, treatment planning, treatment delivery, treatment response monitoring and patient follow-up, almost every step of clinical cancer management involves imaging. It was recently proposed that medical images would in fact carry an immense source of potential data for decoding tumour phenotypes [24]. Medical images routinely aquired in the clinic could therefore be closely tied to the rapid-learning paradigm. As depicted in Figure 1.3, imaging of cancer processes can be performed at the molecular, functional and anatomical levels in order to characterize the genome, proteome, metabolome, physiome and anatome.

In this thesis, the imaging types being investigated are focused on functional and anatomical cancer processes. For one, the combination of functional imaging in positron emission tomography (PET) (more details in section 2.1.1) with the anatomical information in computed tomography (CT) (more details in section 2.1.2) scans provides an efficient tool to accurately localize metabolic abnormalities in the human body. The injection of a radiopharmaceutical tracer in the body such as fluorodeoxyglucose (FDG) allows to reveal regions of significantly increased glucose uptake, a dominant characteristic of tumour cells over normal tissues due to their high metabolic activity in support for rapid growth. FDG-PET/CT imaging thus facilitates detection of primary and metastatic cancers that may not be apparent by routine staging procedures, and has profound impact on clinical management and therapy decision-making [25]. On the other hand, the importance of magnetic resonance imaging (MRI) (more details in section 2.1.3) in the clinical environment has exceeded most hopes of researchers due to its ability to manipulate and adjust tissue contrast with increasingly complex pulse sequences [26]. MR imaging is without a doubt one of the wonders of modern medicine and a beautiful example of how physics and mathematics can be exploited in the medical field, as one can generate contrast images that report a very large number of physical (e.g., proton density, T_1- or T_2-based contrast, etc.) and physiological phenomena (e.g., water diffusion, tissue perfusion, oxygen levels, susceptibility variations, etc.) based on the rich physics of nuclear magnetic resonance [27]. In this section, the importance of medical imaging for decoding tumour phenotypes via the quantification of intratumoural heterogeneity will be explained, a characteristic which has directly contributed to the emergence of the new field of "Radiomics".

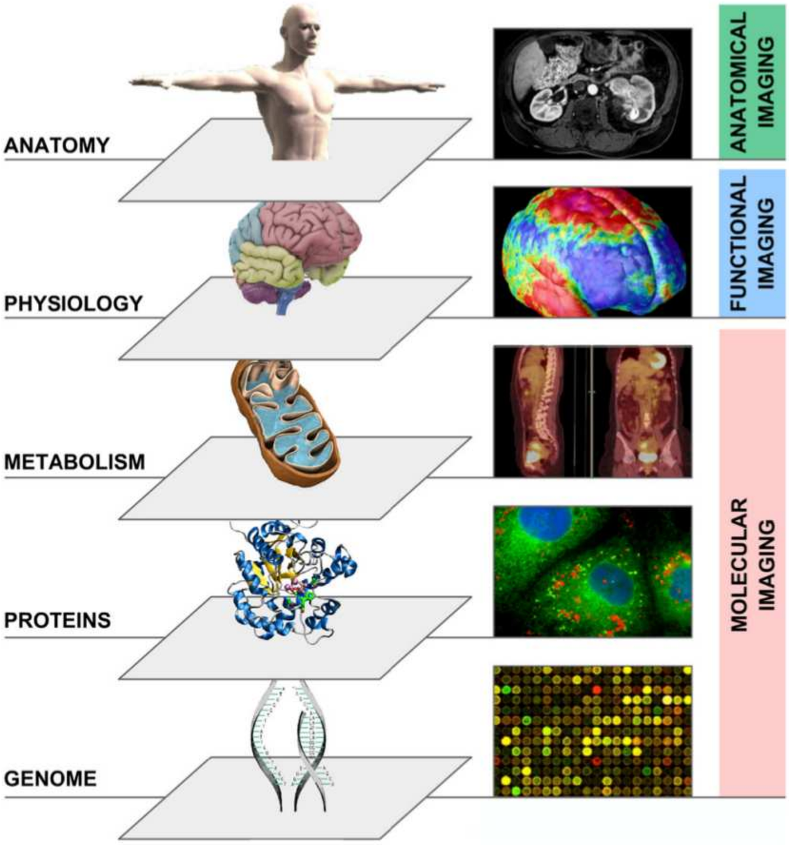


Figure 1.3: Molecular, functional and anatomical imaging of cancer processes. Reprinted with permission from [28]. © 2012 Elsevier. All rights reserved.

1.2.1 Quantification of intratumoural heterogenity

Most tumours do not represent a homogeneous entity, but rather are composed of multiple clonal sub-populations of cancer cells forming complex dynamical systems that exhibit rapid evolution as a result of different therapy perturbations. Differing properties can be attributed to the different sub-populations in terms of growth rate, expression of biomarkers, ability to metastasize, and immunological characteristics [29]. These properties could be explained by the differences in metabolic activity, cell proliferation, oxygenation levels, pH, blood vasculature and necrotic areas observed in the different cell sub-populations within a tumour. These intratumoural variations then in turn create different spatial intensity patterns in different types of medical images. For example, intratumoural variations in soft-tissue sarcoma tumours can be observed from diagnostic images such as MRI T_2-weighted fat-saturated scans as shown in Figure 1.4. For FDG-PET, different tumour sub-regions with different metabolic levels caused by inherent biological differences at the cellular level would also exhibit intratumoural variations in the associated images as we will see later in this thesis. Overall, such intratumoural differences are related to the concept of tumour heterogeneity (i.e., intratumoural heterogeneity), a characteristic that can be observed with substantial differences even amongst tumours of the same histopathological type. In solid cancers, the tremendous extent of heterogeneous characteristics is in fact expressed at multiple scales, as genes, proteins, cellular microenvironments, tissues and anatomical landmarks within tumours exhibit considerable spatial and temporal variations.

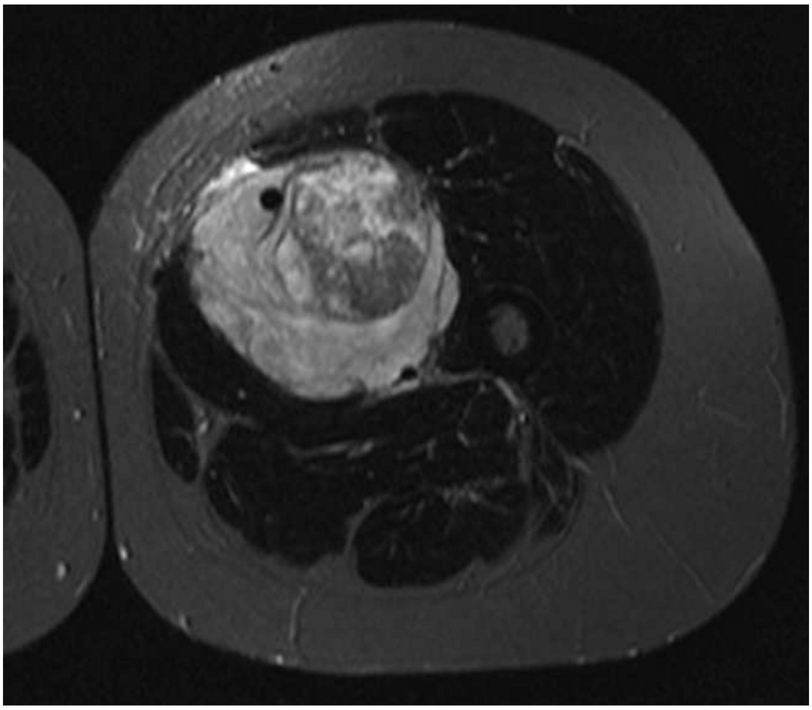


Figure 1.4: Example of intratumoural heterogeneity at the anatomical scale. The image represents a soft-tissue sarcoma of the leg from a MRI T_2-weighted fat-saturated scan.

It is now recognized that tumours exhibiting heterogeneous characteristics are associated with high risk of resistance to treatment, progression, metastasis or recurrence, leading to poor patient outcomes [30–32]. Ideally, the study of tumour heterogeneity should thus provide molecular signatures specific to the patient to be treated such that tumour aggressiveness and sensitivity to therapeutic response can be assessed prior to treatments. However, studying tumour heterogeneity using histopathological samples from invasive tumour biopsies (i.e., the removal of a piece of tissue from a tumour) is difficult, as the information obtained may vary depending on which part of the tumour is sampled, and the knowledge of the characteristics of individual components of a tumour may also not be sufficient to predict the behaviour of the whole [33]. The quantification of intratumoural heterogeneity using imaging biomarkers (i.e., biomarkers measured from medical images) extracted from the entire tumour region would provide a global assessment of intratumoural heterogeneity, and is thus an area of active research in oncology. Specifically, morphological/shape (e.g., volume, eccentricity, compactness, etc.), histogram-based/intensity (e.g., variance, skewness, kurtosis, etc.) and texture features are examples of imaging biomarkers that could be extracted from the region-of-interest (ROI) defining the tumour region in medical images (Appendix A provides the complete description of imaging features used in this work). Texture analysis is probably the most promising method for the characterization of intratumoural heterogeneity, as it involves the quantitative description of the spatial distribution of different gray-levels within a given ROI. As an example, Figure 1.5 shows an image with three different Brodatz textures [34], each represented by different textural properties. It can clearly be seen that the three different regions of that image have different spatial arrangements of gray levels (e.g., some are more heterogeneous than others → texture analysis can quantify this effect). Presently, the most commonly used textural metrics by the medical imaging community are the Gray-Level Co-occurence Matrix (GLCM) features [35], the Gray-Level Run-Length Matrix (GLRLM) features [36–38], the Gray-Level Size Zone Matrix (GLSZM) features [39] and the Neighborhood Gray-Tone Difference Matrix (NGTDM) features [40]. The methodology used to extract these textural metrics is presented in section 2.3.

Overall, different imaging biomarkers could act as surrogates of intratumoural heterogeneity, which in turn could provide a quantitative assessement of the aggressiveness of tumours. As illustrated in Figure 1.6, one of the overall goals that could be pursued in the context of the precision oncology paradigm would be to find the set of imaging biomarkers extracted from pre-treatment medical images that best discriminate between patients responding well to treatment from those who do not. This imaging information could in turn assist physicians in tailoring therapy choices for each patient. In a proof-of-concept retrospective study, El Naqa *et al.* [41] were the first to present a robust methodology dedicated to the prediction of tumour

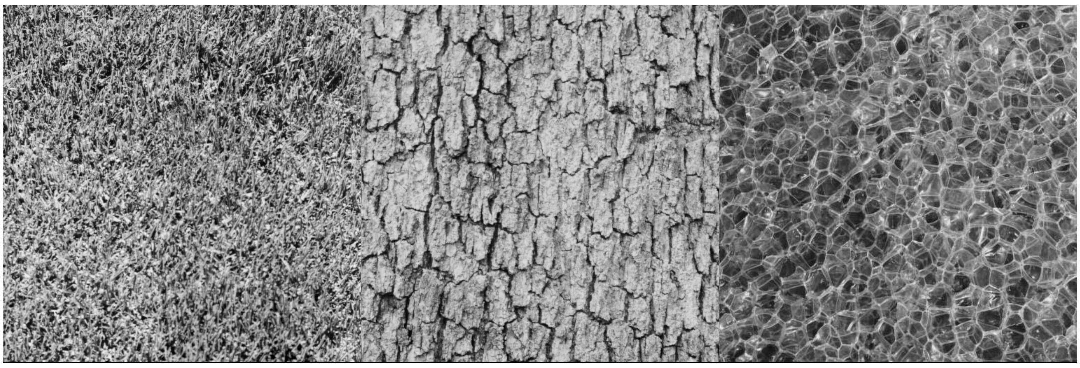


Figure 1.5: Image texture example. The image contains three different Brodatz textures, each with different textural properties. Reprinted from [34]. © 1966 Dover Publications. All rights reserved.

outcomes using texture-based multivariable models. Using logistic regression, the authors combined different image-based features including GLCM textures to predict disease persistence in cervix cancer and overall survival in head-and-neck cancer from pre-chemoradiotherapy FGD-PET scans. Shortly after this baseline study, the first use of the word "radiomics" was reported in the literature, as it will be described in the next section. In the past few years, the use of radiomic analysis – particularly texture analysis – for the assessment of tumour aggressiveness via the quantification of intratumoural heterogeneity has gained a lot of interest in the medical imaging community due to its great potential in extensively characterizing the complexity of spatial variations of gray-level distributions (i.e., spatial intensity patterns) within tumours.

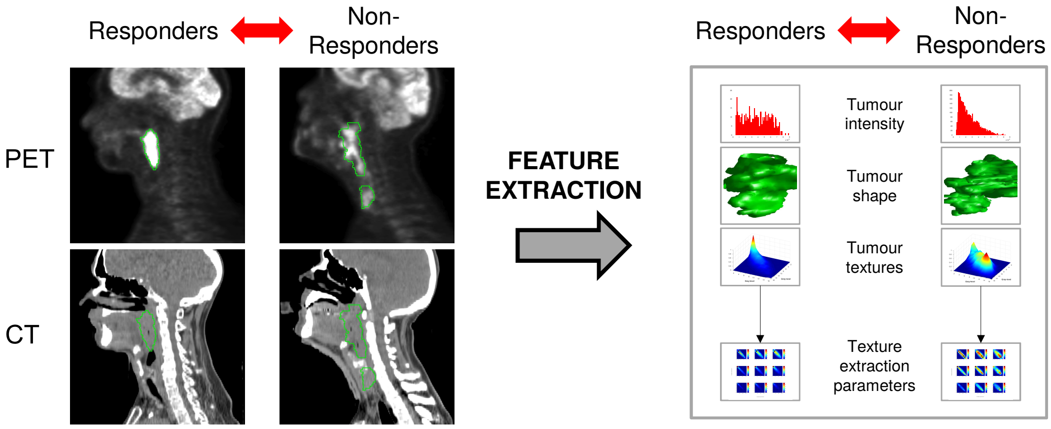


Figure 1.6: Principle of treatment response prediction via extraction of imaging biomarkers. Different types of features can be extracted from the tumour region of pre-treatment medical images to discriminate between patients likely to respond well to treatment from those who are not. The pictures on the left represent pre-treatment FDG-PET and CT images of head-and-neck cancer patients, with the primary tumour and lymph nodes contoured in green.

1.2.2 Definition and hypothesis of radiomics

In 2007 and 2008, respectively, Segal *et al.* [42] and Diehn *et al.* [43] showed that gene-expression signatures and clinical phenotypes could be inferred from tumour imaging features. This concept constitutes the central hypothesis of the field of "radiomics":

> *The genomic heterogeneity of aggressive tumours could translate into heterogeneous tumour metabolism and anatomy, which in turn could be captured using advanced quantitative analysis of medical images.*

Shortly following the study of El Naqa *et al.* [41] in 2009, Gillies *et al.* [44] employed in 2010 the first use of the word "radiomics" (to the best of our knowledge) to describe how imaging features can reflect gene expression. Afterwards, in 2012, Lambin *et al.* [28] and Kumar *et al.* [45] put the grounds on the field with two comprehensive descriptions of the processes and challenges of radiomics. No consensus definition of radiomics exists yet, and the following is an attempted definition of our own:

> *"Radiomics" refers to the characterization of tumour phenotypes via the extraction of high-dimensional mineable data from all types of medical images and whose subsequent analysis aims at supporting clinical decision-making.*

Similarly to computer-aided diagnosis (CAD) systems [46], radiomics consists of a top-to-bottom approach to better understand the underlying biology of tumours. A large number of features are extracted from medical images, and subsequent data mining attempts to identify the features that are associated to different tumour phenotypes. In the last years, this new emerging field of radiomics experienced an exponential growth as detailed in the excellent reviews of Hatt *et al.* [47] and Yip & Aerts [48]. Many researchers now advocates for the integration of radiomics into the panomics framework. Despite being in its early development stage with yet much standardization and validation work to perform, the use of high-order imaging biomarkers dedicated to the quantification of intratumoural heterogeneity holds great promise for better tumour aggressiveness assessment and subsequent treatment personalization.

1.2.3 Major objective of radiomics

In this section, the major objective of radiomics is described in more details. Figure 1.7 illustrates the conceptual objective.

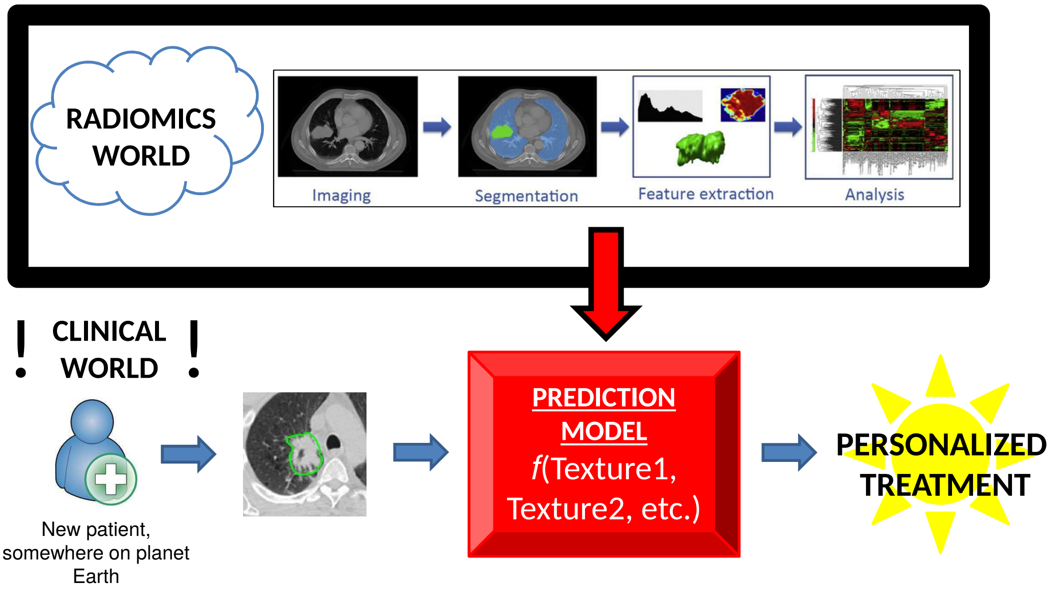


Figure 1.7: Major objective of radiomics. The major goal of radiomics analysis is to construct highly generalizable and predictive tumour outcome prediction models. The ultimate objective is to use these models in day-to-day clinical practice to assist physicians in tailoring cancer treatments to each individual patient, and hopefully improve survival. Parts of the figure are reprinted with permission from [28]. © 2012 Elsevier. All rights reserved.

Radiomics analysis starts with the acquisition of medical images. A region-of-interest (ROI) defining the tumour region is thereafter delineated, either via manual segmentation from an expert physician or semi-automatic methods. A large number of radiomics features are then extracted from the ROI (Appendix A provides the complete description of radiomic features used in this work). In the subsequent analysis part, models combining relevant prognostic factors via machine learning may be constructed to improve outcome prediction performance, as multivariable models are expected to more comprehensively characterize intratumoural heterogeneity than single features. The framework of multivariable model construction usually starts with the identification of imaging biomarkers that are significantly associated to a given tumour clinical phenotype or outcome (e.g., likelihood of development of distant metastases). Some of these features are then selected via a feature selection process and combined using a machine learning algorithm (e.g., logistic regression, random forests, etc.) to form a multivariable model, and its predictive properties are estimated. The search for the best parsimonious model (the simplest model with the best prediction performance) is the crucial step of any multivariable approach. Enough variables need to be selected in the model in order to reach maximum predictive power, but the number of variables must also be kept low such that the model can subsequently be generalized to, ideally, the whole patient population. With small sample size and without the presence of an independent dataset, the estimation of the prediction performance to unseen data can be internally simulated on the local patient population using different resampling techniques such as cross-validation or bootstrapping (more details in section 2.4.3). Once a final model is constructed (e.g., function of texture 1, texture 2, etc.), its predictive properties needs to be further validated onto independent external datasets.

Overall, the major objective of radiomics analysis is to be able to construct a robust prediction model for a given tumour outcome from retrospective medical imaging datasets, and to show that it is highly generalizable and predictive when applied to different independent datasets. If the application of the model is therafter demonstrated to improve patients' survival via the conduction of clinical trials, it could ultimately be used in day-to-day clinical practice to assist physicians from different hospitals in the world in their future choice of precise/personalized treatments for patients afflicted by cancer.

1.2.4 Overview of the role of radiomics in oncology

The workflow of radiomics analysis leading to the extraction of clinically relevant information involves many steps such as medical imaging acquisition, image processing, tumour segmentation, feature extraction, statistical analysis, and development and validation of multivariable models for tumour outcome prediction via statistical or machine learning techniques. The complexity of such workflow opens the door to many interesting development possibilities in the field. Medical physicists could play an important role in the research and development leading to the translation of radiomics anaysis in the clinical environment, for every of the steps mentioned above. Figure 1.8 roughly generalizes the radiomics analysis workflow into four major steps: I) *Medical imaging acquisition*; II) *Radiomics modeling*; III) *Tumour aggressiveness assessment*; and IV) *Personalization of treatments*. In this section, some of the works that have been performed in the field (non-exhaustive list), potential applications as well as possible developments in each of these four general steps will be briefly mentioned.

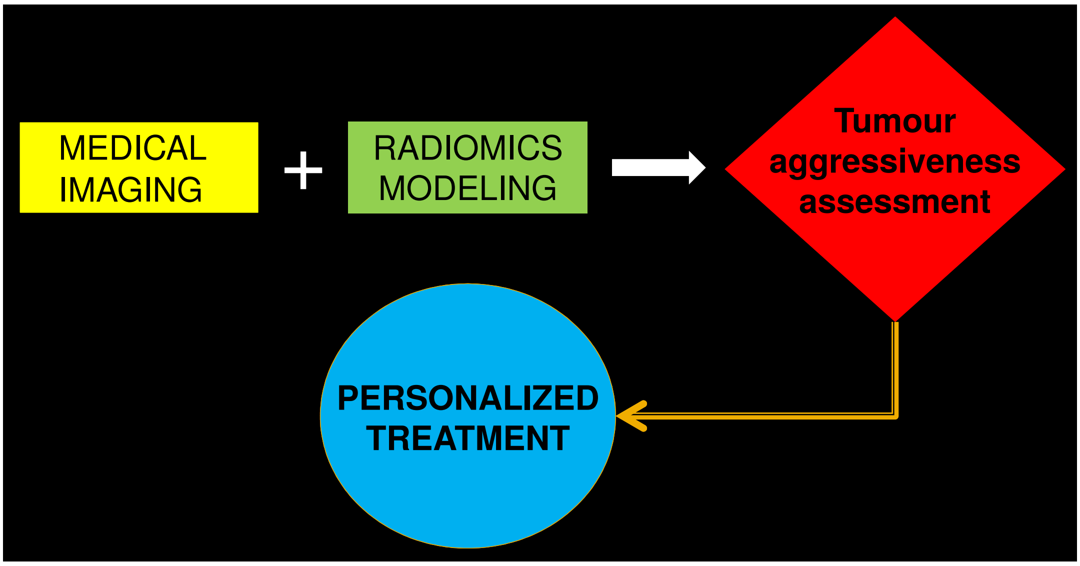


Figure 1.8: Overview of the role of radiomics in oncology. In every step of the general workflow pictured above, medical physicists could play an important role in the research and development leading to the translation of radiomics analysis in the clinical environment.

Medical imaging acquisition

At the very start of radiomics analysis lies the medical images acquired to probe the anatomy and metabolism of tumours. Many studies have investigated the reproducibility of various texture features under test-retest scans (i.e., two scans of the same patient repeated after a short period of time) [49–52]. Others also studied the impact of different image reconstruction parameter variations on texture features [51–55]. Furthermore, some works examined the influence of varying image acquisition protocols on the resulting textures [53, 56–59]. The common denominator of all the studies enumerated here is their main working objective: they aim at identifying the texture features that could be stable and that are presumably able to conserve predictive properties under varying imaging conditions. While the identification of stable features is valuable to build robust and reproducible texture-based predictive models, we hypothesize in this thesis that it is also essential to identify the acquisition settings that would yield optimal use of texture features for a given clinical problem, an exercise which is currently not performed in the radiomics community. This hypothesis is addressed in Chapter 5 of this thesis. Moreover, post-processing methods were also developed by some groups in order to improve the quality of medical images after scanning acquisitions, such as intensity non-uniformity corrections in MRI [60] and partial-volume effect (PVE) corrections in PET [61]. Overall, the optimization of medical imaging acquisition protocols and subsequent image post-processing is an interesting avenue to explore in order to enhance the predictive properties of texture features.

Radiomics modeling

The core of radiomics analysis lies in the extraction of features relevant to tumour aggressiveness assessment, and the subsequent statistical or learning analysis to relate these features (or combination of features) to tumour outcomes. Radiomic feature extraction involves many image pre-processing steps such as tumour segmentation, image interpolation, spatial filtering and image quantization. These steps are described in more details in section 2.2 of this thesis.

Many studies have investigated the impact of contouring variations on texture features [50, 52, 62–64]. The use of semi-automatic segmentation methods such as those described in the works of Hatt *et al.* [65] and Parmar *et al.* [63] would be a good way to improve the stability of radiomic feature extraction. Some studies have also examined the impact of variations in voxel size and image quantization on texture features [64, 66–68]. Most studies notably report the high impact of voxel size on texture measurements. Again, the investigational objective of all similar studies in the literature lies solely in the identification of stable texture features under different extraction parameters. In this thesis, we hypothesize that different texture features better represent the underlying biology of tumours when computed using different extraction parameters (isotropic voxel size, quantization algorithm, number of gray levels)[1], and that is it essential to optimize the set of extraction parameters of different texture features for a given application. We hereby denote this process as "texture optimization". We also hypothesize that the image fusion of different modalities (e.g., FDG-PET and MRI) could create composite textures with better predictive properties. These hypotheses are addressed in Chapter 3 of this thesis. Furthermore, the standardization and development of texture features and image pre-processing methods such as in the colossal work of Zwanenburg *et al.* [69] constitute fundamental prerequisites for the translation of radiomics analysis to the clinic.

[1] An isotropic voxel size refers to voxels of a 3D imaging volume with the same dimension in the three directions of space, (i.e., cubic voxels). Textures can be extracted from imaging volumes with different isotropic voxel sizes. The quantization of an imaging volume is a process used prior to texture computation for reducing the intensity range of the ROI of an imaging volume into a discretised number of gray-level bins (N_g). Different algorithms exist for that purpose. The final number of gray levels N_g in a quantized imaging volume is another parameter influencing the absolute value of texture features.

Moreover, continuous improvements in machine learning techniques for optimal modeling of tumour outcomes via radomics analysis also constitute an area of active research. For example, Parmar *et al.* [70] investigated the prognostic values of radiomic features when combined with different popular choices of feature selection and machine learning algorithms. However, advancements in multivariable modeling processes are required in order to reduce the number of false-positive results [71], as well as to take into account the imbalance in the proportion of positive/negative outcomes (e.g., distant metastases occur in ∼15 % of head-and-neck cancers) and the extraction of texture features using multiple parameters. These topics are addressed in Chapter 3 and Chapter 6 of this thesis.

Tumour aggressiveness assessment

The potential ability of radiomic analysis to decode tumour phenotypes and to subsequently assess tumour aggressiveness brings about many potential applications in the structure of precision oncology [47, 48]. Just to name a few, many works have now used radiomics analysis for the prediction of tumour outcomes (e.g., treatment response, distant metastases, local recurrences, survival, etc.) [72–74], tumour staging [75–77], tissue identification [78–80], and assessment of cancer genetics (a.k.a. radiogeneomics) [73, 81, 82]. The exact processes about how physiological processes translate into imaging phenotypes remain however unclear, and future investigations to elucidate the biological meaning of radiomic features is required [48].

Overall, one of the major challenges in the upcoming years in precision oncology will be to construct models that can comprehensively integrate panomics and clinical information (e.g., tumour stage, grade, sub-type, etc.) with radiomics data. No consensus methodology exists yet, and in Chapter 6 of this thesis, we hypothesize that a random forest algorithm would be very well suited to construct prediction models integrating input prognostic factors of different types such as radiomics (continuous inputs) and clinical information (categorical inputs).

Personalization of treatments

Ultimately, the underlying major goal of radiomics analysis is to provide physicians with additional information that may allow to better personalize cancer treatments. The exact manner in which treatments would be tailored to each patient following cancer risk assessment via radiomics analysis, however, remains to be defined, but possible scenarios could be elaborated. If, for example, a patient would be identified via radiomics analysis of pre-treatment medical images to be at higher risk (than standard levels) of developing distant metastases, the chemotherapy doses could be strengthened. Inversely, for lower risk patients, diminishing or completely removing chemotherapy doses would increase the quality of life of patients, not to mention that it would also reduce the cost of the overall treatment. Another example would be for patients identified to be at higher risks of developing a local recurrence (or metastasis, as we will see in Chapter 4) of their primary tumour. In this case, it could be envisioned to carry out radiotherapy differently by boosting the sub-regions of the tumour that are more radioresistant due to low oxygenation levels, for example. Overall, the dose delivery could be "painted" and modulated accordingly to the different biological processes inherent to the different tumour sub-regions as identified via functional medical imaging, a procedure known as "dose painting" [83]. However, much research efforts are still required before this paradigm is used in routine radiotherapy treatment planning [84]. A more detailed discussion of this topic is provided in section 1.3, and Chapter 4 of this thesis presents a dose painting feasibility study performed in the context of radiomics analysis.

1.3 Paradigm of radiation dose painting

As previously mentioned, the underlying major goal of radiomics analysis is to provide physicians with additional risk assessment information that may allow to better personalize cancer treatments. Radiotherapy may provide an ideal setting to apply the principles of rapid learning in a precision oncology context given the field's high degree of computerisation and long use of predictive models [14, 21, 85]. In this section, a brief introduction to radiation therapy will be presented, followed by a short discussion about how different biological target volumes could be used in the context of dose painting.

1.3.1 Radiation therapy

Radiation therapy involves the use of ionizing radiation to kill malignant cells via the damaging of deoxyribonucleic acid (DNA) strands. The source of the radiation can either be external (and thus penetrates through the body before reaching the tumour) or internal via the insertion of a radiation source inside the tumour (brachytherapy). The overall goal of radotherapy is to provide the maximum radiation dose to cancerous cells to control the tumour while minimizing the dose imparted to normal tissues. External beam radiation therapy (EBRT) is more common due to its non-invasiveness procedure, but the inherent drawback is additional damage to healthy tissues. The absorbed dose is defined as the average energy deposited per unit mass, and is measured in units of Gray (Gy), with 1 Gy = 1 J/kg. Two forms of action of the absorbed dose cause damage to the DNA: I) *Direct* action, where charged particle tracks directly deposit energy to the DNA and cause strand breaks; and II) *Indirect* action, where the deposition of energy to the DNA is caused by water radiolysis and the production of free radicals such as hydroxyl (OH^-) via the ionization of water particles. Due to the latter process, tumour microenvironments with low oxygen levels are more resistant to radiation.

Once it is determined that a patient diagnosed with cancer is to undergo radiation therapy, a first set of CT images around the treatment site is acquired. This set of CT images is denoted as the planning or simulation CT (“CTsim”). The positioning of the patient during this scan is set up to be the same as during the radiation treatment. Targeted tumour volumes and organs at risks are then contoured on the planning CT. In order to achieve maximal dose conformity, careful tumour volume definition is a prerequisite to meaningful 3D treatment planning, and is usually performed by expert radiation oncologists. For treatment planning purposes, four main types of target tumour volumes are contoured as illustrated in Figure 1.9. The International Commission on Radiation Units and Measurements (ICRU) defines these target volumes as:

- Gross Tumour Volume (GTV): “The Gross Tumour Volume (GTV) is the gross palpable or visible/demonstrable extent and location of malignant growth” [86].

- Clinical target volume (CTV): “The clinical target volume (CTV) is the tissue volume that contains a demonstrable GTV and/or sub-clinical microscopic malignant disease, which has to be eliminated. This volume thus has to be treated adequately in order to achieve the aim of therapy, cure or palliation" [86].

- Internal target volume (ITV): Consists of the CTV plus an internal margin designed to take into account variations in position and size of the CTV due to organ motions such as breathing and bladder or rectal contents [87, 88].
- Planning target volume (PTV): "The planning target volume (PTV) is a geometrical concept, and it is defined to select appropriate beam arrangements, taking into consideration the net effect of all possible geometrical variations, in order to ensure that the prescribed dose is actually absorbed in the CTV" [86].

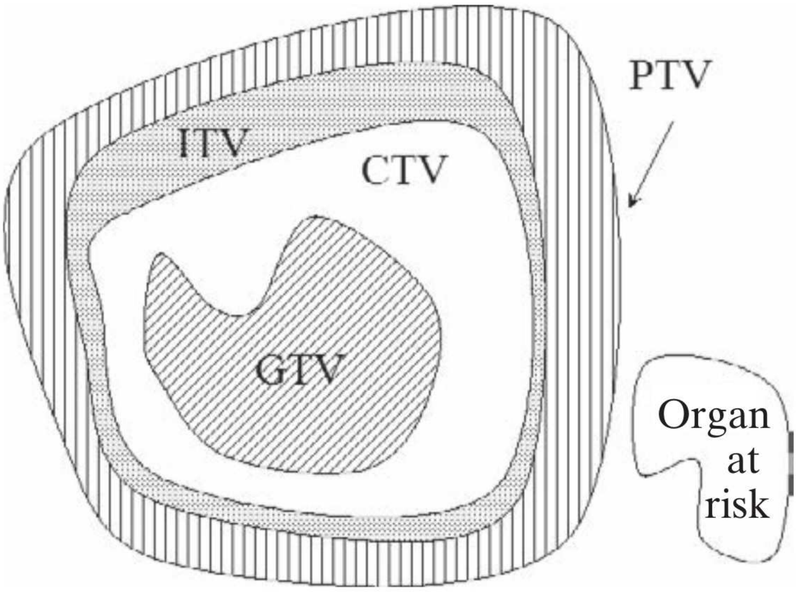


Figure 1.9: Graphical representation of the target contours in radiation therapy. Reprinted with permission from [88]. © 2005 IAEA. All rights reserved.

The contoured CT images are thereafter imported into a treatment planning system to undergo (using dose calculation engines) optimization of a number of one or more beam types of different energies (typically in the range of 4-25 MV for photons, 4-25 MeV for electrons) to be directed at the tumour from one or more directions. Depending on the specific contoured structure (e.g., target volumes, organs at risk, etc.), the associated objective function of the optimization process is usually cast in terms of dose-volume constraints summarized using the D_x and V_x metrics. The D_x metric is the dose in Gy received by at least x % of the volume of the region-of-interest, whereas the V_x metric is the percentage volume of the region-of-interest receiving at least x Gy. For an example prescription dose of 50 Gy to the tumour, example dose-volume constraints to ensure homogeneous coverage at the desired prescription dose could require that a minimum of 50 Gy covers at least 95 % of the PTV, that $> 99\ \%$ of the PTV receives at least 97 % of the prescribed dose and that $< 2\ \%$ of the PTV receives at least 110 % of the prescribed dose. These constraints to the PTV structure would be represented by $D_{95\,\%} \geq 50\ \mathrm{Gy}$, $V_{48.5\,\mathrm{Gy}} > 99\ \%$ and $V_{55\,\mathrm{Gy}} < 2\ \%$, respectively.

For EBRT, the finalized treatment plan is sent as a set of electron or high-energy X-ray beam delivery instructions to a linear accelerator (LINAC) on which the patient to be treated is positioned. Modern treatment planning systems and LINACs can provide radiation treatments in the form of Volumetric Arc Therapy (VMAT), an advanced form of Intensity Modulated Radiation Therapy (IMRT) with a single or multi-arc treatment. In this process, the photon fluence is dynamically modulated by the multi-leaf collimators (MLC) of the LINAC using multiple small and irregular field sizes in order to maximize the conformity of the dose distribution to the target. More details about the whole radiotherapy process can be found in the excellent textbook by Podgorsak [89].

1.3.2 The Biological Target Volume (BTV)

In an important seminal paper in 2000, Ling *et al.* [83] proposed the concept of the "biological target volume" (BTV). The authors hypothesized that the "BTV can be derived from biological images . . . [and that its use] may provide the pertinent information to guide the painting or sculpting of the optimal dose distribution [in radiotherapy]" [83]. Biological images would broadly include the metabolic, biochemical, physiological and functional types of noninvasive images. In the context of radiation therapy, the images providing information about the radiosensitivity of tumours would be regarded as radiobiological images. The concept of the BTV as originally depicted in the work of Ling *et al.* [83] is illustrated in Figure 1.10.

The central "dogma" in radiation therapy has been to strive for a homogeneous dose to the target volume as defined by the GTV, CTV and PTV volumes. Many researchers now challenge this approach and advocate for a inhomogeneous dose distribution inside the GTV to better take into account the underlying biological processes within tumours [90, 91]. As depicted in Figure 1.10, low oxygenation levels (i.e., hypoxia), high tumour burden

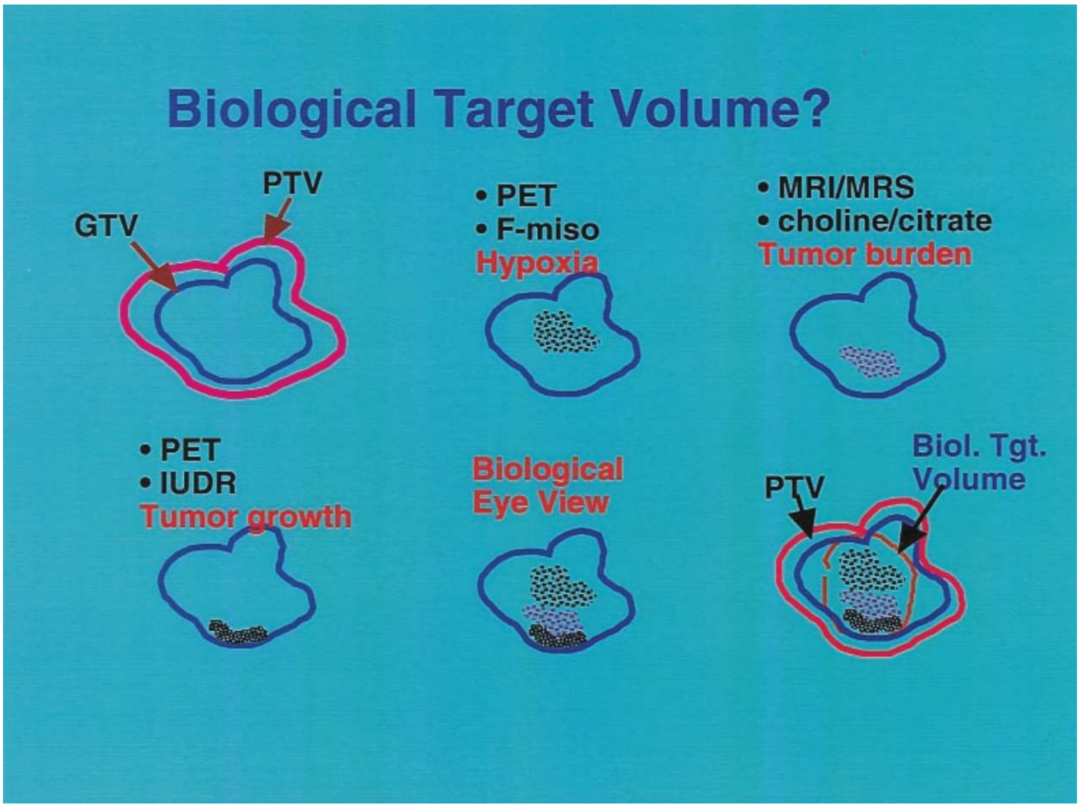


Figure 1.10: Schematic illustration of the concept of biological target volume. Whereas at present the radiotherapy target volume is still characterized by the concepts of GTV, CTV and PTV, biological images may provide information for defining the biological target volume (BTV) to improve dose targeting to certain GTV subregions. For example, regions of low pO_2 levels may be derived from FMISO-PET images, high tumor burden from MRI data of choline/citrate ratio, and high proliferation from ^{124}IUDR-PET measurements. Reprinted with permission from [83]. © 2000 Elsevier. All rights reserved.

and high cellular proliferation could be considered to form a single BTV on which a dose boosting strategy would be applied. Nowadays, dose boosting to hypoxic volumes to counteract radioresistance within tumours is considered one of the major aim of dose painting [92]. The dose painting hypothesis in terms of therapeutic gain is such that [91]: I) local recurrences take place in radioresistant microenvironmental tumour niches; II) molecular and functional imaging allows spatiotemporal mapping of these radioresistant regions; and III) progress in radiation therapy planning and delivery technologies can facilitate dose boosting to such regions, which in turn should lead to improved local tumour control with acceptable side effects. Examples of partial verifications of the dose painting hypothesis include a phase I clinical trial where it was shown that dose escalation guided by FDG-PET sub-volumes appears to be well-tolerated in head-and-neck cancer [93]. In soft-tissue sarcomas (STSs), it was shown that a boost dose to the margin at risk is also well-tolerated by patients [94].

Although improvements in tumour outcomes remains to be verified, the hypothesis that dose painting could be carried out on the basis of biological images is supported by a large body of experimental and clinical data. Briefly, in the case of PET imaging, it has been demonstrated that FDG uptake is dependent on the tumour microenvironment such that different regions of low oxygenation levels (hypoxia), cellular proliferation, blood flow and necrosis correlates either positively or negatively with FDG uptake [95, 96]. Also, fluoromisonidazole (FMISO) continues to be the main radiopharmaceutical used in PET imaging for the evaluation, prognostication and quantification of tumor hypoxia [97]. In the case of MRI, the vast variety of contrasts allowed by the numerous possible acquisition sequences can definitely play a role in the assessment of tumour physiology and the identification of cell sub-populations [98]. In addition to anatomical imaging, MR allows functional imaging of biological processes in the human body. For example, diffusion-weighted magnetic resonance imaging (DW-MRI) quantifies the degree of isotropic water diffusion in extracellular space as affected by the size and the distribution of cellular populations. It has been shown that DW-MRI can be used to assess regional cellularity and the aggressiveness of tumours [99, 100]. On the other hand, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) can provide information about vascular characteristics such as tissue perfusion, plasma volume, and mean transit time. It was recently shown that low-perfusion DCE-MRI information could also be used as a surrogate of tumour hypoxia [101, 102].

Moreover, Bentzen [103] suggested that dose painting should also be performed during the course of radiotherapy, such that adaptative treatments would take into account the temporal response of cancer to radiation. Ideally, the full four-dimensional dose distribution would be constrained, a concept referred to as "theragnostic imaging for radiation therapy". The authors also went further by proposing that biological imaging could one day allow dose painting to be carried out on a voxel-by-voxel basis (in contrast to providing homogeneous boost doses to well-defined sub-regions of the GTV), a concept they denoted as "dose painting by numbers" [91, 103]. Although most likely feasible with current LINAC technology [104], the clinical evidence for this technique belongs to the future. Validation of the biological imaging targets are still under investigations, and the spatial resolution of biological images needs to be improved to accurately probe the sub-millimeter microenvironments within tumours. Also, as tumours significantly evolve in metabolic activity and distribution of well-oxygenated and hypoxic regions during radiotherapy as defined by the 5 R's of radiobiology (repair, redistribution, reoxygenation, repopulation, radiosensitivity) [105], such changes would need to be monitored frequently during radiotherapy. At the moment, our ability to deliver precise and ultra-conformal radiation treatments may have surpassed our ability to efficiently image the tumour microenvironment with both high sensitivity and specificity, and more efforts in biological imaging research is still vital.

In Chapter 4 of this thesis, we consider a hybrid solution between a single boost to a discrete tumour sub-region and a complete dose painting by numbers approach. Our treatment planning feasibility study is carried out with STS patients, a heterogeneous group of malignant neoplasms of mesenchymal cell origin with a metastatic rate of $\sim$50 % in the case of high-grade tumours [106]. We hypothesize that double nested dose boosting to hypermetabolic and hypoxic tumour sub-regions of STS patients would be an improvement as compared to current clinical practice, where a homogeneous dose of 50 Gy to the PTV is prescribed as standard of care. In this study, a first level of boost is planned to the FDG-PET hypermetabolic and potentially more aggressive tumour sub-regions. Then, a second level of boost with higher dose is planned to the hypoxic sub-regions contained only within the hypermetabolic ones, as increasing evidence suggests that FDG accumulates preferentially in hypoxic cancer cells [95, 107–109]. Moreover, given additional evidence that intratumoural hypoxia can drive the metastatic phenotype [110, 111], including in STSs [112], we further hypothesize that higher radiotherapy doses to radioresistant components of the tumour might be a useful strategy to reduce the risk of developing distant metastases in STS cancer. Hence, if we identify, at the time of diagnosis, the STS patients that are at higher risk of developing distant metastases using radiomics analysis, a valuable treatment personalization strategy for these patients could be to incorporate a dose boost to the hypoxic tumour sub-regions (as defined via biological images) into radiotherapy planning.

1.4 Thesis hypotheses, objectives and organization

The focus of this work is on the development of radiomic-based models for the prediction of tumour outcomes. Our main working hypotheses are as follows:

- *Hypothesis 1*: Radiomics features such as textural metrics can assess tumour aggressiveness via the quantification of intratumoural heterogeneity. This information obtained prior to treatments could assist physicians in improving the personalization of cancer therapy.
- *Hypothesis 2*: Different texture features better quantify intratumoural heterogeneity when computed using different extraction parameters such as voxel size, quantization algorithm and number of gray levels. The optimization of texture feature extraction is essential to enhance the predictive properties of image textures for a given clinical endpoint.
- *Hypothesis 3*: The image fusion of different modalities such as FDG-PET and MRI can create composite textures with better predictive properties.
- *Hypothesis 4*: The optimization of medical imaging acquisition protocols can enhance the predictive properties of texture features extracted from the resulting acquired images.
- *Hypothesis 5*: The integration of radiomics data with other panomics and clinical information can enhance the performance of tumour outcome prediction models.
- *Hypothesis 6*: Radiation dose boosting to hypermetabolic and hypoxic tumour sub-regions can improve post-radiotherapy tumour outcomes.

Cancer risk assessment via radiomics analysis provides a valuable rationale for integrating or not dose painting in its general form into radiotherapy delivery for a given patient.

There are five main objectives in this work, all of which fall within the aim of using radiomics analysis to assist physicians in tailoring cancer therapy to each patient:

- *Objective 1*: Develop a robust methodology for the construction of radiomic-based prediction models that takes into account texture optimization and the imbalance between the proportion of patients with positive and negative events for a given tumour outcome (Chapter 3 and Chapter 6).
- *Objective 2*: Create fused FDG-PET/MR images with better textural predictive properties than the separate FDG-PET and MR images (Chapter 3).
- *Objective 3*: Verify the feasibility of double nested dose boosting to hypermetabolic and hypoxic tumour sub-regions inside the GTV in radiotherapy planning, and investigate the practical feasibility and clinical utility of acquiring four different types of biological images (FDG-PET, FMISO-PET, DW-MRI, DCE-MRI) at different time points in the course of radiotherapy management (Chapter 4).
- *Objective 4*: Enhance the predictive properties of a texture-based model by optimizing FDG-PET and MR image acquisition protocols (Chapter 5).
- *Objective 5*: Validate the methodology developed in *Objective 1* using independent external datasets, and develop a complementary methodology for integrating radiomics data with clinical information for better tumour outcome prediction performance (Chapter 6).

Overall, this thesis is organized in seven chapters and one appendix. Being a manuscript-based thesis, each chapter is written in a self-contained manner, and some concepts and references overlap between the different chapters. In more details, the organization of this thesis is as follows:

- *Chapter 1*: Introduction to the concepts of “Precision Medicine”, “Rapid-learning”, “Radiomics” and “Biological Target Volume”, from a clinical perspective.

- *Chapter 2*: Mathematical and computational background on radiomics modeling, notably in terms of image pre-processing, texture feature computation and machine learning algorithms employed in this work.
- *Chapter 3*: Description of the first manuscript. This retrospective study is about the development of texture feature computation and machine learning methods for the construction of radiomic-based prediction models taking into account texture optimization and data imbalance. Ultimately, an optimal radiomic-based model was constructed for the prediction of lung metastases in soft-tissue sarcomas using pre-treatment fused FDG-PET/MR images.
- *Chapter 4*: Description of the second manuscript. In this study, FDG-PET, FMISO-PET, DW-MRI and DCE-MRI images were prospectively acquired at pre-, mid- and post-radiotherapy timepoints for 18 patients at our institution between August 2013 and February 2016. The radiomic-based model developed in Chapter 3 was validated and the technical feasibility of dose painting was investigated onto these prospective patients.
- *Chapter 5*: Description of the third manuscript. In this study, the possibility of enhancing texture-based models (constructed using the methods developed in Chapter 3) via the optimization of PET and MR image acquisition protocols was investigated using computerized simulations.
- *Chapter 6*: Description of the fourth manuscript. In this study, radiomic-based models were constructed (using the methods developed in Chapter 3) for the prediction of locoregional recurrences and distant metastases in head-and-neck cancer. Two different patient cohorts were used for training the models using a new imbalance-adjustment strategy, and two other patient cohorts were used for independent testing. Furthermore, we used a random forest algorithm to combine radiomics data with patient clinical information.
- *Chapter 7*: Summary highlighting the scientific novelty in each manuscript, how the objectives of the research were met, the implications of our findings, as well as related future work that we intend to perform. Finally, a strong call is made for the standardization of radiomics methods, better transparency of radiomics studies, and full radiomics programming code and data sharing.

- *Appendix A*: Mathematical description of all radiomic features employed in this work.

1.5 References

1. ACS. *Global Cancer Facts & Figures* 3rd ed. 61 pp. (American Cancer Society – ACS, Atlanta, 2015).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians* **67,** 7–30 (2017).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
4. *TNM classification of malignant tumours* (eds Sobin, L. H., Gospodarowicz, M. K., Wittekind, C. & International Union against Cancer) 7th ed. (Wiley-Blackwell, Chichester, 2010). 310 pp.
5. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
6. Meldrum, C., Doyle, M. A. & Tothill, R. W. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* **32,** 177–195 (2011).
7. Renfro, L. A., An, M.-W. & Mandrekar, S. J. Precision oncology: a new era of cancer clinical trials. *Cancer Letters. New developments on Targeted Cancer Therapy* **387,** 121–126 (2017).
8. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* **19,** A68–A77 (2015).
9. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
10. Delattre, O. & Bult, C. Editorial overview: characterizing the cancer genome: mechanistic insights and translational opportunities. *Current Opinion in Genetics and Development* **42,** 78–80 (2017).
11. Wang, E. *et al.* Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Seminars in Cancer Biology* **30,** 4–12 (2015).
12. Yu, K.-H. & Snyder, M. Omics profiling in precision oncology. *Mol. Cell Proteomics* **15,** 2525–2536 (2016).
13. Marx, V. Biology: the big challenges of big data. *Nature* **498,** 255–260 (2013).
14. El Naqa, I. Biomedical informatics and panomics for evidence-based radiation therapy. *WIREs Data Mining Knowl. Discov.* **4,** 327–340 (2014).
15. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

16. Ashley, E. A. The precision medicine initiative: a new national effort. *JAMA* **313,** 2119–2120 (2015).

17. Marrone, M., Schilsky, R. L., Liu, G., Khoury, M. J. & Freedman, A. N. Opportunities for translational epidemiology: the important role of observational studies to advance precision oncology. *Cancer Epidemiol. Biomarkers Prev.* **24,** 484–489 (2015).

18. Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J. & Mills, G. B. Building a personalized medicine infrastructure at a major cancer center. *J. Clin. Oncol.* **31,** 1849–1857 (2013).

19. Etheredge, L. M. A rapid-learning health system. *Health Aff.* **26,** w107–w118 (2007).

20. Abernethy, A. P. *et al.* Rapid-learning system for cancer care. *J. Clin. Oncol.* **28,** 4268–4274 (2010).

21. Lambin, P. *et al.* Rapid Learning health care in oncology – an approach towards decision support systems enabling customised radiotherapy. *Radiother. Oncol.* **109,** 159–164 (2013).

22. Shrager, J. & Tenenbaum, J. M. Rapid learning for precision oncology. *Nat. Rev. Clin. Oncol.* **11,** 109–118 (2014).

23. Prasad, V., Fojo, T. & Brada, M. Precision oncology: origins, optimism, and potential. *The Lancet Oncology* **17,** e81–e86 (2016).

24. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278,** 563–577 (2016).

25. Dresel, S. *PET in oncology* 1st ed. (Springer, New York, 2008).

26. Bushberg, J. T., Seibert, J. A., Jr, E. M. L. & Boone, J. M. *The Essential Physics of Medical Imaging* 3rd ed. 1048 pp. (LWW, Philadelphia, 2011).

27. Plewes, D. B. & Kucharczyk, W. Physics of MRI: a primer. *J. Magn. Reson. Imaging* **35,** 1038–1054 (2012).

28. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48,** 441–446 (2012).

29. Heppner, G. H. & Miller, B. E. Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metastasis Rev.* **2,** 5–23 (1983).

30. Fidler, I. J. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res.* **50,** 6130–6138 (1990).

31. Yokota, J. Tumor progression and metastasis. *Carcinogenesis* **21,** 497–503 (2000).

32. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).

33. Longo, D. L. Tumor heterogeneity and personalized medicine. *N. Engl. J. Med.* **366,** 956–957 (2012).

34. Brodatz, P. *Textures: A Photographic Album for Artists and Designers* 1st ed. 112 pp. (Dover Publications, New York, 1966).

35. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).

36. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).

37. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

38. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).

39. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

40. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).

41. El Naqa, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* **42,** 1162–1171 (2009).

42. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25,** 675–680 (2007).

43. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105,** 5213–5218 (2008).

44. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin. Radiol.* **65,** 517–521 (2010).

45. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30,** 1234–1248 (2012).

46. Giger, M. L., Chan, H.-P. & Boone, J. Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Med. Phys.* **35,** 5799–5820 (2008).

47. Hatt, M. *et al.* Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur. J. Nucl. Med. Mol. Imaging,* 1–15 (2016).

48. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61,** R150–R166 (2016).

49. Tixier, F. *et al.* Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *J. Nucl. Med.* **53,** 693–700 (2012).

50. Leijenaar, R. T. H. *et al.* Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncologica* **52,** 1391–1397 (2013).

51. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6,** 23428 (2016).

52. Van Velden, F. H. P. *et al.* Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol. Imaging Biol.* **18,** 788–795 (2016).

53. Galavis, P. E., Hollensen, C., Jallow, N., Paliwal, B. & Jeraj, R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* **49,** 1012–1016 (2010).

54. Nyflot, M. J. *et al.* Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J. Med. Imaging* **2,** 041002 (2015).

55. Yan, J. *et al.* Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J. Nucl. Med.* **56,** 1667–1673 (2015).

56. Mayerhoefer, M. E., Szomolanyi, P., Jirak, D., Materka, A. & Trattnig, S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study. *Med. Phys.* **36,** 1236–1243 (2009).

57. Waugh, S. A., Lerski, R. A., Bidaut, L. & Thompson, A. M. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. *Med. Phys.* **38,** 5058–5066 (2011).

58. Zhao, B., Tan, Y., Tsai, W. Y., Schwartz, L. H. & Lu, L. Exploring variability in CT characterization of tumors: a preliminary phantom study. *Translational Oncology* **7,** 88–93 (2014).

59. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features: *Investigative Radiology* **50,** 757–765 (2015).

60. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging* **17,** 87–97 (1998).

61. Boussion, N., Rest, C. C. L., Hatt, M. & Visvikis, D. Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging. *Eur. J. Nucl. Med. Mol. Imaging* **36,** 1064–1075 (2009).

62. Hatt, M., Tixier, F., Rest, C. C. L., Pradier, O. & Visvikis, D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur. J. Nucl. Med. Mol. Imaging* **40,** 1662–1671 (2013).

63. Parmar, C. *et al.* Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* **9,** e102107 (2014).

64. Orlhac, F. *et al.* Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J. Nucl. Med.* **55,** 414–422 (2014).

65. Hatt, M. *et al.* Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int. J. Radiat. Oncol. Biol. Phys.* **77,** 301–308 (2010).

66. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56,** 38–44 (2015).

67. Bogowicz, M. *et al.* Stability of radiomic features in CT perfusion maps. *Phys. Med. Biol.* **61,** 8736 (2016).

68. Molina, D. *et al.* Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images. *Computers in Biology and Medicine* **78,** 49–57 (2016).

69. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative. *arXiv preprint.* arXiv: 1612.07003 (2016).

70. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* **5,** 13087 (2015).

71. Chalkidou, A., O'Doherty, M. J. & Marsden, P. K. False discovery rates in PET and CT studies with texture features: a systematic review. *PLoS One* **10,** e0124165 (2015).

72. Win, T. *et al.* Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non–small cell lung cancer. *Clin. Cancer. Res.* **19,** 3591–3599 (2013).

73. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

74. Upadhaya, T., Morvan, Y., Stindel, E., Le Reste, P. .-.-J. & Hatt, M. A framework for multimodal imaging-based prognostic model building: preliminary study on multimodal MRI in glioblastoma multiforme. *IRBM* **36,** 345–350 (2015).

75. Ganeshan, B., Abaleke, S., Young, R. C., Chatwin, C. R. & Miles, K. A. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* **10,** 137–143 (2010).

76. Dong, X. *et al.* Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumor stage. *Nucl. Med. Commun.* **34,** 40–46 (2013).

77. Mu, W. *et al.* Staging of cervical cancer based on tumor heterogeneity characterized by texture features on 18 F-FDG PET images. *Phys. Med. Biol.* **60,** 5123 (2015).

78. Mahmoud-Ghoneim, D., Toussaint, G., Constans, J.-M. & de Certaines, J. D. Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas. *Magn. Reson. Imaging* **21,** 983–987 (2003).

79. Way, T. W. *et al.* Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Med. Phys.* **33,** 2323–2337 (2006).

80. Xu, R. *et al.* Texture analysis on (18)F-FDG PET/CT images to differentiate malignant and benign bone and soft-tissue lesions. *Ann. Nucl. Med.* **28,** 926–935 (2014).

81. Nair, V. S., Gevaert, O., Davidzon, G., Plevritis, S. K. & West, R. NF-kB protein expression associates with 18F-FDG PET tumor uptake in non-small cell lung cancer: a radiogenomics validation study to understand tumor metabolism. *Lung Cancer* **83,** 189–196 (2014).

82. Zhou, H. *et al.* MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro. Oncol.* **19,** 862–870 (2017).

83. Ling, C. C. *et al.* Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality. *Int. J. Radiat. Oncol. Biol. Phys.* **47,** 551–560 (2000).

84. Devic, S. Towards biological target volumes definition for radiotherapy treatment planning: Quo vadis PET/CT? *J. Nucl. Med. Radiat. Ther.* **4** (2013).

85. Dokic, I. *et al.* Next generation multi-scale biophysical characterization of high precision cancer particle radiotherapy using clinical proton, helium-, carbon- and oxygen ion beams. *Oncotarget* **7,** 56676–56689 (2016).

86. ICRU. *ICRU Report 50: Prescribing, Recording and Reporting Photon Beam Therapy* 50 (International Commission on Radiation Units and Measurements (ICRU), Bethesda, MD, 1993).

87. ICRU. *ICRU Report 62: Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU Report 50)* (International Commission on Radiation Units and Measurements (ICRU), Bethesda, MD, 1999).

88. Parker, W. & Patrocinio, H. *"Clinical treatment planning in external photon beam radiotherapy", Radiation Oncology Physics – A Handbook for Teachers and Students* (ed Podgorsak, E. B.) 219–272 (IAEA, Vienna, 2005).

89. Podgorsak, E. B. *Radiation Oncology Physics – A Handbook for Teachers and Students* 657 pp. (IAEA, Vienna, 2005).

90. Tanderup, K., Olsen, D. R. & Grau, C. Dose painting: art or science? *Radiother. Oncol.* **79,** 245–248 (2006).

91. Bentzen, S. M. & Gregoire, V. Molecular imaging–based dose painting: a novel paradigm for radiation therapy prescription. *Seminars in Radiation Oncology* **21,** 101–110 (2011).

92. Horsman, M. R., Mortensen, L. S., Petersen, J. B., Busk, M. & Overgaard, J. Imaging hypoxia to improve radiotherapy outcome. *Nat. Rev. Clin. Oncol.* **9,** 674–687 (2012).

93. Madani, I. *et al.* Positron emission tomography-guided, focal-dose escalation using intensity-modulated radiotherapy for head and neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **68,** 126–135 (2007).

94. Tzeng, C.-W. D. *et al.* Preoperative radiation therapy with selective dose escalation to the margin at risk for retroperitoneal sarcoma. *Cancer* **107,** 371–379 (2006).

95. Pugachev, A. *et al.* Dependence of FDG uptake on tumor microenvironment. *Int. J. Radiat. Oncol. Biol. Phys.* **62,** 545–553 (2005).

96. Henriksson, E. *et al.* 2-Deoxy-2-[18F]Fluoro-D-Glucose uptake and correlation to intratumoral heterogeneity. *Anticancer Res.* **27,** 2155–2159 (2007).

97. Rajendran, J. G. & Krohn, K. A. F-18 fluoromisonidazole for imaging tumor hypoxia: imaging the microenvironment for personalized cancer therapy. *Seminars in Nuclear Medicine* **45,** 151–162 (2015).

98. Pathak, A. P. *"Magnetic Resonance Imaging of Tumor Physiology", Magnetic Resonance Imaging: Methods and Biologic Applications* (ed Prasad, P. V.) 279–297 (Humana Press, Totowa, 2006).

99. Padhani, A. R. *et al.* Diffusion-weighted magnetic resonance Imaging as a cancer biomarker: consensus and recommendations. *Neoplasia* **11,** 102–125 (2009).

100. Padhani, A. R. Diffusion magnetic resonance imaging in cancer patient management. *Seminars in Radiation Oncology* **21,** 119–140 (2011).

101. Cho, H. *et al.* Noninvasive multimodality imaging of the tumor microenvironment: registered dynamic magnetic resonance imaging and positron emission tomography studies of a preclinical tumor model of tumor hypoxia. *Neoplasia* **11,** 247–259 (2009).

102. Stoyanova, R. *et al.* Mapping tumor hypoxia in vivo using pattern recognition of dynamic contrast-enhanced MRI data. *Transl. Oncol.* **5,** 437–447 (2012).

103. Bentzen, S. M. Theragnostic imaging for radiation oncology: dose-painting by numbers. *The Lancet Oncology* **6,** 112–117 (2005).

104. Alber, M., Paulsen, F., Eschmann, S. M. & Machulla, H. J. On biologically conformal boost dose optimization. *Phys. Med. Biol.* **48,** N31–N35 (2003).

105. Steel, G. G., McMillan, T. J. & Peacock, J. H. The 5Rs of radiobiology. *Int. J. Radiat. Biol.* **56,** 1045–1048 (1989).

106. Brennan, M. F. Soft tissue sarcoma: advances in understanding and management. *The Surgeon* **3,** 216–223 (2005).

107. Dierckx, R. A. & Van de Wiele, C. FDG uptake, a surrogate of tumour hypoxia? *Eur. J. Nucl. Med. Mol. Imaging* **35,** 1544–1549 (2008).

108. Huang, T. *et al.* Tumor microenvironment–dependent 18F-FDG, 18F-Fluorothymidine, and 18F-Misonidazole uptake: a pilot study in mouse models of human non–small cell lung cancer. *J. Nucl. Med.* **53,** 1262–1268 (2012).

109. Li, X.-F., Du, Y., Ma, Y., Postel, G. C. & Civelek, A. C. 18F-Fluorodeoxyglucose uptake and tumor hypoxia: revisit 18F-Fluorodeoxyglucose in oncology application. *Transl. Oncol.* **7,** 240–247 (2014).

110. Bristow, R. G. & Hill, R. P. Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability. *Nat. Rev. Cancer* **8,** 180–192 (2008).

111. Wouters, B. G. & Koritzinsky, M. Hypoxia signalling through mTOR and the unfolded protein response in cancer. *Nat. Rev. Cancer* **8,** 851–864 (2008).

112. Brizel, D. M. *et al.* Tumor oxygenation predicts for the likelihood of distant metastases in human soft tissue sarcoma. *Cancer Res.* **56,** 941–943 (1996).

Chapter 2

Background on radiomics modeling

2.1 Medical imaging in cancer management 35
2.1.1 Positron Emission Tomography 36
2.1.2 Computed Tomography 39
2.1.3 Magnetic Resonance Imaging 41
2.2 Image processing 51
2.2.1 Spatial filtering and image fusion 51
2.2.2 Tumour segmentation 56
2.2.3 Image interpolation 57
2.2.4 Image quantization 59
2.2.5 Texture extraction parameters 63
2.3 Texture analysis 63
2.3.1 Gray-Level Co-occurence Matrix 66
2.3.2 Gray-Level Run-Length Matrix 67
2.3.3 Gray-Level Size Zone Matrix 68
2.3.4 Neighborhood Gray-Tone Difference Matrix 70
2.4 Machine learning 71
2.4.1 Mathematical notations and concepts 72
2.4.2 Logistic regression 73
2.4.3 Bootstrapping 74
2.4.4 Random forests 75
2.4.5 Performance measures 78
2.4.6 Estimation of prediction performance 80
2.5 References 84

2.1 Medical imaging in cancer management

This section provides a brief introduction to the theoretical backgrounds of PET, CT and MR imaging.

2.1.1 Positron Emission Tomography

Fundamentally, positron emission tomography (PET) imaging starts with the injection of a radiopharmaceutical in the body. The radiopharmaceutical is comprised of a radionuclide that is attached to a chemical compound, which acts like a physiological analog known as the "tracer". The tracer is chosen in order to target the metabolic function of interest of tumours undergoing a certain biological process such as, for instance, glucose uptake. On the other hand, the radionuclide is used in the imaging acquisition process and acts as a source of radiation emission captured by the imaging scanner. Radionuclides are unstable isotopes undergoing transient radioactive decay. Proton-rich isotopes with a low atomic mass number are used in PET imaging as they undergo the following positron decay process:

$$p \rightarrow n + e^+ + \nu \tag{2.1}$$

In the decay process of Equation 2.1, one proton (p) of the unstable nucleus of the radionuclide gets converted into a neutron (n). The energy liberated in the conversion process is transferred to a positron (e^+) and a neutrino (ν), which are ejected from the nucleus with a continuous kinetic energy spectrum. Figure 2.1 depicts the physical principles of PET imaging, starting from the emission of the positron.

Once the positron is emitted at a specific location in the body, it travels a few millimeters in tissues depending on its energy and undergoes several scattering events. At the end of its track, the positron annihilates with an electron (e^-) and the rest mass energy of the two particles is converted into two photons each of energy of 511 keV and nearly anti-parallel to each other. The detection of these two coincident photons along different lines of response (LORs) allows inference about the location of the radiopharmaceutical in the body. The detection of annihilating photons is recorded in time coincidence by several rings of radiation detectors that are placed around the outside of the patient in the PET scanner (and therefore it detects photons escaping from the inside of patients). A cylindrical PET scanner essentially consists of multiple rings of dectectors (Figure 2.1) stacked in the axial direction, with a patient in its center. A single detector on a ring is made of a scintillating crystal that converts the high-energy photons into brief pulses of visible light every time it is struck by an annihilation photon. The crystal is optically coupled to a photomultiplier tube (PMT) that converts and amplifies the scintillation light into an electrical signal.

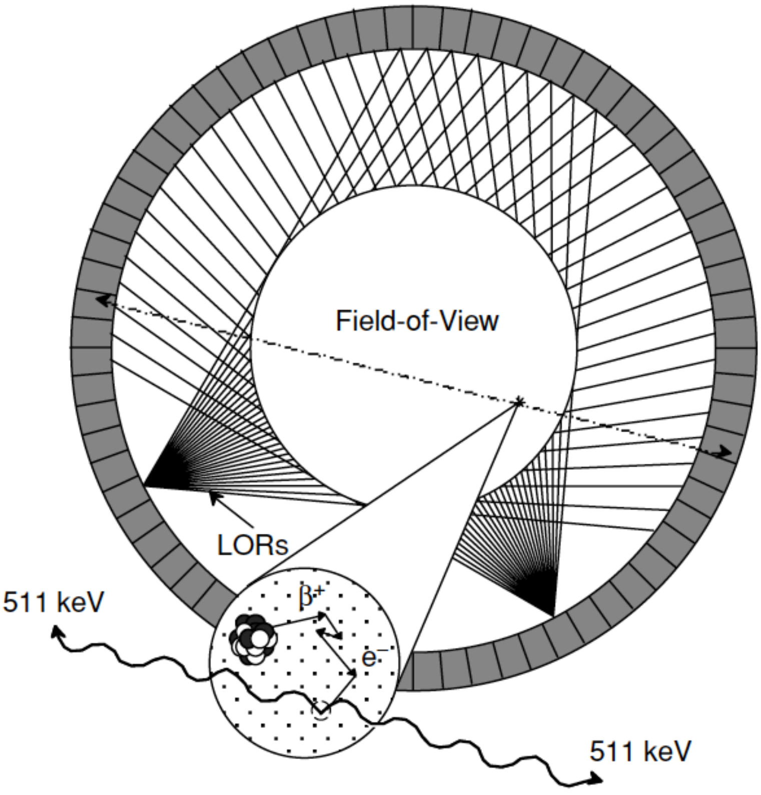


Figure 2.1: Physical principles of PET imaging. An unstable isotope attached to a physiological analog tracer (e.g., FDG for glucose uptake) first emits a positron in a decay process. The positron travels in the body for a few millimiters before annihilating with an electron. The annihilation process in turn creates two anti-parallel photons each of energy of 511 keV that are thereafter recorded in time coincidence by opposing radiation detectors in the cylindrical PET scanner. Reprinted with permission from [1]. © 2006 Springer. All rights reserved.

Coincident annihilating photons along many LORs are thus captured by the dectectors of the whole scanner. The extent of axial coincidence data combined per slice when detectors are allowed to be in coincidence with detectors in neighboring rings is an effect denoted as "span". Higher numbers of span increase slice sensitivity at the expense of a loss of resolution. If we now consider a single ring of detectors as a 2D example: each time a coincidence is recorded between two detectors of the ring, a matrix sinogram **M** is incremented at position (i, j), where i is the relative distance between two detectors and j the azimutal angle between two detectors with respect to the horizontal. From **M**, it is possible to reconstruct a spatial map of the radioactivity concentration [Bq/kg] of the radiopharmaceutical in the body using the "Maximum-Likelihood Expectation Maximization" (ML-EM) iterative (k) algorithm:

$$\mathbf{n}^{k+1} = \frac{\mathbf{n}^k}{\mathbf{A}^{\mathrm{T}}\mathbf{I}}\mathbf{A}^{\mathrm{T}}\left[\frac{\mathbf{m}}{\mathbf{A}\mathbf{n}^k}\right], \tag{2.2}$$

where **n** is the stacked column vector of all the different column vectors of a matrix image **N** to be reconstructed, **m** is the stacked column vector of all the different column vectors of **M**, **A** is the system matrix of the scanner simulating the noise-free transformation of an image to a sinogram space, and **I** is the identity matrix. The ML-EM is usually initialized with an image matrix full of 1′s. A faster variant of the ML-EM algorithm is called the "Ordered Subsets Expectation Maximisation" (OSEM) algorithm, in which only a subset of the measured data **m** is used in a given update of **n**. Typically, the subsets are chosen so as to select only a limited number of azimuthal angles in the sinogram data for each update. During the reconstruction, diverse corrections are applied, taking notably into account the attenuation and scatter of the annihilating photons traveling in the body, as well as random coincidences. For a more detailed description of the underlying physics of PET imaging, of the generalization of the image reconstruction in 3D and of PET imaging corrections, the reader is referred to references [2–5].

Nowadays, the most widely used radiopharmaceutical in the clinic for cancer detection and staging is ^{18}F-Fluorodeoxyglucose (FDG). The FDG tracer is a glucose analog in which the positron-emitting radionuclide fluorine-18 (^{18}F) with half-life of 110 minutes substitutes a normal hydroxyl group in the glucose molecule. Another type of pharmaceutical used in nuclear medicine is the hypoxia tracer ^{18}F-Fluoromisonidazole (FMISO), created from the chemical synthesis of ^{18}F with 2-nitroimidazole. Once the PET scan is performed, the images defining the radioactivity concentration map of the human body are converted into a semi-quantitative measure known as the standard uptake value (SUV) in order to account for injection and body weight variability as defined in Equation 2.3:

$$\mathrm{SUV} = \frac{\text{radioactivity concentration [Bq/kg]}}{\text{injected dose [Bq]}} \times \text{body weight [kg]} \tag{2.3}$$

2.1.2 Computed Tomography

Compute tomography (CT) imaging has proven to be a most useful tool in medical imaging to provide anatomical information of the human body. It is notably used in conjunction to PET scans in dedicated PET/CT scanners (notably for attenuation correction purposes), as well as for radiotherapy treatment planning. In contrast to PET imaging which relies on the *emission* of photons inside the human body to determine the concentration of a radiopharmaceutical, CT imaging relies on the *transmission* of photons through a patient to predominantly determine anatomical information by measuring the linear attenuation coefficient μ along a range of transmission lines. Similarly to PET, a CT scanner is formed of a cylinder bore, as depicted in Figure 2.2.

In CT imaging acquisitions, a source of X-ray radiation, usually in the energy range of 80-140 kVp, is rotated along the exterior of the detector rings to collect a large number of attenuation readings. As the radiation goes through patients, the beam is attenuated and the transmitted radiation is converted by the detectors into an electrical signal. Modern CT scanners use multiple rows of detectors in order to acquire multiple slices at once. In terms of physics, the intensity of radiation I at a point x within a "1D medium" of linear attenuation coefficient $\mu(x)$ is governed by the Beer-Lamber law:

$$I(x) = I(0)e^{-\int_0^x \mu(x')\,dx'}, \tag{2.4}$$

where $I(0)$ is the intensity of the radiation entering the body found by calibrating the scanner with measurements taken without a patient present ($\mu = 0$). The objective of the CT reconstruction is thereafter to determine how much attenuation of the X-ray beam occured in each voxel. This process is performed using a filtered backprojection algorithm [7], which essentially consists of dividing evenly the attenuation measurements along the path of the ray with a filter function convolved to each view before backprokjection.

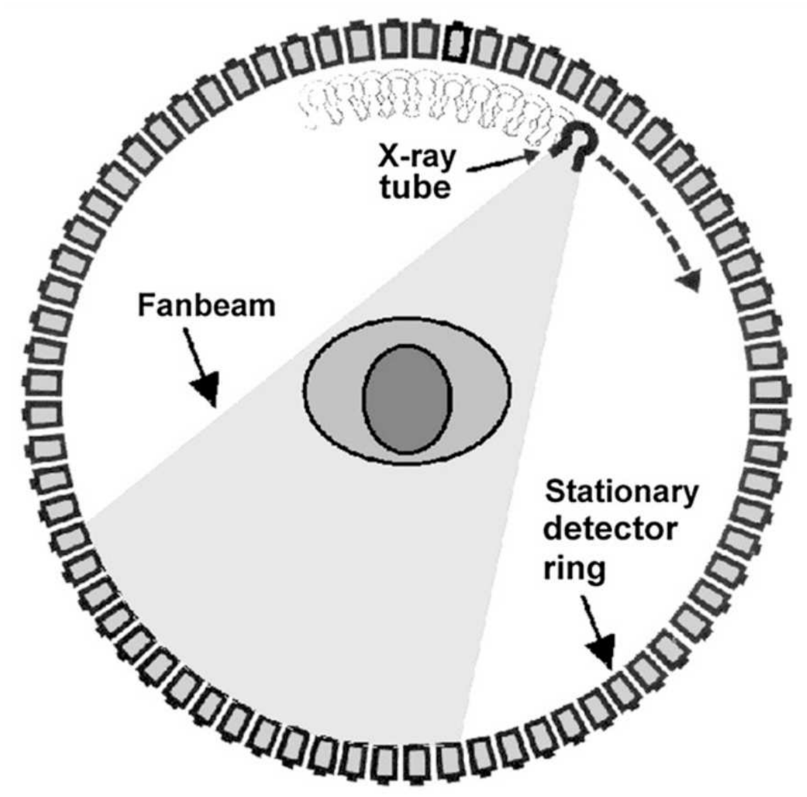


Figure 2.2: Physical principles of CT imaging. In CT imaging acquisitions, a source of X-ray radiation is rotated along the exterior of the detector rings. As the radiation goes through patients, the beam is attenuated and the transmitted radiation is converted by the detectors into an electrical signal. Reprinted with permission from [6]. © 2007 Society of Nuclear Medicine and Molecular Imaging, Inc. All rights reserved.

Once the attenuation map of the patient is reconstructed, the image intensities are converted to Hounsfield units (HU), which yields the attenuation coefficient relative to that of water, normalised according to:

$$\mathrm{HU} = 1000 \times \frac{\mu - \mu_{\mathrm{WATER}}}{\mu_{\mathrm{WATER}} - \mu_{\mathrm{AIR}}}. \tag{2.5}$$

HU values for typical materials include air at -1000, water at 0, soft-tissue in the range [100, 300], and bone in the range [700, 3000].

2.1.3 Magnetic Resonance Imaging

From an anatomical point of view, MR imaging provides a much superior soft-tissue contrast than CT images without loss of spatial resolution. MR scans are also non-invasive, as they do not expose patients to X-ray radiation as in the case of CT imaging. In this section, the underlying physical and imaging principles of MRI are presented.

Physical principles

The nuclear magnetic resonance (NMR) phenomenon occurs in atoms possessing a non-zero nuclear spin angular momentum. Due to its abundance in the human body, the NMR of the hydrogen atom (^{1}H) is most often exploited in MR imaging. In the presence of a large external and constant magnetic field B_0, a net ensemble of proton spins (^{1}H) align in the B_0 direction such that a net magnetization vector M_0 is created in tissues, as depicted in Figure 2.3.

The coupling of the magnetic moment of nuclear spins with the angular momentum of nucleons causes the magnetization vector M_0 to precess around B_0 with an angular frequency known as the Larmor frequency and defined as $\omega_0 = \gamma B_0$, where γ is the gyromagnetic ratio and is nuclei-specific (^{1}H: $\gamma = 42.58$ MHz/Tesla). Now, let B_0 lie in the z-direction, with M_0 precessing around it in its equilibrium position. By using a radiofrequency coil, a radiofrequency pulse (RF) with time-varying (general case) magnetic field $B_1(t)$ tuned to Larmor frequency ω_0 can be applied in the transverse plane (xy-plane). As a result, M_0 is excited into the transverse plane as depicted in Figure 2.4. As a result of the RF pulse, M_0 precesses towards the transverse plane for a duration τ by which $B_1(t)$ is applied. The resulting angular displacement θ by which M_0 is rotated away from the longitudinal axis (z-axis) is given by $\theta = \int_0^\tau \gamma\, B_1(t)\, dt$. A 90° RF pulse is such that the combination of

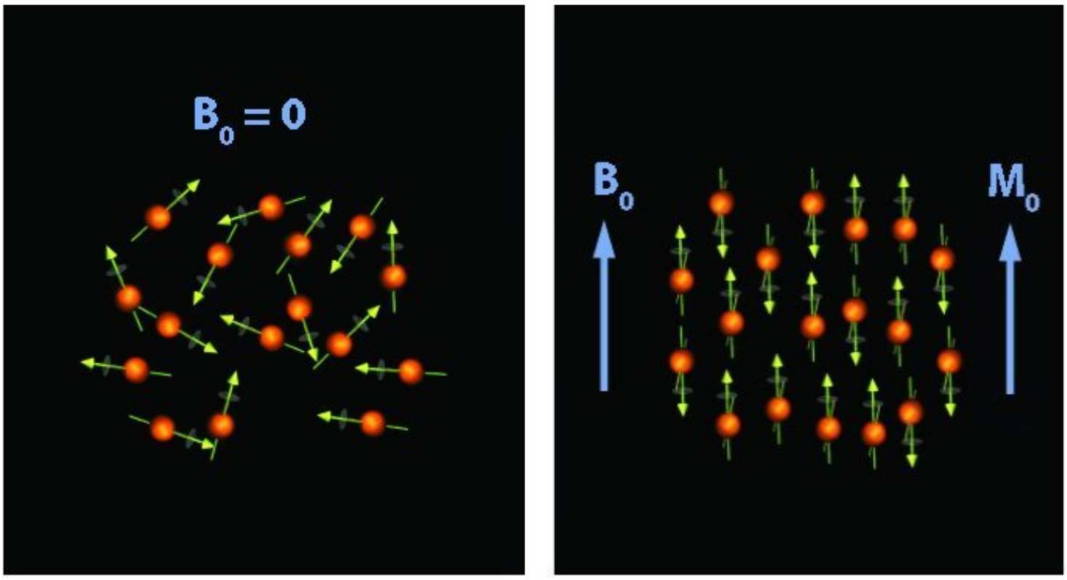


Figure 2.3: Alignment of nuclear spins in the presence of an external field. Without the presence of an external magnetic field B_0, no net magnetization exists in tissues. In the presence of an external field, a net magnetization vector M_0 is created. Reprinted with permission from [8]. © 2011 Eva Alonso Ortiz. All rights reserved.

τ and $B_1(t)$ generates $\theta = 90°$ for a given γ. After the RF pulse is completed, the transverse magnetization (M_{xy}) decays with characteristic relaxation time T_2 (spin-spin relaxation time, caused by a loss of phase coherence across a population of spins), and the longitudinal magnetization (M_z) recovers with characteristic time T_1 (spin-lattice relaxation time, a process whereby spins exchange energy with their surroundings) to its previous equilibrium state M_0 in the z-direction as governed by B_0. The spin-spin and spin-lattice relaxation times are physical quantities characteristics of each tissue in the human body. The rotating magnetization in the transverse plane then induces an oscillating electrical signal that can be captured and demodulated by two amplified radiofrequency coils placed at right angles in the transverse plane.

Imaging principles

In order to generate a MR image, spatial localization is necessary. This is achieved by applying different gradients of magnetic fields in the x-y-z directions in addition to the main field B_0, such that the total field strength varies in space. In this manner, the frequency of precession of spins varies with location since it is proportional to the magnetic field strength ($\omega = \gamma B$). A general formalism known as the Bloch equations [10] describes both the precession of the magnetization vector in different locations of the 3D space

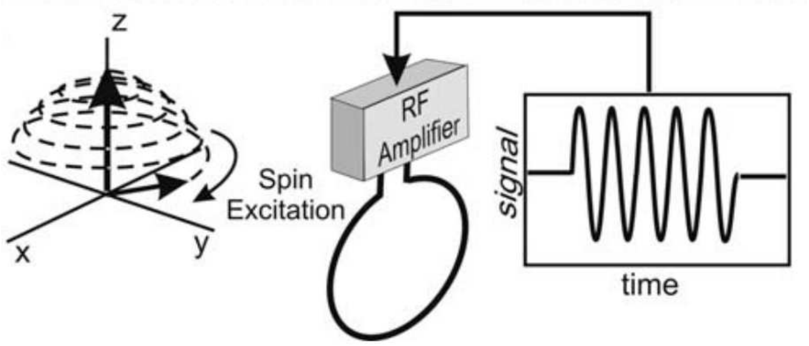


Figure 2.4: Principles of magnetization excitation and signal acquisition. The black arrow represents the net magnetization vector M_0. A constant magnetic field B_0 applied in the z-direction has for effect to have M_0 precessing/rotating around B_0. If another field B_1 tuned to the Larmor frequency is applied in the x-direction via a 90° radiofrequency pulse, this will cause M_0 to be excited and to be completely brought down along the y-axis. The subsequent precession/rotation of M_0 in the xy-plane then induces an oscillating electrical signal that can be captured by radiofrequency coils. Reprinted with permission from [9]. © 2012 John Wiley and Sons. All rights reserved.

due to arbitrary applied magnetic fields as well as the transverse and longitudinal relaxations, such that:

$$\frac{d\overrightarrow{\mathbf{M}}(t)}{dt} = \overrightarrow{\mathbf{M}}(t) \times \gamma\overrightarrow{\mathbf{B}}(t) - \frac{M_x(t)\,\hat{\mathbf{x}} - M_y(t)\,\hat{\mathbf{y}}}{T_2} - \frac{\big(M_z(t) - M_0\big)\,\hat{\mathbf{z}}}{T_1}, \tag{2.6}$$

where $\overrightarrow{\mathbf{M}}(t) = M_x(t)\,\hat{\mathbf{x}} + M_y(t)\,\hat{\mathbf{y}} + M_z(t)\,\hat{\mathbf{z}}$ and $\overrightarrow{\mathbf{B}}(t) = B_x(t)\,\hat{\mathbf{x}} + B_y(t)\,\hat{\mathbf{y}} + B_z(t)\,\hat{\mathbf{z}}$. We now start with the following initial conditions: at $t = 0$, the 90° RF pulse has just been applied inside a MRI scanner with a permanent magnetic field $\overrightarrow{\mathbf{B}} = B_0\,\hat{\mathbf{z}}$. At that particular moment, the magnetization vector M_0 (initially in the z-direction before the RF pulse due to B_0) is now completely defined in the xy-plane such that $\left\|\overrightarrow{\mathbf{M}}_{xy}(t=0)\right\| = M_0$, with $M_{xy} \triangleq M_x + i\,M_y$ defined as a complex quantity and $\left\|\overrightarrow{\mathbf{M}}_{xy}(t)\right\| = \sqrt{M_x(t)^2 + M_y(t)^2}$. From these initial conditions, we can now solve Equation 2.6 for $M_z(t)$ and $\left\|\overrightarrow{\mathbf{M}}_{xy}(t)\right\|$, such that:

$$M_z(t) = M_0\,(1 - e^{t/T_1}), \tag{2.7}$$

$$\left\|\overrightarrow{\mathbf{M}}_{xy}(t)\right\| = M_0\,e^{-t/T_2}. \tag{2.8}$$

Equation 2.7 thus governs the recovery of the longitudinal magnetization to its equilibrium state over time after applying a 90° RF pulse as illustrated in Figure 2.5a, and Equation 2.8 the decay of the transverse magnetization as illustrated in Figure 2.5b. Furthermore, the T_1 and T_2 relaxation times are physical characteristics varying for every voxel of an imaging volume due to different tissue compositions with different molecular microenvironments, such that T_1 and T_2 are functions of the spatial location $\overrightarrow{\mathbf{r}}$. Variations over space of $T_1(\overrightarrow{\mathbf{r}})$ and $T_2(\overrightarrow{\mathbf{r}})$ relaxation times form the fundamental basis of the contrast in MR images, since different tissues with different relaxation times in the body produce different MR signals depending on the time after which the signal is acquired following an excitation pulse. Another source of contrast in MR images comes from the variations in proton (i.e., ^{1}H) density of different tissues over space, which inherently affects the magnitude of the magnetization vector in every voxel.

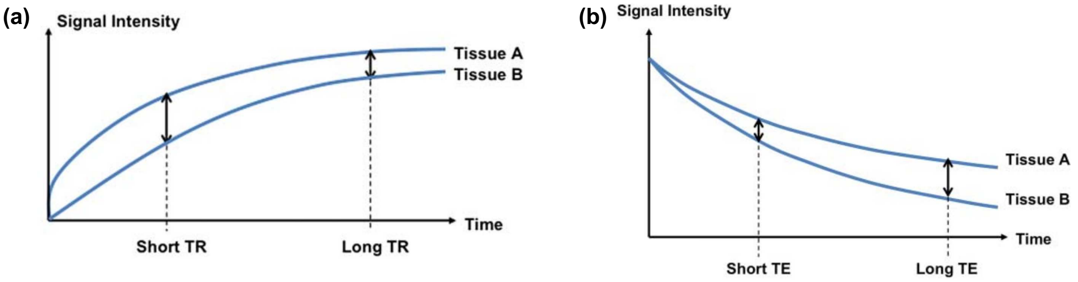


Figure 2.5: Principles of MR image contrast. (a) Recovery of the longitudinal magnetization to its equilibrium state after a 90° excitation pulse as governed by the T_1 relaxation time in Equation 2.7. Depending on the repetition time (TR) of a spin-echo sequence, for example, different tissues will produce different MRI signals, leading to different contrasts in the resulting images. (b) Decay of the transverse magnetization after a 90° excitation pulse as governed by the T_2 relaxation time in Equation 2.8. Depending on the echo time (TE) of a spin-echo sequence, for example, different tissues will produce different MRI signals, leading to different contrasts in the resulting images. Overall, a T_1-weighted image is created with a short TR and a short TE, and a T_2-weighted image is created with a long TR and a long TE. Reprinted with permission from [11]. © 2013 BMJ Publishing Group Ltd. All rights reserved.

Fundamentally, the signal acquired at a given time from the receiving coils contains the contributions of all excited spins of a given imaging volume that are spatially oscillating with different frequencies as governed by the tissue composition and gradient fields applied at that particular time. In other words, the intensity of the MR signal at a given time represents one point of the Fourier space of the MR image, known as the k-space. Taking the inverse Fourier transform of the time signal in turn linearly maps the contribution of each frequency component to its corresponding spatial location. This is the central concept allowing the formation of MR images. For example, a spin-echo sequence can be used to form an image where the contrast is based on the differences in T_1 or T_2 relaxation times of the different tissues of the human body. Figure 2.6 illustrates the formation of a typical spin-echo sequence. A 90° RF pulse is first employed at the beginning of the sequence. Simultaneously, a slice selection gradient (Gss) is applied such that only spins within a slice of interest are excited. At time TE/2, spins have started to dephase by a certain amount and a 180° RF pulse is applied to invert the phase of the spins. The spins then start to rephase such that at the echo time (TE), they are refocused and a high intensity spin-echo (SE) signal is created from the constructive interference of the spins. At time TE, the readout is performed by the frequency-encoding gradient (Gfe) in order to fill one line of the k-space. The process is repeated at the repetition time (TR) for another line of the k-space by using a different phase-encoding gradient (Gpe) strength. The process goes on for many TRs until the whole k-space is sampled. For more details about MRI physics, image formation and pulse sequences, the reader is referred to references [9, 11–13].

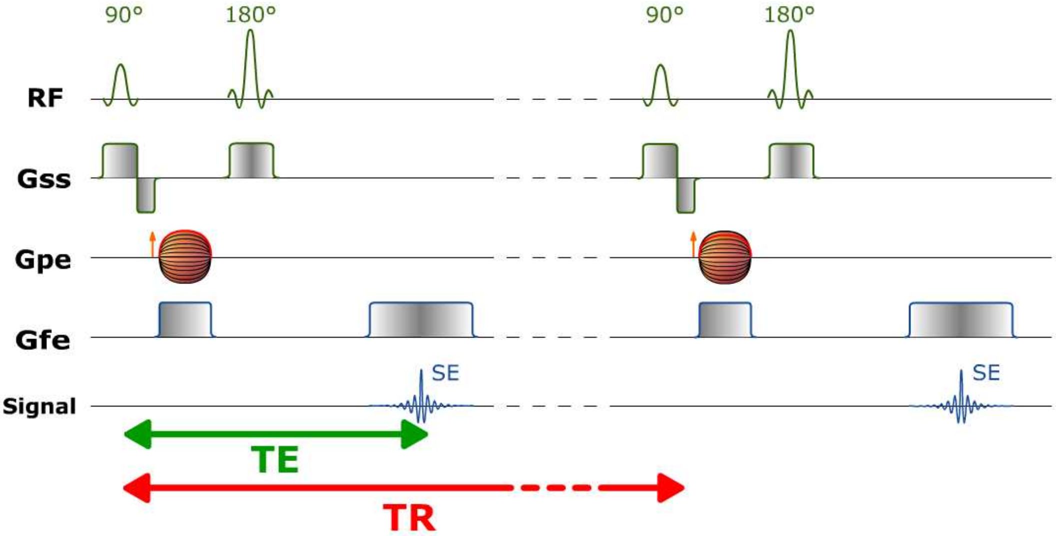


Figure 2.6: Schematic representation of a typical MRI spin-echo sequence. RF: Radiofrequency pulses, Gss: slice selecting gradient, Gpe: phase-encoding gradient, Gfe: frequency encoding gradient, SE: spin-echo, TE: echo time, TR: repetition time. Reprinted with permission from "e-Anatomy, Micheau A, Hoa D, www.imaios.com". © 2008-2017 IMAIOS SAS. All rights reserved.

Anatomical imaging

The vast variety of contrasts offered in MR imaging depends on the timing of data acquisition and strength of the different gradients used in the MRI sequences. Essentially, the repetition time TR of the sequence acquisition governs the time by which the longitudinal magnetization can recover, whereas the echo time TE governs the time by which the transverse magnetization can decay and be reverted back with a 180° RF pulse. Hence, if we go back to Figure 2.5, T_1-weighted images (i.e., emphasis on the contrast in the T_1 relaxation time of different tissue) are formed with a short TR $\sim T_1$ and a short TE. On the other hand, T_2-weighted images (i.e., emphasis on the contrast in the T_2 relaxation time of different tissue) are formed with a long TR and a long TE $\sim T_2$. Inherently, different choices of TE and TR in the spin-echo sequence will affect image contrast. In order to form images with different flavors of contrasts, many different kinds of MRI sequences exist. One class is defined as fat-suppression sequences and its main purpose is to enhance tumour visualization from its surrounding. T_2-weighted fat-saturation (T2FS) and short tau inversion recovery (STIR) sequences are part of this class, and are described in more details below:

- *T2FS*: This form of fat suppression technique exploits the small difference in resonant frequency between fat and water protons, which is related to their different electronic environments (chemical-shift effect). The sequence starts with a spectrally selective 90° pulse that ideally tips only the fat spins into the transverse plane. Only fat spins would contribute to the signal at this point. However, a spoiling gradient is applied immediately after the 90° pulse in order to dephase the fat spins in the transverse plane. As a result, fat signal decays to zero without affecting the water spins in their equilibrium state. The fat signal is then said to be "saturated" such that its contribution is suppressed in the subsequent standard MR sequence. The fat saturation step must then be repeated for every repetition of the MR sequence.

- *STIR*: Inversion-recovery methods exploit the fact that the characteristic time T_1 of fat is shorter than that of water. The sequence first starts with a 180° RF pulse such that the spins become anti-parallel to the main magnetic field. Subsequently to the pulse, the longitudinal magnetization of fat will return to equilibrium faster than the longitudinal magnetization of water. At one point in time, the longitudinal magnetization of fat will be null as it crosses the xy-plane. If a 90° pulse is applied at that time, only the magnetization of water will be transferred to the xy-plane to produce the signal of interest. Hence, for the rest of any subsequent standard MR sequence, the fat spins will not contribute to the signal. The sequence has to be repeated with a long enough TR such that all spins have time to recover.

Diffusion-weighted magnetic resonance imaging

In tissues, the water movement is not entirely random, as it can be influenced or hindered by flow within conduits and interactions with cellular membranes, vascular structures or macromolecules. Using a dedicated MRI sequence, diffusion-weighted magnetic resonance imaging (DW-MRI) quantifies the degree of isotropic water diffusion in extracellular space as affected by the size and the distribution of cellular populations via the apparent diffusion coefficient (ADC) metric, expressed in units of [mm^2/s]. The measured ADC is, therefore, inversely related to the cellularity of tumours.

DW-MRI essentially consists of a T_2-weighted sequence in which two additional and identical diffusion gradients are applied (Figure 2.7). The time between the first gradient and the 180° RF pulse is the same as the time between the pulse and the second gradient. At the start of the sequence, the 90° RF excitation pulse first brings the net magnetization into the transverse plane. The first diffusion gradient is then applied right after this pulse and causes a phase shift (i.e., dephasing) on the spins. Then follows the usual 180° RF pulse causing a phase flip, such that after it is being applied the spins would eventually rephase to create an echo signal. For the static water molecules, the use of the second diffusion gradient (identical to the first one) has for effect to restore the phase shift caused by the first gradient, and thus to restore phase coherence. For the moving water molecules, the acquired phase shift during the first gradient is not completely restored by the second gradient, resulting in residual phase incoherence. As a result, the acquired signal from the moving spins is lower than the one from the static spins.

The signal loss for moving spins is a function of the gyromagnetic ratio γ, the diffusion gradients intensity G, the time of application of each diffusion gradient δ, and the time separation between the two diffusion gradients Δ. All those gradient-dependent terms are gathered into a single factor called the "b-value" expressed in units of [s/mm^2] , such that:

$$b = \gamma^2 \, G^2 \, \delta^2 \, \big(\Delta - \frac{\delta}{3}\big). \tag{2.9}$$

The signal reduction $S(b)$ relative to a signal acquired with $b = 0$ can then be expressed as:

$$S(b) = S(b_0) \, e^{-b \cdot \mathrm{ADC}}. \tag{2.10}$$

By acquiring two imaging sets, one with $b = 0$ and the other with a non-zero b-value, the ADC of each voxel of an imaging volume can be separately

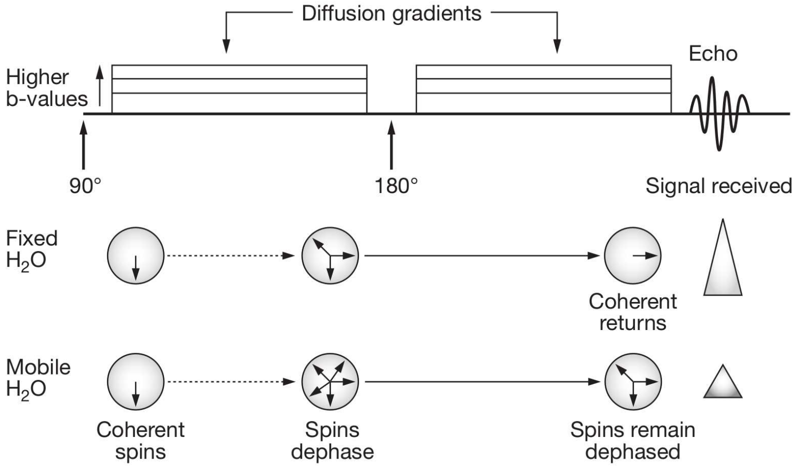


Figure 2.7: Schematic illustration of a DW-MRI sequence. The DW-MRI sequence is acquired using a T_2-weighted sequence, in which two identical diffusion gradients are additionally applied before and after the 180° RF pulse. The dephasing caused by the first gradient will be cancelled by the second gradient for static water molecules, but not for moving ones. As a result, the acquired signal from the moving water molecules is lower than the one from the static water molecules. Higher diffusion weighting (i.e., larger b-values) is usually achieved by increasing the amplitude of the diffusion gradients. Reprinted with permission from Macmillan Publishers Ltd [14]. © 2008 Nature Publishing Group. All rights reserved.

extracted by solving for ADC in Equation 2.10, such that:

$$\mathrm{ADC} = -\frac{1}{b - b_0} \ln \left(\frac{S(b)}{S(b_0)} \right). \tag{2.11}$$

For more information about DW-MRI, the reader is referred to references [14, 15].

Dynamic contrast-enhanced magnetic resonance imaging

One of the hallmarks of cancer is angiogenesis – the creation of new blood vessels. In fact, the growth of malignant cancer depends upon its ability to initiate the formation of new blood vessels to allow the tumour to grow and to supply it with oxygen and nutrients [16]. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) can provide information about vascular characteristics such as tissue perfusion, plasma volume and mean transit time. It thus provides valuable knowledge in oncology to characterize tumour phenotypes.

In DCE-MRI acquisition, a region-of-interest surrounding the tumour is first selected, and sets of 3D MR images are acquired before, during, and after the injection of a contrast agent (e.g., paramagnetic species such as gadolinium-based molecules) into the vein of a patient (Figure 2.8a). DCE-MRI is thus a 4D imaging technique aiming at quantifying the passage over time of the contrast agent from the vascular space of blood vessels to the interstitial space (or extravascular-extracellular space: EES) consisting of tumour cells and extracellular matrix components, thus providing information about tumour microvascular permeability. The passage of the contrast agent in the the interstitial space has for effect to decrease the T_1 relaxation time of its surrounding, leading to an enhanced MR signal proportional to the contrast agent concentration in a given voxel. Eventually, a wash-out effect can be observed if the vascular permeability is high and if there is reflux of contrast agent back to the vascular space. Typically, the signal acquisition is performed with 3D fast gradient-echo sequences acquired every 5-30 s for approximately 3-7 min [17]. Each set of acquired 3D images corresponds to one time point, and each voxel in the 4D set of images then gives rise to its own time-signal intensity-curve, which can thereafter be analyzed with different mathematical models to characterize the perfusion properties of tumour tissues. Different types of time-signal intensity-curves with different enhancement and wash-out characteristics are created depending on the low or high permeability of microvessels of each tissue. A study by van Rijswijk *et al.* [18] proposed a classification of time-signal intensity-curves of tissues into five different types as illustrated in Figure 2.8b.

Once the time-signal intensity-curves are acquired for each voxel of each time point, a generalized kinetic model [20] is used to relate the change in signal intensity to the contrast agent concentration in the tissue over time de noted as $C_t(t)$, such that:

$$\frac{dC_t(t)}{dt} = K^{\text{Trans}} \left(C_p(t) - C_t(t)/\nu_e\right) = K^{\text{Trans}}\, C_p(t) - \kappa_{ep}\, C_t(t), \tag{2.12}$$

where $C_p(t)$ is the arterial blood plasma concentration as a function of time, which needs to be measured for each patient by including, for example, a large vessel in the imaging field of view. Furthermore, K^{Trans} is the volume transfer constant between the blood plasma and the interstitial space per unit volume of tissue (min^{-1}), ν_e is the volume of interstitial space per unit volume of tissue, and κ_{ep} is the rate constant between the insterstitial space and blood plasma (min^{-1}) [21]. By solving Equation 2.12 for the time-signal intensity-curve of each voxel using numerical methods [22], quantitative parametric maps of K^{Trans}, ν_e and κ_{ep} can be obtained for the imaged tumour volume. Finally, maps of the initial area under the time-signal intensity-curves (IAUC$_x$) from the start of injection to a number x of seconds

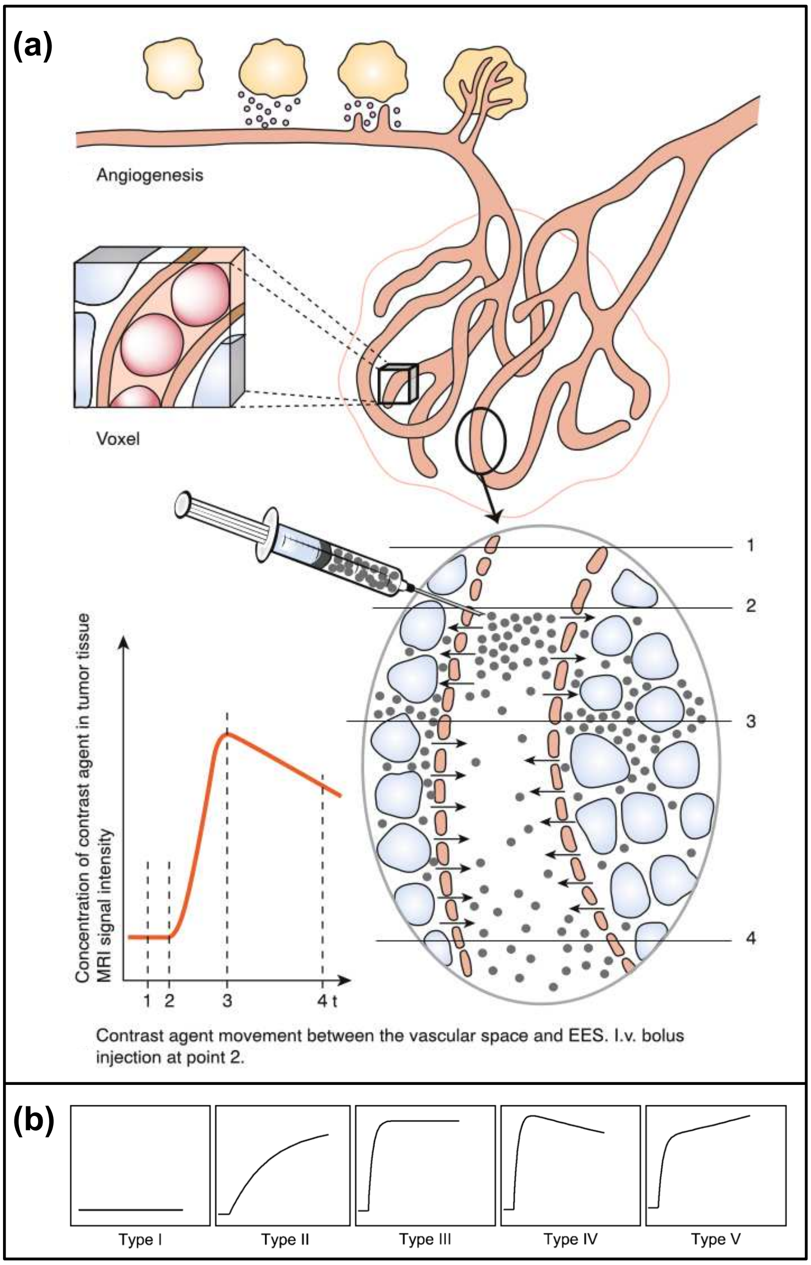

(a)
Angiogenesis
Voxel
1
2
3
4
Concentration of contrast agent in tumor tissue
MRI signal intensity
1 2 3 4 t
Contrast agent movement between the vascular space and EES. I.v. bolus injection at point 2.
(b)
Type I
Type II
Type III
Type IV
Type V

Figure 2.8 : Principles of DCE-MRI.
(a) Malignant cancers initiate the formation of new blood vessels that can grow into the tumour to supply it with oxygen an nutrients – a process called angiogenesis. In DCE-MRI acquisitions, multiple 3D sets of images are acquired over time after the injection of a paramagnetic contrast agent (e.g., gaolinium-based molecules) into the vein of a patient. The passage of the contrast agent (small grey circles in the figure) from the vascular space to the interstitial space of tumours (or extravascular-extracellular space: EES) has for effect to decrease the T_1 relaxation time of its surrounding, leading to an enhanced MR signal proportional to the contrast agent concentration in a given voxel. At first, the contrast agent accumulates in the interstitial space of tissues before it diffuses back into the vasculature from which it is excreted. The kinetics of such process gives rise to different time-signal intensity-curves for each voxel depending on the permeability of microvessels of each tissue. Reprinted with permission from [19]. © 2005 Springer. All rights reserved.
(b) Classification of DCE-MRI time-signal intensity-curves into 5 different tissue types as defined by van Rijswijk *et al.* [18]. I) no enhancement; II) gradual increase of enhancement; III) rapid initial enhancement followed by a plateau phase; IV) rapid initial enhancement followed by a washout phase; and V) rapid initial enhancement followed by sustained late enhancement. Reprinted with permission from [18]. © 2004 RSNA. All rights reserved.

post-injection can also be extracted from the DCE-MRI data [23]. For more information about DCE-MRI acquisition and parametric analysis, the reader is referred to references [16, 24, 25].

2.2 Image processing

Zwanenburg *et al.* [26] have very recently proposed a detailed sequence of image processing steps required prior to the extraction of radiomic features from medical images. Prior to this work, no clear consensus on the exact sequence of processing steps existed. The work of Zwanenburg *et al.* [26] could standardize and bring to consensus the manner in which radiomics features are extracted, and the reader is invited to consult the document currently published on *arXiv*. Nonetheless, it is pointed out in the document that the order of the sequence of operations could be interchanged for some processing steps. In this section, we briefly describe the major image processing steps that were used in this thesis in the context of the work of Zwanenburg *et al.* [26]: I) Spatial filtering; II) Tumour segmentation; II) Image interpolation; and IV) Image quantization. Our starting point is the stack of medical images acquired from a dedicated scanner, from which data conversion (e.g., SUV conversion in PET), imaging corrections (e.g., PVE corrections in PET [27], metal artifact suppresion in CT [28], non-uniformity corrections in MRI [29], etc.) and denoising operations may have been *a priori* applied.

2.2.1 Spatial filtering and image fusion

In radiomics analysis, image filters can be used to enhance different aspects of the image (e.g., edges, specific range of image frequencies, etc.) and reduce noise prior to radiomic feature computation. Laplacian, Gaussian, wavelet, Law's and Gabor filters are commonly used in the medical imaging community. The description of various spatial filters is however outside the scope of this work. In this section, we describe the theory behind the wavelet transform filter, since this filter is at the core of the methodology developed in Chapter 3 to enhance texture features via wavelet band-pass filtering, and to perform FDG-PET and MR image fusion. Due to space constraints, only a brief overview of the theory is provided, and the reader is referred to the comprehensive reviews by Strang & Nguyen [30] and Burrus *et al.* [31] for further details.

General theory of wavelet decomposition

The goal of wavelet analysis is to decompose a signal over a family of wavelets generated from a mother wavelet. Let the signal of interest be in space x and the mother wavelet of interest be $\psi(x)$. The mother wavelet is a squared-integrable function over all space with zero average. The class of expansion functions generated from the mother wavelet are defined as:

$$\psi_k^j(x) = 2^{j/2}\,\psi(2^j x - k). \tag{2.13}$$

This implies that all wavelets $\psi_k^j(x)$ are dilated (or scaled) and translated versions of $\psi(x)$ as defined by the integers j and k, respectively. The goal of wavelet expansion is to generate a set of functions $\psi_k^j(x)$ such that any signal in the space of squared-integrable functions $L^2(\mathbb{R})$ can be represented by the series:

$$f(x) = \sum_k \sum_j w_j^k\, 2^{j/2}\,\psi(2^j x - k). \tag{2.14}$$

In Equation 2.14, the set of expansion coefficients w_j^k is called the discrete wavelet transform (DWT) of $f(x)$. If the expansion is unique, the set of functions $\psi_k^j(x)$ is called a *basis* for the class of functions that can be so described. The power of such a basis is that it can simultaneously express a signal at different scales and spatial locations. However, in wavelet theory, the formulation of such multiresolution analysis is made in terms of two closely related basis functions. In addition to the mother wavelet, we introduce the scaling function $\phi(x)$ that can be expressed in terms of a weighted sum of translated versions of $\phi(2x)$ such that:

$$\phi(x) = \sqrt{2}\sum_{n\in\mathbb{Z}} l(n)\,\phi(2x - n). \tag{2.15}$$

Likewise, the mother wavelet $\psi(x)$ is expressed as:

$$\psi(x) = \sqrt{2}\sum_{n\in\mathbb{Z}} h(n)\,\psi(2x - n). \tag{2.16}$$

Equation 2.15 is governed by "low-pass" coefficients $l(n)$ of the wavelet expansion and Equation 2.16 is governed by "high-pass" coefficients $h(n)$, and the relation between these coefficients is $h(n) = (-1)^n\, l(1-n)$. Figure 2.9 displays an example of the scaling and wavelet functions of a specific class of wavelets known as "*sym8*".

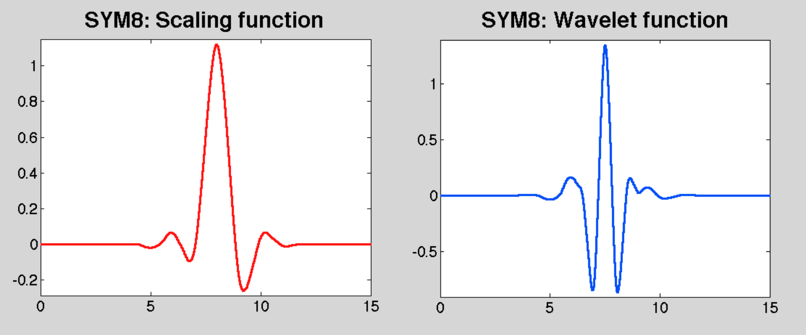


Figure 2.9: Example of a wavelet basis function. The scaling and wavelet functions *sym8* are illustrated.

With such double-basis representation, the decomposition of a signal into a finite number of levels J becomes:

$$f(x) = \sum_{k} a_k^j \, 2^{J/2} \, \phi(2^J x - k) + \sum_{k} \sum_{j=1}^{J} d_k^j \, 2^{j/2} \, \psi(2^j x - k) \tag{2.17}$$

Equation 2.17 implies that the a_k^J coefficients are used to represent the approximation of signal at the lowest level (or scale) J with the scaling function $\phi(x)$. As such, $\phi(x)$ is used to represent the coarse details of the signal, or its low-frequency components. The rest of the decomposition coefficients d_k^j are used to represent the fine details of the signal, or its high-frequency components. These coefficients are obtained at all scales using the family of functions $\psi_k^j(x)$. Finally, all coefficients at scale j can be expressed in terms of the coefficients of the previous scale using the following recursive equations:

$$\begin{aligned} a_k^j &= \sum_{n \in \mathbb{Z}} a_k^{j-1} \, l(n - 2k) \\ d_k^j &= \sum_{n \in \mathbb{Z}} a_k^{j-1} \, h(n - 2k), \quad \text{for } j = 1, 2, \ldots, J \end{aligned} \tag{2.18}$$

2D and 3D discrete wavelet transform

Essentially, the DWT up to level (or scale) J is performed through a cascade-tree of low-pass and high-pass filters followed by downsampling by a factor of 2. The wavelet coefficients a_k^j and d_k^j are obtained by the convolution over space of the proper scaling and wavelet functions defined at each level j. Practically speaking, for a 2D image, performing one level of a 2D wavelet decomposition consists of filtering and downsampling an image $I(x, y)$ both horizontally and vertically with the 1D low-pass filter (*L*) ϕ and the 1D high-pass filter (H) ψ. As a result, the wavelet coefficients of four different subbands are produced: *LL*, *LH*, *HL*, *HH* (Figure 2.10a). Every subband now has half the initial size of I in both the x and y directions. In order to obtain the 2D discrete wavelet decomposition at higher levels (level 2 shown in Figure 2.10c), the same process is performed on the *LL* subband generated from the previous decomposition level (level 1 shown in Figure 2.10b) and is repeated up to the desired level of decomposition. The generalization of the wavelet theory to 3 dimensions is straightforward: 3D wavelet decomposition consists of filtering and downsampling an imaging volume $V(x, y, z)$ in the x, y and z directions with the 1D low-pass filter (*L*) ϕ and the 1D high-pass filter (*H*) ψ. The wavelet coefficients of eight different subbands are then produced: *LLL*, *LHL*, *LHH*, *HLL*, *HHL*, *HLH* and *HHH*. In radiomics analysis, different research groups now compute complete sets of features from the wavelet coefficient intensities of all the different subbands; an example is the work of Aerts *et al.* [32]. In this thesis, new sets of radiomic features were not computed from the different subbands. Instead, as described in Chapter 3, we notably applied different weights to the wavelet coefficients of different subbands before wavelet reconstruction in order to obtain a new "filtered volume" (wavelet band-pass filtering).

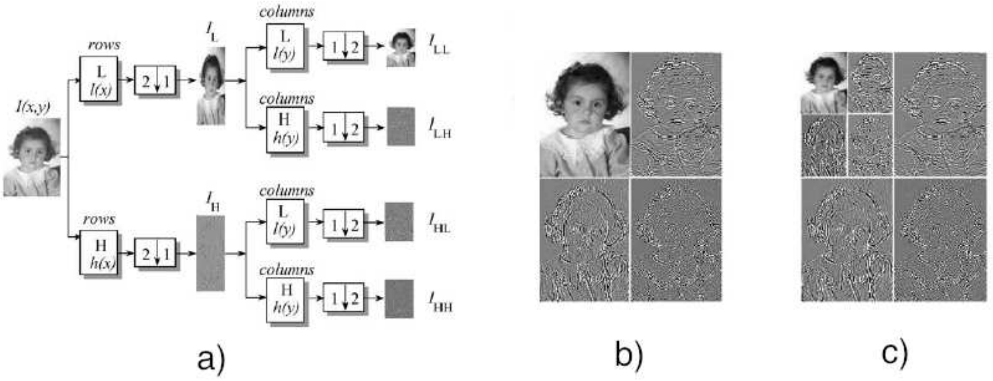


Figure 2.10: Principles of 2D discrete wavelet decomposition. (a) Wavelet decomposition process. For a 2D image, low-pass (*L*) and high-pass (*H*) filters are applied in both the x and y directions, thereby creating 4 wavelet subbands: *LL*, *LH*, *HL* and *HH*. (b) Level 1 of a 2D wavelet decomposition. (c) Level 2 of a 2D wavelet decomposition. Reprinted with permission from [33]. © 2004 Pattern Recognition Society. All rights reserved.

Finally, the procedure used to reconstruct the original signal from the wavelet coefficients is known as the inverse discrete wavelet transform (IDWT), which is simply the reverse process of the DWT. In practice, each subband is first upsampled by a factor of 2 by inserting zeros in-between the wavelet coefficients. Next, each sub-band is convolved with the appropriate reconstruction filters. For example, the *HL* sub-band is first upsampled and convolved horizontally with the 1D high-pass (wavelet) reconstruction filter. Then, it is upsampled and convolved vertically with the 1D low-pass (scaling) reconstruction filter. The reconstruction filters are the original scaling and wavelet filters flipped from left to right about their central position. Once this process has been applied to the four subbands (in 2D analyses), the results are added together to obtain a reconstructed image.

Principles of wavelet image fusion

For the purpose of this thesis (Chapter 3), image fusion can be described as the process of combining information from two different images into a single composite image that is more informative for texture analysis. The concept of image fusion using the DWT was proposed by Li *et al.* [34] and is described in details in the work of Pajares & Manuel de la Cruz [33]. The general framework of wavelet image fusion is depicted in Figure 2.11.

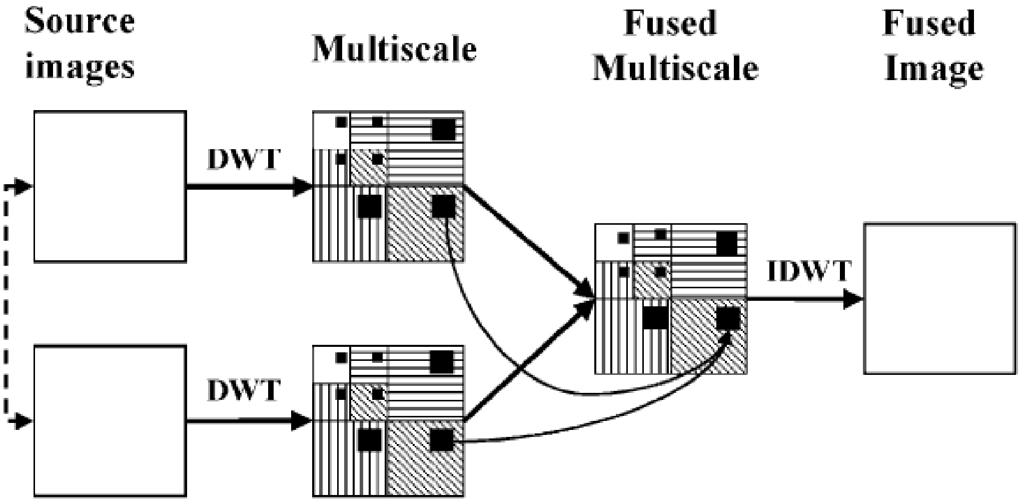


Figure 2.11: Principles of wavelet image fusion. The corresponding wavelet coefficients of two different images are first obtained using the discrete wavelet transform (DWT) and are thereafter merged in a way as to obtain the image characteristics sought. The inverse discrete wavelet transform (IDWT) is then applied on the set of merged coefficients to obtain a new fused image. Reprinted with permission from [33]. © 2004 Pattern Recognition Society. All rights reserved.

Let us assume that the two images to be fused are co-registered (if misregistration occurs, artefacts will be present in the fused image) and have the same resolution. For the fusion scheme presented in Figure 2.11, it means that resampling and registration strategies have to be applied prior to fusion. Then, the fusion process starts with the application of the DWT to both images with a given scaling function of choice. The decomposition can go up to an arbitrary number of levels (2 decomposition levels are shown in Figure 2.11). Afterwards, the respective wavelet coefficients of the two images are merged together. In other words, the *LL* coefficients of image 1 are merged with the *LL* coefficients of image 2, the *HL* coefficients of image 1 are merged with the *HL* coefficients of image 2, and this process is repeated for all subbands. Subsequently to the latter step, only one set of fused wavelet coefficients exists. Finally, the IDWT is applied to the fused wavelet coefficients in order to reconstruct the fused image. Fundamentally, the key step in DWT image fusion is based on how the wavelet coefficients of the two different images are combined. The goal is to merge the wavelet coefficients in an appropriate way in order to obtain the image characteristics sought. Finally, the whole wavelet fusion process described in this section for 2D images can easily be generalized to the fusion of 3D volumes.

2.2.2 Tumour segmentation

The computation of radiomic (including texture) features currently relies on the accurate definition of a region-of-interest (ROI), i.e., segmentation, from which the features are extracted. In the context of this thesis, the ROI represents the gross tumour volume (GTV) used for radiotherapy treatment planning as manually defined on a slice-by-slice basis by expert radiation oncologists using dedicated contouring software.

From a given imaging volume $V(x,y,z)$, tumour segmentation leads to the creation of a "ROI map" (or mask) denoted as $R(x,y,z)$, for which every voxel at position (x,y,z) in R is defined as:

$$R(x,y,z) = \begin{cases} 1 & \text{if } V(x,y,z) \text{ in ROI,} \\ 0 & \text{otherwise.} \end{cases}$$

An example of the process of image segmentation is depicted in Figure 2.12. Note that the schematic image slices in that figure are not composed of isotropic voxels, i.e., they are not composed of voxels with the same dimensions in the three directions of space. This is generally the case for out-of-scanner images in PET, CT and MR imaging.

In radiotherapy, different ROI structures representing target contours and organs at risks are saved in "DICOM RTstruct" format. After segmentation, image registration [35] may occur in order to propagate tumour contours to different imaging frame of references.

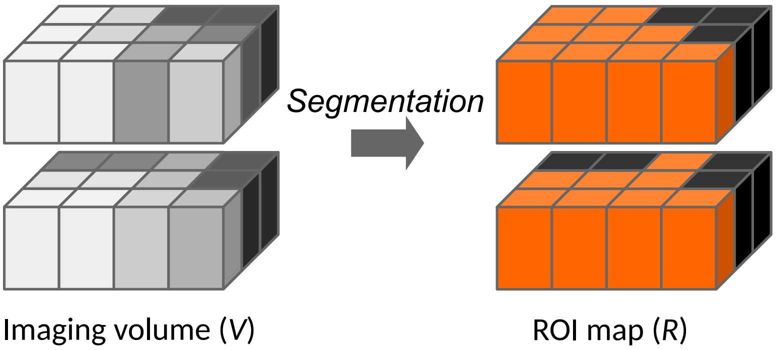


Figure 2.12: Schematic illustration of the image segmentation process. The segmentation of an imaging volume leads to the creation of a "ROI map". In this arbitrary example, the segmentation of the imaging volume on the left is performed by excluding voxels with $> 50\,\%$ grey, thereby leading to orange voxels included and black voxels excluded from the ROI map on the right. Reprinted from [26]. © 2016-2017 IBSI. Creative Commons Attribution 4.0 International License.

2.2.3 Image interpolation

In order to obtain rotationally invariant texture features, interpolation to isotropic voxel size (i.e., same voxel dimension in three directions of space) is imperative. Maintaining isotropic voxel dimensions across different patients and institutions is important for reproducibility of radiomics analyses. Furthermore, the optimization of such extraction parameter can significantly enhance the predictive properties of textures. Generally, voxel size is not isotropic for medical images. The radiomics user may then decide to set the isotropic voxel size to a desired "*scale*" such as the in-plane resolution. In that case, an example imaging volume $V(x, y, z)$ (spatially filtered or not) with voxel size of $2 \times 2 \times 3$ mm^3 would be isotropically resampled to a voxel size of $2 \times 2 \times 2$ mm^3. Likewise, the ROI map $R(x, y, z)$ would also be interpolated to the dimensions of the resampled imaging volume. Here, the interpolated volume is denoted as $V_{\text{I}}(x, y, z)$ and the interpolated ROI map as $R_{\text{I}}(x, y, z)$. Figure 2.13 illustrates this process.

A number of algorithms are available in most software for interpolation, such as, for example, *nearest neighbour* and *cubic interpolation*. One strategy among others could be to interpolate the ROI map using the *nearest neighbour* algorithm to conserve 0's and 1's in the interpolated ROI, and the imaging volume using *cubic interpolation* to produce smooth interpolated images.

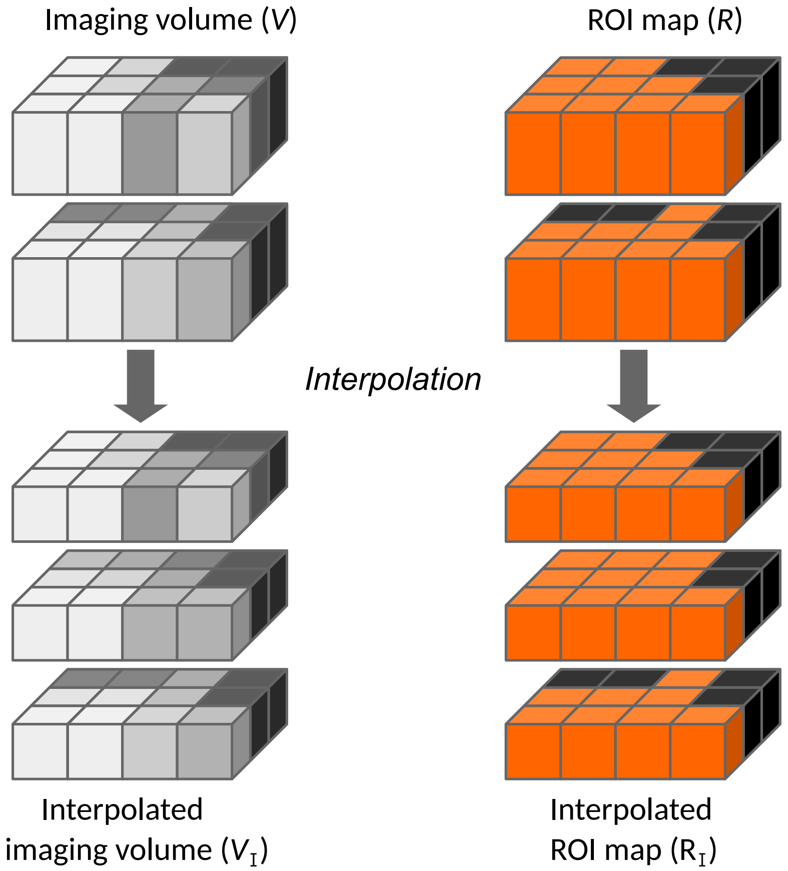

Imaging volume (*V*)
ROI map (*R*)
Interpolation
Interpolated imaging volume (V_I)
Interpolated ROI map (R_I)

Figure 2.13: Schematic illustration of the image interpolation process. The example imaging volume with voxel size of 2 × 2 × 3 mm^3 and associated ROI map are isotropically resampled to a voxel size of 2 × 2 × 2 mm^3, leading to an interpolated imaging volume and an interpolated ROI map. In the process, an extra image slice is created. Reprinted from [26]. © 2016-2017 IBSI. Creative Commons Attribution 4.0 International License.

Following interpolation, different outlier correction methods could subsequently be used to improve the quality of the images prior to texture analysis. One such method was proposed by Collewet *et al.* [36] for making MRI texture measurements more reliable. The method consists of excluding voxels within the tumour region (i.e., the ROI map) with intensities outside the range $\mu \pm 3\sigma$, where μ and σ are the mean and the standard deviation of the imaging voxels part of the tumour region, respectively.

2.2.4 Image quantization

Image quantization is a process used prior to texture computation for reducing the intensity range of the ROI of a medical imaging volume into a discretised number of gray-level bins (N_g). Notably, image quantization helps in reducing the noise dependence in the calculation of textures. In this section, the ROI extraction and quantization processes are detailed.

ROI extraction process

The computation of texture features is performed for the voxels included in a given ROI of the imaging volume, and so does the preliminary quantization step. Prior to image quantization, the ROI of the imaging volume is thus first isolated from the surrounding voxels. The ROI map is used to keep only the voxels of the imaging volume contained within the ROI, and the rest of the voxels are excluded from texture analysis. Excluded voxels are commonly replaced by a placeholder value such as a *NaN*, leading to the creation of a “ROI imaging volume” denoted as $V_\text{R}(x,y,z)$. Figure 2.14 illustrates the ROI extraction process. The ROI imaging volume $V_\text{R}(x,y,z)$ is also defined in terms of the interpolated volume $V_\text{I}(x,y,z)$ and the interpolated ROI map $R_\text{I}(x,y,z)$ from the current image processing workflow such that:

$$V_\text{R}(x,y,z) = \begin{cases} V_\text{I}(x,y,z) & \text{if } R_\text{I}(x,y,z) = 1 \\ NaN & \text{otherwise.} \end{cases}$$

Quantization process

Following ROI extraction, image quantization (or discretisation) can proceed. Let $\mathbf{v}_R$ be the column vector of all imaging intensity values $V_\text{R}(x,y,z)$ for which $V_\text{R}(x,y,z) \neq NaN \, \forall \, (x,y,z)$; i.e., all the imaging intensity values from V_R that are not set to NaN. The quantization process maps $\mathbf{v}_R$ to a finite set

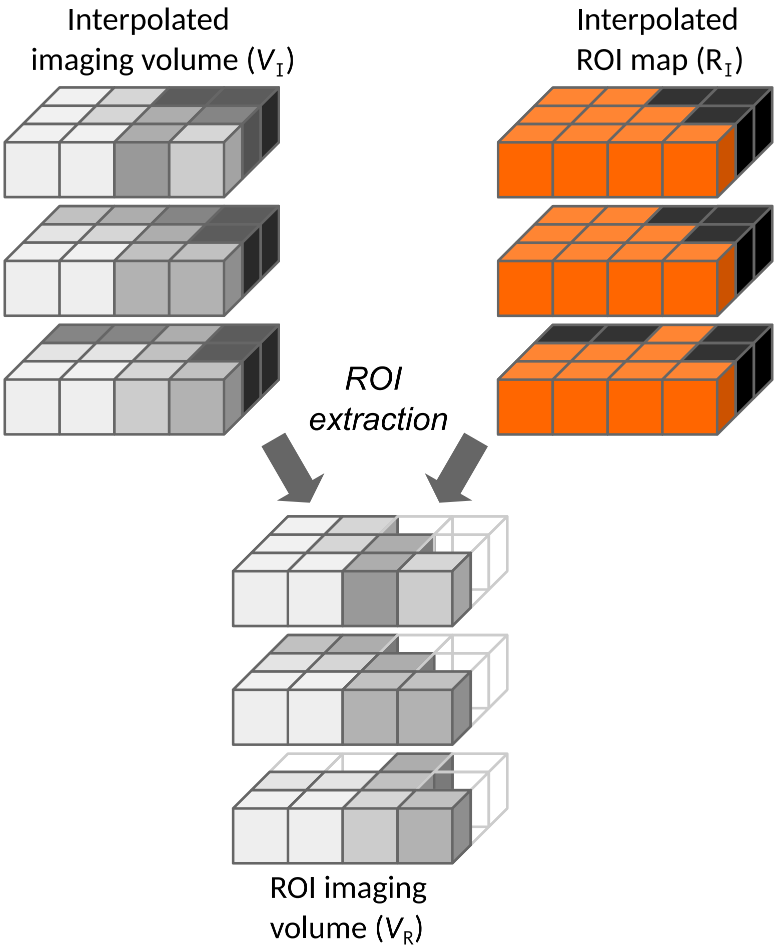

Interpolated imaging volume (V_I)
Interpolated ROI map (R_I)
ROI extraction
ROI imaging volume (V_R)

Figure 2.14: Schematic illustration of the image ROI extraction process. The interpolated ROI map is used to keep only the voxels of the interpolated imaging volume contained within the interpolated ROI, and the rest of the voxels are excluded. Excluded voxels are represented by empty voxels in the figure and are commonly replaced by a placeholder value such a *NaN*, leading the creation of a "ROI imaging volume". Reprinted from [26]. © 2016-2017 IBSI. Creative Commons Attribution 4.0 International License.

of *reconstruction levels* $\mathbf{r}_{N_g} = \{r_g \in [\min(\mathbf{v}_R), \max(\mathbf{v}_R)] : g = 1, 2, \ldots, N_g\}$ by defining a set of decision levels $\mathbf{t}_{N_g} = \{t_g \in [\min(\mathbf{v}_R), \max(\mathbf{v}_R)] : g = 1, 2, \ldots, N_g + 1\}$, with $t_1 = \min(\mathbf{v}_R)$ and $t_{N_g+1} = \max(\mathbf{v}_R)$. The imaging intensity values of the "quantized ROI imaging volume" $V_{\mathrm{Q}}(x, y, z)$ are thereafter set to $g = \{1, 2, \ldots, N_g\}$, such that $V_{\mathrm{Q}}(x, y, z) = g \, \forall \, V_{\mathrm{R}}(x, y, z) \in [t_g, t_{g+1})$. This process is illustrated in Figure 2.15. All quantization algorithms attempt to resolve, for a given number of gray levels N_g, the reconstruction and decision levels of an input vector. Three different types of quantization algorithms are described below.

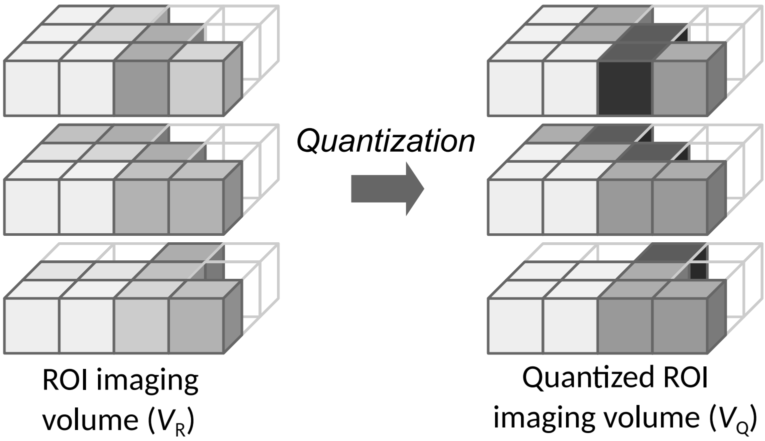

Quantization
ROI imaging volume (V_{R})
Quantized ROI imaging volume (V_{Q})

Figure 2.15: Schematic illustration of the image quantization process. In this example figure, the gray levels from the ROI imaging volume were quantized into $N_g = 3$ bins to create a "quantized ROI imaging volume". Reprinted from [26]. © 2016-2017 IBSI. Creative Commons Attribution 4.0 International License.

Uniform quantization. The simplest and most commonly used quantization scheme in current radiomic studies in the literature is called the "uniform quantizer". It uniformly divides the range of intensities of $\mathbf{v}_R$ into N_g gray levels such that the transition and reconstruction levels become:

$$\begin{aligned} t_g &= \min(\mathbf{v}_R) + \frac{\max(\mathbf{v}_R) - \min(\mathbf{v}_R)}{N_g}, \quad \text{for } g = 1, 2, \ldots, N_g + 1 \\ r_g &= \frac{t_g + t_{g+1}}{2}, \quad \text{for } g = 1, 2, \ldots, N_g \end{aligned} \tag{2.19}$$

Equal-probability quantization. In their original work on GLCM-based texture features, Haralick *et al.* [37] proposed that the quantization of images prior to computation of the GLCM should be done using an equal-probability quantization scheme in order for the extracted textures to be invariant under monotonic gray-tone transformations. This quantization scheme attempts to define decision thresholds in an imaging volume with the goal that the number of voxels with reconstruction level r_g be the same for all gray levels g (i.e., for all quantized bins). Hence, each gray level approximately has an equal probabilty of occurrence in the quantized imaging volume. Similarly to the uniform quantizer, the reconstruction levels are taken as the average of two consecutive decision levels. In this thesis, an equal-probability quantization algorithm similar to the one described by Haralick *et al.* [37] was implemented in an in-house MATLAB® algorithm using the function *histeq.m* to ensure a monotonic transformation of the intensity histograms.

Lloyd-Max quantization. In 1982, Lloyd [38] formulated the concept of Max [39] (of quantizing an input signal to achieve minimal distortion) into a coherent quantization theory now known as the "Lloyd-Max" quantization algorithm. Lloyd enounced his optimization criterion as the minimization of "average quantization noise power". Essentially, this scheme optimally minimizes the mean square quantization error of the output. By taking the formulation of Jain [40], let X be the input data and $Q(X)$ the output of quantization. For N_g gray levels, the mean-squared error ϵ is:

$$\epsilon = E\Big[\big(X - Q(X)\big)^2\Big] = \int_{t_1}^{t_{N_g+1}} \big(X - Q(X)\big)^2\, p_X(X)\, dX, \tag{2.20}$$

where $p_X(X)$ is the amplitude probability density of the input volume data X. The necessary conditions for minimizing ϵ are obtained by differentiating Equation 2.20 with respect to the decision levels t_g and the reconstruction levels r_g. By equating to 0 and from the fact that $t_{g-1} \leq t_g$, we obtain:

$$\begin{aligned} t_g &= \frac{r_g + r_{g-1}}{2}, \quad \text{for } g = 1, 2, \ldots, N_g + 1 \\ r_g = E\Big[X | X \in [t_g, t_{g+1})\Big] &= \frac{\int_{t_g}^{t_{g+1}} X\, p_X(X)\, dX}{\int_{t_g}^{t_{g+1}} p_X(X)\, dX}, \quad \text{for } g = 1, 2, \ldots, N_g \end{aligned} \tag{2.21}$$

Practically speaking, the two parts of Equation 2.21 have to be solved simultaneously (given boundary conditions t_1 and t_{N_g+1}) using an iterative scheme. In this work, this procedure was performed using the function *lloyds.m* of MATLAB®.

2.2.5 Texture extraction parameters

There are multiple image processing steps leading to texture analysis:

- Medical images can be spatially filtered using, for example, wavelet analysis in order to enhance different image frequency components. Furthermore, different imaging modalities (e.g., FDG-PET and MRI) can be fused into a single set of images to create composite textures with better predictive properties.
- Different tumour volumes encompassing or not different anatomical or biological targets can be segmented in order to verify their benefit for tumour aggressiveness assessment via texture analysis.
- Images can be interpolated to different isotropic voxel sizes.
- Different quantization algorithms using different numbers of gray levels can be used prior to texture analysis.

Overall, all these different steps open the door to the optimization of different texture extraction parameters for each different textures depending on the subsequent clinical endgoal, a process we name "texture optimization". An example of how different extraction parameters affect the resulting processed images prior to texture analysis is shown in Figure 2.16. It can clearly be seen that a single FDG-PET image processed with different extraction parameters yields processed images with clear differences in textural characteristics. In this thesis, it will be shown that the optimization of texture extraction parameters for a given clinical task can significantly increase the predictive properties of textures.

2.3 Texture analysis

In radiomics analysis, textures are a central type of features that can be extracted from a tumour ROI. Other types of features include morphological and histogram-based features (Appendix A.1 and Appendix A.2, respectively). However, textures remain the core of radiomic feature computation given their higher-order characterization of spatial patterns in imaging volumes. In this thesis, texture features from four major categories were extracted: I) Gray-Level Co-occurence Matrix (GLCM) features; II) Gray-Level Run-Length Matrix (GLRLM) features; II) Gray-Level Size Zone Matrix (GLSZM) features; and IV) Neighborhood Gray-Tone Difference Matrix

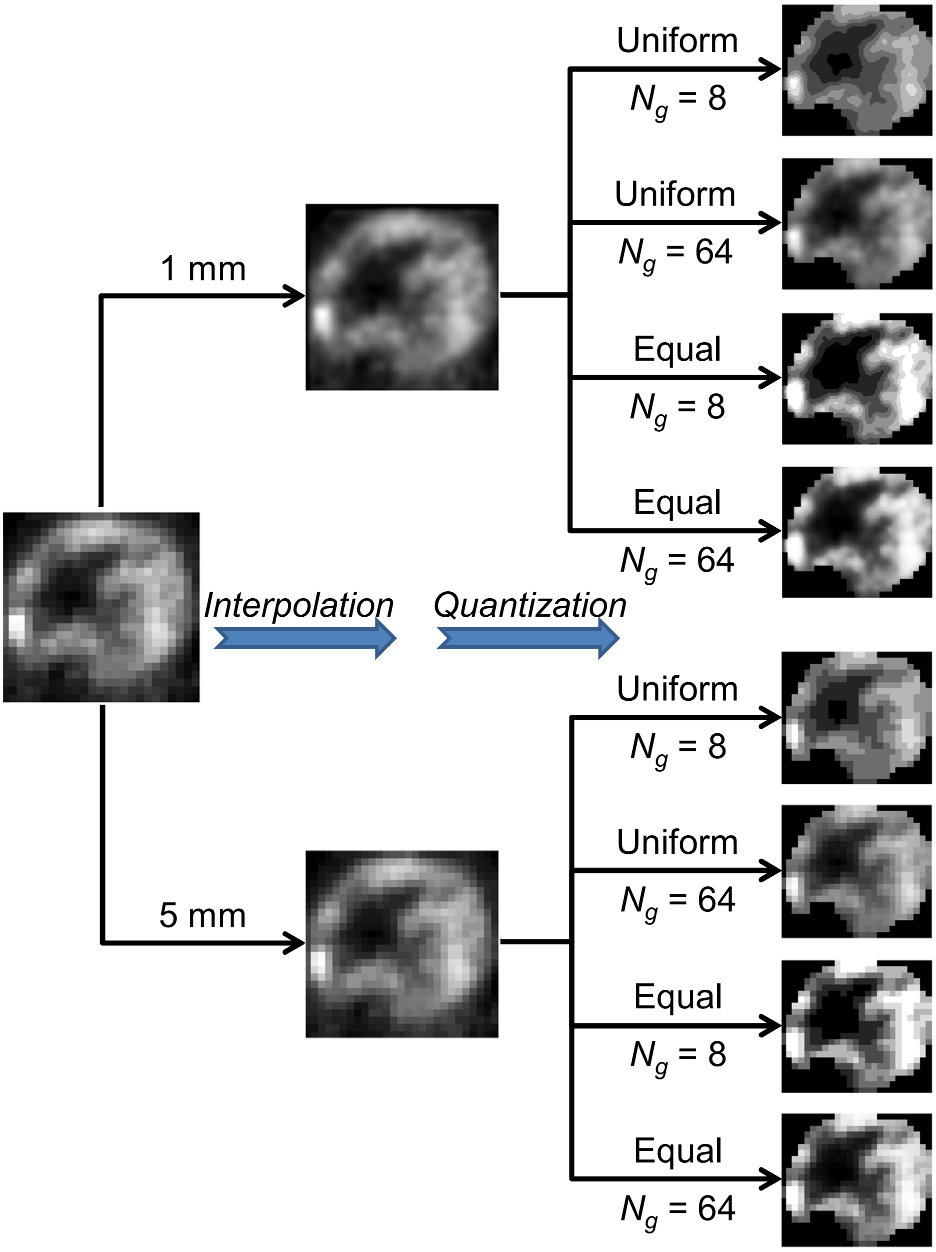


Figure 2.16: Impact of different texture extraction parameters on processed images. The example image on the left is a FDG-PET image of a soft-tissue sarcoma with voxel size of $5.47 \times 5.47 \times 3.27$ mm^3. This image was then interpolated to isotropic voxel sizes of 1 mm and 5 mm using cubic interpolation. Finally, the resulting interpolated images were quantized using uniform ("Uniform") and equal-probabilty ("Equal") quantization algorithms with 8 and 64 numbers of gray levels (N_g). All the different processed images on the right can subsequently be used for texture analysis.

(NGTDM) features. The first and crucial step towards the computation of the different texture features from these four categories is to calculate a matrix $\mathbf{P}$ summarizing the neighborhood properties of interest (differently for each category). Thereafter, different mathematical operations can be applied to the different matrices to obtain the final texture features f.

Texture features were originally designed to assess surface texture in 2D images. One strategy to obtain a global assessment of a 3D tumour ROI using 2D texture analysis could be to average the texture features obtained in each slice of an imaging volume stack. However, a better strategy (as performed in this thesis) would be to generalize the original 2D computation of texture matrices [37, 41–45] to 3D. Another major difference point in this thesis compared to the original definitions of some textures is the concept of directionality. Originally, some textures were defined to be computed in a specific direction of the image space (e.g., the horizontal direction). Features for a given direction could then be used on their own for a given application, or corresponding features computed for all directions could be averaged together to globally characterize an image or imaging volume. In the context of intratumoural heterogeneity quantification, the averaging of texture features from different directions however consists of taking an average of limited texture measurements as we have previously shown [46]. Therefore, in order to obtain the best global assessment of 3D tumour regions, all texture matrices in this thesis were computed only once per quantized ROI imaging volume V_{Q} for the whole 3D space as illustrated in Figure 2.17, by considering that the direct neighborhood of each imaging volume voxel consists of its 26 directly connected neighbors. Hence, the following 13 directions of 3D space were simultaneously considered in the computation of textures matrices: (0,0,1), (0,1,0), (1,0,0), (0,1,1), (0,1,-1), (1,0,1), (1,0,-1), (1,1,0), (1,-1,0), (1,1,1), (1,1,-1), (1,-1,1) and (1,-1,-1). Novel discretization length difference corrections were also defined in this work. In this section, theoretical examples of the calculation of texture matrices (GLCM, GLRLM, GLSZM, NGTDM) using discretization length difference corrections will be provided in 2D for the sake of simplicity. The subsequent 3D generalization and mathematical operations for texture feature computations are provided in the list of 3D radiomic features in Appendix A.

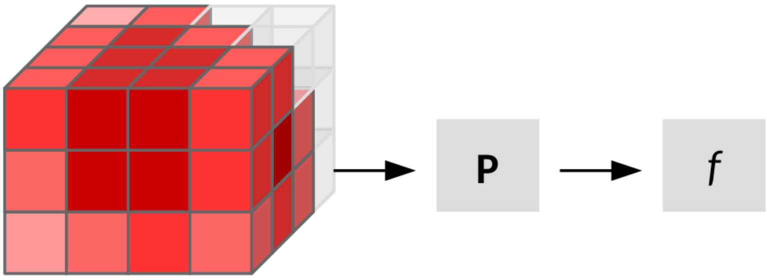


Figure 2.17: Full 3D approach to calculate texture matrices. From a quantized ROI imaging volume V_{Q}, only one texture matrix $\mathbf{P}$ is calculated per texture category (GLCM, GLRLM, GLSZM, NGTDM) by simultaneously taking into account the 26-connected neighbors of each voxel from the 13 directions of 3D space. Placeholder values such as *NaN* (represented by empty voxels in the figure) are always excluded from texture analysis. Single sets of features f are therafter calculated from the texture matrices of each category. Reprinted from [26]. © 2016-2017 IBSI. Creative Commons Attribution 4.0 International License.

2.3.1 Gray-Level Co-occurence Matrix

In 1973, Haralick *et al.* [37] proposed the concept of texture analysis from the GLCM. In their original pioneering work, the investigators took into account the statistical nature of textures, which is based on the assumption that texture information is contained in the overall spatial relationship that the gray levels have to one another.

Let $\mathbf{P}$ define the GLCM of a quantized ROI imaging volume $V_{\mathrm{Q}}(x,y,z)$ with isotropic voxel size. Each entry $P(i,j)$ of $\mathbf{P}$ represents the number of times voxels of gray level i are neighbours with voxels of gray level j in V. The GLCM is thus a symmetric matrix of size $N_g \times N_g$, where N_g represents the pre-defined number of quantized gray levels set in $V_{\mathrm{Q}}(x,y,z)$. Now, let us consider the 2D test image in Table 2.1a. The resulting GLCM of that image is filled in by examining the neighborhood of every pixel in the image, to then increment the GLCM accordingly. For example, the center pixel with gray level 2 is five times neighbor with gray level 1, one time with gray level 2 and two times with gray level 3. Traditionally in the radiomics community, this would lead to the incrementation of $P(2,1)$ by 5, of $P(2,2)$ by 1 and of $P(2,3)$ by 2, respectively. However, this latter incrementation scheme does not take into account the discretization length differences between the different pixels. In our work, we apply corrections for these differences that are valid for both 2D or 3D cases as follows: neighbours at a distance of $\sqrt{3}$ voxels around a center voxel increment the GLCM by a value of $\sqrt{3}$, neighbours at a distance of $\sqrt{2}$ voxels around a center voxel increment the GLCM by a value of $\sqrt{2}$, and neighbours at a distance of 1 voxel around a center voxel increment the GLCM by a value of 1. As shown in Table 2.1b-e, this discretization length difference correction scheme is best viewed when different GLCMs are separately computed in the (1,0), (1,1), (0,1) and (-1,1) directions, respectively. These directions correspond to the 0°, 45°, 90° and 135° directions, respectively. The final step for this example test image is to sum up (or merge) all the GLCMs obtained from all directions (Table 2.1b-e) into the final GLCM of Table 2.1f, the one and only GLCM from which the features described in Appendix A.3 would be computed. In 3D, the 13 directions of 3D space is examined with 26-voxel connectivity.

Table 2.1: GLCM computation example.

1	1	1
1	2	2
3	3	1

(a) 2D image

$$1\times\begin{vmatrix} 4 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix}$$

(b) GLCM – 0°

$$\sqrt{2}\times\begin{vmatrix} 2 & 1 & 0 \\ 1 & 0 & 2 \\ 0 & 2 & 0 \end{vmatrix}$$

(c) GLCM – 45°

$$1\times\begin{vmatrix} 2 & 3 & 1 \\ 3 & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix}$$

(d) GLCM – 90°

$$\sqrt{2}\times\begin{vmatrix} 0 & 3 & 1 \\ 3 & 0 & 0 \\ 1 & 0 & 0 \end{vmatrix}$$

(e) GLCM – 135°

$$\left|\begin{array}{cccc} N_g \times N_g & \rightarrow & i & \rightarrow \\ \downarrow & 8.83 & 9.66 & 3.41 \\ j & 9.66 & 2.00 & 3.83 \\ \downarrow & 3.41 & 3.83 & 2.00 \end{array}\right|$$

(f) GLCM – Merged

2.3.2 Gray-Level Run-Length Matrix

In 1975, Galloway [41] proposed the concept of texture analysis from the GLRLM. The author defined 5 features that could be extracted from a GLRLM. Then, two other features were later defined by Chu *et al.* [43] in 1990, four by Dasarathy & Holder [44] in 1991, and two by Thibault *et al.* [45] in 2009. Essentially, the GLRLM quantifies the frequency of 1D runs of voxels with identical gray levels in a given imaging volume.

Let $\mathbf{P}$ define the GLRLM of a quantized ROI imaging volume $V_{\mathrm{Q}}(x, y, z)$ with isotropic voxel size. Each entry $P(i, j)$ of $\mathbf{P}$ represents the number of runs of gray level i and of length j in $V_{\mathrm{Q}}(x, y, z)$. A run is a 1D line of connected voxels with an identical gray level. The GLRLM is a matrix of size

$N_g \times L_r$, where N_g represents the pre-defined number of quantized gray levels set in $V_{\mathrm{Q}}(x, y, z)$, and L_r the length of the longest run (of any gray level). Now, let us consider the 2D test image in Table 2.2a. The resulting GLRLM of that image is filled in by counting all the possible runs of connnected pixels with identical gray levels for a given direction. The particular directional GLRLM of interest is then incremented accordingly. For example, let us consider runs with a gray level of 1 in the (1,1) – i.e., 45° – direction in the test image. It can be seen that there exists three runs of gray level 1 with length 1, and one run of gray level 1 with length 2. Traditionally in the radiomics community, this would lead to the incrementation of $P(1, 1)$ by 3 and of $P(1, 2)$ by 1, respectively. However, this latter incrementation scheme does not take into account the discretization length differences between the different pixels. In our work, we apply corrections for these differences that are valid for both 2D or 3D cases as follows: runs constructed from voxels separated by a distance of $\sqrt{3}$ increment the GLRLM by a value of $\sqrt{3}$, runs constructed from voxels separated by a distance of $\sqrt{2}$ increment the GLRLM by a value of $\sqrt{2}$, and runs constructed from voxels separated by a distance of 1 increment the GLRLM by a value of 1. Table 2.2b-e presents the resulting GLRLMs of the test image using discretization length corrections for the 0°, 45°, 90° and 135° directions, respectively. The final step for this example test image is to sum up (or merge) all the GLRLMs obtained from all directions (Table 2.2b-e) into the final GLRLM of Table 2.2f, the one and only GLRLM from which the features described in Appendix A.4 would be computed. In 3D, the 13 directions of 3D space is examined with 26-voxel connectivity.

2.3.3 Gray-Level Size Zone Matrix

In 2009, [45] generalized the 1D GLRLM concept to 2D and 3D to thereby create the GLSZM texture analysis method. In addition, the authors created two new texture features from the original set defined by Galloway [41], Chu *et al.* [43], and Dasarathy & Holder [44]. Essentially, the GLRLM quantifies the frequency of 2D or 3D zones of voxels with identical gray levels in a given image or imaging volume.

Let $\mathbf{P}$ define the GLSZM of a quantized ROI imaging volume $V_{\mathrm{Q}}(x, y, z)$ with isotropic voxel size. Each entry $P(i, j)$ of $\mathbf{P}$ represents the number of zones of gray level i and of size j in $V_{\mathrm{Q}}(x, y, z)$. A zone is a 2D or 3D region of connected voxels with an identical gray level. The GLSZM is a matrix of size $N_g \times L_z$, where N_g represents the pre-defined number of quantized

Table 2.2: GLRLM computation example.

1	1	1
1	2	2
3	3	1

(a) 2D image

$$1\times\begin{vmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{vmatrix}$$

(b) GLRLM – 0°

$$\sqrt{2}\times\begin{vmatrix} 3 & 1 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{vmatrix}$$

(c) GLRLM – 45°

$$1\times\begin{vmatrix} 3 & 1 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{vmatrix}$$

(d) GLRLM – 90°

$$\sqrt{2}\times\begin{vmatrix} 5 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{vmatrix}$$

(e) GLRLM – 135°

$$\left|\begin{array}{cccc} N_g \times L_r & \rightarrow & j & \rightarrow \\ \downarrow & 16.3 & 2.41 & 1.00 \\ i & 7.66 & 1.00 & 0.00 \\ \downarrow & 7.66 & 1.00 & 0.00 \end{array}\right|$$

(f) GLRLM – Merged

gray levels set in $V_{\mathrm{Q}}(x, y, z)$, and L_z the size of the largest zone (of any gray level). Connected zones of identical gray levels in a 2D image are identified with 8-voxel connectivity, and with 26-voxel connectivity for a 3D imaging volume. Now, let us consider the 2D test image in Table 2.3a. The resulting GLSZM of that image is filled in by counting all the possible zones of connected voxels with identical gray levels. The GLSZM is then incremented accordingly. For example, let us consider zones of connected voxels with a gray level of 1 in the test image. It can be seen that there exists one zone of gray level 1 with size 1, and one zone of gray level 1 with size 4. This leads to the incrementation of $P(1, 1)$ by 1 and of $P(1, 4)$ by 1, respectively. Table 2.3b shows the final GLSZM for all the possible connected zones of different gray levels in the test image. The features described in Appendix A.5 would then be computed from this final GLSZM. In 3D, 26-voxel connectivity is used to determine connected zones.

Table 2.3: GLSZM computation example.

1	1	1
1	2	2
3	3	1

(a) 2D image

$$\left|\begin{array}{ccccc} N_g \times L_z & \rightarrow & j & \rightarrow & \\ \downarrow & 1 & 0 & 0 & 1 \\ i & 0 & 1 & 0 & 0 \\ \downarrow & 0 & 1 & 0 & 0 \end{array}\right|$$

(b) GLSZM

2.3.4 Neighborhood Gray-Tone Difference Matrix

In 1989, Amadasun & King [42] proposed the concept of texture analysis from the NGTDM. The authors defined 5 features that could be extracted from a NGTDM. Essentially, the NGTDM quantifies the spatial differences between neighboring voxels.

Let $\mathbf{P}$ define the NGTDM of a quantized ROI imaging volume $V_{\mathrm{Q}}(x, y, z)$ with isotropic voxel size. Each entry $P(i)$ of $\mathbf{P}$ represents the summation of the gray-level differences between all center voxels with gray level i and the average gray level of their 8-connected neighbors in 2D space and 26-connected neighbors in 3D space. The NGTDM is a matrix of size $N_g \times 1$, where N_g represents the the pre-defined number of quantized gray levels set in $V_{\mathrm{Q}}(x, y, z)$. Now, let us consider the 2D test image in Table 2.4a. The resulting NGTDM of that image is filled in by examining the absolute intensity difference of each center voxel with its corresponding average neighborhood (thus excluding the center voxel in the average calculation). $P(i)$ is therafter incremented at the corresponding position of the examined gray level i. For example, let us consider gray level 2 in the test image. There are two pixels assigned to a gray level of 2, one at the center and another at the right border of the image. Hence, two incrementations of $P(2)$ will be required to obtain the corresponding NGTDM for gray level 2. Traditionally in the radiomics community, the calculation of $P(2)$ is performed as follows:

$$\begin{aligned} P(2) = &\left|2 - (1+1+1+1+2+3+3+1)/8\right| \\ &+ \left|2 - (1+1+2+3+1)/5\right| \\ &= 0.78 \end{aligned} \tag{2.22}$$

However, this latter incrementation scheme does not take into account the discretization length differences between the different pixels. In our work, we apply corrections for these differences that are valid for both 2D or 3D cases as follows: all averages around a center pixel/voxel are performed such that the neighbours at a distance of $\sqrt{3}$ voxels are given a weight of $1/\sqrt{3}$, the neighbours at a distance of $\sqrt{2}$ voxels are given a weight of $1/\sqrt{2}$, and the neighbours at a distance of 1 voxel are given a weight of 1. Using corrections for discretization length differences, the calculation of $P(2)$ thus becomes:

$$\begin{aligned} P(2) &= \left|2 - \frac{(\frac{3}{\sqrt{2}} \times 1 + 2 \times 1 + 1 \times 2 + \frac{1}{\sqrt{2}} \times 3 + 1 \times 3)}{\frac{3}{\sqrt{2}} + 2 + 1 + \frac{1}{\sqrt{2}} + 1}\right| \\ &+ \left|2 - \frac{(\frac{1}{\sqrt{2}} \times 1 + 2 \times 1 + 1 \times 2 + \frac{1}{\sqrt{2}} \times 3)}{\frac{1}{\sqrt{2}} + 2 + 1 + \frac{1}{\sqrt{2}}}\right| \\ &= 0.81 \end{aligned} \tag{2.23}$$

Table 2.4b shows the final NGTDM calculated using the scheme of Equation 2.23 for all gray levels. The features described in Appendix A.6 would then be computed from this final NGTDM. In 3D, 26-voxel connectivity is used to determine the average neighborhood of each voxel.

Table 2.4: NGTDM computation example.

1	1	1
1	2	2
3	3	1

(a) 2D image

$$\left|\begin{array}{cc} N_g \times 1 & \\ \downarrow & 3.65 \\ i & 0.81 \\ \downarrow & 2.16 \end{array}\right|$$

(b) NGTDM

2.4 Machine learning

One of the overall goal in this thesis is to construct radiomic-based multivariable models that can predict different tumour outcomes. Different types of models can be constructed using different machine learning algorithms such as logistic regression and random forests, for example. A multivariable model is usually composed of few predictive features. The number of features is usually kept low in order for the model to be generalizable to patient cohorts other than the one(s) from which it is trained. The search for the best parsimonious model (the simplest model with best predictive properties) is the crucial step in all machine learning approaches, and it may involve the estimation of prediction performance using different resampling techniques (e.g., cross-validation or bootstrapping) and performance measures (e.g., AUC). In this section, a brief theoretical overview of these concepts is provided. Only the methods used in this work were considered.

2.4.1 Mathematical notations and concepts

In this section, we detail the mathematical notation used throughout this thesis to define input data matrices, input row or column vectors of features, as well as input feature values. Overall in this work, data matrices are denoted with capital letters in bold font, column or row vectors with small letters in bold font, and single variable values with small letters in italic font.

Let $\mathbf{X}$ be the following matrix of input data composed of N rows and M columns, with $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{pmatrix}$$

The matrix $\mathbf{X}$ could represent, for example, a matrix of data to use as an input to a given machine learning algorithm, with N being the number of patients, and M the number of imaging features. Each entry $x_{i,j}$ in the matrix would represent the numerical value of a given imaging feature (i.e., variable) j extracted for patient (i.e., instance) i.

A row vector of input variables from $\mathbf{X}$ representing the set of M imaging feature values over all the different features for a given patient i is then defined as $\mathbf{x}_i = \{x_{i,j} \in \mathbb{R} : j = 1, 2, \ldots, M\}$, such that:

$$\mathbf{x}_i = \begin{pmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,M} \end{pmatrix} .$$

Similarly, a column vector of input variables from $\mathbf{X}$ representing the set of N imaging feature values over all the different patients for a given feature j is then defined as $\mathbf{x}_j = \{x_{i,j} \in \mathbb{R} : i = 1, 2, \ldots, N\}$, such that:

$$\mathbf{x}_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{N,j} \end{pmatrix} .$$

A label $y_i \in \{0, 1\}$ is also defined for each patient i, and it could represent the tumour outcome of a given patient. For example, the class of patients that developed lung metastases would be defined with $y_i = 1$ and denoted as "positive instances", and the class of patients that did not develop lung metastases would be defined with $y_i = 0$ and denoted as "negative instances". The column vector of input labels representing the set of all outcome values y_i for all patients is then defined as $\mathbf{y} = \{y_i \in \{0,1\} : i = 1,2,\ldots,N\}$, such that:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Finally, the overall goal of the machine learning approaches used in this thesis is to find a function f in the form of a multivariable model that best maps the set of input data $\mathbf{X}$ to the set of labels $\mathbf{y}$, such that $f : \mathbf{X} \Rightarrow \mathbf{y}$.

2.4.2 Logistic regression

In logistic regression modeling, we are interested in finding a linear combination of p variables from an input data matrix $\mathbf{X}$ such that the multivariable model of interest takes the form:

$$g(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij}, \text{ for } i = 1,2,\ldots,N. \tag{2.24}$$

In Equation 2.24, the set $\boldsymbol{\beta} = \{\beta_j \in \mathbb{R} : j = 0,1,\ldots,p\}$ is the set of logistic regression coefficients of the model to be determined such that the conditional probability of y_i given the input $\mathbf{x}_i$ is maximized for $i = 1,2,\ldots,N$. This operation is carried out using a logistic regression model (logit transformation) of the form:

$$\pi(\mathbf{x}_i) = P(y_i = 1 \,|\, \mathbf{x}_i) = \frac{\exp\left[g(\mathbf{x}_i)\right]}{1 + \exp\left[g(\mathbf{x}_i)\right]}. \tag{2.25}$$

The form of the logistic regression model shown in Equation 2.25 is commonly used in oncology since it models a sigmoidal relationship between the input variables and the response endpoint within the range [0,1] as shown in Figure 2.18, lending itself to a clinically meaningful probability interpretation of observed responses. To be more specific, $\pi(\mathbf{x}_i)$ expresses the conditional probability that the outcome y_i equals 1 given the input $\mathbf{x}_i$. Consequently, the conditional probability that the outcome y_i equals 0 given the input $\mathbf{x}_i$ is $P(y_i = 0 \,|\, \mathbf{x}_i) = 1 - \pi(\mathbf{x}_i)$. If we assume the N observations to be independent, it follows that a convenient way to express the conditional probability of a set of dichotomous outcome states given the set of input data is by using the "log-likelihood" function:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \Big[y_i \ln\big(\pi(\mathbf{x}_i)\big) + (1 - y_i) \ln\big(1 - \pi(\mathbf{x}_i)\big) \Big]. \tag{2.26}$$

The set of logistic regression coefficients that maximizes the log-likelihood function of Equation 2.26 is found by separately differentiating $L(\boldsymbol{\beta})$ with respect to all β_j coefficients embedded in $\pi(\mathbf{x}_i)$ and then equating to zero. This yields a set of $p + 1$ non-linear equations to be solved simultaneously using an iterative weighted least-square method. The presentation of this methodology goes beyond the scope of this text, but the interested reader is referred to reference [47] for a general description of the methods used by most software.

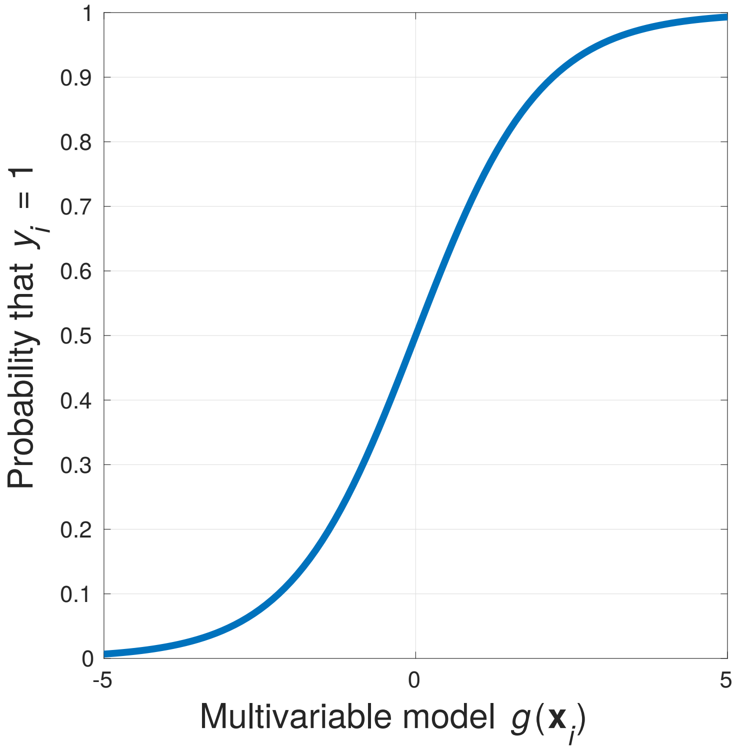


Figure 2.18: Sigmoidal curve of logistic regression models.

2.4.3 Bootstrapping

Bootstrapping is a statistical resampling method introduced by Efron [48] in 1979. The motivation of his pioneering work was to develop a more general yet simple alternative to cross-validation techniques for the estimation of unknown probability distributions of random variables based on the observed data. Bootstrapping is in fact less prone to overestimation of statistical significance as compared to cross-validation techniques, one of the major pitfalls of data mining. Bootstrap tutorials, reviews and applications in medicine can be readily found in the literature [49–51]. In this thesis, bootstrapping is used as the resampling method of choice to estimate the prediction performance of radiomic models constructed using logistic regression as described in section 2.4.6. In this section, we demonstrate how bootstrap resampling leads to the creation of two disjoint training and testing sets from a sample data matrix.

Let a data matrix $\mathbf{X}$ of N instances (e.g., patients) and M variables (e.g., imaging features). A bootstrap sample $\mathbf{X}^*$ is a sample of N randomly drawn instances with replacement from the available sample $\mathbf{X}$ (some of the $\mathbf{x}_i$ may appear 0, 1, 2, 3 ... times in $\mathbf{X}^*$). The bootstrap sample can be thought of as a new training sample different from the observed sample. The set of all original vectors $\mathbf{x}_i$ that do not appear in $\mathbf{X}^*$ is considered as a testing sample and is denoted as $\mathbf{X}^*(0)$. As an example, consider the following situation:

- Let a set of data of 10 instances be $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$.
- 10 instances are randomly drawn from $\mathbf{X}$ with replacement. For example, let the selected instances be $\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_5, \mathbf{x}_9, \mathbf{x}_3, \mathbf{x}_5\}$.
- Instances drawn from $\mathbf{X}$ constitute a bootstrap training sample: $\mathbf{X}^* = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_5, \mathbf{x}_9, \mathbf{x}_3, \mathbf{x}_5\}$.
- Instances not drawn from $\mathbf{X}$ constitute a bootstrap testing sample: $\mathbf{X}^*(0) = \{\mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8, \mathbf{x}_{10}\}$.

The bootstrap resampling process is usually repeated for many bootstrap samples $b = 1, 2, \ldots, B$. Each different bootstrap training and testing sets created from the different bootstrap samples b are denoted $\mathbf{X}^{*b}$ and $\mathbf{X}^{*b}(0)$, respectively.

2.4.4 Random forests

The random forest algorithm was introduced by Breiman [52] in 2001. A random forest consists of an ensemble of fully-grown decision-trees, where two levels of randomness are introduced: I) Creation of each tree using a different bootstrap sample; and II) Creation of each branching step of each tree with randomly chosen variables available from the input data. These two points will be discussed below.

Although decision-trees could also be used for regression, our utilization in this thesis is limited to classifications problems. A decision-tree essentially divides an input space into several branching nodes. As described in Figure 2.19 with an example decision-tree, a testing instance would eventually fall at the end of one of the different branching nodes (i.e., a leaf node). The classification result (e.g., 0 or 1) for that instance in this decision-tree would be the one assigned at the end of that particular leaf node.

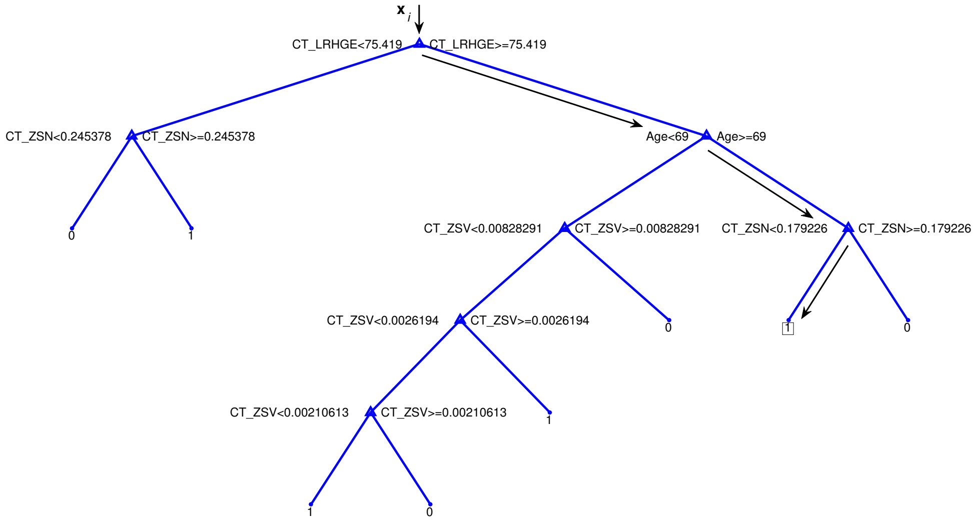


Figure 2.19: Example classification using a decision-tree. Consider a testing patient instance $\mathbf{x}_i$ with the following variable values: $\{\mathrm{CT\_LRHGE} = 80.21, \mathrm{Age} = 75, \mathrm{CT\_LRHGE} = 0.178\}$. The classification result for this patient from the example decision-tree goes as follows: I) From the first branching node on top, CT_LRHGE $= 80.21 \geq 75.419$ results in going to the right branch of the node; II) At the next branching node, Age $= 75 \geq 69$ results in going to the right branch of the node; III) At the next branching node, CT_LRHGE $= 0.178 < 0.179226$ results in going to the left of the node; and IV) Having reached one of the end of the decision-tree (i.e., a leaf node), the instance is classified as the value assigned to that leaf node: 1.

In a random forest, decision-trees are grown until all leaf nodes are pure to only one class (i.e., fully-grown trees), thus until the probability of classification of one class is 100 %. This is in contrast to conventional decision-trees where the classification probability of one of the leaf node could be, for example, 70 % to class "0" and 30 % to class "1". In the context of decision-tree learning, a variable "A" is used to split each node into two daughter branches. The variable choice among other variables is based on how well it can predict a target class "C" based on its information gain defined as [53]:

$$gain(A) = H(A) - H(A\,|\,C), \tag{2.27}$$

where H is the entropy, a classical measure of information in *information theory* that is conveyed by the probability distributions of the variables and the target class. The threshold at each node is also set as to maximize the gain of the tested variables. A particularity of random forests is that the tested variables are potentially different at every node of every decision-tree: a set of $m = \sqrt{p}$ variables are randomly chosen for each test node from an available set of p variables, representing the lower level of randomness introduced in random forests.

The most important concept of a random forest is, again, that it consists of an ensemble of fully-grown decision-trees. Being fully-grown, each decision-tree has no bias with generally a high variance. The power of random forests resides in taking an average of multiple unbiased decision-trees to reduce the variance of classification. The higher level of randomness introduced in random forests results from the training of different decision-trees: from an input data matrix $\mathbf{X}$, each decision-tree is trained with a different bootstrap sample $\mathbf{X}^{*b}$. The number of trees T in the forest is thus conventionally equal to the number of bootstrap samples B that are drawn from $\mathbf{X}$. For binary classification problems such as in this thesis, the final classification of a random forest consists in taking the average of the classification results of all decision-trees in the forest. For example, let a random forest be composed of 100 decision-trees. If 64 decision-trees provide a classification result of "1" and 36 a classification result of "0", the probability of classification of class "1" is 64 %, wherehas it is 36 % for class "0". The overall prediction of the random forest classifier would thus be class "1" since $P(y_i = 1 \,|\, \mathbf{x}_i) > 0.5$. This process is illustrated in Figure 2.20.

We can now formally define the multivariable model response $p_{\text{RF}}(\mathbf{x}_i)$ of a random forest composed of T decision-trees. Let $h_t(\mathbf{x}_i) \in \{0, 1\}$ be the classification result of the decision-tree t in the random forest. The probability of observing outcome $y_i = 1$ given an input $\mathbf{x}_i$ is hereby defined as:

$$p_{\text{RF}}(\mathbf{x}_i) = P(y_i = 1 \,|\, \mathbf{x}_i) = \frac{\sum_{t=1}^{T} h_t(\mathbf{x}_i)}{T}. \tag{2.28}$$

For more information about random forests, the reader is referred to references [54, 55].

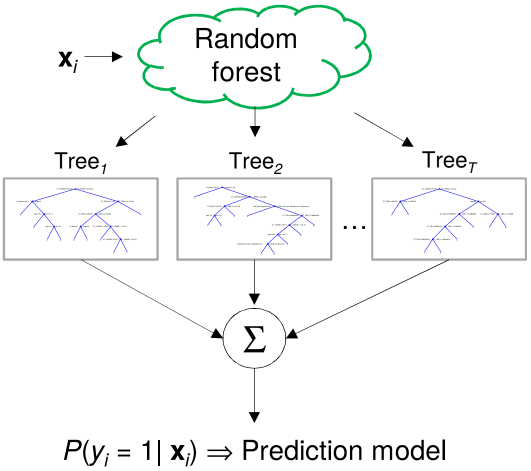


Figure 2.20: Schematic representation of a random forest prediction. A random forest consists of T fully-grown decision-trees. The prediction result of the random forest consists of the average classification result obtained from each decision-tree in the forest, and it is thus expressed as a probability of observing an outcome y_i given a set of input data $\mathbf{x}_i$

2.4.5 Performance measures

In this section, two categories of performance measures are defined: I) Spearman's rank correlation coefficient used in univariate analyses; and II) Receiver operating characteristic (ROC) metrics used in multivariable analyses.

Spearman's rank correlation (univariate anaysis)

In this thesis, the Spearman's rank correlation coefficient r_s is used in univariate analyses to quantify the association of single feature variables $\mathbf{x}_j$ with an outcome vector of interest $\mathbf{y}$. First, the individual values x_{ij} of $\mathbf{x}_j$ and y_i of $\mathbf{y}$ are converted to the set of ranks $\mathbf{r}_j^x$ and $\mathbf{r}_j^y$ with individual values r_{ij}^x and r_{ij}^y that they take in their data vector, respectively. Ties in $\mathbf{y}$ and potential ties in $\mathbf{x}_j$ are assigned a rank equal to the average of their positions in the ascending order of the values. Then $r_s(\mathbf{x}_j, \mathbf{y})$ is defined as:

$$r_s(\mathbf{x}_j, \mathbf{y}) = \frac{\sum_{i=1}^{N}\left(r_{ij}^x - \overline{\mathbf{r}_j^x}\right)\left(r_{ij}^y - \overline{\mathbf{r}_j^y}\right)}{\sqrt{\sum_{i=1}^{N}\left(r_{ij}^x - \overline{\mathbf{r}_j^x}\right)^2 \sum_{i=1}^{N}\left(r_{ij}^y - \overline{\mathbf{r}_j^y}\right)^2}}, \tag{2.29}$$

where $\overline{\mathbf{r}_j^x}$ and $\overline{\mathbf{r}_j^y}$ are the average of $\mathbf{r}_j^x$ and $\mathbf{r}_j^y$, respectively. Spearman's rank correlation describes how well two variables are monotonically related, independently of their linear association as it is the case with Pearson's coefficient. A result of +1 implies perfect positive correlation, a result of -1 implies perfect negative correlation and a result of 0 implies no correlation between the variables. The statistical significance of the correlation (p-value) is thereafter determined using a Student's t-test. When summarizing correlation results for multiple feature vectors $\mathbf{x}_j$ with a given outcome vector $\mathbf{y}$, Bonferonni [56] or Benjamini-Hochberg [57] corrections for multiple testing comparisons can be applied to reduce the false discovery rate (i.e., Type I errors).

ROC metrics (multivariable analysis)

In this thesis, ROC metrics are used in multivariable analyses to assess the prediction performance of radiomic-based models in different types of testing sets (bootstrap testing sets in Chapter 3 and Chapter 5, independent testing sets in Chapter 4 and Chapter 6). Here, the "assessment of prediction performance" refers to the overall efficiency of a classifier in correctly predicting positive instances (i.e., $y_i = 1$) as being of class "1" and negative instances (i.e., $y_i = 0$) as being of class "0".

In binary classification theory, four quantities of interest are first calculated: I) TP: number of true positive instances; II) FP: number of false positive instances; III) TN: number of true negative instances; and IV) FN: number of false negative instances. Table 2.5 summarizes how a testing patient i is to be classified as a TP_i, FP_i, TN_i or FN_i instance in the context of the multivariable model responses of logisitic regression ($g(\mathbf{x}_i)$ and $\pi(\mathbf{x}_i)$ in Equation 2.24 and Equation 2.25) and random forests ($p_{\text{RF}}(\mathbf{x}_i)$ in Equation 2.28).

Table 2.5: Classification of testing instances i: TP, FP, TN and FN.

Multivariable model response		True outcome value	Classification
Logisitic regression	Random forests		
$g(\mathbf{x}_i) \geq 0 \Rightarrow \pi(\mathbf{x}_i) \geq 0.5$	$p_{\text{RF}}(\mathbf{x}_i) \geq 0.5$	$y_i = 1$	TP_i
$g(\mathbf{x}_i) \geq 0 \Rightarrow \pi(\mathbf{x}_i) \geq 0.5$	$p_{\text{RF}}(\mathbf{x}_i) \geq 0.5$	$y_i = 0$	FP_i
$g(\mathbf{x}_i) < 0 \Rightarrow \pi(\mathbf{x}_i) < 0.5$	$p_{\text{RF}}(\mathbf{x}_i) < 0.5$	$y_i = 0$	TN_i
$g(\mathbf{x}_i) < 0 \Rightarrow \pi(\mathbf{x}_i) < 0.5$	$p_{\text{RF}}(\mathbf{x}_i) < 0.5$	$y_i = 1$	FN_i

Given that $\text{TP} = \sum_i \text{TP}_i$, $\text{FP} = \sum_i \text{FP}_i$, $\text{TN} = \sum_i \text{TN}_i$ and $\text{FN} = \sum_i \text{FN}_i$, the "sensitivity", "specificity" and "accuracy" of classification is calculated as:

$$\begin{aligned} \text{sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{accruracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \end{aligned} \tag{2.30}$$

For the majority of tumour outcomes, the proportion of positive and negative instances is not the same due to a low occurrence rate of events (e.g., development of distant metastases). When working with unbalanced datasets, it is crucial to perform model training by using imbalance adjustments in order to obtain a performance of classification balanced between sensitivity and specificity. Different methods for that purpose are developed in Chapter 3 and Chapter 6.

Finally, another metric summarizing prediction performance can be extracted from the ROC curve of a binary classifier. The ROC curve is a plot of the true positive rate (i.e., the sensitivity) against the false positive rate (i.e., $1-$ specificity) when new subsets of $\{\text{TP}, \text{FP}, \text{TN}, \text{FN}\}$ are obtained for varying decision thresholds DT (e.g., classification to class "1" with $g(\mathbf{x}_i) \geq DT$). The metric of interest enabling the assessment of the quality of the classifier is then:

$$\text{AUC} = \text{Area under the ROC curve} \tag{2.31}$$

Figure 2.21 displays an example ROC curve for two classifiers f_1 and f_2. Since the area under the curve of f_1 is higher than f_2, its AUC metric is also higher. An AUC of 0.5 corresponds to a random classifier whereas an AUC of 1 corresponds to a perfect classifier. One way of interpreting this metric is that the greater is the AUC, the better is the separation rank between the positive and negative instances.

2.4.6 Estimation of prediction performance

The estimation of prediction performance is a process occuring during model training. The main goal is to estimate, from a training patient dataset $\mathbf{S}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}\}$, the set of model parameters (e.g., features, logistic regression coefficients, random forest architecture, etc.) that would allow for optimal prediction performance of the complete model on the whole patient population. The set $\mathbf{S}_{\text{train}}$ could be, for example, one or combined sample patient cohort(s) coming from one or more cancer centers. The proof of concept to validate that a complete model generalizes well to the whole patient population involves to test the model, at a later stage, onto independent patient cohorts coming from external cancer centers.

The prediction performance estimation process must thus take place entirely in the available training dataset $\mathbf{S}_{\text{train}}$. The usual process is to subdivide this set into two disjoint sets that we define as $\mathbf{S}_{\text{subTrain}} = \{\mathbf{X}_{\text{subTrain}}, \mathbf{y}_{\text{subTrain}}\}$

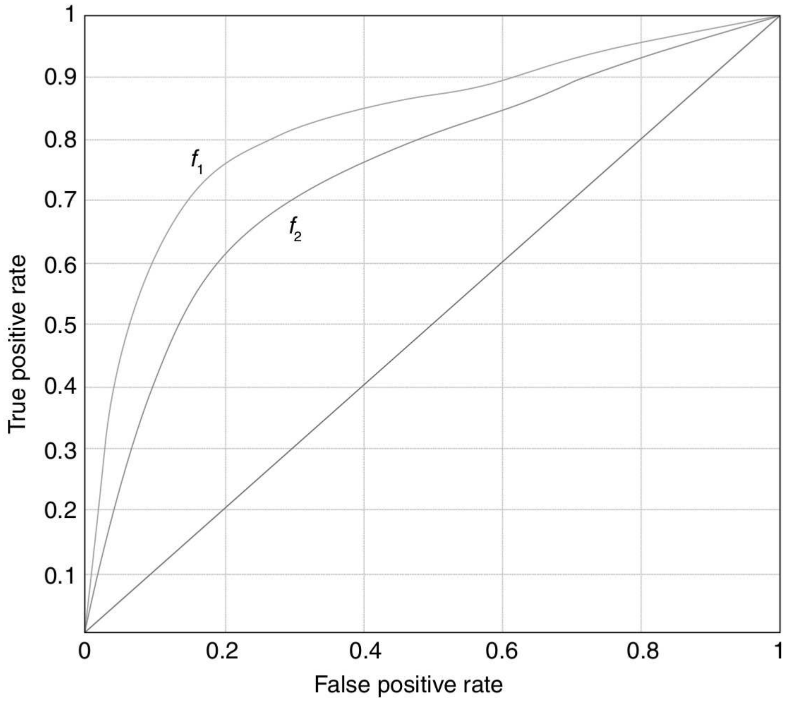


Figure 2.21: Concept of the AUC metric. This figure displays an example of a receiver operating characteristic (ROC) curve. The ROC curve is obtained by computing the true and false positive rates of a classifier under varying decision thresholds. The area under the ROC curve (AUC) is a measure of the prediction performance of a classifier. An AUC of 0.5 corresponds to a random classifier whereas an AUC of 1 corresponds to a perfect classifier. In the example above, the classifer f_1 has a higher AUC than f_2. Reprinted with permission from [58]. © 2015 Springer. All rights reserved.

and $\mathbf{S}_{\text{subTest}} = \{\mathbf{X}_{\text{subTest}}, \mathbf{y}_{\text{subTest}}\}$, such that $\mathbf{S}_{\text{train}} = \{\mathbf{S}_{\text{subTrain}} \cup \mathbf{S}_{\text{subTest}}\}$.[1] Let us then define the estimated AUC that we obtain when a model is trained in $\mathbf{S}_{\text{subTrain}}$ and thereafter tested in $\mathbf{S}_{\text{subTest}}$ as:

$$\widehat{\text{AUC}} = \text{AUC}(\mathbf{S}_{\text{subTrain}}, \mathbf{S}_{\text{subTest}}). \tag{2.32}$$

In this section, we describe below the methods used in this thesis when: I) estimating the prediction performance of logistic regression models in terms of the AUC using bootstrap resampling; and II) estimating the prediction performance of random forests in terms of the AUC using stratified random sub-sampling.

Bootstrap resampling

Let us recall that a sample data matrix $\mathbf{X}$ can be sub-divided into B bootstrap training samples $\mathbf{X}^{*b}$ and bootstrap testing samples $\mathbf{X}^{*b}(0)$, with $b = 1, 2, \ldots, B$. Here, $\mathbf{X}^{*b}$ is to equivalent to $\mathbf{S}_{\text{subTrain}}$, and $\mathbf{X}^{*b}(0)$ to $\mathbf{S}_{\text{subTest}}$.[2] Three AUC estimation methods are described here: I) the ordinary bootstrap method; II) the 0.632 bootstrap method; and III) the 0.632+ boostrap method.

The estimation of the AUC using the ordinary bootstrap method ($\widehat{\text{AUC}}$) goes as follows:

$$\widehat{\text{AUC}} = \frac{1}{B} \sum_{b=1}^{B} \text{AUC}(\mathbf{X}^{*b}, \mathbf{X}^{*b}(0)). \tag{2.33}$$

However, Efron [59] demonstrated in 1983 that the estimation of the AUC using the ordinary bootstrap method (Equation 2.33) is pessimistically biased (i.e., it underestimates the true AUC) because $\mathbf{X}^{*b}(0)$ is farther away from the sample cohort $\mathbf{X}$ than a typical test sample randomly drawn from the true population $\mathbf{X}_{\text{true}}$. On average over an infinite number of bootstrap samples, the ratio of the distance between $\mathbf{X}_{\text{true}}$ and $\mathbf{X}$ to the distance between $\mathbf{X}^{*b}(0)$

[1] Formally in machine learning theory, $\mathbf{S}_{\text{train}}$ should be denoted as the "teaching set", $\mathbf{S}_{\text{subTrain}}$ as the "training set", $\mathbf{S}_{\text{subTest}}$ as the "validation set", and external and independent sample patient cohorts as "testing sets". Therefore, what we denote in this thesis as a "bootstrap training sample" is formally a training set, and what we denote as a "bootstrap testing sample" is formally a validation set. However, for the sake of consistency with previously published material contained in this thesis, we will carry on with our current definitions in the text that follows.

[2] This is not strictly true, as $\mathbf{S}_{\text{subTrain}}$ and $\mathbf{S}_{\text{subTest}}$ include the definition of the labels $\mathbf{y}$. However, in the text that follows and for the sake of consistency with already published material contained in this thesis and from other research groups, $\text{AUC}(\mathbf{S}_1, \mathbf{S}_2)$ is assumed to always contain the definition of the labels in $\mathbf{S}_1$ and $\mathbf{S}_2$, even when $\text{AUC}(\mathbf{X}^{*b}, \mathbf{X}^{*b}(0))$ is used (computing the AUC always requires a set of labels).

and $\mathbf{X}$ is $1 - 1/e \approx 0.632$. To correct for that bias, the estimation of the AUC using the 0.632 bootstrap method is defined by Efron [59] as:

$$\widehat{\mathrm{AUC}}_{0.632} = (1-0.632)\times\mathrm{AUC}(\mathbf{X},\mathbf{X})-0.632\times\frac{1}{B}\sum_{b=1}^{B}\mathrm{AUC}(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)), \quad (2.34)$$

where the AUC calculated on the whole training cohort $\mathrm{AUC}(\mathbf{X},\mathbf{X})$ (i.e, the model is both trained and tested on $\mathbf{S}_{\text{train}}$) provoks an optimistic bias (i.e., it overestimates the true AUC) on the estimated AUC to correct for the pessimistic bias of the ordinary bootstrap method.

Then, in 1997, Efron & Tibshirani [60] designed a method to balance the pessimistic and optimistic bias of the ordinary and 0.632 bootstrap methods, respectively: the 0.632+ method. Sahiner *et al.* [61] slightly modified that method to obtain an estimated AUC under the 0.632+ method defined as:

$$\begin{aligned}
\widehat{\mathrm{AUC}}_{0.632+} &= \frac{1}{B}\sum_{b=1}^{B}\left[(1-\alpha(b))\cdot\mathrm{AUC}\big(\mathbf{X},\mathbf{X}\big)+\alpha(b)\cdot\mathrm{AUC}'\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)\right],\\
\mathrm{AUC}'\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big) &= \max\left\{0.5,\mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)\right\},\\
\alpha(b) &= \frac{0.632}{1-0.368\cdot R(b)},\\
R(b) &= \begin{cases} 1 & \text{if } \mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big) \leq 0.5,\\ \dfrac{\mathrm{AUC}\left(\mathbf{X},\mathbf{X}\right)-\mathrm{AUC}\left(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\right)}{\mathrm{AUC}\left(\mathbf{X},\mathbf{X}\right)-0.5} & \text{if } 2 > \dfrac{\mathrm{AUC}\big(\mathbf{X},\mathbf{X}\big)}{\mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)} > 1,\\ 0 & \text{otherwise}. \end{cases}
\end{aligned} \quad (2.35)$$

In this work, the $\widehat{\mathrm{AUC}}_{0.632+}$ is the main metric used for the estimation of the prediction performance of radiomics-based models constructed using logistic regression.

Stratified random sub-sampling

In Chapter 6, prediction performance estimation for random forests is performed using stratified random sub-sampling – this method is preferred to bootstrapping for random forests since this machine learning algorithm inherently uses bootstrap resampling to train different decision-trees. The strategy employed here essentially consists of sub-dividing $\mathbf{S}_{\text{train}}$ into S different splits. Each split $s = 1, 2, \ldots, S$ contains two disjoint sets that are randomly sampled from $\mathbf{S}_{\text{train}}$ without replacement such that $\mathbf{S}_{\text{train}} = \{\mathbf{S}^s_{\text{subTrain}} \cup \mathbf{S}^s_{\text{subTest}}\}$, where $\mathbf{S}^s_{\text{subTrain}}$ is chosen to be twice the size of $\mathbf{S}^s_{\text{subTest}}$. The "stratification" part of the process implies that the random sub-sampling is such that $\mathbf{S}_{\text{train}}$, $\mathbf{S}^s_{\text{subTrain}}$ and $\mathbf{S}^s_{\text{subTest}}$ all contain the same proportion of positive and negative instances. For example, consider the following situation:

- A example training set is defined as:
  - $\mathbf{S}_{\text{train}} = \{\mathbf{x}^-_1, \mathbf{x}^-_2, \mathbf{x}^-_3, \mathbf{x}^-_4, \mathbf{x}^-_5, \mathbf{x}^+_6, \mathbf{x}^-_7, \mathbf{x}^-_8, \mathbf{x}^+_9, \mathbf{x}^-_{10}, \mathbf{x}^-_{11}, \mathbf{x}^+_{12}\}$
  - Instances $i = \{1, 2, 3, 4, 5, 7, 8, 10, 11\}$ are negative
  - Instances $i = \{6, 9, 12\}$ are positive
- Stratified random sub-sampling creates one split s with:
  - $\mathbf{S}^s_{\text{subTrain}} = \{\mathbf{x}^-_1, \mathbf{x}^-_3, \mathbf{x}^+_6, \mathbf{x}^-_7, \mathbf{x}^-_8, \mathbf{x}^-_{10}, \mathbf{x}^-_{11}, \mathbf{x}^+_{12}\}$
  - $\mathbf{S}^s_{\text{subTest}} = \{\mathbf{x}^-_2, \mathbf{x}^-_4, \mathbf{x}^-_5, \mathbf{x}^+_9\}$

Finally, after randomly producing S splits, the estimation of the AUC is performed such that:

$$\widehat{\text{AUC}} = \frac{1}{S}\sum_{s=1}^{S} \text{AUC}(\mathbf{S}^s_{\text{subTrain}}, \mathbf{S}^s_{\text{subTest}}). \tag{2.36}$$

2.5 References

1. Zaidi, H. & Hasegawa, B. *"Overview of Nuclear Medical Imaging: Physics and Instrumentation", Quantitative Analysis in Nuclear Medicine Imaging* (ed Zaidi, H.) 1–34 (Springer, Singapore, 2006).
2. Ollinger, J. M. & Fessler, J. A. Positron-emission tomography. *IEEE Signal Processing Magazine* **14,** 43–55 (1997).
3. Fahey, F. H. Data acquisition in PET imaging. *J. Nucl. Med. Technol.* **30,** 39–49 (2002).
4. Zaidi, H. *Quantitative Analysis in Nuclear Medicine Imaging* 583 pp. (Springer, Singapore, 2006).
5. Cherry, S. R., Sorenson, J. A. & Phelps, M. E. *Physics in Nuclear Medicine* 4th ed. 544 pp. (Saunders, Philadelphia, 2012).
6. Goldman, L. W. Principles of CT and CT technology. *J. Nucl. Med. Technol.* **35,** 115–128 (2007).
7. Herman, G. T. *Fundamentals of Computerized Tomography : Image Reconstruction from Projections* 2nd ed. 276 pp. (Springer, London, 2009).

8. Ortiz, E. A. *Quantitative Functional MRI Based Evaluation of Caffeine's Effects on Brain Physiology.* Master thesis (Medical Physics Unit, McGill University, 2011).

9. Plewes, D. B. & Kucharczyk, W. Physics of MRI: a primer. *J. Magn. Reson. Imaging* **35,** 1038–1054 (2012).

10. Bloch, F. Nuclear induction. *Phys. Rev.* **70,** 460–474 (1946).

11. Currie, S., Hoggard, N., Craven, I. J., Hadjivassiliou, M. & Wilkinson, I. D. Understanding MRI: basic MR physics for physicians. *Postgrad. Med. J.* **89,** 209–223 (2013).

12. Bernstein, M. A., King, K. F. & Zhou, X. J. *Handbook of MRI Pulse Sequences* 1st ed. 1040 pp. (Elsevier Academic Press, Amsterdam, 2004).

13. Nishimura, D. G. *Principles of Magnetic Resonance Imaging* ed. 1.2. 238 pp. (Stanford Univ, San Francisco, 2016).

14. Patterson, D. M., Padhani, A. R. & Collins, D. J. Technology insight: water diffusion MRI—a potential new biomarker of response to cancer therapy. *Nat. Clin. Pract. Onco.* **5,** 220–233 (2008).

15. Hagmann, P. *et al.* Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *RadioGraphics* **26,** S205–S223 (2006).

16. Jackson, A., Buckley, D. L. & Parker, G. J. M. *Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology* 311 pp. (Springer, Heidelberg, 2005).

17. Fayad, L. M., Jacobs, M. A., Wang, X., Carrino, J. A. & Bluemke, D. A. Musculoskeletal tumors: how to use anatomic, functional, and metabolic MR techniques. *Radiology* **265,** 340–356 (2012).

18. Van Rijswijk, C. S. P. *et al.* Soft-tissue tumors: value of static and dynamic gadopentetate dimeglumine–enhanced MR imaging in prediction of malignancy. *Radiology* **233,** 493–502 (2004).

19. Gribbestad, I. S. *et al. "An Introduction to Dynamic Contrast-Enhanced MRI in Oncology", Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology* (eds Jackson, A., Buckley, D. L. & Parker, G. J. M.) 3–22 (Springer, Heidelberg, 2005).

20. Tofts, P. S. *et al.* Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. *J. Magn. Reson. Imaging* **10,** 223–232 (1999).

21. Hylton, N. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *J. Clin. Oncol.* **24,** 3293–3298 (2006).

22. Horsfield, M. A. & Morgan, B. Algorithms for calculation of kinetic parameters from T1-weighted dynamic contrast-enhanced magnetic resonance imaging. *J. Magn. Reson. Imaging* **20,** 723–729 (2004).

23. Evelhoch, J. L. Key factors in the acquisition of contrast kinetic data for oncology. *J. Magn. Reson. Imaging* **10,** 254–259 (1999).

24. Lavini, C. *et al.* Pixel-by-pixel analysis of DCE MRI curve patterns and an illustration of its application to the imaging of the musculoskeletal system. *Magn. Reson. Imaging* **25,** 604–612 (2007).

25. Yankeelov, T. E. & Gore, J. C. Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples. *Curr. Med. Imaging Rev.* **3,** 91–107 (2009).

26. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative. *arXiv preprint.* arXiv: 1612.07003 (2016).

27. Boussion, N., Rest, C. C. L., Hatt, M. & Visvikis, D. Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging. *Eur. J. Nucl. Med. Mol. Imaging* **36,** 1064–1075 (2009).

28. Gjesteby, L. *et al.* Metal artifact reduction in CT: where are we after four decades? *IEEE Access* **4,** 5826–5849 (2016).

29. Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging* **17,** 87–97 (1998).

30. Strang, G. & Nguyen, T. *Wavelets and filter banks* Rev. ed. 520 pp. (Wellesley-Cambridge Press, Wellesley, MA, 1997).

31. Burrus, C. S., Gopinath, R. A. & Guo, H. *Introduction to wavelets and wavelet transforms: a primer* 268 pp. (Prentice Hall, Upper Saddle River, N.J., 1998).

32. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

33. Pajares, G. & Manuel de la Cruz, J. A wavelet-based image fusion tutorial. *Pattern Recognition* **37,** 1855–1872 (2004).

34. Li, H., Manjunath, B. S. & Mitra, S. K. Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing* **57,** 235–245 (1995).

35. Goshtasby, A. A. *2-D and 3-D Image Registration: for Medical, Remote Sensing, and Industrial Applications* 1 edition. 284 pp. (Wiley-Interscience, Hoboken, NJ, 2005).

36. Collewet, G., Strzelecki, M. & Mariette, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn. Reson. Imaging* **22,** 81–91 (2004).

37. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).

38. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28,** 129–137 (1982).

39. Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory* **6,** 7–12 (1960).

40. Jain, A. K. *Fundamentals of Digital Image Processing* 592 pp. (Prentice Hall, Upper Saddle River, NJ, 1989).

41. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).

42. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).

43. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

44. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).

45. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

46. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56,** 38–44 (2015).

47. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* 2nd ed. 532 pp. (Chapman and Hall, London, 1989).

48. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7,** 1–26 (1979).

49. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application* 582 pp. (Cambridge University Press, Cambridge, 1997).

50. Wehrens, R., Putter, H. & Buydens, L. M. C. The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems* **54** (2000).

51. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statist. Med.* **19,** 1141–1164 (2000).

52. Breiman, L. Random forests. *Machine Learning* **45,** 5–32 (2001).

53. El Naqa, I., Li, R. & Murphy, M. J. *Machine Learning in Radiation Oncology: Theory and Applications* 1st ed. 336 pp. (Springer International Publishing, Cham, Switzerland, 2015).

54. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. 745 pp. (Springer, New York, 2009).

55. Louppe, G. *Understanding Random Forests: From Theory to Practice.* PhD thesis (Department of Electrical Engineering & Computer Science, University of Liège, 2014).

56. Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association* **56,** 52–64 (1961).

57. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57,** 289–300 (1995).

58. Japkowicz, N. & Shah, M. *"Performance Evaluation in Machine Learning", Machine Learning in Radiation Oncology: Theory and Applications* (eds Naqa, I. E., Li, R. & Murphy, M. J.) 41–56 (Springer International Publishing, Cham, Switzerland, 2015).

59. Efron, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316–331 (1983).

60. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560 (1997).

61. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.* **35,** 1559–1570 (2008).

Chapter 3

Development of a radiomic model for the early prediction of lung metastases

3.1 Foreword 90
3.2 Abstract 90
3.3 Introduction 91
3.4 Materials and Methods 94
3.4.1 Data 95
3.4.2 Tumour volume definition 97
3.4.3 Imaging data pre-processing 97
3.4.4 FDG-PET/MRI volume fusion 98
3.4.5 Feature extraction 99
3.4.6 Univariate analysis 103
3.4.7 Multivariable analysis 105
3.5 Results 109
3.5.1 Univariate analysis 109
3.5.2 Multivariable analysis 113
3.6 Discussion 116
3.7 Conclusion 119
3.8 Online resources 120
3.9 Acknowledgments 120
3.10 Supplementary Material 120
3.10.1 Definition of texture features 120
3.10.2 Texture features with highest predictive value 121
3.10.3 Degrees of freedom on texture extraction parameters . 121
3.10.4 Uncertainty analysis 123
3.10.5 Permutation tests 126
3.11 References 126

3.1 Foreword

This Chapter presents a study published as the following paper: Martin Vallières, Carolyn R. Freeman, Sonia Skamene & Issam El Naqa. "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities". *Phys. Med. Biol.* **60,** 5471–5496 (2015).

In this study, the groundwork for the construction of radiomic-based prediction models via logistic regression is presented. This process notably involves texture optimization, feature set reduction, feature selection, prediction performance estimation via bootstrapping, and imbalance-adjusted learning. Furthermore, we also explored a novel approach based on the fusion of FDG-PET and MR imaging volumes to better quantify intratumoural heterogeneity using texture analysis. Ultimately, as single and complete multivariable model composed of four features was constructed for the prediction of lung metastases in soft-tissue sarcomas.

3.2 Abstract

This study aims at developing a joint FDG-PET and MRI texture-based model for the early evaluation of lung metastasis risk in soft-tissue sarcomas (STSs). We investigate if the creation of new composite textures from the combination of FDG-PET and MR imaging information could better identify aggressive tumours. Towards this goal, a cohort of 51 patients with histologically proven STSs of the extremities was retrospectively evaluated. All patients had pre-treatment FDG-PET and MRI scans comprised of T_1-weighted and T_2-weighted fat-suppression sequences (T2FS). Nine non-texture features (SUV metrics and shape features) and forty-one texture features were extracted from the tumour region of separate (FDG-PET, T1 and T2FS) and fused (FDG-PET/T1 and FDG-PET/T2FS) scans. Volume fusion of the FDG-PET and MRI scans was implemented using the wavelet transform. The influence of six different extraction parameters on the predictive value of textures was investigated. The incorporation of features into multivariable models was performed using logistic regression. The multivariable modeling strategy involved imbalance-adjusted bootstrap resampling in the following four steps leading to final prediction model construction: 1) feature set reduction; 2) feature selection; 3) prediction performance estimation; and 4) computation

of model coefficients. Univariate analysis showed that the isotropic voxel size at which texture features were extracted had the most impact on predictive value. In multivariable analysis, texture features extracted from fused scans significantly outperformed those from separate scans in terms of lung metastases prediction estimates. The best performance was obtained using a combination of four texture features extracted from FDG-PET/T1 and FDG-PET/T2FS scans. This model reached an area under the receiver-operating characteristic curve (AUC) of 0.984 ± 0.002, a sensitivity of 0.955 ± 0.006, and a specificity of 0.926 ± 0.004 in bootstrapping evaluations. Ultimately, lung metastasis risk assessment at diagnosis of STSs could improve patient outcomes by allowing better treatment adaptation.

3.3 Introduction

Soft-tissue sarcomas (STSs) constitute a heterogeneous group of malignant neoplasms of mesenchymal cell origin. More than 50 sub-types are recognized by the World Health Organization (WHO). STSs are relatively uncommon, representing approximately 0.7 % of new adult malignancies in the United States [1]. The majority of new cases are either intermediate or high-grade tumours and may arise in virtually all sites, with the extremities as the most common site of origin [2]. In general, the different forms of therapy lead to excellent local control of STSs of the extremities. However, approximately 25 % of all patients with STSs develop distant metastases [3]. In the case of high-grade tumours specifically, the metastatic rate goes up to approximately 50 % [2]. The lungs are the main site of metastases with approximately 80 % of metastatic cases in STSs of the extremities [4]. The prognosis of patients who develop lung metastases is generally poor, with a 3-year survival rate of 46 % for patients who have undergone surgical resection of lung metastases, and 17 % for patients who did not [5]. Better systemic therapies at earlier stages are thus needed for the management of STSs of the extremities with risk for lung metastases [2]. In this situation, more aggressive chemotherapy regimens or targeted cancer therapy adapted to the histopathology of the tumour could be considered [6]. The specific and early evaluation of lung metastasis risk (or prediction of lung metastases) in the course of STS management is therefore of great interest since it could potentially allow for better adapted treatments and consequently, improve overall survival.

Most tumours do not represent a homogeneous entity, but rather are composed of multiple clonal sub-populations of cancer cells forming complex dynamical systems that exhibit rapid evolution as a result of different therapy perturbations. In solid cancers such as STSs, the tremendous extent of heterogeneous characteristics is expressed at multiple levels. Genes, proteins, cellular microenvironments, tissues and anatomical landmarks within tumours exhibit considerable spatial and temporal variations that could potentially yield valuable information about tumour aggressiveness. However, studying tumour heterogeneity using histopathological samples from biopsies is very difficult since the procedure is invasive and the information obtained may vary depending on which part of the tumour is sampled [7]. This issue is addressed by the new emerging field of "radiomics", which refers to the extraction and analysis of large amounts of information from medical images using advanced quantitative feature analysis [8, 9]. The central hypothesis of *radiomics* is that the genomic heterogeneity of aggressive tumours could translate into heterogeneous tumour metabolism and anatomy, a concept demonstrated by Segal *et al.* [10] and Diehn *et al.* [11], and recently verified by Aerts *et al.* [12]. Diagnostic images could thus reveal important prognostic information about disease risk. In this work, we attempt to quantify intratumoural heterogeneity in STSs using texture analysis performed on 2-deoxy-2-[18F]fluoro-D-glucose (FDG) positron emission tomography (PET) and magnetic resonance imaging (MRI). Figure 3.1 depicts how functional FDG-PET and anatomical MR imaging information together reflect the heterogeneous sub-region characteristics of aggressive STSs.

The texture of an image could be globally defined as the spatial arrangement of pixels of different intensities (i.e., gray levels). Texture analysis is concerned with the quantitative description of the spatial distributions of different gray levels within a region of interest from the extraction of different imaging features. The interest for this method by the *radiomics* community has grown up rapidly in the last few years, as it is considered to have the potential to extensively characterize the complexity of imaging intensity patterns within tumours (i.e., intratumoural heterogeneity). The association of different texture features with different clinical endpoints (i.e., predictive value) have been previously reported in different cancer sites on single patient cohorts: El Naqa *et al.* [13], Tixier *et al.* [14] and Cook *et al.* [15] using texture features extracted from FDG-PET scans, and Vaidya *et al.* [16] using texture features separately extracted from FDG-PET and CT scans. More recently, Aerts *et al.* [12] assessed the prediction performance of a prognostic

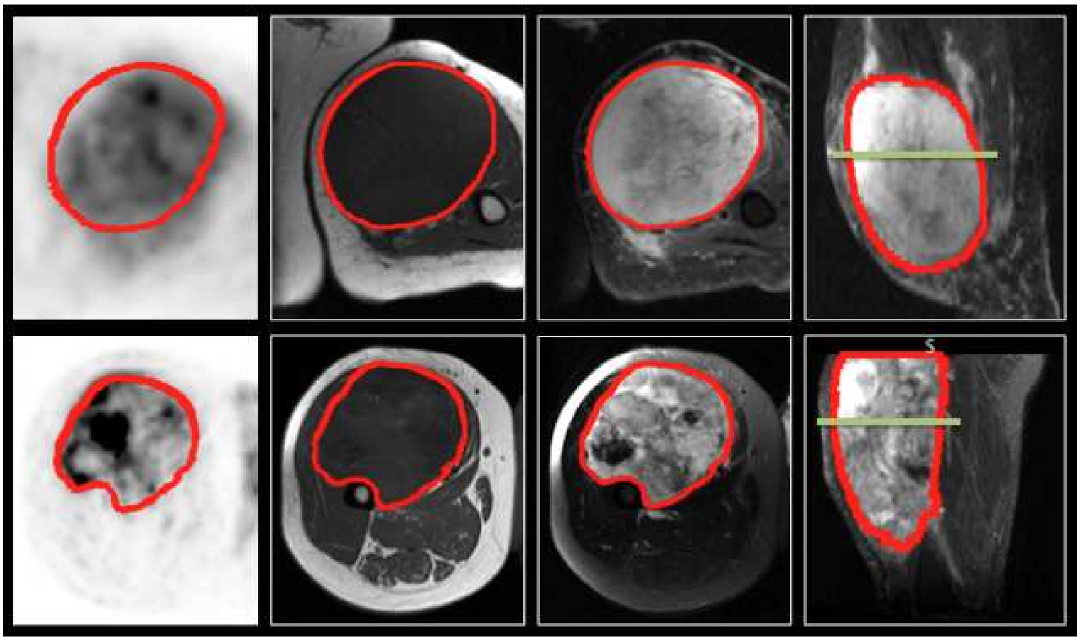


Figure 3.1: FDG-PET and MRI diagnostic images of two patients with soft-tissue sarcomas of the extremities. Top row: patient that did not develop lung metastases. Bottom row: patient that eventually developed lung metastases. 1st column: FDG-PET images, axial plane. 2nd column: T_1-weighted images, axial plane. 3rd column: T_2-weighted fat-saturated images, axial plane. 4th column: short tau inversion recovery images, sagittal plane. The green lines in the two images of the 4th column correspond to the plane shown in the three other respective images.

radiomics signature extracted from CT scans on multiple patient cohorts of different cancer types, whereas Hatt *et al.* [17] evaluated the complementarity of a few texture features with the metabolically active tumour volume extracted from FDG-PET scans.

With the emergence of individualized medicine, a growing need exists for the development of clinically-integrated prediction models that support treatment decision-making [18]. Once useful imaging biomarkers (e.g., texture features) are identified to be relevant prognostic factors of a given tumour outcome, models combining those factors may be constructed to improve outcome prediction performance. Although studies based on single feature response (univariate analysis) can be informative, multivariable models are expected to more comprehensively characterize intratumoural heterogeneity. Considering the high risk of lung metastases in STSs of the extremities and the resulting poor prognosis, the main objective of this work is to develop a joint texture-based multivariable model from pre-treatment FDG-PET and MRI scans for the evaluation of lung metastasis risk at the time of diagnosis of primary STSs of the extremities. This information could eventually assist physicians in their choice of treatment and potentially improve patient survival. Towards this goal, we first investigate if the creation of new composite textures from the combination of FDG-PET and MR imaging information via volume fusion could better identify aggressive tumours. We then develop multivariable modeling strategies for the construction of texture-based models with optimal predictive and generalizability properties from a large number of *radiomics* features. To our knowledge, this is the first study that explores the potential of texture features for the prediction of lung metastases in STS cancer, and the first study that explores the potential of joint FDG-PET and MRI texture features for the assessment of biological properties of any type of cancer.

3.4 Materials and Methods

To ease reading, some acronyms and definitions frequently used in the text are described in Table 3.1.

Table 3.1: Acronyms/definitions used in this study.

Acronym/Definition	Description
STS	Soft-tissue sarcoma
T1	T_1-weighted
T2FS	T_2-weighted fat-supression
LungMets	Patients that developed lung metastases
NoLungMets	Patients that did not develop lung metastases
Separate scans	FDG-PET, T1 and T2FS scans
Fused scans	FDG-PET/T1 and FDG-PET/T2FS scans
SUV	Standard uptake value
Global	First-order histogram
GLCM	Gray-level co-occurence matrix
GLRLM	Gray-level run-length matrix
GLSZM	Gray-level size zone matrix
NGTDM	Neighbourhood gray-tone difference matrix
MRI Inv.	Inversion of MRI intensities
MRI weight	Weight given to MRI sub-bands in the fusion process
WBPF	Wavelet band-pass filtering
Scale	Isotropic voxel size
Quant. algo.	Quantization algorithm
N_g	Number of gray levels
Predictive value	Degree to which a feature is associated to tumour outcome
Degree of freedom	Combination of texture extraction parameter types allowed to vary
r_s	Spearman's rank correlation coefficient
AUC	Area under the receiver operating characteristic curve

3.4.1 Data

Patient cohort

Subsequent to research ethics board (REB) approval, a database of 51 patients with histologically proven primary STSs of the extremities (denoted as "STSs" only for the rest of the text) was retrospectively retrieved. Patients with metastatic and/or recurrent STSs at presentation were excluded from the study. The patient cohort was divided into two groups: i) 32 patients that did not develop lung metastases (denoted as "NoLungMets"; median follow-up time of 33 months, range 12-70 months); and ii) 19 patients that developed lung metastases (denoted as "LungMets"; median follow-up time of 20 months, range 4-31 months) during the follow-up period. Patients from the *NoLungMets* group with a follow-up time smaller than 12 months were excluded from the study. Lung metastases were either proven by biopsy or diagnosed by an expert physician from the appearance of typical pulmonary lesions on CT and/or FDG-PET images. Table 3.2 provides summary characteristics of the patient cohort.

Table 3.2: Characteristics of STS patient cohort.

Characteristic	Type	No. of patients (%), $n = 51$
Sex	Male	24 (47)
	Female	27 (53)
Age (y)	Range	16–83
	Mean ± STD	55 ± 17
Histology	Liposarcoma	11 (21)
	Malignant fibrous histiocytomas	17 (33)
	Leiomyosarcoma	10 (20)
	Synovial sarcoma	5 (10)
	Fibrosarcoma	1 (2)
	Extraskeletal bone sarcoma	4 (8)
	Other	3 (6)
Extremity site	Lower	47 (92)
	Upper	4 (8)
Grade	High	28 (55)
	Intermediate	15 (29)
	Low	5 (10)
	Ungraded	3 (6)
Recurrence/Spread	Distant – Lungs	19 (37)
	Distant – Other than Lungs	6 (12)
	Locoregional	4 (8)
	None	24 (47)
Treatment	Radiotherapy + Surgery	30 (59)
	Surgery + Chemotherapy	7 (14)
	Radiotherapy + Surgery + Chemotherapy	14 (27)

Imaging data

All 51 eligible patients had pre-treatment FDG-PET/CT and MRI scans between November 2004 and November 2011. All FDG-PET/CT scans were performed on a PET/CT scanner (Discovery ST, GE Healthcare, Waukesha, WI) at the McGill University Health Centre (MUHC). For the PET portion of the scans, a median of 420 MBq (range: 210-620 MBq) of FDG was injected intravenously. Approximately 60 min following the injection, whole-body 2D imaging acquisition was performed using multiple bed positions, with a median of 180 s (range: 160-300 s) per bed position. PET attenuation corrected images were reconstructed (axial plane) using an ordered subset expectation maximization (OSEM) iterative algorithm. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 5.47×5.47 mm^2 (range: 3.91-5.47 mm).

The MRI scans resulted from clinical acquisitions with non-uniform protocols across patients. Twelve patients had their images acquired at the MUHC, and 39 in an outside institution. Three types of MRI sequences routinely used in clinical protocols were selected for the study, namely T_1-weighted (T1), T_2-weighted fat-saturated and short tau inversion recovery (STIR) sequences. Overall, the median in-plane resolution was 0.74×0.74 mm^2, 0.63×0.63 mm^2 and 0.86×0.86 mm^2 (range: 0.23-1.64 mm, 0.23-1.64 mm and 0.23-1.72 mm pixel width), and the median slice thickness was 5.5 mm, 5.0 mm and 5.0 mm (range: 3.0-10.0 mm, 3.0-8.0 mm and 3.0-10.0 mm) for T_1-weighted, T_2-weighted fat-saturated and STIR scans, respectively. T_1-weighted sequences were acquired in the axial plane for all patients. On the other hand, patients were scanned in different planes with either or both T_2-weighted fat-saturated and STIR sequences, which macroscopically are both T_2-weighted sequences aiming to supress the fat signal in the body. From a texture point of view, T_2-weighted fat-saturated and STIR images are considered similar, and they were thus combined in the same scan category with only one of the two sequences used per patient. T_2-weighted fat-saturated scans were selected by default due to their higher axial scan availability ($n = 26$). When T_2-weighted fat-saturated scans were not available, STIR scans were used ($n = 25$). For the rest of the text, this category of scans is referred to as T2FS (T_2-weighted fat-suppression) scans

3.4.2 Tumour volume definition

Contours defining the 3D tumour region for each patient were manually drawn slice-by-slice on T2FS scans by an expert radiation oncologist. For patients with visible edema in the vicinity of the tumours ($n = 32$), two contours were drawn: one incorporating the visible edema and one excluding it, as shown in Figure 3.2. Contours were propagated to FDG-PET and T1 scans using rigid registration with the commercial software MIM® (MIM software Inc., Cleveland, OH). The results presented in this work were obtained from texture analysis performed on the volume of interest of each patient as defined by the contour containing no edema.

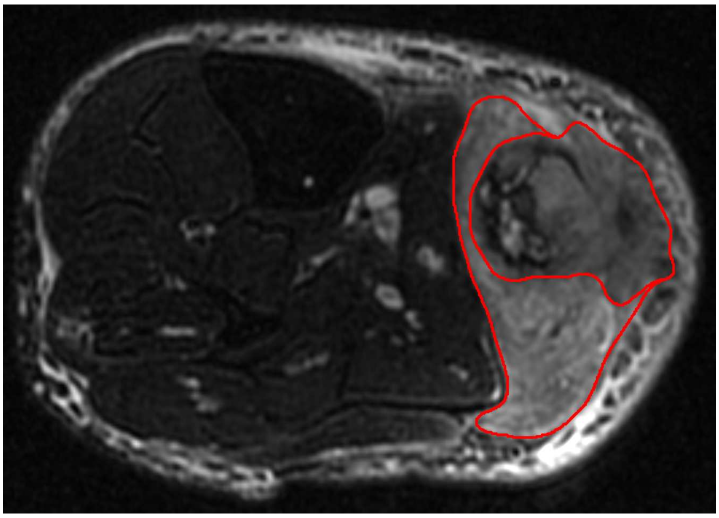


Figure 3.2: Example of soft-tissue sarcoma tumour volume definition performed on the T_2-weighted fat-saturated scan of a patient of the *LungMets* group. The inner contour exclude visible edema in the vicinity of the tumour.

3.4.3 Imaging data pre-processing

Prior to texture analysis, FDG-PET and MRI DICOM data were transferred into MATLAB® (The MathWorks Inc., Natick, MA) format using the software CERR [19]. All subsequent analyses were performed in MATLAB®. FDG-PET scans were first converted to standard uptake value (SUV) maps, followed by the application of a square-root transform to help in the stabilization of the PET noise in the images. MRI scans were kept in raw data form, and voxels within the tumour region with intensities outside the range $\mu \pm 3\sigma$ were rejected and not considered in subsequent texture computations, as suggested by Collewet *et al.* [20] for making MRI texture measurements more reliable.

3.4.4 FDG-PET/MRI volume fusion

The fusion of FDG-PET and MRI volumes first starts with the registration of the scans as described in section 3.4.2. The 3D discrete wavelet transform (DWT) is then used to combine the spatial and frequency characteristics of the two modalities as follows:

1. Downsample the MRI volume (raw data, no pre-processing) to the resolution of the FDG-PET volume (pre-processed, see section 3.4.3) using cubic interpolation. Normalize the intensity range of FDG-PET and MRI tumour regions between 0 and 255. Invert MRI intensities if needed.

2. Apply the 3D DWT to the FDG-PET and MRI volumes up to one decomposition level using the wavelet basis function *symlet8*.

3. Apply the $\mu \pm 3\sigma$ normalization scheme (see section 3.4.3) respectively to the wavelet coefficients of the tumour region of the different MRI sub-bands. The rejected MRI wavelet coefficients are then replaced by the spatially corresponding coefficients of the FDG-PET sub-bands.

4. Perform a weighted average of the spatially corresponding wavelet coefficients of all PET and MRI sub-bands to obtain a single set of fused wavelet coefficients. If the weight given to MRI wavelet coefficients is denoted as "MRI weight" and the weight given to PET wavelet coefficients is denoted as "PET weight", *MRI weight* ranges from 0 to 1 and $\textit{PET weight} = 1 - \textit{ MRI weight}$.

5. Apply the 3D inverse DWT to the set of fused wavelet coefficients using the reconstruction wavelet basis function *symlet8* to obtain a fused FDG-PET/MRI tumour volume.

The choice of the wavelet basis function *symlet8* is based on our previous work [21], in which that basis function was shown to produce fused textures with best predictive value. This could be explained from the fact that *symlets* is a family of orthogonal and compactly supported wavelets with the least asymmetry and highest number of vanishing moments for a given support width, which would help in the local preservation of spatial characteristics of images.

The fusion process yields two new types of scans: FDG-PET/T1 and FDG-PET/T2FS. We also tested if the inversion of MRI intensities prior to fusion with FDG-PET could enhance texture characteristics in the fused volumes. Figure 3.3 shows an example of the fusion of FDG-PET and T2FS scans for a patient of the *LungMets* group.

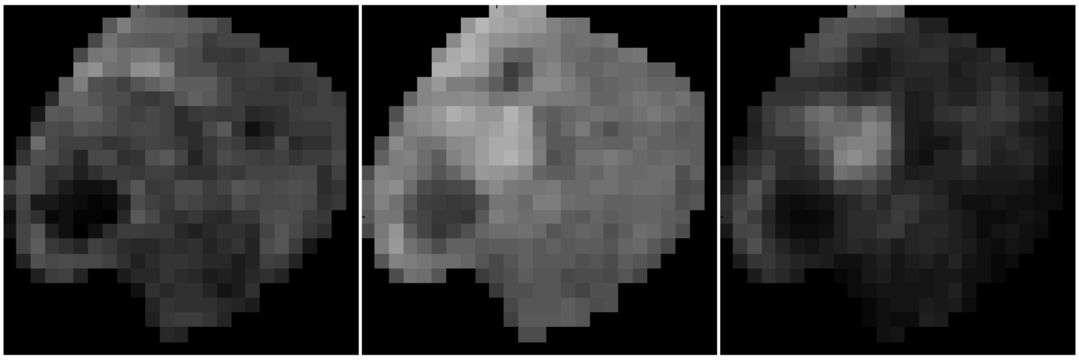


Figure 3.3: Example fusion of PET and MR images. Fusion (middle) of a T_2-weighted fat-saturated scan (left) and a FDG-PET scan (right) with a *MRI weight* value of 0.5, for a patient of the *LungMets* group. The T2-weighted fat-saturated scan was registered and downsampled to the FDG-PET scan resolution and is presented in raw data format (no pre-processing). The FDG-PET scan is presented in pre-processed format. The intensity range of the 3D tumour region of the three scans was normalized between 0 and 255.

3.4.5 Feature extraction

The methodology used to extract the imaging features from the tumour region of the pre-treatment FDG-PET and MRI scans is described below.

Non-texture features

In total, nine non-texture features were extracted for completeness.

SUV metrics (5). Five non-texture features were extracted from the tumour region of the FDG-PET scans.

1. *SUVmax*: Maximum SUV of the tumour region.
2. *SUVpeak*: Average of the voxel with maximum SUV within the tumour region and its 26 connected neighbours.
3. *SUVmean*: Average SUV value of the tumour region.
4. *AUC-CSH*: Area under the curve of the cumulative SUV-volume histogram describing the percentage of total tumour volume above a percentage threshold of maximum SUV [22].

5. *Percent Inactive*: Percentage of the tumour region that is inactive. A threshold of $0.005 \times (SUVmax)^2$ followed by closing and opening morphological operations were used to differentiate active and inactive regions on FDG-PET scans.

Volume. Number of voxels in the tumour region extracted from T2FS scans multiplied by the dimension of voxels.

Size. Longest diameter of the tumour region extracted from T2FS scans.

Solidity. Ratio of the number of voxels in the tumour region to the number of voxels in the 3D convex hull of the tumour region (smallest polyhedron containing the tumour region). This metric is extracted from T2FS scans.

Eccentricity. The ellipsoid that best fits the tumour region is first computed using the framework of Li & Griffiths [23]. The eccentricity is then given by $[1 - a \times b/c^2]^{1/2}$, where c is the longest semi-principal axes of the ellipsoid, and a and b are the second and third longest semi-principal axes of the ellipsoid.

Texture features

In total, 41 texture features were extracted from of the tumour regions of 5 different types of scans: FDG-PET, T1 and T2FS scans (“separate scans”), and FDG-PET/T1 and FDG-PET/T2FS scans (“fused scans”). Table 3.3 presents the list of texture features used in this study. *Global* features are extracted from the intensity histogram of the tumour region, whereas GLCM, GLRLM, GLSZM and NGTDM textures are matrix-based features. In this work, histograms with 100 bins were used for the computation of *Global* features. GLCMs, GLRLMs, GLSZMs and NGTDMs were constructed using 3D analysis of the tumour region with 26-voxel connectivity. Only one GLCM, GLRLM, GLSZM and NGTDM was computed per scan by simultaneously taking into account the neighbouring properties of voxels in the 13 directions of 3D space. However, the 6 voxels at a distance of 1 voxel, the 12 voxels at a distance of $\sqrt{2}$ voxels, and the 8 voxels at a distance of $\sqrt{3}$ voxels around center voxels were treated differently in the calculations of the GLCMs, the GLRLMs and the NGTDMs in order to take into account discretization length differences (assuming that resampling to isotropic voxel size is applied beforehand, see next sub-section). Detailed description and methodology employed to extract the 41 texture features is available in Supplementary Material 3.10.1.

Texture extraction parameters

The influence of the following six extraction parameters on the predictive value of textures was investigated.

Fusion parameters (2). The following two parameters apply only to fused scans (FDG-PET/T1 and FDG-PET/T2FS scans):

1. Inversion of MR imaging intensities in the FDG-PET/MRI fusion process (see section 3.4.4). This parameter is denoted as "“MRI Inv.” *MRI Inv.* of 0 and 1 (no inversion/inversion) were tested.

2. Weight applied to MRI wavelet sub-bands in the FDG-PET/MRI fusion process (see section 3.4.4). This parameter is denoted as “MRI weight”. *MRI weight* values of $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, $\frac{2}{3}$ and $\frac{3}{4}$ were tested in this work.

Wavelet band-pass filtering (1). This parameter is denoted as “WBPF”. This operation is carried out by applying a different weight to band-pass sub-bands (LHL, LHH, LLH,HLL, HHL and HLH) of the tumour region as compared to low- and high-frequency sub-bands (LLL and HHH) in the wavelet domain. The ratio of the weight applied to band-pass sub-bands to the weight applied to the other sub-bands is defined by R. Ratios of $\frac{1}{2}$, $\frac{2}{3}$, 1, $\frac{3}{2}$ and 2 were tested.

Isotropic voxel size (1). This parameter is denoted as “Scale”. Prior to the computation of texture features, all volumes were resampled to an isotropic voxel size set to a desired resolution using cubic interpolation. *Scale* values of 1 mm, 2 mm, 3 mm, 4 mm, 5 mm and initial in-plane resolution (denoted “in-pR”) were tested. For example, if the desired *Scale* was set to 5 mm, a FDG-PET volume with voxel size of $5.47 \times 5.47 \times 3.27$ mm^3 would be isotropically resampled to a voxel size of $5{\times}5{\times}5$ mm^3. If the desired Scale was set to *in-pR*, a MRI volume with voxel size of $0.86 \times 0.86 \times 5$ mm^3 would be isotropically resampled to a voxel size of $0.86 \times 0.86 \times 0.86$ mm^3. Note that *Global* texture features are extracted after isotropically resampling to the in-plane resolution without further processing.

Table 3.3 : Texture features used in this study.

Texture type	Reference(s)	Texture name
Global	–	Variance Skewness Kurtosis
GLCM[a]	Haralick *et al.* [24]	Energy Contrast Correlation Homogeneity Variance Sum Average Entropy
GLRLM[b]	Galloway [25]	Short Run Emphasis (SRE) Long Run Emphasis (LRE) Gray-Level Non-uniformity (GLN) Run-Length Non-uniformity (RLN) Run Percentage (RP)
	Chu *et al.* [26]	Low Gray-Level Run Emphasis (LGRE) High Gray-Level Run Emphasis (HGRE)
	Dasarathy & Holder [27]	Short Run Low Gray-Level Emphasis (SRLGE) Short Run High Gray-Level Emphasis (SRHGE) Long Run Low Gray-Level Emphasis (LRLGE) Long Run High Gray-Level Emphasis (LRHGE)
	Thibault *et al.* [28]	Gray-Level Variance (GLV) Run-Length Variance (RLV)
GLSZM[c]	Galloway [25] Thibault *et al.* [28]	Small Zone Emphasis (SZE) Large Zone Emphasis (LZE) Gray-Level Non-uniformity (GLN) Zone-Size Non-uniformity (ZSN) Zone Percentage (ZP)
	Chu *et al.* [26] Thibault *et al.* [28]	Low Gray-Level Zone Emphasis (LGZE) High Gray-Level Zone Emphasis (HGZE)
	Dasarathy & Holder [27] Thibault *et al.* [28]	Small Zone Low Gray-Level Emphasis (SZLGE) Small Zone High Gray-Level Emphasis (SZHGE) Large Zone Low Gray-Level Emphasis (LZLGE) Large Zone High Gray-Level Emphasis (LZHGE)
	Thibault *et al.* [28]	Gray-Level Variance (GLV) Zone-Size Variance (ZSV)
NGTDM[d]	Amadasun & King [29]	Coarseness Contrast Busyness Complexity Strength

[a] GLCM: Gray-level co-occurence matrix.

[b] GLRLM: Gray-level run-length matrix.

[c] GLSZM: Gray-level size zone matrix.

[d] NGTDM: Neighbourhood gray-tone difference matrix.

Quantization of gray levels (2). Prior to the computation of texture features, the full intensity range of the tumour region was quantized to a smaller number of gray levels N_g. The quantization process maps the voxel values of a volume to a finite set $\mathbf{r} = \{r_k \in \mathbb{R} : k = 1, 2, \ldots, N_g\}$ of reconstruction levels by defining a set $\mathbf{t} = \{t_k \in \mathbb{R} : k = 1, 2, \ldots, N_g + 1\}$ of decision levels. Two extraction parameters are related to the quantization of gray levels in the volumes:

1. Quantization algorithm. This parameter is denoted as "Quant. algo." Equal-probability and Lloyd-Max quantization algorithms were implemented in this work using the functions *histeq* and *lloyds* of MATLAB, respectively. Equal-probability quantization attempts to define decision thresholds in the volume such that the number of voxels with reconstructed level r_k is the same in the quantized volume for all k (i.e., for all gray levels), whereas Lloyd-Max quantization attempts to minimize the mean-squared quantization error of the output [30, 31].

2. Number of gray levels (N_g) in the quantized volume. N_g of 8, 16, 32 and 64 were tested.

Texture extraction summary. Considering the full set of texture extraction parameters of *Global* features and higher-order texture features (GLCM, GLRLM, GLSZM and NGTDM), a total of $27\,405$ and $182\,700$ scan-texture-parameter combinations were computed in this work for separate and fused scans, respectively. Figure 3.4 presents a summary of the workflow of extraction of texture features.

3.4.6 Univariate analysis

Univariate association between the whole set of features (9 non-texture features and $210\,105$ scan-texture-parameter features) and lung metastases development in STSs was investigated using Spearman's rank correlation (r_s). To correct for multiple test comparisons, the Bonferroni correction method was applied: the significance level was lowered to a value $p < \alpha/K$, where K is the number of comparisons and α the significance level set to 0.05.

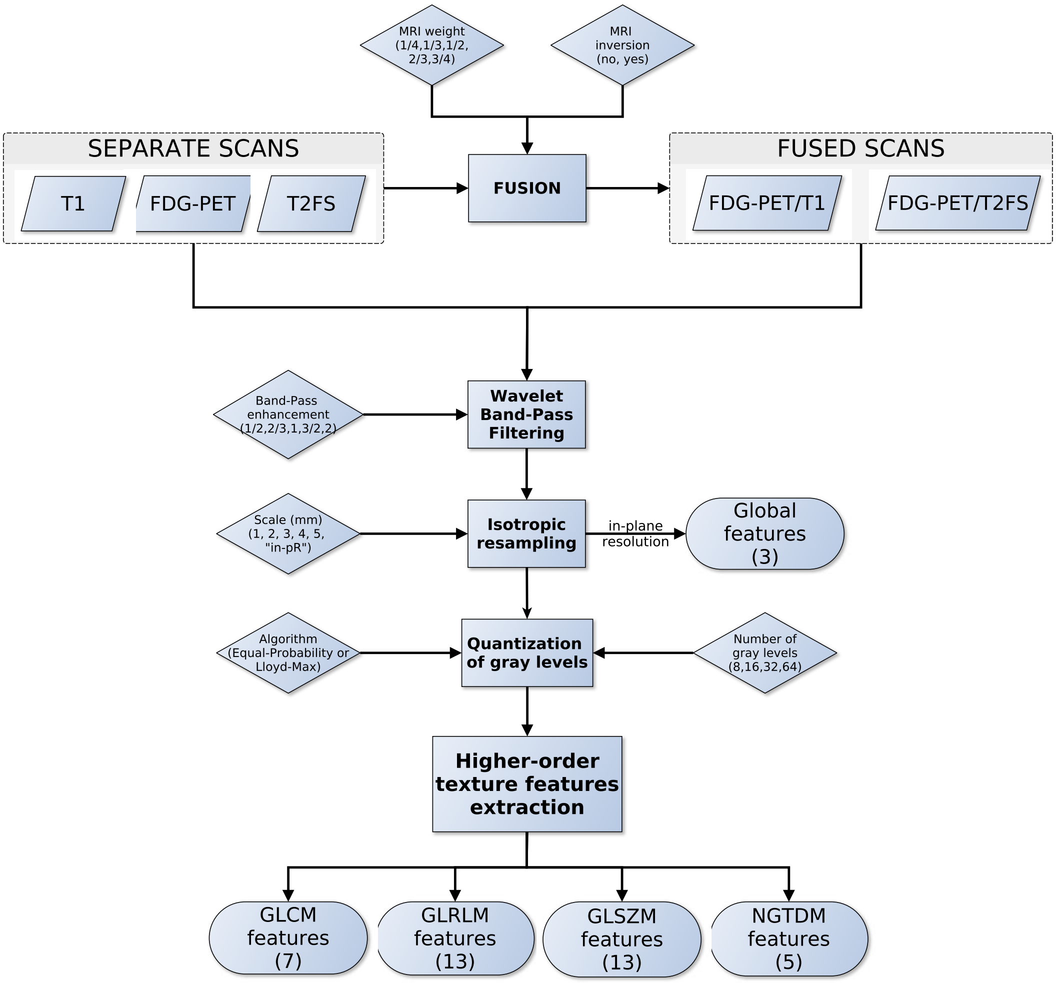

MRI weight (1/4,1/3,1/2, 2/3,3/4)
MRI inversion (no, yes)
SEPARATE SCANS
T1
FDG-PET
T2FS
FUSION
FUSED SCANS
FDG-PET/T1
FDG-PET/T2FS
Band-Pass enhancement (1/2,2/3,1,3/2,2)
Wavelet Band-Pass Filtering
Scale (mm) (1, 2, 3, 4, 5, "in-pR")
Isotropic resampling
in-plane resolution
Global features (3)
Algorithm (Equal-Probability or Lloyd-Max)
Quantization of gray levels
Number of gray levels (8,16,32,64)
Higher-order texture features extraction
GLCM features (7)
GLRLM features (13)
GLSZM features (13)
NGTDM features (5)

Figure 3.4: Workflow of extraction of texture features.

3.4.7 Multivariable analysis

The process of combining features into a multivariable model was achieved using the logistic regression utilities of the software DREES [32]. We are interested in finding a linear combination of p variables such that the multivariable model of interest takes the form:

$$g(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } i = 1, 2, \ldots, N, \tag{3.1}$$

where the vector of input variables (imaging data) of the i^{th} patient is $\mathbf{x}_i = \{x_{ij} \in \mathbb{R} : j = 1, 2, \ldots, p\}$, for a number N of patients. The set $\boldsymbol{\beta} = \{\beta_j \in \mathbb{R} : j = 0, 1, \ldots, p\}$ is the set of regression coefficients of the model to be determined such that the conditional probability of the set of outcome states {0,1} given the input data $\mathbf{x}_i$ is maximized for $i = 1, 2, \ldots, N$. This operation is carried out using a logistic regression model (logit transformation) of the form:

$$\pi(\mathbf{x}_i) = \mathrm{P}\left(y_i = 1 | \mathbf{x}_i\right) = \frac{\exp\left[g(\mathbf{x}_i)\right]}{1 + \exp\left[g(\mathbf{x}_i)\right]}, \quad \text{for } i = 1, 2, \ldots, N. \tag{3.2}$$

Following the work by [33], we adopted the 0.632+ bootstrap method and the area under the receiver operating characteristic curve (AUC) metric to estimate which models learned from our patient cohort would best predict lung metastases on new prospective data from the whole (or true) STS population. Let $\mathrm{AUC}(\mathbf{S}_{\text{train}}, \mathbf{S}_{\text{test}})$ denote the value of the test AUC obtained when the classifier is trained on set $\mathbf{S}_{\text{train}}$ (computing logistic regression coefficients) and tested in set $\mathbf{S}_{\text{test}}$ (testing $g(\mathbf{x}_i)$). Also, let the observed sample (imaging data of our patient cohort) be denoted as the matrix $\mathbf{X} = \{\mathbf{x}_i : i = 1, 2, \ldots, N\}$. A bootstrap sample denoted as $\mathbf{X}^* = \{\mathbf{x}_i^* : i = 1, 2, \ldots, N\}$ is a sample of input variables $\mathbf{x}_i$ of N patients randomly drawn with replacement from the available sample $\mathbf{X}$. The set of original data vectors that do not appear in $\mathbf{X}^*$ is denoted as $\mathbf{X}^*(0)$. The generation of a large number B of randomly drawn bootstrap samples $\mathbf{X}^{*b}$ for $b = 1, 2, \ldots, B$ is used to estimate a statistical quantity of interest on the unknown true population distribution. With the 0.632+ bootstrap method, the estimated AUC is then calculated as:

$$
\begin{aligned}
\widehat{\mathrm{AUC}}_{0.632+} &= \frac{1}{B}\sum_{b=1}^{B}\left[(1-\alpha(b))\cdot \mathrm{AUC}\big(\mathbf{X},\mathbf{X}\big) + \alpha(b)\cdot \mathrm{AUC}'\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)\right],\\
\mathrm{AUC}'\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big) &= \max\left\{0.5, \mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)\right\},\\
\alpha(b) &= \frac{0.632}{1-0.368\cdot R(b)},\\
R(b) &= \begin{cases} 1 & \text{if } \mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big) \le 0.5,\\ \dfrac{\mathrm{AUC}\left(\mathbf{X},\mathbf{X}\right)-\mathrm{AUC}\left(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\right)}{\mathrm{AUC}\left(\mathbf{X},\mathbf{X}\right)-0.5} & \text{if } 2 > \dfrac{\mathrm{AUC}\big(\mathbf{X},\mathbf{X}\big)}{\mathrm{AUC}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)} > 1,\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}
\tag{3.3}
$$

In this work, each time a bootstrap sample $\mathbf{X}^{*b}$ was drawn from $\mathbf{X}$ in the multivariable analysis, the probability of choosing a negative instance (*NoLungMets* patient group class) was made equal to the probability of choosing a positive instance (*LungMets* patient group class), a procedure hereby denoted as “imbalance-adjusted bootstrap resampling”.

Prediction models were constructed for three different types of initial feature sets: i) 9 non-texture features + 9 135 FDG-PET scan-texture-parameter features; ii) 9 non-texture features + 27 405 separate FDG-PET and MRI scan-texture-parameter features; and iii) 9 non-texture features + 182 700 fused FDG-PET/MRI scan-texture-parameter features. First, feature set reduction was performed through a stepwise forward feature selection scheme in order to create reduced feature sets containing 25 different scan-texture features from larger initial sets, a procedure carried out using the *Gain* equation:

$$
\begin{aligned}
\widehat{\mathrm{Gain}}_j &= \gamma\cdot|\widehat{r}_s(\mathbf{x}_j,\mathbf{y})|\\
&\quad + \delta_a\cdot\left[\sum_{k=1}^{f}\left(\frac{2(f-k+1)}{f(f+1)}\right)\widehat{\mathrm{PIC}}(\mathbf{x}_k,\mathbf{x}_j)\right]\\
&\quad + \delta_b\cdot\left[\frac{1}{F}\sum_{l=1}^{F}\widehat{\mathrm{PIC}}(\mathbf{x}_l,\mathbf{x}_j)\right],\\
&\quad \text{where}\quad \widehat{r}_s(\mathbf{x}_j,\mathbf{y}) = \frac{1}{B}\sum_{b=1}^{B} r_s(\mathbf{x}_j^{*b},\mathbf{y}),\\
&\quad \text{and}\quad \widehat{\mathrm{PIC}}(\mathbf{x}_k,\mathbf{x}_j) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{PIC}(\mathbf{x}_k^{*b},\mathbf{x}_j^{*b}).
\end{aligned}
\tag{3.4}
$$

In Equation 3.4, $r_s(\mathbf{x}_j, \mathbf{y})$ is the Spearman's rank correlation computed between feature j defined as $\mathbf{x}_j = \{x_{ij} \in \mathbb{R} : i = 1, 2, \ldots, N\}$ and the outcome vector $\mathbf{y} = \{y_i \in \{0 : NoLungMets, 1 : LungMets\} : i = 1, 2, \ldots, N\}$. $\text{PIC}(\mathbf{x}_k, \mathbf{x}_j)$ is the potential information coefficient defined as $\text{PIC}(\mathbf{x}_k, \mathbf{x}_j) = 1 - \text{MIC}(\mathbf{x}_k, \mathbf{x}_j)$, where $\text{MIC}(\mathbf{x}_k, \mathbf{x}_j)$ is the maximal information coefficient between feature j and k as defined by Reshef *et al.* [34]. The sum over k is a sum over all f features that have already been chosen to be part of the reduced feature set (employed in forward selection schemes), whereas the sum over l is a sum over all F features that have not yet been removed from a larger initial set (employed in backward selection schemes). The sum over the k features is always done in order of appearance of the different features in the reduced set in order to favour the features from the larger initial set with the least dependence with the features chosen first in the reduced set. In this work, γ was set to 0.5, δ_a to 0.5 and δ_b to 0. Every time a new feature had to be chosen in the reduced set from a larger initial set, a new *Gain* was calculated for all remaining features in the larger initial set using imbalance-adjusted bootstrap resampling (1000 samples). Note that Equation 3.4 allows to rank specific scan-texture-parameter features, as part 1 of the *Gain* equation uses Spearman's rank correlations varying over the whole set of texture extraction parameters. However, to speed up calculations, average scan-texture features over all texture extraction parameters were used in part 2 (and 3 if needed) of the *Gain* equation.

From the reduced feature sets, stepwise forward feature selection was then carried out by maximizing the 0.632+ bootstrap AUC. For a given model order and a given reduced feature set, the feature selection step was divided into 25 experiments. In each of these experiments, a different feature from the reduced set was used as a different "starter". For a given *starter*, 1000 logistic regression models $g(\mathbf{x}_i)$ or order 2 were first created using imbalance-adjusted bootstrap resampling (1000 samples) for each of the remaining features in the reduced feature set. Then, the single remaining feature that maximized the 0.632+ bootstrap AUC of the models was chosen, and the process was repeated up to model order 10. Finally, for each model order, the experiment that yielded the highest 0.632+ bootstrap AUC was identified, and combination of features were chosen for model orders of 1 to 10 (only features of the models were selected, but logistic regression coefficients were not yet computed).

The feature reduction and feature selection processes were repeated for all possible combinations of texture extraction parameter types being allowed to vary or not (i.e., degrees of freedom): 2^4 combinations in the case of the initial feature set containing textures extracted from FDG-PET scans and the initial feature set containing textures extracted from separate scans, and 2^6 combinations in the case of the initial feature set containing textures extracted from fused scans. For example, in an experiment where only *MRI weight* and *Quant. algo.* texture extraction parameters were allowed to vary, the FDG-PET initial feature set contained 9 non-texture features + 79 scan-texture-parameter features, the separate scan initial feature set contained 9 non-texture features + 237 scan-texture-parameter features, and the fused scan initial feature set contained 9 non-texture features + 790 scan-texture-parameter features. Then, separately for each model order of 1 to 10 for the three feature sets, the degree of freedom on texture extraction parameters yielding the multivariable models with the highest 0.632+ bootstrap AUC was found. For the experiments in which specific extraction parameters were not allowed to vary, baseline parameters had to be defined. Table 3.4 presents the baseline texture extraction parameters used for the five different types of scans.

Table 3.4: Baseline texture extraction parameters.

MRI Inv.	*MRI weight*	*WBPF*	*Scale*	*Quant. algo.*	N_g
No	$\frac{1}{2}$	$R = 1$	*in-pR*	Lloyd-Max	32

MRI Inv.: MRI Inversion, *MRI weight*: weight in FDG-PET/MR fusion process, *WBPF*: wavelet band-pass filtering, *Scale*: isotropic voxel size, *Quant. algo.*: quantization algorithm, N_g: Number of gray-levels, R: ratio of the weight applied to band-pass sub-bands to the weight applied to low- and high-frequency sub-bands. *in-pR*: in-plane resolution.

Once optimal combination of features were identified for model orders of 1 to 10 for the three different types of feature sets, imbalance-adjusted bootstrap resampling (1000 samples) was again performed for all models. Prediction performance was then estimated using the 0.632+ bootstrap method in terms of AUC as defined in Equation 3.3, and in terms of sensitivity and specificity metrics as defined in Equation 3.5. Using the prediction estimates for the three initial feature sets, a single combination of features possessing the best parsimonious properties was then determined.

$$\begin{aligned}
\widehat{\mathrm{S}}_{0.632+} &= \frac{1}{B}\sum_{b=1}^{B}\left[(1-\alpha(b))\cdot \mathrm{S}\big(\mathbf{X},\mathbf{X}\big) + \alpha(b)\cdot \mathrm{S}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)\right],\\
\alpha(b) &= \frac{0.632}{1-0.368\cdot R(b)},\\
R(b) &= \begin{cases} \dfrac{\mathrm{S}\left(\mathbf{X},\mathbf{X}\right)-\mathrm{S}\left(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\right)}{\mathrm{S}\left(\mathbf{X},\mathbf{X}\right)} & \text{if } \dfrac{\mathrm{S}\big(\mathbf{X},\mathbf{X}\big)}{\mathrm{S}\big(\mathbf{X}^{*b},\mathbf{X}^{*b}(0)\big)} > 1,\\ 0 & \text{otherwise,}\end{cases}\\
&\text{for } \mathrm{S}:\mathrm{Sensitivity},\mathrm{Specificity}.
\end{aligned} \tag{3.5}$$

The last step in the construction of the final prediction model was to compute the coefficients of the optimal combination of features using imbalance-adjusted bootstrap resampling (1000 samples). Let the logistic regression coefficient of feature j computed in a bootstrap sample $\mathbf{X}^{*b}$ and modeling an outcome vector $\mathbf{y}$ be denoted as $\beta_j(\mathbf{X}^{*b},\mathbf{y})$ for $j = 0, 1, \ldots, p$, where p is the multivariable model order and $j = 0$ refers to the offset of the model $g(\mathbf{x}_i)$. The computation of the different coefficient estimates of the final prediction model was then performed as follows:

$$\widehat{\beta_j} = \frac{1}{B}\sum_{b=1}^{B}\beta_j(\mathbf{X}^{*b},\mathbf{y}), \quad \text{for } j = 0, 1, \ldots, p. \tag{3.6}$$

Figure 3.5 summarizes the workflow of multivariable analysis.

3.5 Results

3.5.1 Univariate analysis

Table 3.5 presents the Spearman's rank correlation (r_s) between the nine non-texture features and lung metastases development in STSs. Table 3.6 presents the Spearman's rank correlation between the 205 different scan-texture features and lung metastases development in STSs. For each entry in Table 3.6, two values appear: the r_s in the case of texture features extracted using baseline parameters as defined in Table 3.4 (left), and the maximal r_s in the case of texture features extracted using the optimal set of extraction parameters when all parameters are allowed to vary, i.e., with full degrees of freedom (right). In Table 3.5 and Table 3.6, the values in italic font indicate features for

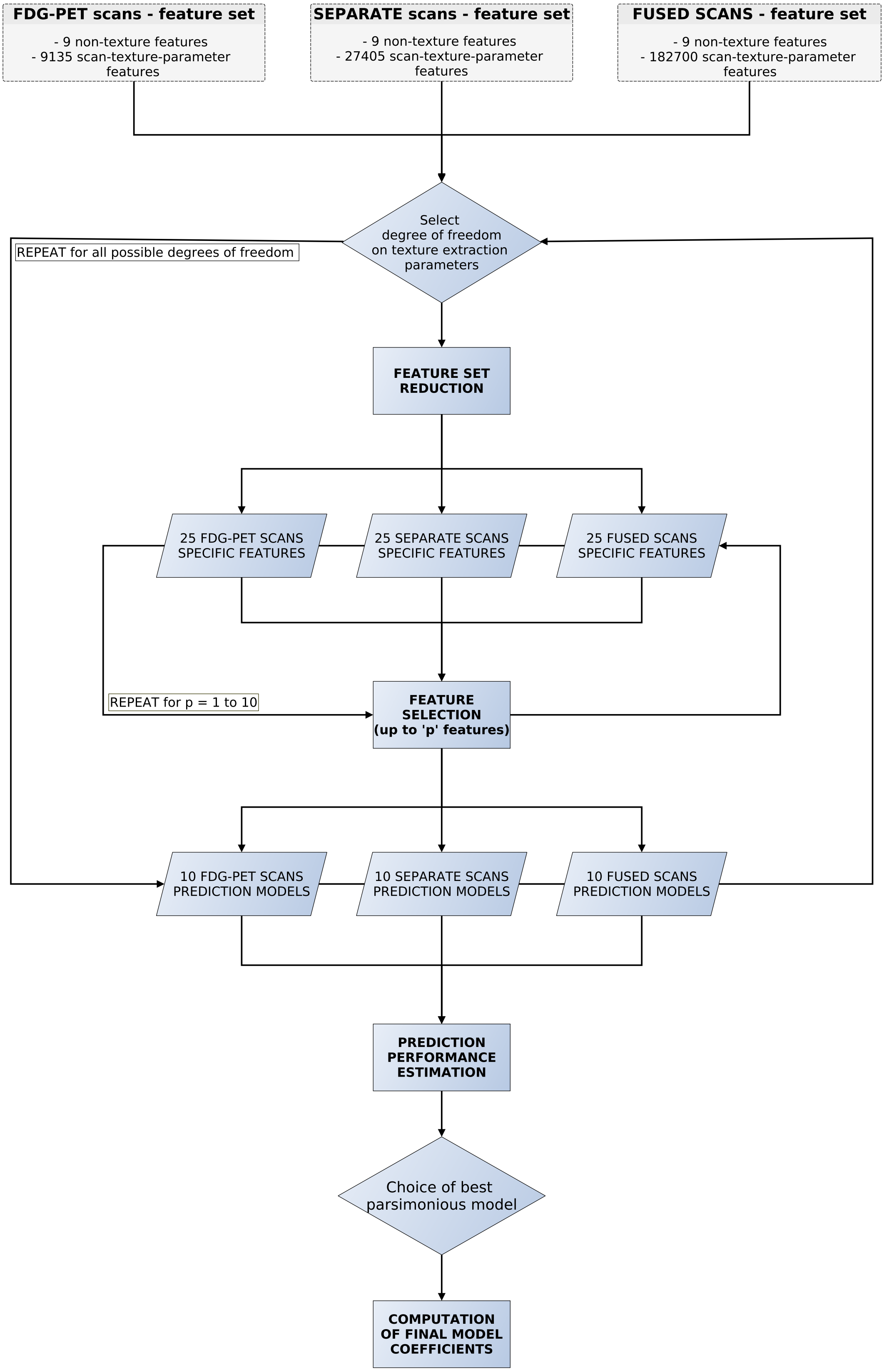

FDG-PET scans - feature set
- 9 non-texture features
- 9135 scan-texture-parameter features
SEPARATE scans - feature set
- 9 non-texture features
- 27405 scan-texture-parameter features
FUSED SCANS - feature set
- 9 non-texture features
- 182700 scan-texture-parameter features
Select degree of freedom on texture extraction parameters
REPEAT for all possible degrees of freedom
FEATURE SET REDUCTION
25 FDG-PET SCANS SPECIFIC FEATURES
25 SEPARATE SCANS SPECIFIC FEATURES
25 FUSED SCANS SPECIFIC FEATURES
REPEAT for p = 1 to 10
FEATURE SELECTION (up to 'p' features)
10 FDG-PET SCANS PREDICTION MODELS
10 SEPARATE SCANS PREDICTION MODELS
10 FUSED SCANS PREDICTION MODELS
PREDICTION PERFORMANCE ESTIMATION
Choice of best parsimonious model
COMPUTATION OF FINAL MODEL COEFFICIENTS

Figure 3.5: Workflow of multivariable analysis.

which $p < \alpha/K < 0.05/(9$ non-texture features + 5 scans × 41 textures × 2 extraction parameter degrees of freedom) ≈ 0.00012 according to the correction for multiple testing comparisons.

Non-texture features

Table 3.5: Spearman's rank correlation (r_s) between non-texture features and lung metastases development in STSs. The values in italic font indicate features for which $p < 0.05/419 \approx 0.00012$.

Feature	r_s	p-value
SUVmax	*0.52*	*0.0001*
Percent Inactive	*0.51*	*0.0001*
SUVpeak	0.5	0.0002
AUC-CSH	−0.29	0.04
Volume	0.28	0.04
SUVmean	0.28	0.04
Solidity	0.24	0.09
Size	0.18	0.19
Eccentricity	−0.17	0.25

AUC-CSH: Area under the curve of the cumulative SUV-volume histogram.

In Table 3.5, it can be seen that the non-textural features that are highly correlated with lung metastases are *SUVmax* and *Percent Inactive*. Note the positive signs of r_s for these two features.

Texture features

In Table 3.6, it can be seen that texture features extracted from FDG-PET scans generally have a higher predictive value than the texture features extracted from MRI scans. The results in Table 3.6 also reveal that textures extracted from fused scans generally have a higher predictive value than those extracted from separate scans. Moreover, it can be seen that different extraction parameters significantly impact the predictive value of the resulting textures. The extraction parameters used to produce the texture features with the highest predictive value for the five different types of scans are presented in Supplementary Material 3.10.2.

Texture extraction parameter effect

The Wilcoxon rank sum test was performed between the set of absolute Spearman's rank correlation coefficients obtained from each of the 41 different texture features extracted using baseline extraction parameters, and the sets of

Table 3.6: Spearman's rank correlation between texture features and lung metastases development in STSs using *baseline* | *optimal* texture extraction parameters. The values in italic font indicate features for which $p < 0.05/419 \approx 0.00012$.

Type	Texture	FDG-PET		T1		T2FS		FDG-PET/T1		FDG-PET/T2FS	
Global	Variance	0.12	0.13	0.21	0.21	0.03	0.05	−0.06	−0.31	−0.44	−0.49
	Skewness	0.23	0.25	−0.36	−0.36	0.16	0.17	0.13	0.28	0.28	0.39
	Kurtosis	−0.02	0.06	−0.28	−0.31	−0.15	−0.15	−0.11	−0.24	0.33	0.42
GLCM	Energy	−0.01	0.49	−0.22	−0.44	0.14	0.23	−0.04	*−0.51*	0.37	*0.53*
	Contrast	−0.14	−0.44	−0.08	−0.33	−0.15	0.33	−0.19	*−0.52*	−0.36	*−0.51*
	Entropy	0.01	−0.43	0.12	0.41	−0.15	0.22	−0.02	−0.46	−0.38	*−0.55*
	Homogeneity	0.28	0.49	0.04	0.26	0.18	0.26	0.23	*0.53*	0.42	*0.51*
	Correlation	0.28	0.42	0.18	0.32	0.15	0.25	0.26	0.42	0.10	0.38
	Sum Average	−0.35	*−0.52*	0.28	*0.51*	−0.11	−0.29	−0.18	*−0.52*	−0.27	*−0.52*
	Variance	0.31	−0.43	0.32	0.49	−0.07	−0.26	−0.20	−0.50	−0.32	*−0.59*
GLRLM	SRE	−0.33	*−0.53*	−0.04	−0.31	−0.18	−0.34	−0.25	*−0.54*	−0.41	*−0.53*
	LRE	0.34	*0.53*	0.02	0.34	0.20	0.33	0.25	*0.53*	0.36	*0.53*
	GLN	0.15	0.32	0.25	0.29	0.25	0.30	0.12	0.33	0.18	0.33
	RLN	0.13	0.32	0.26	0.30	0.22	0.30	0.14	0.33	0.14	0.33
	RP	−0.34	*−0.54*	−0.02	−0.33	−0.19	−0.34	−0.25	*−0.54*	−0.39	*−0.52*
	LGRE	0.29	−0.48	0.06	−0.37	−0.08	−0.33	0.11	−0.50	−0.16	*−0.58*
	HGRE	−0.23	−0.32	0.23	−0.47	−0.12	−0.28	−0.12	0.48	−0.21	0.43
	SRLGE	0.26	−0.49	0.08	−0.44	−0.10	−0.38	0.09	−0.48	−0.17	*−0.58*
	SRHGE	−0.25	−0.35	0.10	−0.44	−0.17	−0.31	−0.14	−0.40	−0.25	−0.44
	LRLGE	0.34	*0.51*	0.09	0.37	0.08	0.40	0.10	0.50	−0.11	*0.54*
	LRHGE	−0.21	*0.55*	0.27	0.50	0.16	0.36	−0.11	*0.52*	−0.12	*0.55*
	GLV	−0.33	*−0.52*	−0.10	−0.41	−0.28	−0.39	−0.25	*−0.53*	0.26	*−0.55*
	RLV	−0.31	*−0.57*	−0.18	−0.38	−0.27	−0.40	−0.27	*−0.56*	−0.35	*−0.55*
GLSZM	SZE	−0.18	−0.42	−0.05	*−0.51*	−0.05	−0.35	−0.41	*−0.63*	−0.41	*−0.60*
	LZE	0.19	0.50	0.23	0.33	0.21	0.36	0.19	*0.53*	0.30	*0.52*
	GLN	0.15	0.33	0.20	0.31	0.21	0.29	0.10	0.32	0.13	0.33
	ZSN	0.15	0.32	0.22	0.31	0.18	0.29	0.06	0.31	0.04	0.34
	ZP	−0.20	*−0.51*	−0.17	−0.41	−0.15	−0.34	−0.25	*−0.58*	−0.40	*−0.56*
	LGZE	−0.09	−0.48	0.31	0.36	−0.01	0.36	0.12	−0.46	−0.16	*−0.55*
	HGZE	−0.12	0.38	−0.19	−0.40	−0.04	−0.29	−0.17	0.50	−0.06	*0.52*
	SZLGE	−0.22	−0.46	0.28	−0.39	−0.01	0.34	0.10	−0.47	−0.19	*−0.58*
	SZHGE	−0.07	0.27	−0.17	0.47	−0.04	0.29	−0.31	*0.52*	−0.21	0.50
	LZLGE	0.32	*0.52*	0.20	0.34	0.21	0.35	0.25	*0.55*	0.35	*0.55*
	LZHGE	0.01	0.47	0.27	0.44	0.20	0.34	0.12	0.47	0.22	0.46
	GLV	−0.18	−0.45	−0.25	−0.34	−0.26	−0.33	−0.21	*−0.53*	−0.24	*−0.53*
	ZSV	−0.21	−0.48	−0.07	−0.43	−0.05	−0.33	−0.09	*0.53*	−0.01	−0.48
NGTDM	Coarseness	−0.06	−0.26	−0.22	−0.28	−0.21	−0.30	−0.08	−0.29	−0.13	−0.34
	Contrast	−0.14	−0.39	0.16	0.36	0.02	0.33	−0.11	−0.46	−0.33	*−0.51*
	Busyness	0.23	0.39	0.22	0.28	0.22	0.30	0.20	0.39	0.17	0.39
	Complexity	0.21	*−0.55*	−0.16	0.39	−0.13	0.40	0.14	*0.52*	0.22	−0.48
	Strength	0.04	−0.25	−0.24	−0.29	−0.21	−0.29	−0.09	−0.37	−0.07	−0.33
Average absolute values		0.20	0.42	0.18	0.37	0.15	0.31	0.16	0.46	0.24	0.49

maximal absolute Spearman's rank correlation coefficients obtained from optimal texture-parameters for each of the 41 different textures when one extraction parameter was allowed to vary, and the others set to baseline. The same process was repeated for all parameters of all scans, and also for a Wilcoxon rank sum test comparing baseline extraction parameters to full degrees of freedom on extraction parameters (ALL PARAMs). Table 3.7 presents the p-value of the Wilcoxon rank sum tests, with multiple testing corrections ($\alpha = 0.05$, $K = 29$). The results point out that the optimization of the *Scale* texture extraction parameter has the highest impact on the predictive value for lung metastases development in STSs. In general, each extraction parameter seems to positively impact the predictive value of textures.

Table 3.7: p-value of the Wilcoxon rank sum tests asserting the significance of the effects of texture extraction parameters on the correlation of texture features with lung metastases development in STSs. The values in italic font indicate features for which $p < 0.05/29 \approx 0.0017$.

Scan	*MRI Inv.*	*MRI weight*	*WBPF*	*Scale*	*Quant. algo*	N_g	ALL PARAMs
FDG-PET	–	–	0.0727	*< 0.0010*	0.0017	0.0260	*< 0.0010*
T1	–	–	0.2406	*< 0.0010*	0.1159	0.0369	*< 0.0010*
T2FS	–	–	0.2105	*0.0013*	0.0066	0.0352	*< 0.0010*
FDG-PET/T1	0.1085	0.0019	0.0034	*< 0.0010*	0.0024	0.0143	*< 0.0010*
FDG-PET/T2FS	0.5937	0.1259	0.3076	0.0024	0.4143	0.3909	*< 0.0010*

MRI Inv.: MRI Inversion, *MRI weight*: weight in FDG-PET/MRI fusion process, *WBPF*: wavelet band-pass filtering, *Scale*: isotropic voxel size, *Quant. algo.*: quantization algorithm, N_g: Number of gray-levels, ALL PARAMs: all texture extraction parameters allowed to vary.

3.5.2 Multivariable analysis

We compared the prediction performance estimation of multivariable models constructed using three different types of initial feature sets: i) FDG-PET textures + non-texture features; ii) separate FDG-PET and MRI textures + non-texture features; and iii) fused FDG-PET/MRI textures + non-texture features. We performed experiments for all degrees of freedom on texture extraction parameters. Figure 3.6 presents the prediction performance estimation of multivariable models with optimal degrees of freedom on texture extraction parameters, obtained separately for each model order of each initial feature set. Supplementary Material 3.10.3 also provides detailed comparison between prediction estimates obtained in the experiments using baseline, full and optimal degrees of freedom on texture extraction parameters. Results show that multivariable models constructed with texture features extracted from separate scans provide no significant prediction estimation improvements as compared to multivariables models constructed with texture features extracted from FDG-PET scans only. On the other hand, multivariable models constructed from fused scans significantly improve the prediction performance estimation compared to FDG-PET scans alone.

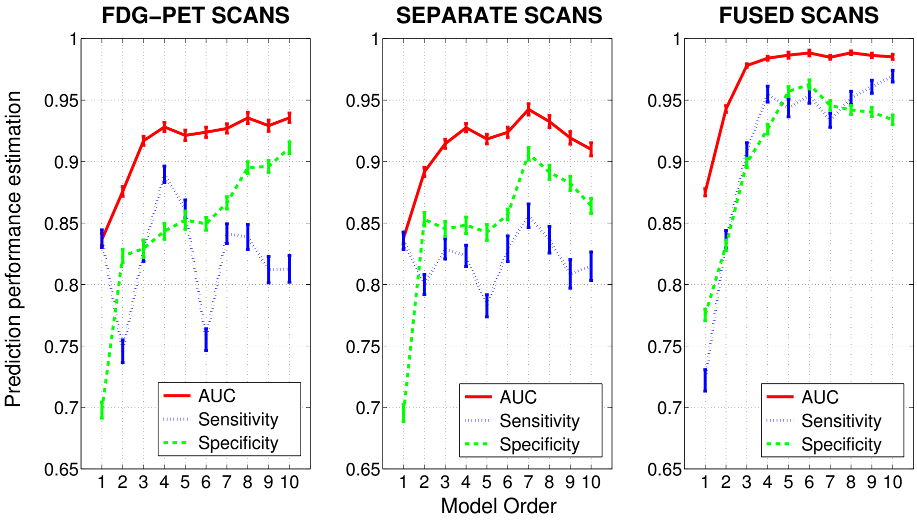


Figure 3.6: Estimation of prediction performance of multivariable models constructed from FDG-PET scans, SEPARATE scans, and FUSED scans using optimal degrees of freedom on texture extraction parameters, for model orders of 1 to 10. The optimal degrees of freedom were found in terms of maximum 0.632+ bootstrap AUC, separately for each model order. Error bars represent the standard error of the mean on a 95 % confidence interval.

By inspecting the curves in Figure 3.6, we determined that the simplest multivariable model with best predictive properties (best parsimonious model) is obtained by linearly combining 4 texture features extracted from fused FDG-PET/MRI scans. The associated prediction performance estimation of this optimal combination of features using 1000 bootstrap samples yielded an AUC of 0.984 ± 0.002, a sensitivity of 0.955 ± 0.006 and a specificity of 0.926 ± 0.004. These last results were obtained using the 0.632+ bootstrap method, and as a comparison, the same model reached an AUC of 0.976 ± 0.002, a sensitivity of 0.938 ± 0.008 and a specificity of 0.892 ± 0.006 using the ordinary bootstrap method ($\widehat{\mathrm{AUC}} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{AUC}\big(\mathbf{X}^{*b}, \mathbf{X}^{*b}(0)\big)$). Next, the logistic regression coefficients of the final prediction model were computed using 1000 bootstrap samples. We hence propose the following complete multivariable model response $g(\mathbf{x}_i)$ to be computed from fused FDG-PET/MRI scans at the time of diagnosis of STSs for the prediction of future lung metastases development:

$$
\begin{aligned}
&g(\mathbf{x}_i) = \\
&-256 \times \text{FDG-PET/T2FS}(\textit{MRI Inv.} = \text{No Inv.}, \textit{MRI weight} = 1/2, \textit{R} = 3/2, \textit{Scale} = 3\text{ mm}, \textit{Quant. algo.} = \text{Lloyd-Max}, \textit{Ng} = 64) \text{ - - GLSZM/SZE} \\
&\quad + \\
&5360 \times \text{FDG-PET/T1}(\textit{MRI Inv.} = \text{Inv.}, \textit{MRI weight} = 1/2, \textit{R} = 1/2, \textit{Scale} = \textit{in-pR}, \textit{Quant. algo.} = \text{Lloyd-Max}, \textit{Ng} = 16) \text{ - - GLSZM/ZSV} \\
&\quad + \\
&1.75 \times \text{FDG-PET/T1}(\textit{MRI Inv.} = \text{Inv.}, \textit{MRI weight} = 3/4, \textit{R} = 1, \textit{Scale} = 2\text{ mm}, \textit{Quant. algo.} = \text{Lloyd-Max}, \textit{Ng} = 8) \text{ - - GLSZM/HGZE} \\
&\quad + \\
&3.16 \times \text{FDG-PET/T2FS}(\textit{MRI Inv.} = \text{Inv.}, \textit{MRI weight} = 3/4, \textit{R} = 2, \textit{Scale} = 1\text{ mm}, \textit{Quant. algo.} = \text{Equal}, \textit{Ng} = 8) \text{ - - GLRLM/HGRE} \\
&\quad + 26.7
\end{aligned}
\tag{3.7}
$$

In order to evaluate the precision of the proposed model, we calculated how its response changes using texture features extracted from tumour contours that include surrounding edema. Supplementary Material 3.10.4 details the calculations. Overall, an absolute value of ± 4.89 was estimated as the uncertainty of the model due to contouring variations. This uncertainty is constant across all values of $g(\mathbf{x}_i)$. Then, to summarize how the model can separate the instances of the two patient classes (*LungMets* versus *NoLungMets*), the vector $\mathbf{g} = \{g(\mathbf{x}_i) \in \mathbb{R} : i = 1, 2, \ldots, N\}$ was computed for all patients using the multivariable model response of Equation 3.7 and was transformed into the posterior probability $\pi(\mathbf{x}_i)$ of observing outcome $y_i = 1$ (i.e., developing lung metastases) given the input $\mathbf{x}_i$ by using the logit transform of Equation 3.2. Figure 3.7 displays the plot of $\pi(\mathbf{x}_i)$ versus $g(\mathbf{x}_i)$, along with the associated $95\ \%$ confidence intervals (CIs) on $g(\mathbf{x}_i)$ for $i = 1, 2, \ldots, N$. For each bootstrap sample b used to calculate the final logistic regression coefficients of Equation 3.7, a new value of $g(\mathbf{x}_i^{*b})$ was calculated for $i = 1, 2, \ldots, N$ from the new coefficients computed on $\mathbf{x}_i^{*b}$. Then, the lower and upper CI bounds were estimated for each point i by calculating the 2.5 and 97.5 percentiles from the bootstrap distribution of $g(\mathbf{x}_i^{*b})$ for $b = 1, 2, \ldots, B$. In Figure 3.7, the dots represent patients who eventually developed lung metastases, and the crosses those who did not develop lung metastases. The uncertainty due to contouring variations around the classification threshold $g(\mathbf{x}_i) = 0$ is also shown, and Supplementary Material 3.10.4 provides the data (lung mets status, $g(\mathbf{x}_i)$ and CIs) used to construct the figure. It can be seen that the multivariable model of Equation 3.7 can clearly separate the patients of the two risk groups. Note that the Spearman's rank correlation between the model response vector **g** and the outcome vector **y** reached $r_s = 0.84, p < 0.001$.

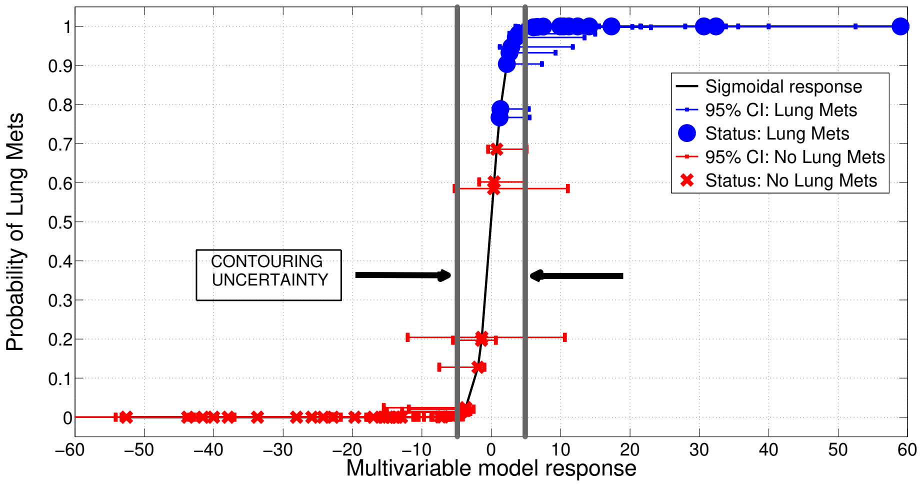


Figure 3.7: Probability of developing lung metastases as a function of the response of the multivariable model proposed in this work, for all patients of the cohort.

Finally, further validation of the proposed model was performed using permutation tests [35]. This operation was carried out by randomly shuffling the real outcome vector **y** of our patient cohort (i.e., keeping the same proportion of 0's and 1's). In order to have a direct comparison with the model of Equation 3.7, only the prediction performance of multivariable models of order 4 for the feature set comprised of textures extracted from fused scans were analyzed. For each of the 1000 permutation tests, a different multivariable model was constructed and its prediction performance was assessed. Table 3.8 presents a summary of the results of the permutation tests, where the p-values were estimated using the Monte Carlo sampling approach described by Ernst [36]. Supplementary Material 3.10.5 also provides a display of the permutation distributions. Overall, results in Table 3.8 show that the null hypothesis can be rejected. Very few tests yielded prediction estimates higher than the true observed values. The results give strong evidence that the effect observed on the sample data is most likely present in the general STS population.

3.6 Discussion

In this work, an imaging model was identified for the prediction of future lung metastases development at the time of diagnosis of STSs. This multivariable model is composed of four texture features extracted from fused

Table 3.8: Summary of permutation tests. Comparison between the single value of the performance metrics estimates of the multivariable model proposed in this work found using the real outcome vector ($\widehat{\mathrm{TRUE}}$), and the distribution of the performance metrics estimates of models of order 4 from fused scans found in 1000 permutation tests using shuffled outcome vectors ($\widehat{\mathrm{PERM}}$). SD: Standard deviation.

Metric	$\mathrm{Value}_{\widehat{\mathrm{TRUE}}}$	$\mathrm{Mean}_{\widehat{\mathrm{PERM}}}$	$\mathrm{SD}_{\widehat{\mathrm{PERM}}}$	$\mathrm{Range}_{\widehat{\mathrm{PERM}}}$	$\widehat{\mathrm{PERM}} > \mathrm{Value}_{\widehat{\mathrm{TRUE}}}$	$\hat{p}$
AUC	0.984	0.895	0.04	$[0.745, 0.988]$	3 out of 1000	0.004
Sensitivity	0.955	0.798	0.05	$[0.634, 0.938]$	0 out of 1000	0.001
Specificity	0.926	0.812	0.05	$[0.666, 0.948]$	6 out of 1000	0.007

FDG-PET/MRI scans. The use of texture features extracted from fused FDG-PET/MRI scans constitutes a new technique proposed in this work that revealed to be promising for tumour heterogeneity quantification. Our approach focused on standard-of-care medical images in order to strengthen its clinical impact. We believe that the methodology developed in this work could be generalized to other types of cancer and tumour outcomes.

First, the association of nine non-texture features with lung metastases development in STSs was presented in Table 3.5. As expected, *SUVmax* was significantly associated with lung metastasis risk in STS cancer, a result in agreement with other studies [37–39]. A significant positive association was also found between *Percent Inactive* and lung metastases development, meaning that the larger the volume of inactive FDG-PET regions in reference to patient-specific *SUVmax* values, the higher the risk of developing lung metastases in STS cancer. In Table 3.6, the strong and positive associations of *Homogeneity* and *LZE*, and the strong and negative association of *Complexity*, for example, corroborate this last assertion. This could be explained from the fact that patients from the *LungMets* group often possess large and uniform low-uptake regions (as compared to maximum SUV) in the inner portion of their tumour on FDG-PET scans, most likely representing necrotic areas. The presence of these inner, low-uptake uniform regions suggests that the tumour is rapidly increasing in size and might be more at risk to metastasize. As demonstrated, texture analysis can however reveal more information about the tumour underlying biology than simple imaging metrics: for example, the strong and positive association of the *LZLGE* metric with lung metastases confirms that large low-uptake regions in FDG-PET have significant predictive value, but the positive association of the *LZHGE* metric (although to a lesser extent) suggest that large high-uptake regions may also play an important role in the characterization of aggressive tumours. Note that this last information would not have been captured by textures extracted with standard baseline parameters (defined in Table 3.4) only. This suggests that texture optimization is a desirable property to enhance the predictive value of textures, and Table 3.7 revealed that *Scale* is the extraction parameter that has the most influence on texture definition. An image with a given intrinsic resolution but resampled to different resolutions will produce different texture measurements, as imaging patterns are better captured for a certain range of resolutions. In this sense, an optimal scale at which textures offer best discrimination between the two classes of STS patients likely exists. In general, different texture features will better represent the underlying tumour biology using different extraction parameters, and the optimal set of parameters to use is application-specific and will depend on many factors such as the clinical endpoint studied and the imaging modalities employed.

Furthermore, the results presented in Figure 3.6 showed that MRI textures alone are generally not useful in comparison to FDG-PET textures. However, the addition of the MR imaging information to FDG-PET in the fusion process seems to significantly improve and even stabilize the prediction performance estimation of FDG-PET textures. Although some information may be lost in the fusion process, the fusion of FDG-PET and MRI scans may create new textural properties that can better characterize intratumoural heterogeneity than what separate FDG-PET and MRI scans can provide. Figure 3.7 then illustrated how the prediction model proposed in this work could be clinically used for the evaluation of the risk of future lung metastases development in STSs. A patient diagnosed with STS cancer would present in a hospital and undergo FDG-PET and MRI scans (with both T1 and T2FS sequences). A single value of the form of Equation 3.7 could then be obtained by extracting specific textures features from fused FDG-PET/MRI scans. Using the logit transform of Equation 3.2, this value could be transformed into the probability of developing lung metastases. This probability could then provide useful insights to physicians into risk assessment and treatment personalization. Ultimately, provided a given decision threshold and confidence interval, standard treatments could be strengthened for high risk patients and lessened for low risk patients. However, it can be seen from Figure 3.7 that the statistical uncertainty on $g(\mathbf{x}_i)$ measurements (bootstrap CIs) inherent to logistic regression coefficient estimates is significant and constitutes the principal limiting factor of the proposed model. A larger patient cohort is first needed to improve its precision. Moreover, a constant uncertainty of ± 4.89 on $g(\mathbf{x}_i)$ measurements due to contouring variations was found. This uncertainty on $g(\mathbf{x}_i) = 0$ incorporates 19 patients of our cohort (9 from the

LungMets group, 10 from the *NoLungMets* group), and no definitive conclusion could be drawn for these patients in a clinical decision-support system. This emphasizes the need to identify a model that can yet better separate the two patient classes, but also to construct texture-based prediction models from tumours delineated using automatic segmentation methods.

Overall, the multivariable model proposed in this work was found to possess high predictive potential for lung metastases in STS cancer. However, a larger patient cohort is needed to create a more robust model, and independent testing on external datasets is required to confirm its predictive properties. Once this step is cleared, we should investigate how this texture-based model could complement clinical prognostic factors for optimal prediction. The extraction of texture features from medical images and the construction of prediction models are complex processes that need proper validation, and much effort is still required in order to achieve clinical implementation of a texture-based decision-support system. First, a consensus on techniques used for imaging acquisition, data pre-processing, tumour delineation, texture analysis and multivariable modeling is needed in the *radiomics* community. Assuredly, standardization and full transparency on data and methods is the key for the progression of the field [18].

3.7 Conclusion

Textural biomarkers as an intratumoural heterogeneity quantification tool hold great promise for the early prediction of tumour outcomes. In this work, we explored a novel approach based on the fusion of FDG-PET and MRI volumes to better quantify intratumoural heterogeneity using texture analysis. Innovative texture extraction techniques and multivariable modeling strategies were also developed for the construction of tumour outcome prediction models from a large number of *radiomics* features. The results showed that FDG-PET and MRI texture features could act as strong prognostic factors of STSs and could provide insights about their underlying biology. FDG-PET texture features were shown to generally possess a higher predictive value than MRI texture features for lung metastases in STS cancer, but their predictive value was significantly enhanced by the addition of the MR imaging information to FDG-PET via the fusion process. The results also pointed out the importance of the optimization of texture extraction parameters to enhance their predictive value and to better understand the relation between textures and biology. Ultimately, we identified a model combining four texture features extracted from fused FDG-PET/MRI pre-treatment scans to predict future lung metastases development in STSs. This model reached high prediction performance estimates in bootstrapping evaluations and was validated using permutation tests. However, further validation on independent datasets is required to confirm its predictive properties. We believe that the methodology presented in this work could be generalized to other types of cancers and that it could eventually lead to improvements in treatment personalization and patient survival.

3.8 Online resources

Clinical information and imaging data analyzed in this work are available on The Cancer Imaging Archive (TCIA) website under the following DOI: `http://dx.doi.org/10.7937/K9/TCIA.2015.7GO2GSKS`. All software code implemented in this work is freely shared under the GNU General Public License at: `https://github.com/mvallieres/radiomics`.

3.9 Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under the scholarship CGSD3-426742-2012, as well as it was partly supported by the Canadian Institutes of Health Research (CIHR) under grants MOP-114910 and MOP-136774, and the NSERC of Canada under the grant RGPIN 397711-11.

3.10 Supplementary Material

3.10.1 Definition of texture features

In this thesis, please see Appendix A.

3.10.2 Texture features with highest predictive value

Table 3.9: Texture features with the highest Spearman's rank correlation (r_s) with lung metastases in STS cancer and their associated extraction parameters, for the five types of scans investigated in this study. The values in italic font indicate features for which $p < 0.05/419 \approx 0.00012$.

SCAN	TEXTURE	r_s	*MRI Inv.*	*MRI weight.*	*WPBF*	*Scale*	*Quant. algo.*	N_g
FDG-PET	RLV	*-0.57*	–	–	$R = 2/3$	1 mm	Lloyd-Max	16
T1	SZE	*-0.51*	–	–	$R = 2/3$	5 mm	Lloyd-Max	64
T2FS	LRLGE	0.40	–	–	$R = 2/3$	2 mm	Equal	64
FDG-PET/T1	SZE	*-0.63*	Inv.	3/4	$R = 2/3$	5 mm	Lloyd-Max	64
FDG-PET/T2FS	SZE	*-0.60*	No Inv.	1/2	$R = 3/2$	3 mm	Lloyd-Max	64

MRI Inv.: MRI Inversion, *MRI weight*: weight in FDG-PET/MRI fusion process, *WBPF*: wavelet band-pass filtering, *Scale*: isotropic voxel size, *Quant. algo.*: quantization algorithm, N_g: Number of gray-levels, *R*: ratio of weight given to band-pass sub-bands to weight given to low- and high-frequency sub-bands.

3.10.3 Degrees of freedom on texture extraction parameters

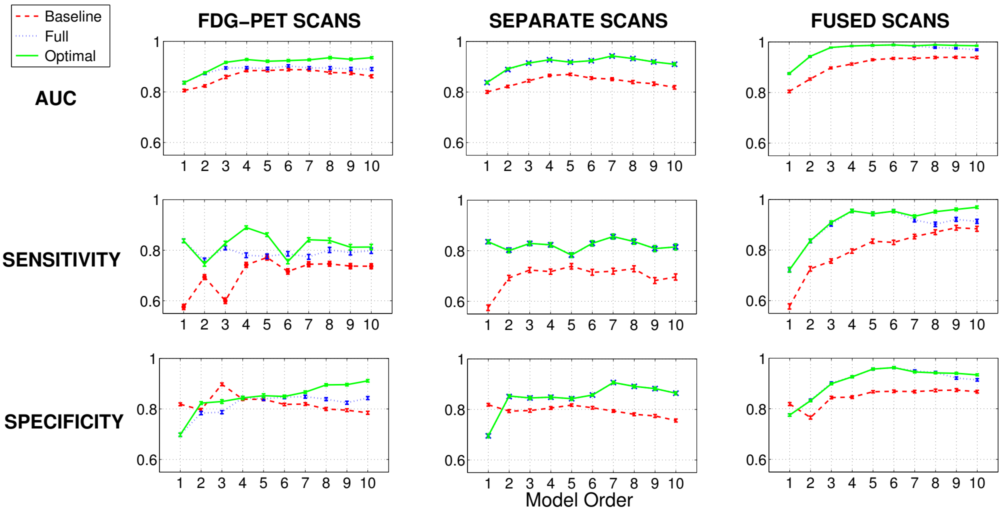


Figure 3.8: Estimation of prediction performance as a function of model order for multivariable models constructed from the FDG-PET, SEPARATE, and FUSED scans for three types of experiment: i) experiments in which no texture extraction parameter was allowed to vary (baseline parameters); ii) experiments in which all texture extraction parameters were allowed to vary (full degree of freedom); and iii) experiments using the particular set of texture extraction parameters allowed to vary yielding the highest 0.632+ bootstrap AUC, separately for each model order of each initial feature set (optimal degree of freedom). Error bars represent the standard error of the mean on a 95 % confidence interval.

There exists a certain degree of model complexity for which prediction models have optimal parsimonious properties. An additional factor to model order that increases model complexity comes from the whole range of texture extraction parameters used in this work, as too specific extraction parameters may reduce the generalizability of prediction models. One way of reducing model complexity is to fix specific extraction parameters by preventing them to fully vary, or in other words, to restrict the degree of freedom on texture extraction parameters. In principle, higher-order models should be coupled to more restricted extraction parameter degrees of freedom in order to reach optimal parsimonious properties, and vice-versa. The results presented in Figure 3.8 confirm that there exists an optimal combination of model order and texture extraction parameter degree of freedom in terms of predictive properties. It can be seen that the higher the model order, the more distance there is between the curves obtained using full and optimal degree of freedom on texture extraction parameters, especially in the case of the FDG-PET feature set. On the other hand, for lower model orders, the curves obtained using full and optimal degree of freedom on texture extraction parameters are generally identical. Table 3.10 shows which extraction parameters were allowed to vary in the experiments yielding the optimal prediction performance estimation in Figure 3.8, for model orders of 1 to 10 and for the three different types of initial feature sets.

Table 3.10: Texture extraction parameters allowed to vary in the experiments yielding optimal prediction performance estimation.

Model order	FDG-PET				SEPARATE				FUSED					
	WBPF	*Scale*	*Quant. algo*	N_g	*WBPF*	*Scale*	*Quant. algo*	N_g	*MRI Inv.*	*MRI weight*	*WBPF*	*Scale*	*Quant. algo*	N_g
1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓			✓	✓	✓		✓	✓	✓	✓		✓
3	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
4	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓		✓
8	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓
9	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓		✓
10		✓		✓	✓	✓	✓	✓	✓		✓	✓		✓

MRI Inv.: MRI Inversion, *MRI weight*: weight in FDG-PET/MR fusion process, *WBPF*: wavelet band-pass filtering, *Scale*: isotropic voxel size, *Quant. algo.*: quantization algorithm, N_g: Number of gray-levels

It can now be directly seen from Table 3.10 that an optimal degree of freedom on texture extraction parameters generally allows less parameters to vary for higher model orders, except in the case of the feature set using separate scans. However, in light of the results we obtained, it is not possible to identify a trend about which texture extraction parameters should be used as a function of model order to obtain optimal prediction performance estimation.

3.10.4 Uncertainty analysis

Contouring variations

In this section, an estimation of the uncertainty of the multivariable model proposed in this work $g(\mathbf{x}_i)$ due to contouring variations is evaluated. Thirty-two patients of the cohort had visible edema that could be clearly identified in the vicinity of their tumours. For these patients, two types of contours were drawn: one incorporating the visible edema (denoted as *Edema*), and one excluding it (denoted as *Mass*). The proposed model and its associated logistic regression coefficients were obtained using textures computed with the *Mass* contours, and we now evaluate how its response changes using textures computed with the *Edema* contours (but with the same logistic regression coefficients). Table 3.11 first presents the volumes of the tumours and the textures features differences of the four features of the proposed model extracted with the two contours, for all patients of the cohort.

In table D.1, the average $\overline{\Delta x_j}$ of the absolute difference between the values x_{ij} of the four features $j = 1, 2, 3, 4$ of the proposed model extracted from the *Mass* and *Edema* contours over all patients $i = 1, 2, \ldots, N$ was computed such that:

$$\Delta x_{ij} = \left| \Delta x_{ij}\left[Edema\right] - \Delta x_{ij}\left[Mass\right] \right|, \quad \text{for } j = 1, 2, 3, 4$$

$$\overline{\Delta x_j} = \frac{1}{N}\sum_{i=1}^{N} \Delta x_{ij}, \quad \text{for } j = 1, 2, 3, 4$$

For the 19 patients without the *Edema* contour, $\overline{\Delta x_j}$ was considered to be 0 and was still incorporated in the calculation of $\overline{\Delta x_j}$. Then, the global contribution of contouring variations on the uncertainty of $g(\mathbf{x}_i)$ was obtained by adding in quadrature the contribution of each average texture variation $\overline{\Delta x_j}$ such that (recalling that $g(\mathbf{x}_i) = -256 \times x_{i1} + 5360 \times x_{i2} + 1.75 \times x_{i3} + 3.16 \times x_{i4} + 26.7$):

$$\epsilon_{\text{contour}} = \sqrt{\sum_{j=1}^{p}\left[\left(\frac{\partial g(\mathbf{x}_i)}{\partial x_{ij}}\right)^2 \cdot \left(\overline{\Delta x_j}\right)^2\right]}$$

$$\epsilon_{\text{contour}} = \sqrt{(-256)^2 \cdot (0.0065)^2 + (5360)^2 \cdot (0.000675)^2 + (1.75)^2 \cdot (1.46)^2 + (3.16)^2 \cdot (0.39)^2}$$

$$\epsilon_{\text{contour}} \approx 4.89$$

Table 3.11: Volumes, texture features and texture features differences of the four features forming the multivariable model proposed in this work in the case where the tumour region was outlined with the contour excluding surrounding edema (*Mass*), and the case where the tumour was outlined with the contour including the surrounding edema (*Edema*).

Patient	Volume	Volume	ΔVolume	Feature	Feature		Feature	Feature		Feature	Feature		Feature	Feature	
number	(cm^3)	(cm^3)	(cm^3)	x_{i1}	x_{i1}	Δx_{i1}	x_{i2}	x_{i2}	Δx_{i2}	x_{i3}	x_{i3}	Δx_{i3}	x_{i4}	x_{i4}	Δx_{i4}
(i)	*Mass*	*Edema*		*Mass*	*Edema*		*Mass*	*Edema*		*Mass*	*Edema*		*Mass*	*Edema*	
1	281	446	165	0.6459	0.6370	0.0089	0.000963	0.000473	0.000490	27.06	24.45	2.61	24.94	25.09	0.15
2	48	55	7	0.6374	0.6174	0.0200	0.003078	0.002921	0.000157	20.64	20.81	0.17	25.37	25.46	0.09
3	69	136	67	0.6708	0.6499	0.0209	0.001563	0.001054	0.000509	21.79	14.85	6.94	22.99	25.21	2.22
4	247	–	0	0.6723	–	0	0.001068	–	0	22.68	–	0	25.42	–	0
5	480	521	41	0.6234	0.5915	0.0319	0.002408	0.001814	0.000594	23.46	24.81	1.35	24.37	24.46	0.09
6	186	–	0	0.6184	–	0	0.003741	–	0	24.35	–	0	27.30	–	0
7	544	–	0	0.6199	–	0	0.001342	–	0	19.83	–	0	24.12	–	0
8	22	–	0	0.7411	–	0	0.001408	–	0	19.67	–	0	24.46	–	0
9	1905	1980	75	0.5866	0.5945	0.0079	0.001618	0.001814	0.000196	22.98	22.35	0.63	24.72	25.07	0.35
10	149	301	152	0.6417	0.6337	0.0080	0.001941	0.002530	0.000589	25.73	24.20	1.53	25.55	25.89	0.34
11	609	752	143	0.5990	0.5987	0.0003	0.001812	0.002577	0.000765	22.02	23.12	1.10	24.94	24.87	0.07
12	67	95	28	0.6702	0.6656	0.0046	0.002115	0.001755	0.000360	20.55	20.25	0.30	23.66	24.28	0.62
13	1378	–	0	0.6312	–	0	0.000839	–	0	30.79	–	0	24.32	–	0
14	362	459	97	0.5984	0.6066	0.0082	0.002784	0.003122	0.000338	22.71	23.36	0.65	25.10	24.81	0.29
15	70	73	3	0.6959	0.6903	0.0056	0.002368	0.001876	0.000492	21.55	21.72	0.17	24.31	24.29	0.02
16	406	623	217	0.6182	0.6203	0.0021	0.002081	0.004367	0.002286	21.44	22.26	0.82	24.90	25.35	0.45
17	438	–	0	0.6139	–	0	0.001980	–	0	25.03	–	0	26.16	–	0
18	117	–	0	0.6367	–	0	0.003734	–	0	24.25	–	0	24.62	–	0
19	13	–	0	0.7545	–	0	0.001689	–	0	21.30	–	0	24.86	–	0
20	2024	3371	1347	0.6478	0.6608	0.0130	0.001151	0.000678	0.000473	25.15	21.31	3.84	27.72	28.81	1.09
21	530	588	58	0.6000	0.5969	0.0031	0.002035	0.001945	0.000090	26.90	34.27	7.37	24.95	25.51	0.56
22	789	1078	289	0.6159	0.6320	0.0161	0.006121	0.002721	0.003400	29.51	24.00	5.51	24.97	25.86	0.89
23	153	265	112	0.6151	0.6152	0.0001	0.002104	0.001664	0.000440	24.47	22.42	2.05	24.97	25.37	0.40
24	65	99	34	0.6222	0.6106	0.0116	0.001204	0.002488	0.001284	22.87	24.85	1.98	21.87	23.22	1.35
25	25	–	0	0.7225	–	0	0.001456	–	0	21.08	–	0	23.27	–	0
26	326	–	0	0.6369	–	0	0.000835	–	0	23.20	–	0	24.17	–	0
27	751	860	109	0.5950	0.5872	0.0078	0.000434	0.002055	0.001621	27.45	29.88	2.43	25.74	25.54	0.20
28	176	–	0	0.5977	–	0	0.001983	–	0	25.51	–	0	26.42	–	0
29	240	458	218	0.6394	0.6194	0.0200	0.001673	0.001755	0.000082	25.71	25.95	0.24	25.05	25.79	0.74
30	300	2618	2318	0.6107	0.6054	0.0053	0.000912	0.001460	0.000548	18.48	25.14	6.66	24.38	26.58	2.20
31	389	433	44	0.6127	0.6069	0.0058	0.002153	0.005309	0.003156	22.99	27.32	4.33	25.22	24.82	0.40
32	89	127	38	0.6463	0.6349	0.0114	0.004610	0.004613	0.000003	26.85	22.99	3.86	24.52	24.13	0.39
33	340	599	259	0.6091	0.6106	0.0015	0.001997	0.000894	0.001103	22.96	20.18	2.78	25.17	25.73	0.56
34	538	652	114	0.6107	0.6020	0.0087	0.001638	0.001438	0.000200	23.95	23.84	0.11	25.23	25.27	0.04
35	27	–	0	0.7080	–	0	0.001360	–	0	15.09	–	0	21.60	–	0
36	287	503	216	0.6365	0.6029	0.0336	0.001849	0.000916	0.000933	22.31	24.06	1.75	25.41	25.59	0.18
37	593	772	179	0.5901	0.5982	0.0081	0.010574	0.004361	0.006213	24.20	18.45	5.75	26.69	27.77	1.08
38	365	467	102	0.5974	0.5922	0.0052	0.002043	0.002079	0.000036	27.25	27.97	0.72	24.93	25.65	0.72
39	567	713	146	0.6131	0.6027	0.0104	0.002710	0.001349	0.001361	29.93	28.39	1.54	24.52	24.64	0.12
40	299	–	0	0.6262	–	0	0.008294	–	0	23.82	–	0	24.72	–	0
41	50	58	8	0.6929	0.7041	0.0112	0.000420	0.000414	0.000006	20.01	21.65	1.64	23.89	23.18	0.71
42	31	33	2	0.6916	0.7160	0.0244	0.003538	0.005487	0.001949	17.42	17.05	0.37	23.05	23.67	0.62
43	143	–	0	0.6997	–	0	0.001437	–	0	29.47	–	0	24.90	–	0
44	583	–	0	0.6382	–	0	0.000879	–	0	23.29	–	0	24.77	–	0
45	348	–	0	0.6546	–	0	0.000982	–	0	23.82	–	0	24.81	–	0
46	563	653	90	0.6511	0.6493	0.0018	0.002120	0.000682	0.001438	28.20	26.89	1.31	26.00	25.87	0.13
47	179	–	0	0.6979	–	0	0.000721	–	0	21.56	21.56	0	24.26	–	0
48	293	337	44	0.6596	0.6509	0.0087	0.001632	0.003431	0.001799	29.05	29.35	0.30	24.75	24.92	0.17
49	1065	–	0	0.6033	–	0	0.000694	–	0	22.91	–	0	24.35	–	0
50	122	307	185	0.6296	0.6226	0.0070	0.001886	0.000358	0.001528	27.30	23.64	3.66	25.07	27.61	2.54
51	146	–	0	0.6851	–	0	0.003453	–	0	20.67	–	0	25.45	–	0

x_{i1}: FDG-PET/T2FS - - GLSZM/SZE
x_{i2}: FDG-PET/T1 - - GLSZM/ZSV
x_{i3}: FDG-PET/T1 - - GLSZM/HGZE
x_{i4}: FDG-PET/T2FS - - GLRLM/HGRE

Overall in our patient cohort, an absolute value of ± 4.89 was calculated for $\epsilon_{\text{contour}}$. This uncertainty is constant across all values of $g(\mathbf{x}_i)$.

Bootstrap confidence intervals

Table 3.12: Response of the multivariable model proposed in this work on the entire patient cohort, and its associated 95 % confidence interval (CI) bounds.

Patient number (i)	Outcome	Model response $g(\mathbf{x}_i)$	CI Lower bound	CI Upper bound	ΔCI
35	0	-51.7	-91.8	-30.9	60.9
8	0	-42.9	-74.7	-25.7	49.0
19	0	-40.7	-71.9	-23.8	48.1
25	0	-39.1	-69.4	-23.9	45.5
41	0	-37.0	-65.8	-23.4	42.4
47	0	-32.8	-58.3	-20.6	37.7
42	0	-27.2	-51.2	-14.7	36.6
3	0	-25.0	-46.2	-14.9	31.3
15	0	-23.3	-42.3	-13.7	28.5
12	0	-21.9	-39.7	-13.1	26.6
4	0	-18.8	-34.5	-11.0	23.5
24	0	-16.1	-36.3	-5.74	30.6
45	0	-14.6	-28.0	-9.47	18.6
30	0	-14.5	-28.8	-7.75	21.1
26	0	-14.0	-28.0	-8.96	19.0
43	0	-13.6	-27.7	-6.69	21.0
7	0	-13.0	-25.7	-7.36	18.3
51	0	-12.7	-28.7	-5.12	23.6
44	0	-12.0	-24.1	-7.86	16.2
1	0	-6.42	-15.4	-2.90	12.5
49	0	-6.10	-16.2	-1.43	14.7
36	0	-6.09	-13.0	-2.91	10.1
48	0	-3.46	-10.7	1.07	11.8
16	0	-3.30	-8.65	-1.19	7.46
29	0	-2.97	-8.02	-1.31	6.71
2	0	-2.79	-11.8	0.522	12.3
5	0	-1.02	-5.58	0.926	6.51
10	0	-0.506	-4.12	2.07	6.19
20	0	-0.460	-10.7	12.0	22.7
13	0	1.24	-5.63	10.7	16.3
11	0	1.31	-2.16	4.44	6.60
34	0	1.68	-1.24	4.30	5.54
33	1	2.09	-0.913	4.31	5.22
31	1	2.22	-0.915	4.08	4.99
23	1	3.14	0.517	5.05	4.53
50	1	3.53	0.110	6.63	6.52
46	1	3.79	-1.67	8.86	10.5
9	1	4.43	0.333	9.91	9.58
18	1	4.86	-1.35	11.0	12.4
27	1	6.98	1.08	16.9	15.9
17	1	7.52	2.73	13.7	11.0
14	1	8.39	4.00	14.1	10.1
21	1	10.8	5.31	18.0	12.7
32	1	11.3	3.00	23.4	20.4
38	1	12.1	6.01	20.2	14.2
28	1	13.3	6.64	23.1	16.5
39	1	15.0	7.15	25.8	18.7
6	1	18.2	8.13	35.2	27.1
40	1	31.5	15.1	64.8	49.7
22	1	33.3	19.1	61.0	41.9
37	1	59.9	37.5	115	77.8

3.10.5 Permutation tests

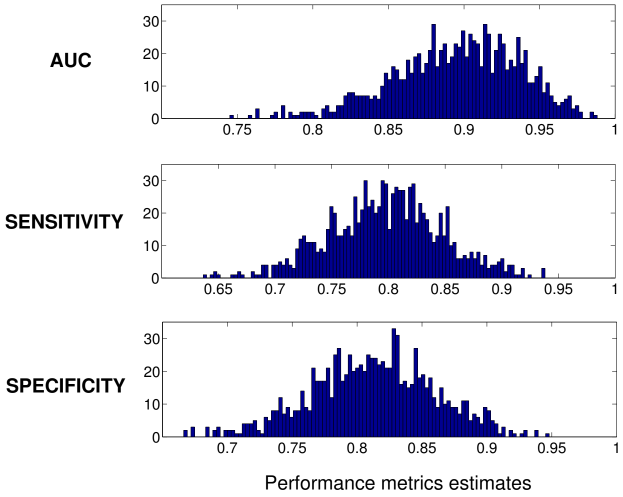


Figure 3.9: Distributions of performance metrics estimates of multivariable models of order 4 found in 1000 permutation tests using shuffled outcome vectors. All histograms were constructed with 100 bins.

3.11 References

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA A Cancer Journal for Clinicians* **64,** 9–29 (2014).
2. Brennan, M. F. Soft tissue sarcoma: advances in understanding and management. *The Surgeon* **3,** 216–223 (2005).
3. Billingsley, K. G. *et al.* Pulmonary metastases from soft tissue sarcoma. *Ann. Surg.* **229,** 602 (1999).
4. Lewis, J. J. & Brennan, M. F. Soft tissue sarcomas. *Curr. Probl. Surg.* **33,** 839–872 (1996).
5. Billingsley, K. G. *et al.* Multifactorial analysis of the survival of patients with distant metastasis arising from primary extremity sarcoma. *Cancer* **85,** 389–395 (1999).
6. Komdeur, R. *et al.* Metastasis in soft tissue sarcomas: prognostic criteria and treatment perspectives. *Cancer Metastasis Rev.* **21,** 167–183 (2002).
7. Longo, D. L. Tumor heterogeneity and personalized medicine. *N. Engl. J. Med.* **366,** 956–957 (2012).
8. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48,** 441–446 (2012).
9. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30,** 1234–1248 (2012).

10. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25,** 675–680 (2007).

11. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105,** 5213–5218 (2008).

12. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

13. El Naqa, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* **42,** 1162–1171 (2009).

14. Tixier, F. *et al.* Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J. Nucl. Med.* **52,** 369–378 (2011).

15. Cook, G. J. R. *et al.* Are pretreatment 18F-FDG PET tumor textural features in non–small cell lung cancer associated with response and survival after chemoradiotherapy? *J. Nucl. Med.* **54,** 19–26 (2013).

16. Vaidya, M. *et al.* Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiother. Oncol.* **102,** 239–245 (2012).

17. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56,** 38–44 (2015).

18. Lambin, P. *et al.* Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat. Rev. Clin. Oncol.* **10,** 27–40 (2013).

19. Deasy, J. O., Blanco, A. I. & Clark, V. H. CERR: A computational environment for radiotherapy research. *Med. Phys.* **30,** 979–985 (2003).

20. Collewet, G., Strzelecki, M. & Mariette, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn. Reson. Imaging* **22,** 81–91 (2004).

21. Vallières, M. *FDG-PET/MR imaging for prediction of lung metastases in soft-tissue sarcomas of the extremities by texture analysis and wavelet image fusion* Master Thesis (McGill University, Montreal, Canada, 2013).

22. Van Velden, F. H. P. *et al.* Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur. J. Nucl. Med. Mol. Imaging* **38,** 1636–1647 (2011).

23. Li, Q. & Griffiths, J. G. *Least squares ellipsoid specific fitting. Proceedings of the Geometric Modeling and Processing 2004.* International Conference on Geometric Modeling and Processing (GMP 04) (Beijing, China, 2004), 335–340.

24. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).

25. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).

26. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

27. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).

28. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

29. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).

30. Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory* **6,** 7–12 (1960).

31. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28,** 129–137 (1982).

32. El Naqa, I. *et al.* Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys. Med. Biol.* **51,** 5719–5735 (2006).

33. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.* **35,** 1559–1570 (2008).

34. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334,** 1518–1524 (2011).

35. Fisher, R. A. *The Design of Experiments* (Oliver and Boyd, Edinburgh, 1935).

36. Ernst, M. D. Permutation methods: a basis for exact inference. *Statistical Science* **19,** 676–685 (2004).

37. Schwarzbach, M. H. M. *et al.* Clinical Value of [18-F] Fluorodeoxyglucose Positron Emission Tomography Imaging in Soft Tissue Sarcomas. *Ann. Surg.* **231,** 380–386 (2000).

38. Eary, J. F. *et al.* Sarcoma tumor FDG uptake measured by PET and patient outcome: a retrospective analysis. *Eur. J. Nucl. Med.* **29,** 1149–1154 (2002).

39. Skamene, S. R. *et al.* Metabolic activity measured on PET/CT correlates with clinical outcomes in patients with limb and girdle sarcomas. *J. Surg. Oncol.* **109,** 410–414 (2014).

Chapter 4

A strategy for treatment personalization

4.1 Foreword 130
4.2 Abstract 131
4.3 Introduction 132
4.4 Materials and Methods 133
4.4.1 Patients 134
4.4.2 Standard-of-care radiotherapy planning and treatment 134
4.4.3 Study design 134
4.4.4 Image analysis 136
4.4.5 Dose painting feasibility study 137
4.5 Results 137
4.5.1 Prediction of lung metastases development 138
4.5.2 General imaging findings 138
4.5.3 Tumour sub-region analysis 139
4.5.4 Dose painting 141
4.6 Discussion 143
4.7 Acknowledgments 145
4.8 Supplementary Material 146
4.8.1 Supplementary results 146
4.8.2 Image acquisition protocols 146
4.8.3 Image registration 150
4.9 References 150

4.1 Foreword

This Chapter presents a study submitted as the following paper: Martin Vallières, Monica Serban, Ibtissam Benzyane, Zaki Ahmed, Shu Xing, Issam El Naqa, Ives R. Levesque, Jan Seuntjens & Carolyn R. Freeman. “The role of FDG-PET, FMISO-PET, DW-MRI and DCE-MRI in the management of soft-tissue sarcomas of the extremities with pre-operative radiotherapy and surgery: a feasibility study". *Radiother. Oncol.* [submitted March 16, 2017].

In this work, the overall results obtained from a prospective study protocol initiated at the McGill University Health Centre are presented. Eighteen soft-tissue sarcoma patients were recruited at our institution between August 2013 and February 2016. Under the study protocol, patients were to receive FDG-PET, FMISO-PET, DW-MRI and DCE-MRI scans at pre-, mid- and post-radiotherapy. Following study termination, different characteristics of images at all time points were analyzed. Furthermore, the radiomic-based model developed in Chapter 3 was validated and the technical feasibility of dose painting was investigated onto these prospective patients, which constitutes a treatment personalization strategy for patients identified to be at higher risk of developing lung metastases. Finally, please note that this paper was submitted to a journal considered as "clinical" (*Radiotherapy and Oncology*) with a short number of words allowed to convey our message, hence explaining the different tone in the writing as compared to the other papers in this thesis.

4.2 Abstract

Background and Purposes: Management of extremity soft-tissue sarcomas (STS) is complex due to the heterogeneity of histologies occurring at different sites in the body. In this work, we validate a previously described FDG-PET/MRI texture-based model for the prediction of lung metastases in STS. We use anatomical and functional imaging at different treatment time points and explore the feasibility of dose painting as a treatment strategy.

Material and Methods: We acquired FDG-PET, FMISO-PET, DW-MRI and DCE-MRI data for 18 patients with extremity STS before, during, and after pre-operative radiotherapy. We tested our lung metastases prediction model using pre-treatment images. We evaluated the feasibility of dose painting using a prescription of 50 Gy to the PTV ($\text{PTV}_{50\,\text{Gy}}$) along with boost doses of 60 Gy to the FDG hypermetabolic GTV ($\text{GTV}_{60\,\text{Gy}}$) and of 65 Gy to the low-perfusion DCE-MRI hypoxic GTV contained within the $\text{GTV}_{60\,\text{Gy}}$ ($\text{GTV}_{65\,\text{Gy}}$) using volumetric arc therapy (VMAT).

Results: The texture-based model for lung metastases prediction reached an

AUC of 0.71, a sensitivity of 0.75, a specificity of 0.85 and an accuracy of 0.82. Descriptive imaging analysis suggested that DW-MRI and DCE-MRI provide complementary information to FDG-PET in STS; however, FMISO-PET did not bring substantial additional value. Dose painting resulted in adequate coverage and homogeneity within the different tumour sub-volumes: i) $D_{95\%}$ to the $PTV_{50\,Gy}$, $GTV_{60\,Gy}$ and $GTV_{65\,Gy}$ were 50.0 Gy, 60.3 Gy and 65.4 Gy, respectively; ii) the homogeneity index (HI) (calculated as the ratio of the $D_{5\%}$ to $D_{95\%}$) for the difference volume of $GTV_{60\,Gy}$ and $GTV_{65\,Gy}$, and for $GTV_{65\,Gy}$, were 1.09 and 1.06, respectively.

Conclusions: Textural biomarkers extracted from pre-treatment images could be used to identify patients that might benefit from dose escalation. The feasibility of this was shown in this patient population, with dose levels of 60 Gy and 65 Gy to intratumoural GTV functional sub-volumes. Moreover, DW-MRI and DCE-MRI techniques are a practical and reliable way to monitor the changing microenvironment of STS during radiotherapy.

4.3 Introduction

Soft tissue sarcomas (STS) comprise a heterogeneous group of tumours arising from mesenchymal tissues. They occur at all ages and in all sites, most commonly the lower extremities. Treatment of STS in adult patients often consists of wide local resection and radiotherapy. Increasingly, pre-operative radiotherapy is favored because of the smaller treated volume and lower dose used in this setting that results in better long-term function compared with postoperative radiotherapy (RT) [1, 2]. With such treatment, local control is over 85 %. However, about 50 % of patients with high grade tumours will develop metastatic disease and require additional treatment, typically chemotherapy or now, for some tumour types, targeted agents [3].

Several studies, including some from our institution, have demonstrated that positron emission tomography (PET) could be used in STS for predicting prognosis, staging the disease and assessing response to therapy [4–7]. ^{18}F-Fluorodeoxyglucose (FDG) PET may be of interest for dose escalation in radiotherapy of STS by targeting the most hypermetabolic and/or regions at high risk for positive margins. STS are heterogeneous tumours that may contain hypoxic regions, and ^{18}F-Fluoromisonidazole (FMISO) uptake could be a useful tool to identify these potential radioresistant regions that could also benefit from a dose boost.

The imaging modality of choice for STS is, however, magnetic resonance imaging (MRI): MRI provides information about the size and location of the tumour as well as its relationship to other structures such as the neurovascular bundle which is important for planning for both surgery and radiotherapy. More complex MRI techniques as compared to standard-of-care sequences may provide valuable information regarding tumour biology: diffusion-weighted magnetic resonance imaging (DW-MRI), for example, could provide information about cellular density, and dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) about vascular characteristics such as blood flow, plasma volume, and mean transit time [8]. In the case of STS, it has been shown that an increase in apparent diffusion coefficients (ADC) obtained from DW-MRI scans is positively correlated with response to therapy [9] and with a decrease in tumour cellularity [10]. Furthermore, it has been suggested that heterogeneity in DCE-MRI pharmacokinetic maps holds potential as a biomarker for STS response to therapy [11, 12].

We have previously developed a retrospective model combining textures from FDG-PET and MRI pre-treatment images to assess tumor aggressiveness in newly diagnosed STS of the extremities using a retrospective cohort of 51 patients [7]. This model was found to possess high potential in predicting the future development of lung metastases, with an area under the receiver operating characteristic curve (AUC) of 0.98 in bootstrapping evaluations. In this work, we tested these findings on a prospective cohort of patients with STS. We then explored the feasibility of using PET and MR imaging information for dose painting as a strategy to improve local control and reduce the risk of developing metastatic disease.

4.4 Materials and Methods

Our primary objective was to collect both anatomic and functional imaging information before, during, and after pre-operative radiotherapy in patients with extremity STS to better understand the underlying biology of STS and predict outcome in terms of distant metastases. Secondary objectives were to evaluate the spatial overlap between tumour sub-regions of FDG-PET, FMISO-PET, DW-MRI and DCE-MRI that could provide a biological rationale for dose painting and to test the feasibility of dose escalation in this patient population.

4.4.1 Patients

Eligible patients were those age $\geq$ 18 years with histologically confirmed primary STS of the extremities without lymph node or distant metastases who were deemed suitable for limb preservation surgery. Patients with rhabdomyosarcoma, Ewing sarcoma/PNET, osteosarcoma or Kaposi sarcoma, or those with contraindications for MRI (e.g., MR unsafe metallic foreign body in the brain or eye, cochlear implant, some types of pacemakers) were not eligible. The study was approved by the Research Ethics Board of the Research Institute of the McGill University Health Centre and all patients provided signed informed consent prior to study entry.

4.4.2 Standard-of-care radiotherapy planning and treatment

Image-guided intensity modulated radiotherapy was applied per our standard practice. The GTV MRI was delineated on the MRI co-registered to the planning CT scan. The CTV margin was +3 cm proximal and distal and +1.5 cm radially, anatomically confined, i.e., not extending into bone or beyond an intact facial barrier or the skin surface. The PTV margin was +5 mm, cropped at 5 mm from the skin. Dose prescription was as follows: minimum 50 Gy in 25 fractions to cover 95 % of the PTV, $>$ 99 % of the PTV to receive $>$ 97 % of the prescribed dose, and $<$ 2 % of the PTV to receive $>$ 110 % of the prescribed dose

4.4.3 Study design

FDG-PET, FMISO-PET, DW-MRI and DCE-MRI images were to be collected pre-radiotherapy (“pre-RT”), mid-radiotherapy (“mid-RT”) and post-radiotherapy (“post-RT”) (Figure 4.1). Image acquisition and registration protocols are provided in Supplementary Material 4.8.2 and 4.8.3, respectively. Standard-of-care MR images acquired for anatomical tumor definition were also collected including T1-weighted (“T1w”), T2-weighted fat-saturated (“T2FS”) and T1-weighted post-injection of a gadolinium contrast agent (“T1w post-gado”) images. We planned a maximum accrual of 20 patients with the expectation that at least 15 would complete all required studies as planned.

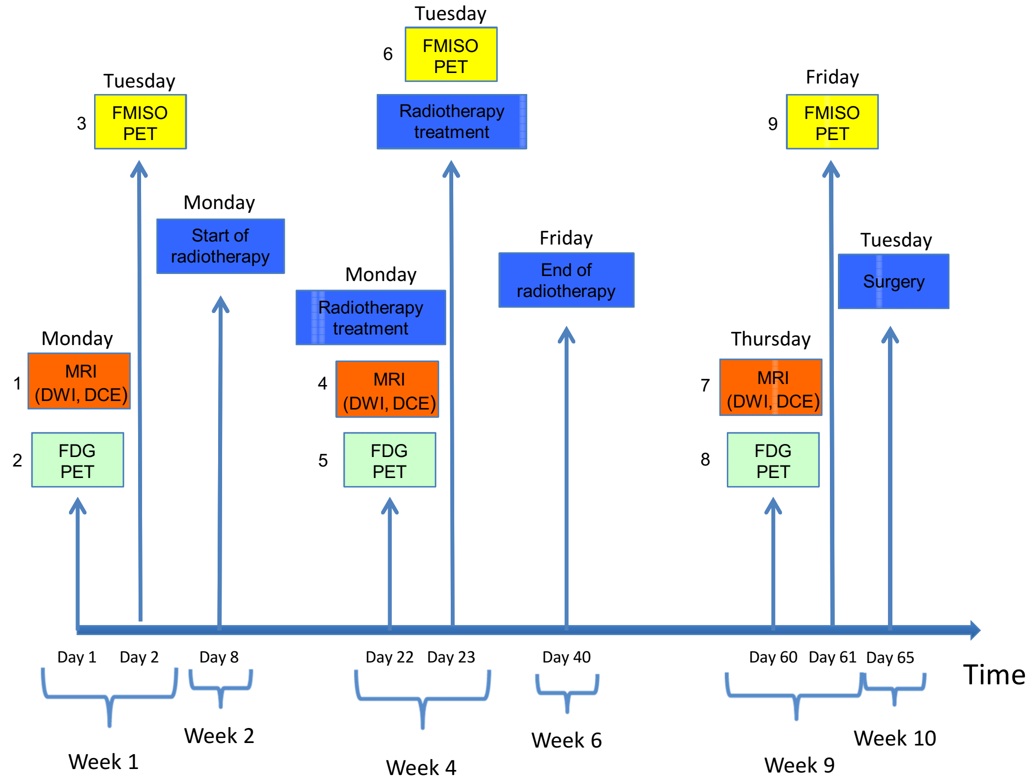


Figure 4.1: Timeline of the study protocol. Imaging studies scheduled at three time-points were performed over two days to accommodate the use of the two PET tracers. Standard-of-care exams included MRI and FDG-PET scans before and after radiation therapy (RT). The mid-RT MRI and FDG-PET scans as well as all DCE-MRI, DWI-MRI and FMISO-PET scans were additional exams.

4.4.4 Image analysis

Validation of texture model

We applied the prediction model developed in our previous work [7] to the patient cohort of this work by extracting and linearly combining texture features from the FDG-PET and MR (T1w and T2FS) images. The performance of the multivariable model response for predicting the future development of lung metastases was assessed using receiver-operating characteristic curve metrics.

Descriptive statistics

FDG-PET and FMISO-PET data were converted into standard SUV maps using injected tracer dose and patient body weight. Apparent diffusion coefficient (ADC) maps were calculated from three DW-MRI sequences acquired with b-values of $100, 500$ and $800\ \mathrm{s/mm^2}$ assuming a standard mono-exponential signal decay model and using a linear fit to the natural logarithm of the pixel data. DCE-MRI data were processed using the Tofts model [13] with a population-based model for the arterial input function [14] to produce maps of the permeability constant $\mathrm{K^{Trans}}$ and interstitial volume v_e. Maps of the initial area under the signal enhancement curve (IAUC) from the injection to 60 seconds post-injection were also extracted from the DCE-MRI data [15]. Descriptive statistics (mean, 25th and 75th percentile) were extracted for each tumour from all images at all available time-points.

Tumour sub-region analysis

Thresholds of the percentage of maximum intensity on imaging scans were manually defined to create discrete high-intensity sub-region contours for each patient. For FDG-PET, FMISO-PET, ADC maps from DW-MRI, and IAUC maps from DCE-MRI, the average thresholds used were (44 ± 8) %, (45 ± 7) %, (47 ± 7) % and (38 ± 8) %, respectively. All images were brought to a common space (MRI) using rigid registration. To analyze the overlap or complementarity of high intensities on the different pre-RT imaging studies, Dice coefficients [16] were calculated between the high-intensity tumour sub-region masks of the different modalities. For longitudinal analysis, high-intensity tumour sub-region contours were created for the mid-RT and post-RT time points using the same thresholds and methods as for the pre-RT time point. This means that the contours could evolve between different imaging time-points and potentially act as markers of tumour response to RT. The percentage volume of high-intensity tumour sub-regions relative to the whole tumour volume was calculated for each imaging modality and for each time point, and results over all patients were summarized using box plots.

4.4.5 Dose painting feasibility study

T1w post-gado, T2FS, DCE-MRI, and FDG-PET/CT were registered to the planning CT scan. The anatomical MRI GTV was contoured on a T1w post-gado and/or T2FS axial MRI. In addition to the standard *anatomical* GTV, we defined a *metabolic* FDG GTV and a *hypoxic* low perfusion DCE-MRI GTV using threshold percentages of maximum values of SUV maps (30 %) and low-perfusion DCE-MRI maps [17] (50 %), respectively. The cumulative margin on the MRI GTV for CTV and PTV expansion ($\mathrm{PTV}_{50\,\mathrm{Gy}}$) was planned to receive the standard prescription dose of $D_{95\%} = 50$ Gy with a maximum dose of 53.5 Gy. The MRI GTV ($\mathrm{GTV}_{53.5\,\mathrm{Gy}}$) was prescribed to a dose of $D_{95\%} \geq 53.5$ Gy. The FDG GTV and low-perfusion DCE-MRI GTV subvolumes of the MRI $\mathrm{GTV}_{53.5\,\mathrm{Gy}}$ were used for dose boosting as follows: i) the FDG GTV ($\mathrm{GTV}_{60\,\mathrm{Gy}}$) was to receive a boost dose of $D_{95\%} \geq 60$ Gy with a maximum dose of 65 Gy; and ii) the low-perfusion DCE-MRI GTV ($\mathrm{GTV}_{65\,\mathrm{Gy}}$) contained within the FDG GTV was to receive a boost dose of $D_{95\%} \geq 65$ Gy with a maximum dose of 70 Gy.

Fourteen patients of the cohort were re-planned with the two levels of dose boosting (60 Gy and 65 Gy) using volumetric arc therapy (VMAT) using an Eclipse treatment planning system (V11.0). The treatment technique consisted of three 6 MV arcs, of which two arcs were designed to cover the entire $\mathrm{PTV}_{50\,\mathrm{Gy}}$ whereas the third arc was designed to cover the boosted volumes only, specifically, the $\mathrm{GTV}_{60\,\mathrm{Gy}}$ and the $\mathrm{GTV}_{65\,\mathrm{Gy}}$ volumes.

4.5 Results

Clinical characteristics of the 18 patients accrued to the study between 2013 and 2016 are given in Supplementary Table 4.2. There were 10 males and 8 females with a median age of 57.5 years (range: 27-80 years). Nine of the 18 patients had tumors in the thigh, four in the shoulder girdle, three in the arm, and two in the leg. Eleven tumors were $>$ 10 cm in size, six were 5-10 cm, and one was $<$ 5 cm. Thirteen of the 18 patients had high-grade tumors. One patient developed local recurrence and this patient and 6 others developed metastatic disease (lung: 5, lymph nodes: 2, bone: 3, liver: one, soft tissue: one, adrenal: one). One patient has died of disease. Twelve remain free of disease at a median of 15 months (range 7-37 months).

Complete imaging data comprising FDG-PET, FMISO-PET, DW-MRI and DCE-MRI were obtained pre-RT in only 14 of the 18 patients. This was due to technical issues with DW-MRI in 3 patients, and with FMISO-PET in one of these and one additional patient. Only 7 of the 18 patients completed all planned imaging studies, mainly due to patient-related factors (practical difficulties with scheduling, claustrophobia, refusal for other reasons).

4.5.1 Prediction of lung metastases development

Our texture-based model performed well when applied to the current patient cohort of this work: the AUC was 0.71, the sensitivity 0.75, the specificity 0.85 and the accuracy 0.82 (Figure 4.2). Development of lung metastases (positive and negative) was correctly predicted for 14 out of 17 patients. Supplementary Figure 4.7 shows FDG-PET imaging examples over the three time points of 4 patients; in two of these the model correctly predicted lung metastases (one positive, one negative) and in two it did not (one positive, one negative). These imaging examples suggest that different tumour sub-regions as defined by the FDG uptake considerably influence the response of the FDG-PET/MRI texture-based model.

4.5.2 General imaging findings

Tumour imaging data generally demonstrated elevated ADC, heterogeneous tumor perfusion elevated in comparison to nearby tissues, and elevated FDG uptake. Qualitatively, FMISO-PET images generally provided little supplementary information as compared to FDG-PET. The progression of relevant prognostic imaging metrics for FDG-PET (75th percentile of SUV distribution), FMISO-PET (75th percentile of SUV distribution), DW-MRI (25th percentile of ADC distribution) and DCE-MRI (mean of K^{Trans} distribution) over the course of RT is shown in Figure 4.3. The progression of these metrics over the course of RT allowed separation of the patients into three groups (increasing, stable, decreasing). For FDG-PET and FMISO-PET (Figure 4.3a and Figure 4.3b) the patient groups are largely overlapping, indicating that FMISO-PET provides only modest supplementary information as compared to FDG-PET. By contrast, for ADC (Figure 4.3c) and K^{Trans} (Figure 4.3d) the patient groups are considerably different from the patient groups observed

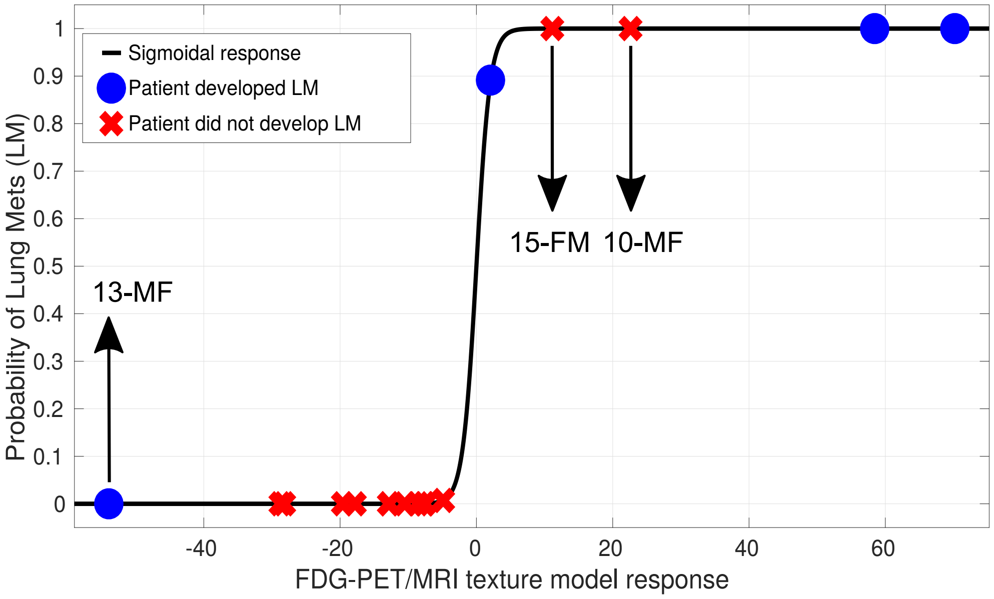


Figure 4.2: Response of the FDG-PET/MRI texture-based model developed in a previous work on a different retrospective cohort of patients [7] when directly applied on the patient cohort of this study. The blue dots represent patients who eventually developed lung metastases and the red crosses those who did not. Lung metastasis development was correctly predicted for all patients except patients 10-MF, 13-MF and 15-FM. The abbreviations MF and FM stand for myxofibrosarcoma and fibromyxoid sarcoma, respectively.

on PET imaging. Also, patients with an overall decrease in mean K^{Trans} distribution after RT seem to first experience an increase in perfusion characteristics at mid-RT, possibly due acute hyperemic response [18]. Furthermore, the radiation response for patients of the same STS subtype (e.g., myxofibrosarcoma) is not always similar (Figure 4.3e). Example images are shown for DW-MRI and DCE-MRI in Supplementary Figure 4.6, and for FDG-PET and FMISO-PET in Supplementary Figure 4.7.

4.5.3 Tumour sub-region analysis

High-intensity tumour sub-region contours for the FDG-PET, FMISO-PET, DW-MRI (ADC map) and DCE-MRI (IAUC map) pre-RT scans of patient 15 are shown in Figure 4.4 (left). Images in Figure 4.4a and Figure 4.4b suggest that FMISO-PET sub-regions do not bring supplementary information as compared to FDG-PET. On the other hand, as expected, ADC and IAUC sub-regions (Figure 4.4c and Figure 4.4d) differ considerably from FDG-PET for that patient. Of note, we observe that the FDG-PET high-intensity subregion seems to be distinctly separated into high- and low-perfusion regions

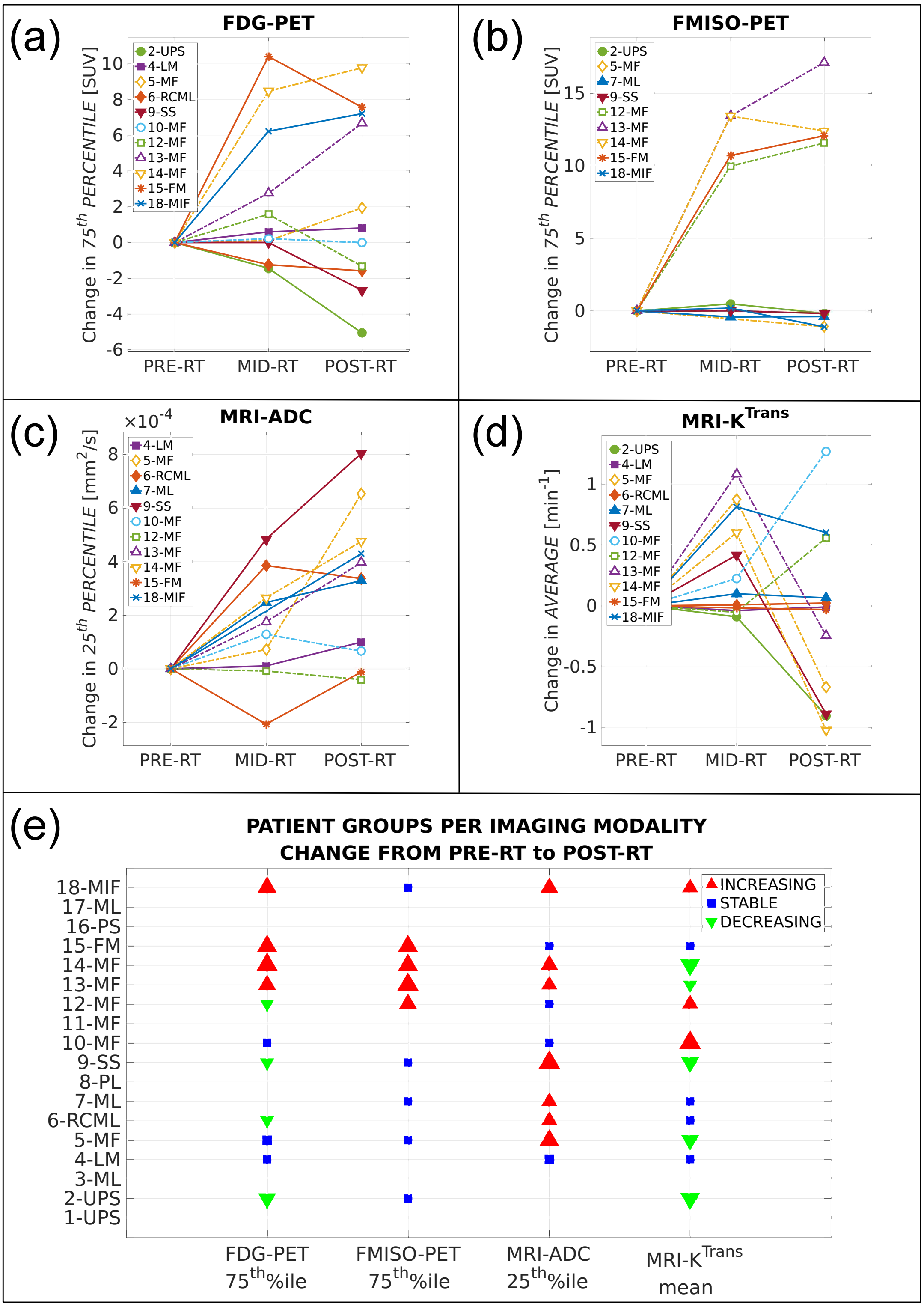


Figure 4.3: Changes in relevant prognostic metrics of FDG-PET, FMISO-PET, DW-MRI and DCE-MRI intratumoural data at 3 time points during radiotherapy (PRE-RT, MID-RT, POST-RT) for patients with available images. (a) 75th percentile of FDG-PET SUV; (b) 75th percentile of FMISO-PET SUV; (c) 25th percentile of apparent diffusion coefficients (ADC); (d) Average of volume transfer coefficients (K^{Trans}); (e) Patient grouping per imaging modality with respect to metric changes from PRE-RT to POST-RT as seen in panels a-b-c-d. The size of the markers is proportional to the relative change. The abbreviations UPS, ML, LM, MF, RCML, PL, SS, FM, PS and MIF stand for undifferentiated pleomorphic spindle cell sarcoma, myxoid liposarcoma, leiomyosarcoma, myxofibrosarcoma, round cell/myxoid liposarcoma, pleomorphic liposarcoma, synovial sarcoma, fibromyxoid sarcoma, pleomorphic spindle cell sarcoma and myxoinflammatory fibroblastic sarcoma, respectively.

obtained from IAUC maps, the latter indicated by the yellow arrows in Figure 4.4d. Using Dice coefficient (s) analysis, we then determined that: i) there is high overlap or low potential of complementarity between high-intensity regions of FDG-PET and FMISO-PET ($s = 0.76 \pm 0.13$); ii) there is medium overlap or medium potential of complementarity between high-intensity regions of FDG-PET and ADC ($s = 0.52 \pm 0.23$), FDG-PET and IAUC ($s = 0.51 \pm 0.15$), FMISO-PET and ADC ($s = 0.58 \pm 0.20$), and FMISO-PET and IAUC ($s = 0.55 \pm 0.15$); and iii) there is low overlap or high potential of complementarity between high-intensity regions of ADC and IAUC ($s = 0.39 \pm 0.17$). In Figure 4.4 (right), the box plots show that the percentage of high-intensity sub-regions generally decreases for all imaging studies as RT progresses, except for ADC for which the trend is unclear.

4.5.4 Dose painting

An example image of low-perfusion DCE-MRI blended with CT of patient 5 is shown in Figure 4.5a, with contours extracted from T1w post-gado ($\mathrm{GTV}_{53.5\,\mathrm{Gy}}$), SUV maps ($\mathrm{GTV}_{60\,\mathrm{Gy}}$) and low-perfusion DCE-MRI ($\mathrm{GTV}_{65\,\mathrm{Gy}}$). An example image of the different boost levels of the dose painting distribution on an axial view of the planning CT of patient 5 is shown in Figure 4.5b. The dose-volume parameters in Table 4.1 demonstrate that adequate coverage and homogeneity can be achieved within the individual tumour sub-volumes. The $\mathrm{D}_{95\%}$ to the $\mathrm{PTV}_{50\,\mathrm{Gy}}$, $\mathrm{GTV}_{60\,\mathrm{Gy}}$ and $\mathrm{GTV}_{65\,\mathrm{Gy}}$ were 50.0 Gy, 60.3 Gy and 65.4 Gy, respectively. The homogeneity index HI (calculated as the ratio of the $\mathrm{D}_{5\%}$ to $\mathrm{D}_{95\%}$) for the difference volume of $\mathrm{GTV}_{60\,\mathrm{Gy}}$ minus $\mathrm{GTV}_{65\,\mathrm{Gy}}$, and for $\mathrm{GTV}_{65\,\mathrm{Gy}}$, were 1.09 and 1.06, respectively.

Table 4.1: Dose-volume parameters for the different dose painting boost levels averaged over 14 patients. The homogeneity index (HI) was calculated as the ratio of $\mathrm{D}_{5\%}$ to $\mathrm{D}_{95\%}$ for differential volumes (i) $\mathrm{PTV}_{50\,\mathrm{Gy}} - \mathrm{GTV}_{53.5\,\mathrm{Gy}}$, (ii) $\mathrm{GTV}_{53.5\,\mathrm{Gy}} - \mathrm{GTV}_{60\,\mathrm{Gy}}$, (iii) $\mathrm{GTV}_{60\,\mathrm{Gy}} - \mathrm{GTV}_{65\,\mathrm{Gy}}$; and volume (iv) $\mathrm{GTV}_{65\,\mathrm{Gy}}$.

Region of interest	$\mathrm{D}_{95\%}$ (Gy)	$\mathrm{D}_{\mathrm{Mean}}$ (Gy)	HI
$\mathrm{PTV}_{50\,\mathrm{Gy}}$	49.99 ± 0.03	52.05 ± 0.45	$1.09 \pm 0.02^{(\mathrm{i})}$
$\mathrm{GTV}_{53.5\,\mathrm{Gy}}$	54.44 ± 2.50	56.70 ± 1.49	$1.17 \pm 0.03^{(\mathrm{ii})}$
$\mathrm{GTV}_{60\,\mathrm{Gy}}$	60.29 ± 0.20	62.65 ± 0.52	$1.09 \pm 0.02^{(\mathrm{iii})}$
$\mathrm{GTV}_{65\,\mathrm{Gy}}$	65.44 ± 0.41	67.59 ± 0.29	$1.06 \pm 0.01^{(\mathrm{iv})}$

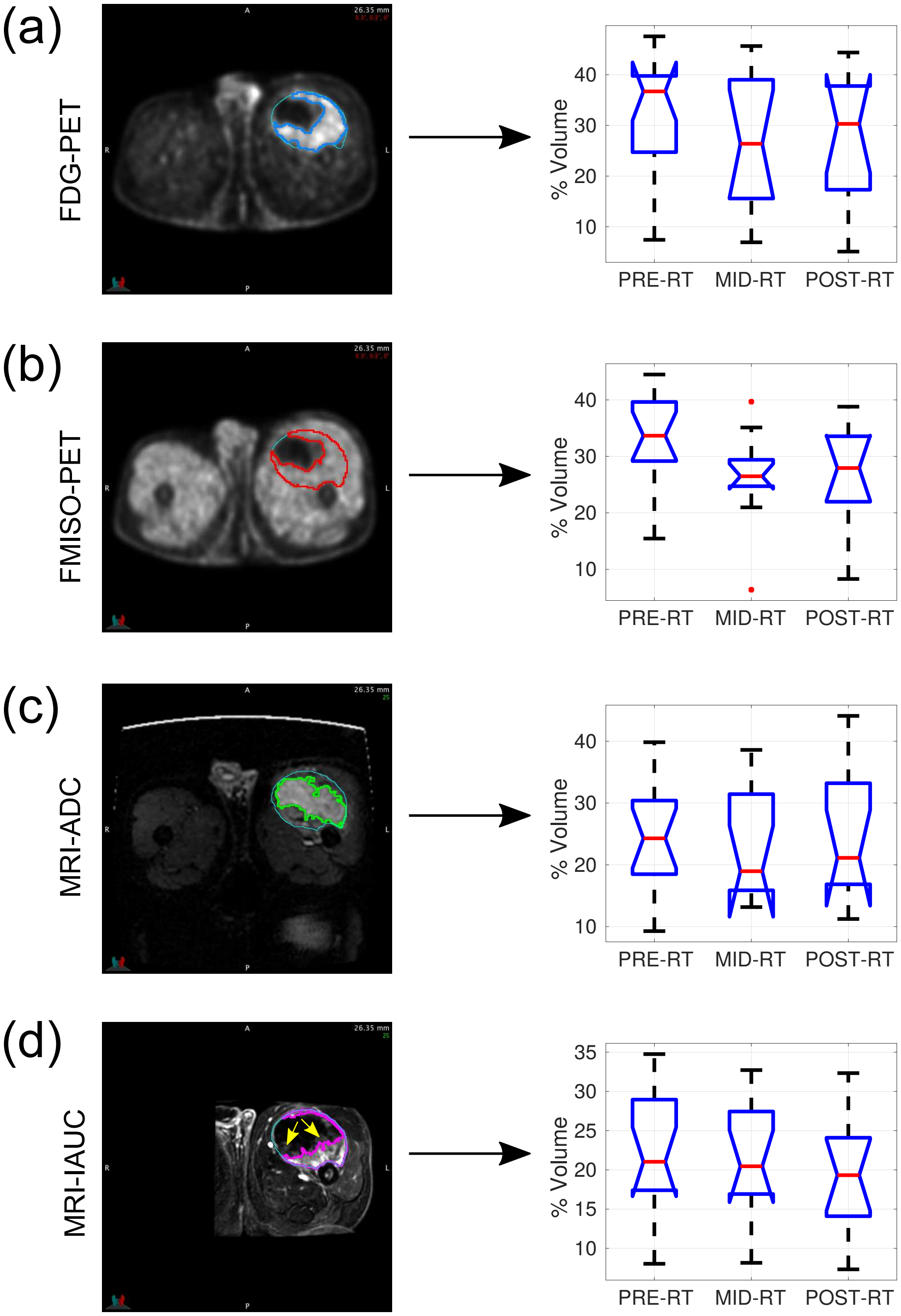


Figure 4.4: Analysis of high-intensity tumour sub-regions over time. (a) FDG-PET scans; (b) FMISO-PET scans; (c) Apparent diffusion coefficient (ADC) maps computed from DW-MRI scans; (d) Maps of the initial area under the signal enhancement curve (IAUC) from the injection to 60 seconds post-injection computed from DCE-MRI scans, with example low perfusion areas identified by yellow arrows. The left column shows example images and high-intensity tumour sub-region contours for the same slice of patient 15 (fibromyxoid sarcoma). The high-intensity tumour sub-region contour is shown in blue, red, green and magenta for FDG-PET, FMISO-PET, ADC and IAUC, respectively. The contour of the whole tumour is shown in cyan in all images. The right column shows notched box plots summarizing the distribution of volume percentages of high-intensity tumour sub-regions over all patients of the cohort, for the three radiotherapy (RT) treatment time points: pre-, mid and post-RT. For each box plot, the median is represented by the red line, the blue box specifies the 25th and 75th percentiles and the whiskers specify the range of the distribution.

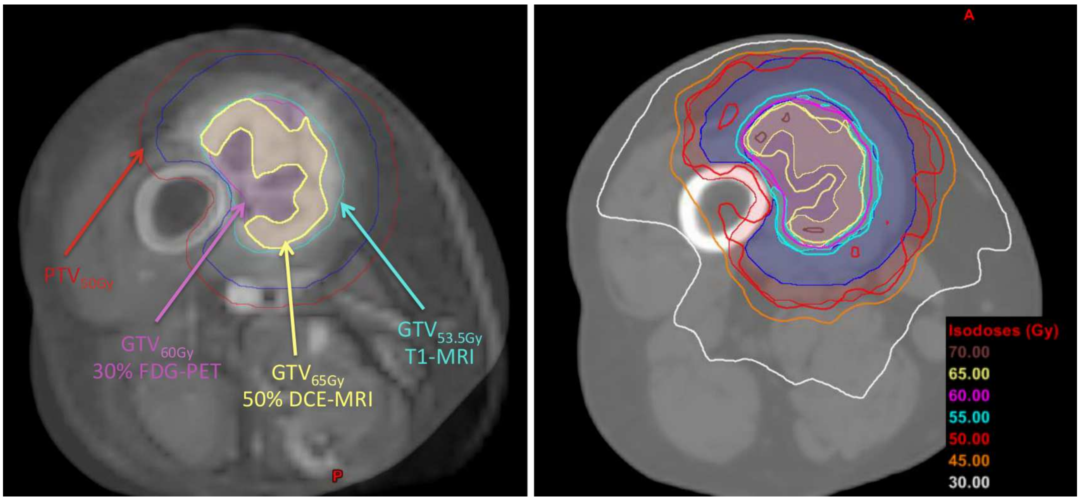


Figure 4.5: Dose painting example. (a) Low-perfusion DCE-MRI and CT blended images of patient 5; (b) Dose painting dose distribution with different levels of dose boosting of patient 5.

4.6 Discussion

Treatment of soft tissue sarcoma has become increasingly standardized over the past two decades, with most patients now being treated at diagnosis with a combination of limb-preserving surgery and radiotherapy. While outcomes in terms of local tumor control and function are generally better than in the past, systemic control has not improved significantly and remains an obstacle to cure, particularly for patients with large, high-grade tumors [3]. This prospective study was thus designed with the objective of using multimodality imaging to improve treatment personalization from the outset. Higher radiotherapy doses to radioresistant components of the tumor might be a useful strategy to reduce the risk of developing metastatic disease, given evidence that intratumoural hypoxia can drive the metastatic phenotype [19, 20]. With this in mind, the goal of this work was to investigate the feasibility of dose escalation to tumour sub-regions inside the GTV based on functional information obtained from PET and MR imaging.

The texture-based model using FDG-PET and MRI that we previously developed [7] performed well in predicting lung metastases development prior to radiotherapy in the new cohort of this study. This information could be useful to identify patients that would benefit the most from dose escalation to different sub-volumes within the GTV.

Changes in anatomical and functional imaging data over the course of radiotherapy were also analyzed. FMISO uptake within the tumours was overall not substantially different from the FDG uptake and generally on the same level as the FMISO uptake within muscles. FDG-PET, DW-MRI and DCE-MRI data displayed considerable inter-patient variability over time, indicating potential for treatment adaptation and personalization. We also observed that progression of simple prognostic metrics from pre- to mid- to post-RT is not consistent between the different imaging modalities. No clear patient grouping using a single or combination of imaging modalities could be defined to predict response to radiotherapy or other outcomes. Exceptions could be myxofibrosarcomas, for which 3 out of 5 patients experienced an overall increase in FDG and FMISO uptake over the course of radiotherapy. This attests to the complexity of STS treatment response management and suggests investigating the use of more complex imaging metrics such as texture features to characterize STS phenotypes.

With regards to dose painting, our goal had first been to achieve an initial level of dose boost to the hypermetabolic and potentially more aggressive tumour sub-regions as seen on the FDG-PET scans. We then attempted to achieve a second level boost with higher dose to the hypoxic volume contained within the hypermetabolic tumour sub-regions, as increasing evidence suggests that glucose demands in hypoxic portions of large tumours are significantly higher than in normoxic cancer cells [21] and that the degree of FDG uptake may indirectly reflect the level of hypoxia [22]. Since in our experience FMISO-PET proved not useful in defining the level of hypoxia in STS as compared to nearby muscles, one approach was to investigate instead, as a surrogate for hypoxia, the use of a low-perfusion DCE volume [17, 23] contained within the high-activity FDG volume, as we observed that in most patients high-activity FDG tumour regions (i.e., excluding the inactive or necrotic part of the tumour) could be distinctly separated into high- and low-perfusion sub-volumes.

An important finding from this work is that dose escalation with VMAT boosting to multiple GTV sub-volumes is technically feasible. It was previously shown that higher radiotherapy doses lead to better local control in retroperitoneal sarcomas [24, 25], and that a boost dose of 57.5 Gy to the margin at risk is well-tolerated in STS patients [26]. In this work, despite the complexity of the multiple targets, it was possible to achieve two levels of dose boost of 60 Gy and 65 Gy within the planning GTV using state-of-the-art radiotherapy systems. However, even if shown feasible from a planning perspective, the question remains as to whether it is desirable to deliver an inhomogeneous dose across the tumour, knowing that most STS change in size during treatment and likely also change significantly in metabolic activity and distribution of well-oxygenated and hypoxic regions. Such changes would need to be monitored frequently, even daily, during radiotherapy, and the treatment to be adapted in real time. In this context, MRI techniques such as DCE-MRI would be preferable to PET, particularly with the advent of novel MR-linac based technologies and knowledge-based planning.

This work represents a unique experience but there are several caveats. First, our cohort was small and heterogeneous with respect to tumour location, size and pathological type. Secondly, uncertainties in PET and MR image registration may have impacted the analysis of the complementarity of the different tumour sub-regions defined from the different imaging modalities. Finally, in our experience, we found that it was not feasible to acquire three imaging scans (FDG-PET, FMISO-PET, MRI) at three different time-points of radiotherapy (pre-RT, mid-RT, post-RT), primarily for practical reasons (mostly patient acceptance). Therefore, our plan going forward will be to focus on the use of longitudinal DW- and DCE-MRI (potentially combined with a pre-treatment FDG-PET) to develop prognostic imaging biomarkers for extremity STS and correlate these studies with genomic biomarkers [27].

In conclusion, FDG-PET and MRI texture features as routinely obtained in our centre prior to radiotherapy could predict development of lung metastases in STS. We plan to test our model in a larger cohort of patients in a prospective multicentre study that will include radio-genomic biomarker analysis and test the hypothesis that DW-MRI and DCE-MRI techniques will prove useful to replace FDG-PET in the model. Finally, despite the complexity of the multiple targets, dose escalation with two levels of boost dose within the planning GTV using a combination of FDG-PET and DCE-MRI is technically feasible. We will continue to explore this interesting approach that could lead to new strategies to improve tumour control.

4.7 Acknowledgments

This work was supported by the Canadian Institutes of Health Research (CIHR) under grants MOP-114910 and MOP-136774. We would also like to thank Asha K. Jeyaseelan and Seema Ambereen for help with patient recruitment and study management, Dr. Lara Hathout for help in writing the initial proposal, Ralf Schirrmacher for supplying the FDG and FMISO, and Drs. Marc Hickeson, Tom Powell and Robert Turcotte for providing expert

knowledge. Special thanks to all the PET and MRI technicians for making this study possible.

4.8 Supplementary Material

4.8.1 Supplementary results

Clinical characteristics of the 18 patients accrued to the study between 2013 and 2016 are given in Supplementary Table 4.2. Example images are shown for DW-MRI and DCE-MRI in Supplementary Figure 4.6, and for FDG-PET and FMISO-PET in Supplementary Figure 4.7.

4.8.2 Image acquisition protocols

^{18}F-Fluorodeoxyglucose PET/CT scan

FDG-PET studies are performed on a hybrid PET/CT scanner (Discovery ST, General Electric Medical Systems, Waukesha, WI, USA), which combines a dedicated, full-ring PET scanner with a 16-slice spiral CT scanner. Patients are required to fast for at least 6 h before the time of their appointment. Blood glucose levels are recorded immediately prior to FDG administration. If the serum glucose level is greater than 11.1 mmol/l (200 mg/dl) the study is rescheduled. A volume of 400 ml of barium sulfate oral contrast is administered and between 370 and 500 MBq (10 and 13.5 mCi) of FDG is injected intravenously. Sixty minutes following FDG injection, CT and PET images are consecutively acquired from the base of the skull to the upper thighs, with additional images acquired as needed according to the STS location.

For the CT scan portion of the study, the settings are the following: 120-140 kVp, 90-110 mA (depending on the body weight), a rotation time of 0.8 s, a table speed of 17 mm per gantry rotation, a pitch of 1.75:1, and a 6 × 0.625 mm detector row configuration. For the PET portion of the study, 2-D acquisition is performed and images are acquired using 4-5 min per bed position (depending on the body weight) and 5 to 6 bed positions (depending on the patient's height). PET attenuation-corrected, PET non-attenuation-corrected, CT, and fused images are reconstructed in the transaxial plane with an ordered subset expectation maximization (OSEM) iterative algorithm.

Table 4.2: Clinical characteristics of patients.

Patient #	Gender	Age	Location (depth)	Size (cm)	Type (grade)	RT dose (Gy/#Fx)	Local failure (time-to-event)	Lung mets (time-to-event)	Other mets (time-to-event)	Follow-up (status)
S001	M	69	Shoulder (superficial)	> 10	UPS (high)	14/7	No	Yes (8 months)	No	10 months (PD)
S002	F	76	Arm (deep)	5 to 10	UPS (high)	50/25	Yes (12 months)	Yes (2 months)	LN/Bone	12 months (PD)
S003	M	46	Leg (deep)	> 10	ML (low)	50/25	No	No	Bone (36 months)	37 months (AWD)
S004	M	55	Thigh (superficial)	> 10	LM (intermediate)	50/25	No	No	No	29 months (NED)
S005	F	61	Thigh (deep)	< 5	MF (high)	50/25	No	No	No	25 months (NED)
S006	F	28	Thigh (deep)	> 10	RCML (high)	50/25	No	No	AW/PV (11/23 months)	29 months (NED)
S007	M	73	Thigh (deep)	> 10	ML (low)	50/25	No	No	No	26 months (NED)
S008	M	80	Shoulder (superficial)	> 10	PL (intermediate)	50/25	No	Yes (9 months)	Bone/Liver (2/9 months)	10 months (AWD)
S009	F	58	Arm (deep)	5 to 10	SS (NA)	50/25	No	No	No	23 months (NED)
S010	M	73	Chest wall (deep)	> 10	MF (high)	50/25	No	No	No	21 months (NED)
S011	F	73	Thigh (deep)	> 10	MF (high)	50/25	No	Yes (9 months)	No	17 months (AWD)
S012	F	62	Thigh (deep)	> 10	MF (high)	50/25	No	No	No	15 months (NED)
S013	M	57	Thigh (superficial)	5 to 10	MF (high)	50/25	No	Yes (6 months)	Reg. LN (5 months)	9 months (DOD)
S014	M	56	Thigh (deep)	5 to 10	MF (high)	50/25	No	No	No	11 months (NED)
S015	M	52	Thigh (deep)	> 10	FM (low)	50/25	No	No	No	14 months (NED)
S016	M	29	Shoulder (deep)	> 10	PS (high)	50/25	No	No	No	7 months (NED)
S017	F	34	Leg (deep)	5 to 10	ML (low)	0	No	No	No	9 months (NED)
S018	F	27	Arm (superficial)	5 to 10	MIF (NA)	50/25	No	No	No	7 months (NED)

* The tumour type abbreviations UPS, ML, LM, MF, RCML, PL, SS, FM, PS and MIF stands for undifferentiated pleomorphic spindle cell sarcoma, myxoid liposarcoma, leiomyosarcoma, myxofibrosarcoma, round cell/myxoid liposarcoma, pleomorphic liposarcoma, synovial sarcoma, fibromyxoid sarcoma, pleomorphic spindle cell sarcoma and myxoinflammatory fibroblastic sarcoma, respectively.

* AWD: Alive with Disease, DOD: Dead of Disease, PD = Patient Deceased, NED: Alive with No Evidence of Disease.

* AW: Abdominal Wall, PV: Paravertebral, LN: Lymph Nodes, NA: Non-Available

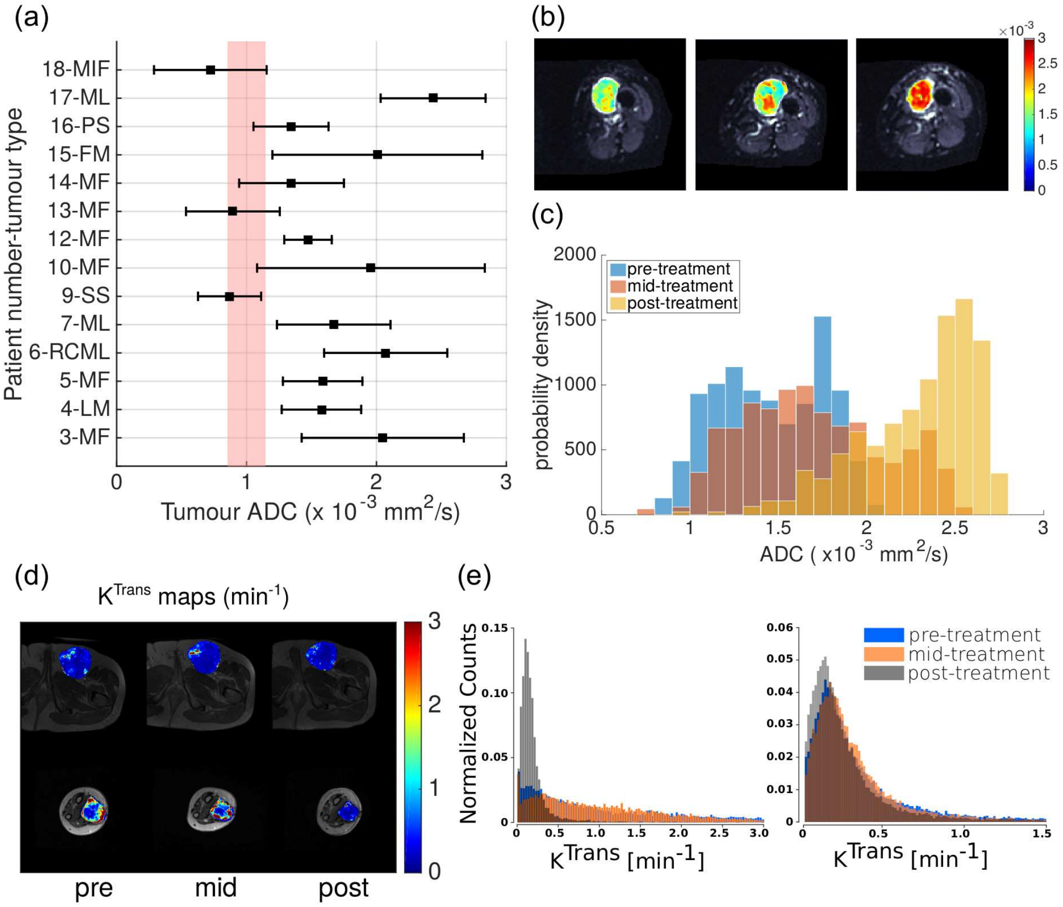


Figure 4.6: General observations from MRI.
(a) The ADC (mean ± standard deviation) for each tumour. The pink box represents the ADC range (one standard deviation) of muscle, computed from muscle regions across all patients. Patient numbers and tumour type abbreviations are given in Table 4.2.
(b) Top, L to R: Example ADC maps (tumour only, colour overlay on axial DW-MRI) from a patient with myxofibrosarcoma (patient 5) show the increase in ADC from pre- to post-radiotherapy.
(c) The therapy-induced increase in ADC in patient 5 is highlighted in the ADC distribution at pre-, mid- and post-treatment exams.
(d) Example K^{Trans} maps in two patients, showing a central slice of the tumour at pre-, mid-, and post-treatment. Top row: patient 4 (leiomyosarcoma in the groin) showing little or no K^{Trans} changes during treatment. Bottom row: patient 9 (synovial sarcoma in the arm), showing evolution of tumour K^{Trans} during treatment.
(e) The histograms show the evolution of K^{Trans} distributions over the treatment course for each of the two patients (left: patient 9, right: patient 4).

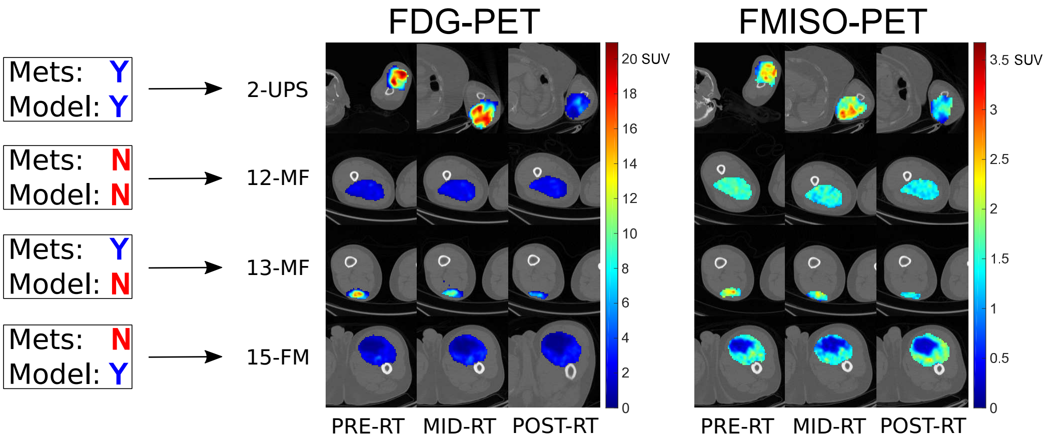


Figure 4.7: Evolution of FDG-PET and FMISO-PET intratumoural uptake of the central slice of the tumour of four example patients over the 3 time points during radiotherapy treatment management (PRE-RT, MID-RT, POST-RT). Note that different colorbars are used for FDG-PET and FMISO-PET. Two patients eventually developed lung metastases (Mets Y: 2-UPS, 13-MF) and two patients did not develop lung metastases (Mets N: 12-MF, 15-FM). The previously developed FDG-PET/MRI texture-based model [7] correctly predicted the lung metastases development status for two patients shown here (2-UPS, 12-MF) and was incorrect for the other two (13-MF, 15-FM). The tumour type abbreviations UPS, MF and FM stands for undifferentiated pleomorphic spindle cell sarcoma, myxofibrosarcoma and fibromyxoid sarcoma, respectively.

^{18}F-Fluoromisonidazole PET/CT scan

FMISO-PET studies are performed on a hybrid PET/CT scanner (Discovery ST, General Electric Medical Systems, Waukesha, WI, USA), which combines a dedicated, full-ring PET scanner with a 16-slice spiral CT scanner. Patients are required to fast for at least 2 hours. Oral contrast (400 ml of barium sulfate) is administered and between 370 and 500 MBq (10 and 13.5 mCi) of FMISO is injected intravenously. One hundred and twenty minutes following FMISO injection, CT and PET images are consecutively acquired from the base of the skull to the upper thighs, with additional images acquired as needed according to the STS location. A period of one day separates the administration of FDG and FMISO. For the CT scan portion of the study, the settings are as for the FDG-PET studies.

Diffusion-weighted MRI

In order to obtain ADC maps, standard echo-planar imaging MRI sequences are acquired using three different diffusion b-values of 0, 100 and 800 s/mm^2. DWI is obtained using the following parameters: FOV of 26 cm, matrix size of 160-256, TR $>$ 3500, TE – minimum, NSA (number of signal averages) of 6, section thickness of 5 mm/1 mm gap with fat suppression.

Dynamic contrast-enhanced MRI

MRI perfusion images are obtained after administration of Gadolinium contrast using 3D FSPGR sequence with the following parameters: TE and TR minimum, Flip angle 25, Bandwidth 42, Matrix size 256 x 128, single excitation and FOV of 24-28 cm.

4.8.3 Image registration

In order to investigate the spatial overlap of PET and MR imaging findings, rigid registration (rotations and translations) is used to spatially transform the MR scans into the reference frame of the PET scans. To perform co-registration, we use the commercial software MIM® (MIM software Inc., Cleveland, OH). MIM® provides an assisted alignment tool that uses normalized mutual information (NMI) as the similarity measure. Practically speaking, to achieve co-registration of MR scans onto PET scans, we first rigidly register the MR images onto the CT images of the combined PET/CT scans. Subsequently to this spatial transformation, we can directly overlay the MR images onto the PET images since the PET and CT images come from the same combined PET/CT scans and are thus in the same reference frame.

4.9 References

1. O'Sullivan, B. *et al.* Preoperative versus postoperative radiotherapy in soft-tissue sarcoma of the limbs: a randomised trial. *Lancet* **359,** 2235–2241 (2002).
2. Davis, A. M. *et al.* Late radiation morbidity following randomization to preoperative versus postoperative radiotherapy in extremity soft tissue sarcoma. *Radiother. Oncol.* **75,** 48–53 (2005).
3. Brennan, M. F. Soft tissue sarcoma: advances in understanding and management. *The Surgeon* **3,** 216–223 (2005).
4. Schwarzbach, M. H. M. *et al.* Prognostic significance of preoperative [18-F] fluorodeoxyglucose (FDG) positron emission tomography (PET) imaging in patients with resectable soft tissue sarcomas. *Ann. Surg.* **241,** 286–294 (2005).
5. Toner, G. C. & Hicks, R. J. PET for sarcomas other than gastrointestinal stromal tumors. *The Oncologist* **13,** 22–26 (2008).

6. Skamene, S. R. *et al.* Metabolic activity measured on PET/CT correlates with clinical outcomes in patients with limb and girdle sarcomas. *J. Surg. Oncol.* **109,** 410–414 (2014).

7. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60,** 5471–5496 (2015).

8. Fayad, L. M., Jacobs, M. A., Wang, X., Carrino, J. A. & Bluemke, D. A. Musculoskeletal tumors: how to use anatomic, functional, and metabolic MR techniques. *Radiology* **265,** 340–356 (2012).

9. Einarsdóttir, H., Karlsson, M., Wejde, J. & Bauer, H. C. F. Diffusion-weighted MRI of soft tissue tumours. *Eur. Radiol.* **14,** 959–963 (2004).

10. Schnapauff, D. *et al.* Diffusion-weighted echo-planar magnetic resonance imaging for the assessment of tumor cellularity in patients with soft-tissue sarcomas. *J. Magn. Reson. Imaging* **29,** 1355–1359 (2009).

11. Van Rijswijk, C. S. P. *et al.* Dynamic contrast-enhanced MR imaging in monitoring response to isolated limb perfusion in high-grade soft tissue sarcoma: initial results. *Eur. Radiol.* **13,** 1849–1858 (2003).

12. Alic, L. *et al.* Heterogeneity in DCE-MRI parametric maps: a biomarker for treatment response? *Phys. Med. Biol.* **56,** 1601–1616 (2011).

13. Tofts, P. S. *et al.* Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. *J. Magn. Reson. Imaging* **10,** 223–232 (1999).

14. Parker, G. J. M. *et al.* Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. *Magn. Reson. Med.* **56,** 993–1000 (2006).

15. Evelhoch, J. L. Key factors in the acquisition of contrast kinetic data for oncology. *J. Magn. Reson. Imaging* **10,** 254–259 (1999).

16. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26,** 297–302 (1945).

17. Stoyanova, R. *et al.* Mapping tumor hypoxia in vivo using pattern recognition of dynamic contrast-enhanced MRI data. *Transl. Oncol.* **5,** 437–447 (2012).

18. Li, S. P. & Padhani, A. R. Tumor response assessments with diffusion and perfusion MRI. *J. Magn. Reson. Imaging* **35,** 745–763 (2012).

19. Bristow, R. G. & Hill, R. P. Hypoxia and metabolism. Hypoxia, DNA repair and genetic instability. *Nat. Rev. Cancer* **8,** 180–192 (2008).

20. Wouters, B. G. & Koritzinsky, M. Hypoxia signalling through mTOR and the unfolded protein response in cancer. *Nat. Rev. Cancer* **8,** 851–864 (2008).

21. Li, X.-F., Du, Y., Ma, Y., Postel, G. C. & Civelek, A. C. 18F-Fluorodeoxyglucose uptake and tumor hypoxia: revisit 18F-Fluorodeoxyglucose in oncology application. *Transl. Oncol.* **7,** 240–247 (2014).

22. Dierckx, R. A. & Van de Wiele, C. FDG uptake, a surrogate of tumour hypoxia? *Eur. J. Nucl. Med. Mol. Imaging* **35,** 1544–1549 (2008).

23. Cho, H. *et al.* Noninvasive multimodality imaging of the tumor microenvironment: registered dynamic magnetic resonance imaging and positron emission tomography studies of a preclinical tumor model of tumor hypoxia. *Neoplasia* **11,** 247–259 (2009).

24. Fein, D. A. *et al.* Management of retroperitoneal sarcomas: does dose escalation impact on locoregional control? *Int. J. Radiat. Oncol. Biol. Phys.* **31,** 129–134 (1995).

25. Feng, M. *et al.* Long-term outcomes after radiotherapy for retroperitoneal and deep truncal sarcoma. *Int. J. Radiat. Oncol. Biol. Phys.* **69,** 103–110 (2007).

26. Tzeng, C.-W. D. *et al.* Preoperative radiation therapy with selective dose escalation to the margin at risk for retroperitoneal sarcoma. *Cancer* **107,** 371–379 (2006).

27. Ybarra, N. *et al.* Correlation of molecular imaging and biomarkers expression in the prediction of metastatic capacity of soft tissue sarcomas. *Int. J. Radiat. Oncol. Biol. Phys.* **96,** E705–E706 (2016).

Chapter 5

Enhancement of radiomic-based prediction models via the optimization of imaging acquisition protocols

5.1 Foreword 154
5.2 Abstract 154
5.3 Introduction 155
5.4 Methods 160
5.4.1 Imaging dataset 160
5.4.2 Construction of a radiomic prediction model 161
5.4.3 Simulations of PET imaging acquisitions 163
5.4.4 Simulations of MR imaging acquisitions 168
5.4.5 STAMP: a software tool for texture optimization 172
5.4.6 Optimization of a texture-based model 174
5.5 Results 175
5.5.1 Texture variations with acquisition parameters 175
5.5.2 Evaluation on a texture-based model 181
5.6 Discussion 185
5.7 Conclusion 189
5.8 Acknowledgments 189
5.9 Supplementary Material 190
5.9.1 Description of 3D radiomic features 190
5.9.2 Construction of radiomic models 190
5.9.3 Software integration for texture analysis of PET and MR simulated images 191
5.9.4 Comparison of clinical and simulated textures 194
5.10 References 195

5.1 Foreword

This Chapter presents a study submitted as the following paper: Martin Vallières, Sébastien Laberge, André Diamant & Issam El Naqa. "Enhancement of multimodality texture-based prediction models via optimization of PET and MR image acquisition protocols: a proof of concept". *Phys. Med. Biol.* [submitted May 15, 2017].

In this study, a new texture-based model for the prediction of lung metastases in soft-tissue sarcomas was constructed using the methods developed in Chapter 3, this time using a subset of the initial patient cohort. The possibility of enhancing that texture-based model via the optimization of PET and MR image acquisition protocols was investigated using computerized simulations.

5.2 Abstract

Texture-based radiomic models constructed from medical images have the potential to support cancer treatment management via personalized assessment of tumour aggressiveness. While the identification of stable texture features under varying imaging settings is crucial for the translation of radiomics analysis into routine clinical practice, we hypothesize in this work that a complementary optimization of image acquisition parameters prior to texture feature extraction could enhance the predictive performance of texture-based radiomic models. As a proof of concept, we evaluated the possibility of enhancing a model constructed for the early prediction of lung metastases in soft-tissue sarcomas by optimizing PET and MR image acquisition protocols via computerized simulations of image acquisitions with varying parameters. Simulated PET images from 30 STS patients were acquired by varying the extent of axial data combined per slice ("span"). Simulated T_1-weighted and T_2-weighted MR images were acquired by varying the repetition time (TR) and echo time (TE) in a spin-echo pulse sequence, respectively. We analyzed the impact of the variations of PET and MR image acquisition parameters on individual textures, and we also investigated how these variations can enhance the global response and the predictive properties of a texture-based model. Our results suggest that it is feasible to identify

an optimal set of image acquisition parameters to improve prediction performance. The model constructed with textures extracted from simulated images acquired with a standard *clinical* set of acquisition parameters reached an average AUC of 0.84 ± 0.01 in bootstrap testing experiments. In comparison, the model performance significantly increased using an *optimal* set of image acquisition parameters ($p = 0.04$), with an average AUC of 0.89 ± 0.01. Ultimately, specific acquisition protocols optimized to generate superior radiomics measurements for a given clinical problem could be developed and standardized via dedicated computer simulations and thereafter validated using clinical scanners.

5.3 Introduction

Medical imaging is foreseen to play a central role in the near future to better assess tumour aggressiveness in the context of cancer treatment management, as radiological images are routinely acquired for almost every patient with cancer [1]. Medical image acquisitions such as 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) positron emission tomography (PET), computed tomography (CT) or magnetic resonance imaging (MRI) are minimally invasive and they would carry an immense source of data that could serve as useful complementary tools to histopathological information for decoding tumour phenotypes [2]. The demonstration that gene-expression signatures could be inferred from tumour imaging features [3, 4] has led to an exponential growth of the new emerging field of "radiomics" in the past few years [5–8]. The fundamental hypothesis of radiomics is that the microscopic genomic heterogeneity of aggressive tumours would translate into macroscopic heterogeneous tumour metabolism and anatomy. In essence, radiomics thus refers to the extraction of high-dimensional mineable data (morphological and histogram-based features, textures, etc.) from all types of medical images, whose subsequent analysis aims at supporting clinical decision-making. In particular, textural metrics such as Gray-Level Co-Occurrence Matrix (GLCM) features [9], Gray-Level Run-Length Matrix (GLRLM) features [10–12], Gray-Level Size Zone (GLSZM) features [13] and Neighborhood Gray-Tone Difference Matrix (NGTDM) features [14], could extensively characterize the complexity of imaging intensities within tumours. Tumours exhibiting heterogeneous characteristics are thought to be associated with high risk of resistance to treatment, progression, metastasis or recurrence

[15–17], and these textural metrics are thus considered to have great potential for the assessment of tumour aggressiveness via the quantification of intratumoural heterogeneity. In the era of personalized medicine, the translation of radiomics analysis into standard cancer care involves the development of multivariable prediction models that can assess the risk of specific tumour outcomes [18]. Once useful imaging biomarkers are identified to be relevant prognostic factors of a given tumour outcome, models combining these factors may be constructed to improve outcome prediction performance, as multivariable models are expected to more comprehensively characterize intratumoural heterogeneity than single features. Overall, radiomics or texture-based models could soon complement other prognostic models currently used in routine clinical practice only if they are trusted to be highly robust, reproducible, generalizable and yet also highly predictive.

The workflow of radiomics analysis leading to the extraction of clinically relevant information involves many steps such as medical imaging acquisition, image processing, tumour segmentation, feature extraction, statistical analysis, and development and validation of multivariable models for tumour outcome prediction via statistical or machine learning techniques. The complexity of such workflow opens the door to many interesting development possibilities in the field, but it can also considerably affect the reproducibility potential of different radiomics studies and the possible use of radiomics in routine clinical settings [19, 20]. Many studies have investigated how procedural variations in single of multiple steps of this workflow may impact texture measurements for different imaging modalities (PET, CT, MRI). For example, Bogowicz *et al.* [21] and Molina *et al.* [22] studied the impact of variations in voxel size and image quantization on texture features extracted from patient images. Hatt *et al.* [23], Leijenaar *et al.* [24], Parmar *et al.* [25] and Van Velden *et al.* [26] studied the impact of contouring variations on texture features extracted from patient images, and Hatt *et al.* [23] also studied the impact of partial volume effect (PVE) corrections in PET. Orlhac *et al.* [27] performed a comprehensive analysis of the relationships of texture measurements with commonly extracted metrics in PET and of the robustness of textures with different quantization schemes and contouring variations using patient images of three different cancer types. Tixier *et al.* [28], Leijenaar *et al.* [24], Zhao *et al.* [29] and Van Velden *et al.* [26] performed test-retest scans (e.g., two scans of the same patient repeated after a short period of time) on patient images to study the reproducibility of texture measurements. Galavis *et al.* [30], Nyflot *et al.* [31], Yan *et al.* [32], Zhao *et al.* [29] and Van Velden *et* *al.* [26] studied the impact of variations in different image reconstruction parameters on texture features extracted from patient images. Mayerhoefer *et al.* [33], Waugh *et al.* [34], Zhao *et al.* [35] and Mackin *et al.* [36] performed phantom studies to explore the variability of textures under different scanning conditions. The greater flexibility in image acquisition settings by using phantoms instead of real patient scans allowed Mackin *et al.* [36] to further evaluate variations in textures from images acquired on different CT scanners (inter-scanner dependence), whereas Mayerhoefer *et al.* [33] and Waugh *et al.* [34] were able to investigate the influence of different MR image acquisition protocols (e.g., echo time, repetition time, etc.) on textures. Galavis *et al.* [30] were also able to investigate the influence of different PET image acquisition protocols (2D versus 3D acquisitions) using a group of patients with solid tumours. Last but not least, Nyflot *et al.* [31] performed an interesting Monte-Carlo simulation analysis using the NEMA image quality phantom to study (among other parameters) the impact of stochastic effects on textural features in PET. While all the studies enumerated above are informative in terms of how texture measurements vary in different settings, the inherent disadvantage of test-retest scans, for instance, is that the reproducibility effect could be confounded with setup errors and/or organ deformations. Phantom studies allow for much greater flexibility in understanding scanning conditions and for negligible positioning differences, but the drawback is that tumour phantoms may not realistically reflect the intratumoural heterogeneity seen in patients. On the other hand, computerized simulations of medical image acquisitions using realistic tumour models would offer a fully controlled environment to study the effects of different acquisition parameters on textural measurements of intratumoural heterogeneity, but the resulting simulated images are only an estimated representation of images acquired using clinical scanners.

Overall, most reproducibility studies in the literature report the high impact of voxel size on texture measurements. However, the common denominator of all those studies is their main working objective: they aim at identifying the texture features that could be stable and that are presumably able to conserve predictive properties under varying imaging conditions. While the identification of stable features is valuable to build robust and reproducible texture-based predictive models, it is also important to identify the settings that would yield optimal use of texture features for a given clinical problem, an exercise which is currently under-reported in the literature. In our previous study, we hypothesized and thereafter verified that the optimization of voxel size and image quantization parameters could enhance the predictive properties of the resulting texture-based models [37]. In this work, we hypothesize that the optimization of image acquisition parameters prior to texture feature extraction could not only assess robustness of texture features but also enhance the performance of multivariable prediction models. As a proof of concept, we evaluate the possibility of enhancing a combined PET and MRI texture-based model constructed for the early prediction of lung metastases in soft-tissue sarcomas (STSs) by optimizing PET and MR image acquisition protocols via computerized simulations of clinical image acquisitions with varying parameters. Realistic digital tumour models are first constructed from clinical images for both PET and MRI simulations with the intent of conserving intratumoural heterogeneity in simulated images. Simulated PET images are then acquired by varying the extent of axial data combined per slice when detectors are allowed to be in coincidence with detectors in neighboring rings (Figure 5.1a), an effect denoted as "span" and that has for consequence to increase slice sensitivity at the expense of a loss of resolution. Simulated T_1-weighted and T_2-weighted MR images are acquired using a spin-echo sequence with standard clinical parameters as typically used at our institution. The repetition time (TR) and the echo time (TE) are then varied in the acquisition of T_1-weighted and T_2-weighted images (Figure 5.1b), respectively, a procedure that would change the contrast in the resulting images. We then analyze the impact of the variations of these PET and MR image acquisition parameters on single textures, and we also investigate how these variations can enhance the global response and the predictive properties of the texture-based model. Our results suggest that different sets of PET and MR image acquisition parameters can substantially affect the resulting extracted textures, and that it could be possible to identify an optimal set of acquisition parameters yielding best prediction performance for a texture-based model. To our knowledge, this is the first study that would explore the potential of varying image acquisition parameters to optimize the performance of texture features and enhance texture-based predictive models. Overall, the simulations of medical imaging acquisitions using realistic digital tumour models would provide a useful and effective framework to study how texture measurements may vary in different acquisition settings and how they could be optimized for a particular application.

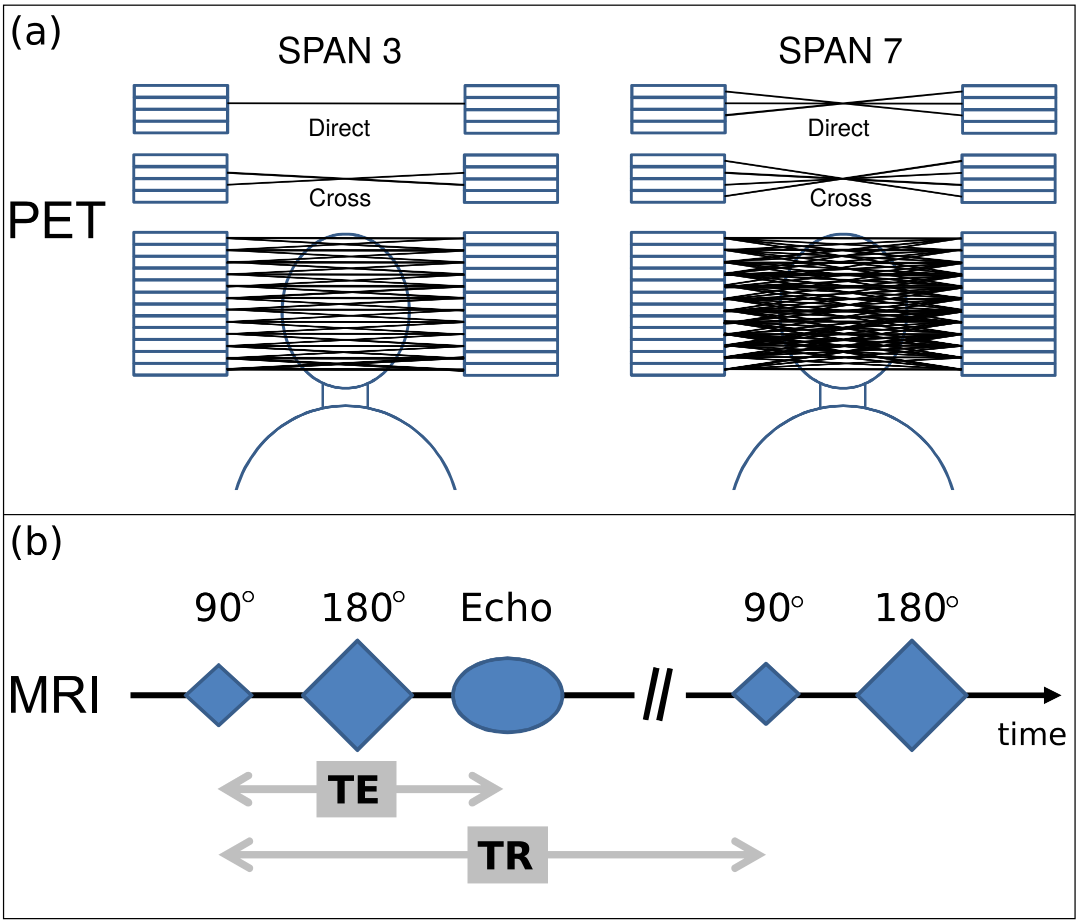


Figure 5.1: Imaging acquisition parameters varied in this study. (a) Simulated PET images are acquired by varying the extent of axial data combined per slice when detectors are allowed to be in "direct" or "cross" coincidence with detectors in neighboring rings, an effect denoted as "span". For example, a span of 3 has the effect to combine 1 plane of direct coincidences and 2 planes of cross coincidences, whereas a span of 7 has the effect to combine 3 planes of direct coincidences and 4 planes of cross coincidences. Spans of 3, 5, 7, 9, 11, 13, 15 and 17 are tested in this study. This image is adapted from Fahey [38]. (b) Simulated T_1-weighted and T_2-weighted MR images are acquired using a spin-echo sequence with standard clinical parameters (schematic view of the 90° excitation pulse, the 180° refocusing pulse and the signal echo used for data acquisition shown here), by varying the repetition time (TR) of the T_1-weighted sequence and the echo time (TE) of the T_2-weighted sequence. Repetition and echo times of $\frac{1}{3}$, $\frac{1}{2}$, 1, 2 and 3 times the values used in the original clinical sequences of each corresponding patient are tested in this study.

5.4 Methods

5.4.1 Imaging dataset

An imaging dataset with histologically proven STSs of the extremities was downloaded from The Cancer Imaging Archive (TCIA) [39]: LINK. This 51 patients dataset has been described in details in our previous work [37]. Briefly, all patients received: 1) pre-treatment FDG-PET/CT scans; and 2) pre-treatment MRI scans consisting of T_1-weighted clinical sequences (hereby denoted as "T1"), and either T_2-weighted fat-saturated clinical sequences (hereby denoted as "T2FS") or short tau inversion recovery (STIR) sequences. In this study, only the subset of 30 patients for whom both the T1 and T2FS MRI sequences were acquired were retained.

From the 30 STS patients used in this study, 11 patients developed lung metastases (hereby denoted as "*Lung Mets*" patients) during the follow-up period (median: 25 months, range: 4–70 months). Patients that did not develop lung metastases (hereby denoted as "*No Lung Mets*" patients) and that had a follow-up period smaller than 12 months were excluded from the study, as well as patients with metastatic and/or recurrent STS at presentation. Lung metastases were either proven by biopsy or diagnosed by an expert physician from the appearance of typical pulmonary lesions on CT and/or FDG-PET images.

All FDG-PET/CT scans were performed on a PET/CT scanner (Discovery ST, GE Healthcare, Waukesha, WI) at the McGill University Health Centre (MUHC). For the PET portion of the scans, a median of 420 MBq (range: 210–620 MBq) of FDG was injected intravenously. Approximately 60 min following the injection, whole-body imaging acquisition was performed using multiple bed positions, with a median of 180 s (range: 160–300 s) per bed position. PET attenuation corrected images were reconstructed (axial plane) using an ordered subset expectation maximization (OSEM) iterative algorithm. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 5.47×5.47 mm^2 (range: 3.91–5.47 mm).

The MRI scans resulted from clinical acquisitions with non-uniform protocols across patients. Twelve patients had their images acquired at the MUHC, and 18 in outside institutions. All images were acquired on a scanner with a 1.5 Tesla (T) magnet. Overall, the median in-plane resolution was 0.74 $\times$ 0.74 mm^2 and 0.63 $\times$ 0.63 mm^2 (range: 0.23–1.64 mm and 0.23–1.64 mm pixel width), and the median slice thickness was 5.5 mm and 5.0 mm (range:

3.0–10.0 mm and 3.0–8.0 mm) for T1 and T2FS scans, respectively. T1 sequences were acquired in the axial plane for all patients. T2FS sequences were acquired in the axial plane for 25 patients, and in the sagittal plane for 5 patients.

Contours defining the 3D tumour region for each patient were manually drawn slice-by-slice on T2FS scans by an expert radiation oncologist. Contours were propagated to FDG-PET and T1 scans using rigid registration with the software MIM® (MIM software Inc., Cleveland, OH).

5.4.2 Construction of a radiomic prediction model

The process of constructing a radiomic model for the prediction of lung metastases in STSs from the set of PET and MR images of the 30 patients of this cohort closely follows the work of Vallières *et al.* [37] and is depicted in Figure 5.2a. First, radiomic features were extracted from the tumour region of PET, T1 and T2FS images. These features can be divided into three different groups: I) 10 first-order statistics features (intensity); II) 5 morphological features (shape); and III) 40 texture features each computed using 40 different combinations of extraction parameters. The 40 texture features were extracted for each scan using all 40 possible combinations of 5 isotropic voxel sizes (“scale”), 2 quantization algorithms (“algo”) and 4 number of gray-levels (“Ng”). Then, feature set reduction was performed from the total set of radiomic features in order to create a reduced feature set balanced between predictive power and non-redundancy. From the reduced feature set, stepwise forward feature selection was carried out to automatically select combinations of 1 to 10 features (i.e., model orders of 1 to 10). The process of combining *p* radiomic features was achieved using the logistic regression utilities of the software DREES [40] such that the multivariable model investigated in this work takes the following form:

$$g(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \text{ for } i = 1, 2, \ldots, N. \tag{5.1}$$

In Equation 5.1 the vector of input variables *j* (imaging data) of the i^{th} patient is $\mathbf{x}_i = \{x_{ij} \in \mathbb{R} : j = 1, 2, \ldots, p\}$, and the set $\boldsymbol{\beta} = \{\beta_j \in \mathbb{R} : j = 0, 1, \ldots, p\}$ is the set of regression coefficients of the model to be determined such that the conditional probability of the set of outcome states {0,1} given the input data $\mathbf{x}_i$ is maximized for $i = 1, 2, \ldots, N$. The model response $g(\mathbf{x}_i)$ can be transformed into the posterior probability $\pi(\mathbf{x}_i)$ of observing outcome $y_i = 1$

(i.e., developing lung metastases) given the input $\mathbf{x}_i$ by using the following logit transform:

$$\pi(\mathbf{x}_i) = \mathrm{P}\left(y_i = 1|\mathbf{x}_i\right) = \frac{\exp\left[g(\mathbf{x}_i)\right]}{1 + \exp\left[g(\mathbf{x}_i)\right]}. \tag{5.2}$$

Then, prediction performance was estimated for the 10 combinations of features using the 0.632+ bootstrap AUC metric [41, 42] with 1000 bootstrap samples. By inspecting the prediction estimates, the simplest model with best predictive properties was chosen, yielding the following four features: I) *SUVpeak* extracted from PET ($\mathsf{SUVpeak}$); II) *High Gray-Level Zone Emphasis (HGZE)* extracted from PET ($\mathsf{PET - HGZE}_{\mathsf{GLSZM}}$); III) *Zone Size Variance (ZSV)* extracted from T1 ($\mathsf{T1 - ZSV}_{\mathsf{GLSZM}}$); and IV) *Long Run Low Gray-level Emphasis (LRLGE)* extracted from T2FS ($\mathsf{T2FS - LRLGE}_{\mathsf{GLRLM}}$). The final logistic regression coefficients of this radiomic prediction model were found by averaging all coefficients computed from another set of 1000 bootstrap samples. Throughout the whole model building process, bootstrap resampling was performed using imbalance-adjustments in order to construct a model with equivalent sensitivity and specificity properties. Detailed descriptions and methods used for radiomic feature extraction, feature set reduction and feature selection are provided in sections 5.9.1, 5.9.2 and 5.9.2 of Supplementary Material, respectively.

The complete model $g(\mathbf{x}_i)$ obtained in this work from clinical scans for the prediction of lung metastases in STSs is detailed in Equation 5.3 with regression coefficients and texture extraction parameters. The model responses for each of the 30 patients of the cohort along with associated bootstrap confidence intervals (95 %) are shown in Figure 5.2b. In this figure, the blue dots represent patients who eventually developed lung metastases, and the red crosses those who did not develop lung metastases. It can be seen that the multivariable model of Equation 5.3 can appreciably separate the patients of the two risk groups, as the average AUC obtained in 1000 bootstrap testing samples after training the model in the corresponding bootstrap training samples ("bootstrap AUC" or ordinary AUC) is 0.85. The significance of each variable in the model was also assessed using the Wald's test implemented in DREES, and a *p*-value of 0.048, 0.088, 0.16 and 0.28 was obtained for the $\mathsf{SUVpeak}$, $\mathsf{PET - HGZE}_{\mathsf{GLSZM}}$, $\mathsf{T1 - ZSV}_{\mathsf{GLSZM}}$ and $\mathsf{T2FS - LRLGE}_{\mathsf{GLRLM}}$ features, respectively.

$$\begin{aligned} g(\mathbf{x}_i) = & \\ & + 1.312 \times \text{SUVpeak} \\ & + 0.0193 \times \text{PET(scale=5mm,algo=Equal,Ng=64)} - \text{HGZE}_{GLSZM} \\ & - 357600 \times \text{T1(scale=4mm,algo=Uniform,Ng=64)} - \text{ZSV}_{GLSZM} \\ & + 117.3 \times \text{T2FS(scale=2mm,algo=Equal,Ng=64)} - \text{LRLGE}_{GLRLM} \\ & - 40.34 \end{aligned} \tag{5.3}$$

5.4.3 Simulations of PET imaging acquisitions

PET tumour models

Papadimitroulas *et al.* [43] recently showed that partial volume effect (PVE) corrections of intratumoural activity of FDG-PET clinical scans as input to GATE simulations is necessary to conserve intratumoural heterogeneity in simulated images as quantified via imaging profiles and textural features. Overall, PVE corrections reduce the effect of 3D blurring in PET images caused by the convolution of the source and the PSF of the imaging system. In this work, the method used for the creation of realistic FDG-PET tumour models is based on the work of Boussion *et al.* [44], where wavelet-based denoising is incorporated into an iterative deconvolution algorithm for PVE correction of the activity data of FDG-PET clinical images. Figure 5.3a shows and example of the creation of a FDG-PET tumour model from a clinical image via PVE correction and wavelet-based denoising.

The general approach involves an iterative deconvolution process aided by a wavelet-based threshold procedure in order to reduce noise and obtain a better representation of the underlying activity distribution of a given PET clinical image. First, the general deconvolution framework is based on the following equation:

$$I(\vec{r}) = O(\vec{r}) \bigotimes PSF(\vec{r}) + N(\vec{r}), \tag{5.4}$$

where I is here defined as the observed "out of the scanner" activity distribution calculated from whole-body PET clinical images, O is the "real" or corrected activity distribution (that we attempt to retrieve using PVE corrections), PSF is the degrading PSF of the scanner and N is an additive noise

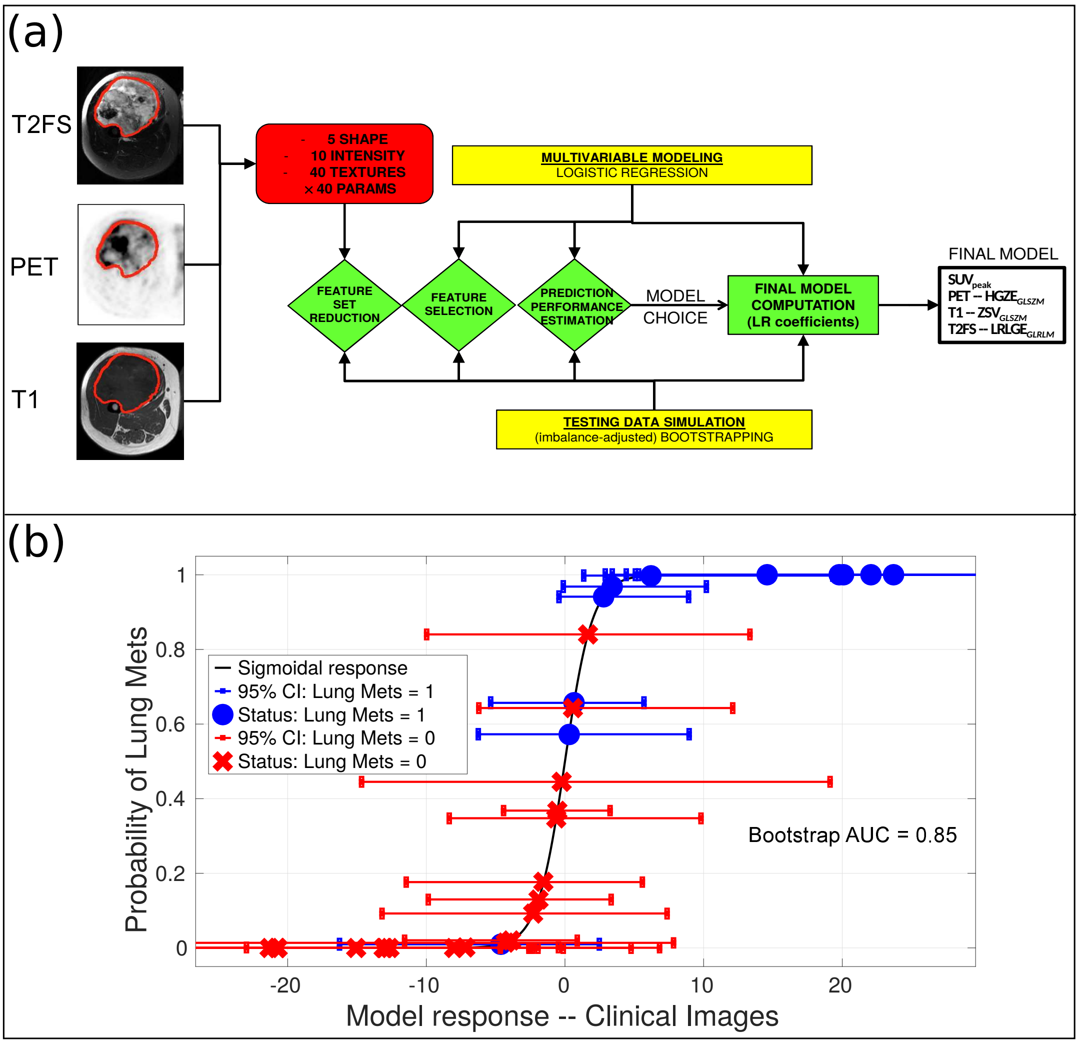


Figure 5.2: Construction of a radiomic model from pre-treatment clinical images for the prediction of lung metastases in soft-tissue sarcomas. (a) Radiomic features (intensity, shape, textures) are first extracted from PET, T_1-weighted (T1) and T_2-weighted fat-saturated (T2FS) clinical images. Feature set reduction is then performed to obtain a reduced feature set balanced between predictive power and non-redundancy. Feature selection and prediction performance processes are then performed using imbalance-adjusted bootstrap resampling in order to estimate the generalizability properties of tested models constructed with equivalent sensitivity and specificity properties. The single combination of variables with best parsimonious properties is then chosen based on prediction performance estimations evaluated with the AUC_{632+} metric, yielding the following final radiomic model investigated in this work: I) *SUVpeak* extracted from PET; II) *High Gray-Level Zone Emphasis (HGZE)* extracted from PET; III) *Zone Size Variance (ZSV)* extracted from T1; and IV) *Long Run Low Gray-level Emphasis (LRLGE)* extracted from T2FS. Variables are combined using logistic regression. (b) Probability of developing lung metastases as a function of the response of the final radiomic model constructed using PET, T1 and T2FS clinical images, for all patients of the cohort. The blue dots represent patients who eventually developed lung metastases, and the red crosses those who did not develop lung metastases. Confidence intervals (95 %) on the model response for each patient were calculated using bootstrapping.

term. In this work, the Lucy-Richardson algorithm [45, 46] was used to iteratively retrieve the object O from the observed data I using a multiplicative regularization step instead of additive [44]:

$$O^{n+1}(\vec{r}) = O^{n}(\vec{r}) \left[\frac{I^{n}(\vec{r}) + Res^{n}(\vec{r})}{I^{n}(\vec{r})} \bigotimes PSF(-\vec{r}) \right], \tag{5.5}$$

where $I^{n}(\vec{r}) = PSF(\vec{r}) \bigotimes O^{n}(\vec{r})$ and $Res^{n}(\vec{r}) = I(\vec{r}) - PSF(\vec{r}) \bigotimes O^{n}(\vec{r})$ is the residual term that converges towards noise. A number of four iterations and an isotropic 5-mm PSF were used in this work. In order to further reduce noise propagation, the residual term is modified before each iteration of the Lucy-Richardson algorithm. For this process, a soft threshold was applied in each subband of the wavelet domain using the biorthogonal 3.5 basis function and a 3-level decomposition of the residual, and the inverse wavelet transform was applied to obtain the denoised residual. To ensure translation-invariance, a 2D undecimated wavelet transform was applied on the three planes of the space (axial, coronal, sagittal) and the final thresholded residual image was obtained by averaging the three sets of data on a voxel-by-voxel basis. The threshold used was the data-driven, subband (b) dependent "BayesShrink" threshold defined as $T_b = \sigma^2/\sigma_X$ [47]. Here, σ^2 denotes the noise variance estimated using the median operator in the first subband of a given decomposition level, such that $\sigma = \frac{Median(|w_{1,l}|)}{0.6745}$ with wavelet coefficients $w_{1,l}$ in the first subband of decomposition level l. Then, σ_X is subband dependent and is defined as $\sigma_X = \sqrt{\max(\sigma_w^2 - \sigma^2, 0)}$, where $\sigma_w^2 = \frac{1}{n^2}\sum_{i=1}^{n^2} w_{i,l}^2$ with wavelet coefficients $w_{i,l}$ in the i^{th} subband of decomposition level l and $n \times n$ is the size of the subband under consideration. This whole procedure ultimately results in PVE-corrected and wavelet-denoised heterogeneous activity distributions to be used in subsequent simulations of PET image acquisitions.

Monte-Carlo simulations

In this study, the general framework for the simulations of PET image acquisitions involved the Geant4 applications for tomography emission (GATE) Monte Carlo toolkit [48, 49]. This software can be used to simulate the transport of radiation from an emitting source (e.g., FDG intratumoural activity) inside human tissues and PET scanner models using Monte Carlo methods. It is optimized for nuclear medicine applications and offers large flexibility in using voxelized phantoms, voxelized sources, and different scanner geometries. We used GATE v7.1 along with Geant4.10.01 code, and all the physical

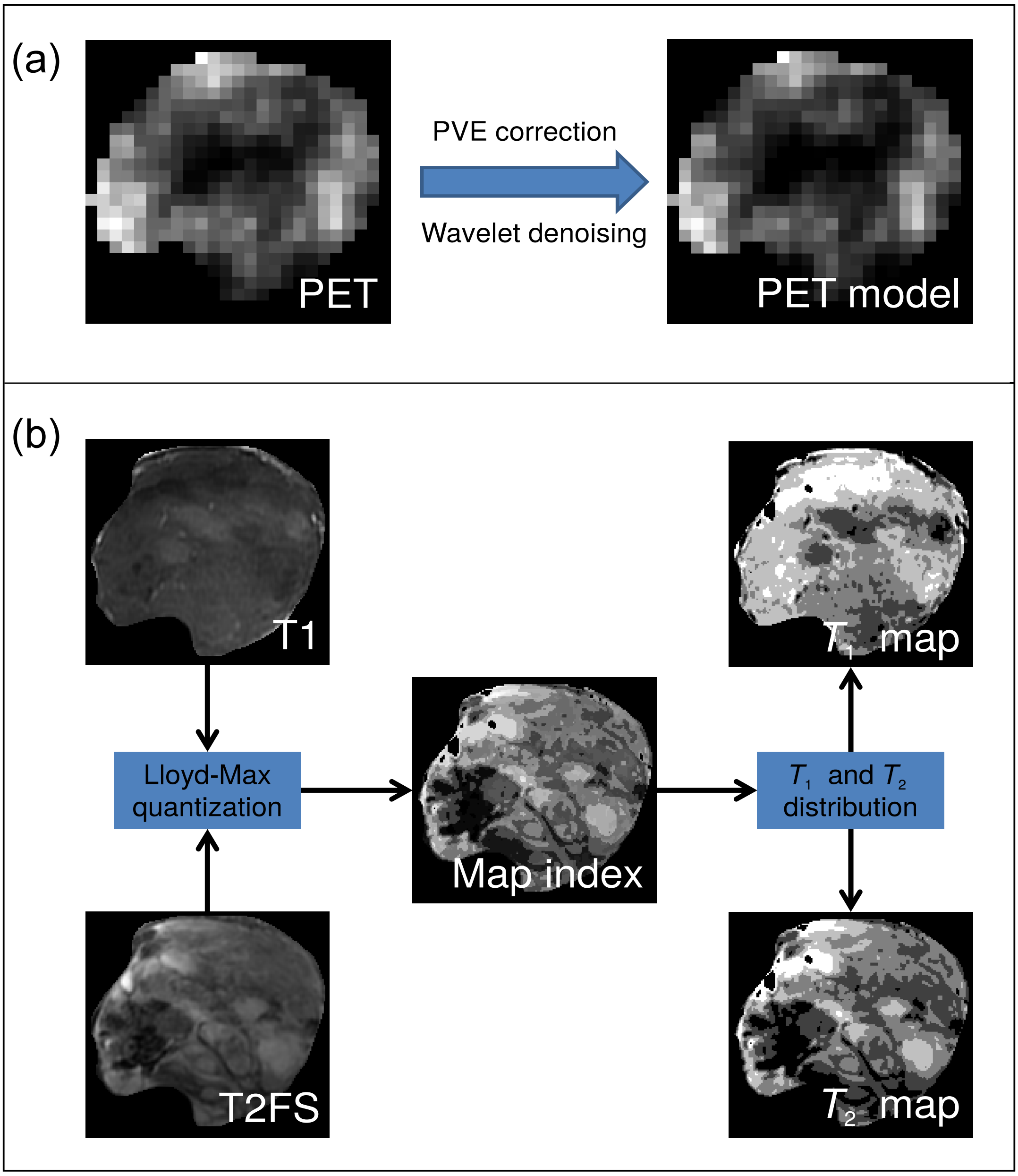


Figure 5.3: Example of the creation of PET and MRI digital tumour models used for image acquisition simulations. (a) Partial-volume effect (PVE) corrections combined with wavelet denoising are applied to the activity maps of PET clinical images to obtain tumor models for PET simulations. (b) The imaging intensities of the T_1-weighted (T1) and T_2-weighted fat-saturated (T2FS) clinical images are first separately discretized into 5 distinct regions per scan using Lloyd-Max quantization. This creates a combined set of $5 \times 5 = 25$ regions used as a map index for MRI simulations. Typical T_1 and T_2 relaxation times for soft-tissue sarcomas at 1.5 T are then distributed throughout the different regions, thereby creating T_1 and T_2 maps used for MRI simulations.

processes appropriate for realistic simulations were modeled using the "standard model".

The simulations were carried out using an approximation of the GE Discovery ST scanner [50] that we designed for GATE simulations in order to match as closely as possible the clinical acquisitions. This scanner consists of 280 detector blocks arranged in 35 modules of 2×4 blocks. Each block contains 6×6 detector BGO crystals of $6.3 \times 6.3 \times 30\ \mathrm{mm}^3$. The scanner contains 24 detector rings with a total of 420 crystals per ring. The scanner has an axial field of view of 15.7 cm and a radial FOV of 70 cm (detector ring diameter of 88.6 cm). In this work, the distance between crystal edges was filled with teflon material and was defined as the average distance left to fill the axial field of view after subtracting the distance filled by all detector rings, i.e., $(157/24 - 6.3)\ \mathrm{mm} = 0.2417\ \mathrm{mm}$. The digitizer was modeled using an energy window between 375 and 650 keV, with an energy blurring set to 15 % of 511 keV and a coincidence time window width of 11.7 ns. The different materials in the scanner were simulated with the materials provided by the GATE Materials Database (*GateMaterials.db*).

We used the PVE-corrected/wavelet-denoised heterogeneous activity distributions constructed from the PET clinical images as input voxelized sources in the GATE software. The axial extent of the input activity source of each patient was set to the axial extent of the scanner (15.7 cm), and the radial extent of the activity source was defined as a circle of 25 cm diameter, all centered on the geometrical center of mass of the tumour. A cylinder of input activity source centered on the tumour of each patient was thus inserted in the geometry of the scanner in GATE, by making sure that the extent of that cylinder fully encompassed the actual extent of all tumours in this study. The dimensions of the voxels of that cylinder were set to the dimensions of the voxels of the original PET clinical images of each patient. The axial center of the activity cylinder was positioned at the axial center of the geometry of the scanner in GATE, but a radial offset was applied to the center of the cylinder accordingly to the original position of the tumours in the clinical scans (i.e., the position in the body of the patient being scanned).

The anatomy of the patients in the simulations was modeled using a voxelized phantom geometry corresponding to the Hounsfield Units (HU) of the CT volumes of the FDG-PET/CT clinical scans. A 3D Gaussian filter was first applied to the whole-body CT volumes with a FWHM of 2.5 mm. The voxel dimensions of the CT volumes were then downsampled to the voxel dimensions of the activity sources of each patient (i.e., of the PET clinical scans) using cubic interpolation. The HU of the CT scans were translated to GATE materials indexes using translator files provided by the GATE Materials Database (*patient-HUmaterials.db* and *patient-HU2mat.txt*). Finally, a CT phantom geometry covering the full axial and radial FOV of the scanner was inserted in the GATE simulations according to original positions in the clinical scans, with the axial center of the CT phantom centered on the axial center of the tumour for each patient (i.e., positioned at the axial center of the geometry of the scanner in GATE).

Finally, 3D PET acquisitions were simulated in GATE using one bed position per study case. Acquisition times were set according to the procedure of each clinical scan of each patient (median of 180 s, range of 160–300 s).

PET imaging reconstruction

The simulated images were reconstructed using the Software for Tomographic Image Reconstruction (STIR) release 2 [51]. The OSMAPOSL 3D iterative algorithm was used with four iterations and 13 subsets, a maximum ring difference of 23, and varying numbers of span (3,5,7,9,11,13,15,17). Similarly to the original GE Discovery ST scanner, 47 image planes were reconstructed per acquisition with an axial sampling interval of 3.27 mm, and an in-plane resolution set to the one observed in each clinical scan of each patient (median of 5.47 × 5.47 mm^2, range of 3.91–5.47 mm). Random and scatter coincidences as well as attenuation corrections were applied as in-loop corrections in the reconstruction algorithm. For attenuation correction purposes, the HU maps of the CT scans were converted to linear attenuation coefficient maps using bilinear-scaling: I) $-1000 < \text{HU} \leq 0$ were linearly converted from 0 to 0.096 cm^{-1}; and 2) HU $>$ 0 were linearly converted from 0.096 to 0.15 cm^{-1}. Finally, reconstructed images were post-processed using an isotropic Gaussian filter with a FWHM of 5 mm.

5.4.4 Simulations of MR imaging acquisitions

MRI tumour models

In this work, we designed an empirical method for the creation of MRI tumour models as inputs to MRI simulation experiments, with the overall goal of preserving intratumoural heterogeneity. Figure 5.3b shows and example of the creation of a MRI tumour model constructed from T1 and T2FS clinical images. Three main inputs are required to proceed with MRI simulations:

maps of the physical T_1 and T_2 relaxation times ("T_1 map", "T_2 map") of the simulated tissues and a map of the proton density.

From the T1 and T2FS clinical scans of each patient, voxels within the tumour region with intensities outside the range $\mu \pm 3\sigma$ were excluded, as suggested by Collewet *et al.* [52] for making MRI texture measurements more reliable. Then, the imaging intensities of the T1 and T2FS clinical images were separately discretized into 5 distinct regions per scan using "Lloyd-Max quantization" [53, 54]. The quantization process maps the voxel values of a volume to a finite set $\mathbf{r} = \{r_k \in \mathbb{R} : k = 1, 2, \ldots, N_g\}$ of N_g reconstruction levels by defining a set $\mathbf{t} = \{t_k \in \mathbb{R} : k = 1, 2, \ldots, N_g + 1\}$ of decision levels, and Lloyd-Max quantization specifically attempts to choose the decision levels in order to minimize the mean-squared quantization error of the output via a clustering method. The separate discretization of the T1 and T2FS clinical scans into 5 regions in turn created a combined set of 25 distinct regions used as a map index for MRI simulations (index 1: T1 region 1 and T2FS regions 1, index 2: T1 region 1 and T2FS region 2, . . . , index 25: T1 region 5 and T2FS region 5).

We used typical T_1 and T_2 relaxation times as reported in the literature for for soft-tissue tumours at 1.5 T in order to obtain a distribution of relaxation times throughout all regions of the map index: central values of 1054 and 62 ms were used for the T_1 and T_2 relaxation times of STSs, respectively [55]. Then, the coefficient of variations of the imaging intensities in the tumour regions of the T1 (2σ) and T2FS (1σ) clinical images were found, and these factors were used to defined the range of T_1 and T_2 relaxation times to be assigned to the different map indexes: $[1054 - 2\sigma/\mu, 1054 + 2\sigma/\mu]$ with increments of $4\sigma/5\mu$ for the five regions with different T_1 values, and $[62 - \sigma/\mu, 62 + \sigma/\mu]$ with increments of $2\sigma/5\mu$ for the five regions with different T_2 values. Lower imaging intensities in T1 scans were assigned to higher T_1 relaxation times, and higher imaging intensities in T2FS scans were assigned to higher T_2 relaxation times [56]. Finally, a constant proton density percentage value of 0.82 relative to water [57] was assumed throughout the tumor region of all patients.

Numerical simulations

In this study, the general framework for the simulations of MR image acquisitions involved the use of the Jülich Extensible MRI Simulator (JEMRIS) [58], version 2.8.1. This open-source software numerically solves the Bloch equations to a series of spin models arranged in 2D or 3D grids. Parallelization of the execution is done using the MPI library and the output consists of the complete MRI signal acquired over the course of a given MRI sequence. This software is written in the C++ language and offers a series of graphical user interface in MATLAB to help in defining the various parameters of the simulations.

MRI simulations were carried out using the T_1 and T_2 maps constructed from the T1 and T2FS clinical images, and by assuming a constant proton density percentage value of 0.82 relative to water. Chemical shifts and T_2^* effects were not modeled in this work. A standard fast spin-echo (FSE) 2D sequence was constructed to simulate the acquisition of T_1-weighted and T_2-weighted images by using the *tse.xml* sequence template of JEMRIS. With this sequence, all 2D images of the 3D tumour regions were acquired and reconstructed separately. The number of averages was set to 1 for all imaging acquisition simulations. A fat saturation spoiling gradient used to dephase the lipid signal as in typical T_2-weighted fat-saturated sequences was not modeled in this work.

In order to match as closely as possible the clinical MR image acquisitions, several parameters from the original clinical sequences were used to define our simulated FSE sequence in JEMRIS. These parameters were retrieved from the DICOM headers of both the T1 and T2FS clinical images and are defined as follows:

- $\textit{nPoints}_{\text{(dicom)}}$: Number of rows and columns in the final image. Retrieved from the “Rows” or “Columns” DICOM fields.

- $\textit{matrixSize}_{\text{(dicom)}}$: Dimensions of the acquired *k*-space frequency/phase data before reconstruction. Retrieved from the “AcquisitionMatrix” DICOM field.

- $\textit{FOV}_{\text{(dicom)}}$: Diameter in mm of the region from within which data were used in creating the reconstruction of the image. Retrieved from the “ReconstructionDiameter” DICOM field.

- $\textit{phaseFOV}_{\text{(dicom)}}$: Ratio of field of view dimension in the phase encoding direction to field of view dimension in the frequency encoding direction. Retrieved from the “PercentPhaseFieldOfView” DICOM field.

- $\textit{pSampling}_{\text{(dicom)}}$: Fraction of acquisition matrix lines acquired. Retrieved from the “PercentSampling” DICOM field.

- *echoTrain*$_{\text{(dicom)}}$: Number of lines in *k*-space acquired per excitation per image. Retrieved from the "EchoTrainLength" DICOM field.
- *flipAngle*$_{\text{(dicom)}}$: Steady state angle in degrees to which the magnetic vector is flipped from the magnetic vector of the primary field. Retrieved from the "FlipAngle" DICOM field.
- *pixelBW*$_{\text{(dicom)}}$: Reciprocal of the total sampling period, in hertz per pixel. Retrieved from the "PixelBandwidth" DICOM field.
- TEc: Time in ms between the middle of the excitation pulse and the peak of the echo produced ($k_x = 0$) in the clinical acquisition ("c"). Retrieved from the "EchoTime" DICOM field.
- TRc: The period of time in ms between the beginning of a pulse sequence and the beginning of the succeeding (essentially identical) pulse sequence in the clinical acquisition ("c"). Retrieved from the "RepetitionTime" DICOM field.

The parameters above were used to set the following parameters in the *tse.xml* JEMRIS sequence module:

- *FOVx*$_{\text{(jemris)}}$: Field of view in the frequency encoding direction. Set to *FOV*$_{\text{(dicom)}}$.
- *FOVy*$_{\text{(jemris)}}$: Field of view in the phase encoding direction. Set to *FOV*$_{\text{(dicom)}}$ × *phaseFOV*$_{\text{(dicom)}}$.
- *Nx*$_{\text{(jemris)}}$: Number of points acquired in the *k*-space in the frequency encoding direction. Set to *matrixSize*$_{\text{(dicom)}}$.
- *Ny*$_{\text{(jemris)}}$: Number of points acquired in the *k*-space in the phase encoding direction. Set to *matrixSize*$_{\text{(dicom)}}$ × *phaseFOV*$_{\text{(dicom)}}$ × *pSampling*$_{\text{(dicom)}}$.
- *Repetitions*$_{\text{(jemris)}}$: Number of lines in *k*-space acquired per excitation per image. Set to *echoTrain*$_{\text{(dicom)}}$.
- *FlipAngle*$_{\text{(jemris)}}$: Flip angle used for the "90°" pulse. Set to *flipAngle*$_{\text{(dicom)}}$.
- *FlatTopTim*$_{\text{(jemris)}}$: Reciprocal of the readout bandwidth. Set to 1/*pixelBW*$_{\text{(dicom)}}$ × 1000.
- *TE*$_{\text{(jemris)}}$: Echo time in the FSE sequence. Set to TEc for the simulation of T_1-weighted images. Values of $\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}$ × TEc were tested for the simulation of T_2-weighted images.

- $TR_{\text{(jemris)}}$: Repetition time in the FSE sequence. Set to TRc for the simulation of T_2-weighted images. Values of $\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\} \times \mathrm{TRc}$ were tested for the simulation of T_1-weighted images.

In this study, the average repetition time used in clinical acquisitions (TRc) was (492 ± 81) ms over the 30 patients of the cohort. The average echo time used in clinical acquisitions (TEc) was (77 ± 12) ms.

MR imaging reconstruction

From the acquired MRI signal in simulations, *k*-space data was generated using the JEMRIS utilities in MATLAB. Prior to image reconstruction, zero padding of the *k*-space data was performed to obtain the right number of points in the final image (if necessary, as it depends on $nPoints_{\text{(dicom)}}$, $matrixSize_{\text{(dicom)}}$, $FOV_{\text{(dicom)}}$ and $phaseFOV_{\text{(dicom)}}$ values for each patient), followed by the application of a fermi filter to prevent sharp transitions in the *k*-space. The total number of zeros to add in the frequency ($nPad_{\text{FR}}$) and phase ($nPad_{\text{PE}}$) encoding directions in the *k*-space is governed by Equations (5.6) and (5.7), respectively. Finally, the final simulated images were separately reconstructed by taking the inverse 2D Fourier transform of the *k*-space data.

$$\begin{aligned} nPad_{\text{FR}} &= 2 \times \frac{\frac{\pi \times nPoints_{\text{(dicom)}}}{FOV_{\text{(dicom)}}} - \frac{\pi \times Nx_{\text{(jemris)}}}{FOV_{\text{(dicom)}}}}{2\pi / FOV_{\text{(dicom)}}} \\ &= nPoints_{\text{(dicom)}} - Nx_{\text{(jemris)}} \end{aligned} \tag{5.6}$$

$$\begin{aligned} nPad_{\text{PE}} &= 2 \times \frac{\frac{\pi \times nPoints_{\text{(dicom)}}}{FOV_{\text{(dicom)}}} - \frac{\pi \times Ny_{\text{(jemris)}}}{FOV_{\text{(dicom)}} \times phaseFOV_{\text{(dicom)}}}}{2\pi / (FOV_{\text{(dicom)}} \times phaseFOV_{\text{(dicom)}})} \\ &= nPoints_{\text{(dicom)}} \times phaseFOV_{\text{(dicom)}} - Ny_{\text{(jemris)}} \end{aligned} \tag{5.7}$$

5.4.5 STAMP: a software tool for texture optimization

To support optimization of image acquisition parameters for texture analyses, a software solution was developed in MATLAB with three main objectives:

1. Integrate programming tools for Monte-Carlo simulations of PET image acquisitions with different acquisition and reconstruction parameters.

2. Integrate programming tools for numerical simulations of MR image acquisitions with different acquisition and reconstruction parameters.

3. Integrate programming tools for textural analysis of clinical and simulated PET and MR images.

The integrated software has been called STAMP, which stands for Simulator for Texture Analysis in MRI and PET. As described below, this platform could facilitate the investigation of PET and MR image acquisition protocol variations on the texture features of simulated images. Example pictures of the three main graphical user interfaces (GUIs) of STAMP with real sample cases are shown in Supplementary Material section 5.9.3.

The simulations of PET image acquisitions in STAMP are achieved using an integrated version of the GATE simulator. The STAMP platform provides a PET simulation GUI that allows the user to specify the geometry of a cylindrical PET scanner from a small set of structural requirements, as well as digitizer parameters and the physical processes used in the simulations. An arbitrary voxelized activity source (e.g., FDG-PET tumour model) can then be imported and visualized into the simulation platform. The GUI also allows the user to choose between adding a water cylinder of a selected radius on top of the voxelized source or adding an arbitrary voxelized phantom with specific chemical composition for each voxels in the overall simulation geometry. The GUI then generates a sample file and a macro file that can be sent to a computer cluster (using the GUI) or simulated locally. Several GATE processes are then started in parallel. Coincidences, scatter and random events occurring within the scanner's detectors in each process are stored into a common sinogram file that can be subsequently used to reconstruct an image of the voxelized source via the GUI.

The simulations of MR image acquisitions in STAMP are achieved using an integrated version of the JEMRIS simulator. MRI sequences are programmed using a GUI already provided by the JEMRIS software developers and integrated to STAMP, and the specifications of the sequences are saved in a XML file. The STAMP platform provides a MRI simulation GUI that allows the user to select a given MRI simulation model (M_0, T_1, T_2, T_2^* and chemical shift maps), a sequence with specific MRI parameters, as well as imaging coils specifications used in the simulations. The GUI then generates a set of simulation files that can be sent to a computer cluster or simulated locally. Parallelization of the execution is done using the MPI library. The

MRI signal is read by the GUI to generate a *k*-space and reconstruct an image of the MRI simulation model.

A third GUI has also been implemented in STAMP in order to visualize PET and MRI simulation results, specify reconstruction algorithms and compute textural features. In order to facilitate data storage for each study case, a dedicated data structure format has been designed to keep track of the original PET and MRI scans in DICOM format, the simulation models developed from these scans, as well as the data obtained from all simulations. From clinical and simulated PET and MR images, GLCM, GLRLM, GLSZM and NGTDM 3D texture features can be computed using STAMP. Additional software features also exist to perform optimal texture extraction via the optimization of intensity quantization, spatial resolution and wavelet filtering. The STAMP platform also provides tools to perform fast texture computations in batch mode using our data structure.

5.4.6 Optimization of a texture-based model

Figure 5.4 presents an overview of the study workflow. In this work, PET simulated images (hereby denoted as "PETsim") were acquired for all patients using the following numbers of span: $3, 5, 7, 9, 11, 13, 15$ and 17. A span of 3 was considered as the parameter used for clinical acquisitions. For MRI, T_1-weighted simulated images (hereby denoted as "T1sim") were acquired using repetition times equal to $\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}$ × the different repetition times set in the FSE sequence of the clinical acquisitions of T1 scans of each patient (TRc), and T_2-weighted simulated images (hereby denoted as "T2sim") were acquired using echo times equal to $\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\}$ × the different echo times set in the FSE sequence of the clinical acquisitions of T2FS scans of each patient (TEc). All other possible user-defined simulation parameters were set to parameters used in clinical imaging acquisitions.

From the whole set of PETsim, T1sim and T2sim images, the $HGZE_{\text{GLSZM}}$ (PETsim – HGZE$_{GLSZM}$), ZSV_{GLSZM} (T1sim – ZSV$_{GLSZM}$) and $LRLGE_{\text{GLRLM}}$ (T2sim – LRLGE$_{GLRLM}$) texture features, respectively, were computed from the tumour region using the same extraction parameters (*scale*, *algo*, *Ng*) as detailed in Equation 5.3. As the quantization process of these three texture features involves a fixed number of bins per ROI, their computation is not dependent on the absolute imaging intensity values of the simulated images. These features were thus calculated from the raw intensity output of simulated images.

Using the texture data from the simulated images acquired using various parameters, the possibility of enhancing the predictive properties of the model of Equation 5.3 was estimated using bootstrapping experiments. The SUVpeak feature was considered as acquired from a "clinical" PET acquisition protocol and was thus obtained from PET clinical images for all experiments in the subsequent optimization analysis. For every combination of different PET, T_1-weighted and T_2-weighted acquisition parameters enumerated above, new logistic regression model responses (i.e., coefficients) were trained in 1000 bootstrap training samples for the following set of four features $\{\mathsf{SUVpeak}, \mathsf{PETsim-HGZE}_{\mathsf{GLSZM}}, \mathsf{T1sim-ZSV}_{\mathsf{GLSZM}}, \mathsf{T2sim-LRLGE}_{\mathsf{GLRLM}}\}$, and these model responses were thereafter directly tested in the corresponding bootstrap testing samples. Re-training of logistic regression coefficients for every different situation mimics how a predictive model would be trained in reality using clinical images acquired with a single set of parameters. The predictive performance of the models trained for every combination of acquisition parameters was then assessed by taking the mean AUC computed in the 1000 bootstrap testing samples ("bootstrap AUC", or ordinary bootstrap). Standard error of the mean was also calculated using a 95 % confidence interval.

5.5 Results

5.5.1 Texture variations with acquisition parameters

Qualitative image analysis

Figure 5.5 first presents example PET and MR simulated images (PETsim, T1sim, T2sim) acquired using different parameters: I) span 3, 9 and 17 for PETsim; II) TR of $\{\frac{1}{3}, 1, 3\} \times \mathrm{TRc}$ for T1sim; and III) TE of $\{\frac{1}{3}, 1, 3\} \times \mathrm{TEc}$ for T2sim.

For PET image acquisitions, it can be observed that an increasing span has the effect to increase image smoothing. Increasing the extent of axial data combined per slice with multiple neighboring detector coincidences has the benefit to augment slice sensitivity, however at the expense of a loss of resolution. In terms of image characteristics, high-intensity and low-intensity regions appear to be better defined, an effect which could be beneficial in terms of textural analysis for a better assessment of zone characteristics within the tumour (e.g., $HGZE_{\mathrm{GLSZM}}$).

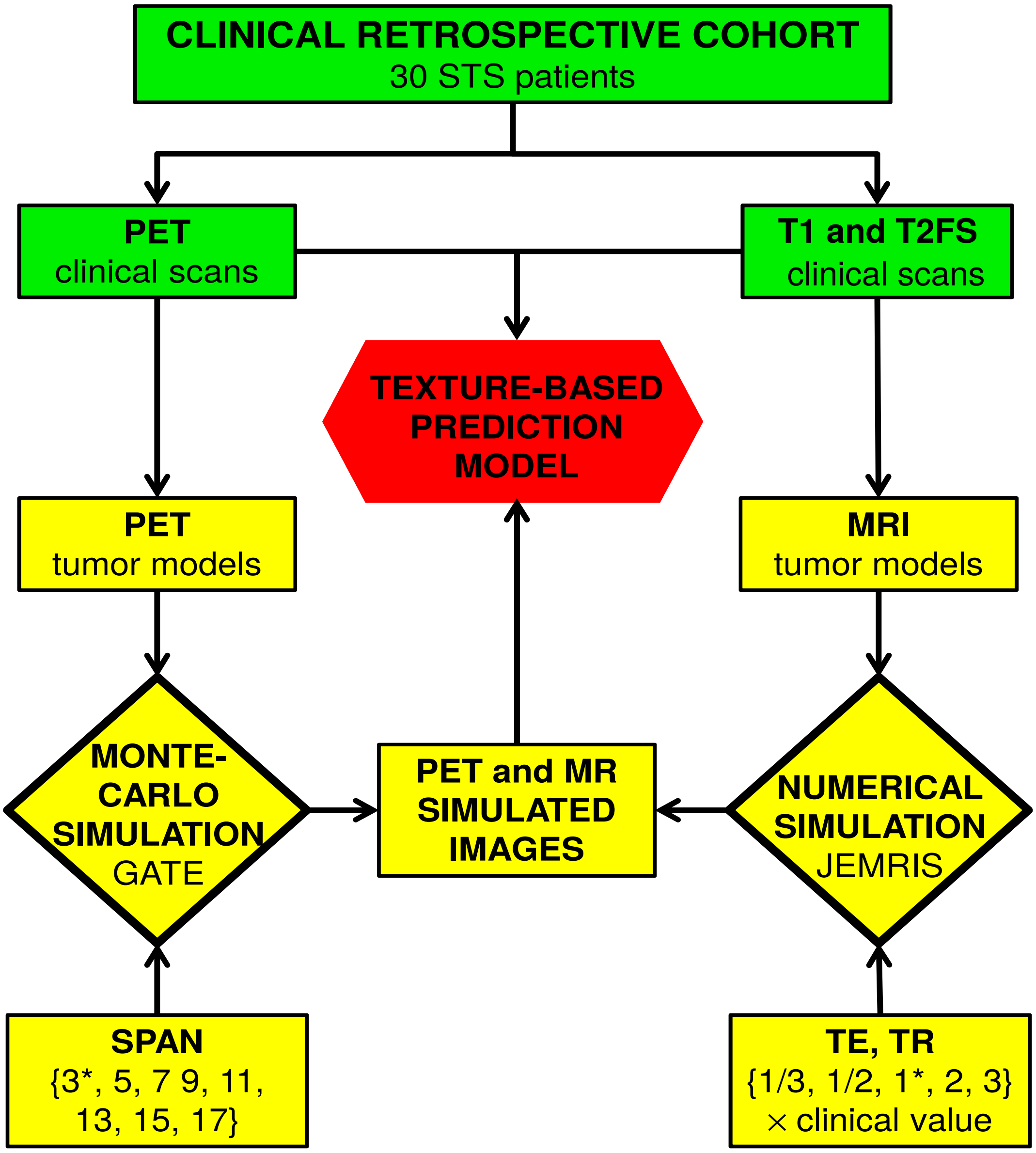


Figure 5.4: Workflow of this study. Digital tumour models were created from the PET, T_1-weighted (T1) and T_2-weighted fat-saturated (T2FS) clinical images of a retrospective cohort of 30 soft-tissue sarcoma patients. These tumour models were used as inputs for the simulations of PET, T_1-weighted and T_2-weighted image acquisitions using the GATE and JEMRIS software. The simulated PET images (PETsim) were acquired with different numbers of span, the simulated T_1-weighted images (T1sim) with different multiples of the repetition times (TR) used in clinical acquisitions for each patient, and the simulated T_2-weighted images (T2sim) with different multiples of the echo times used in clinical acquisitions for each patient. Asterisks (*) in the figure represents parameters used in clinical acquisitions (span3, TRc, TEc). Texture features previously selected in a multivariable model constructed from clinical scans for the prediction of lung metastases in soft-tissue sarcomas were extracted for the whole set of simulated images. The possibility of enhancing the predictive properties of the texture-based model by optimizing PET and MR acquisition protocols was then estimated using bootstrapping experiments with textures extracted from simulated images.

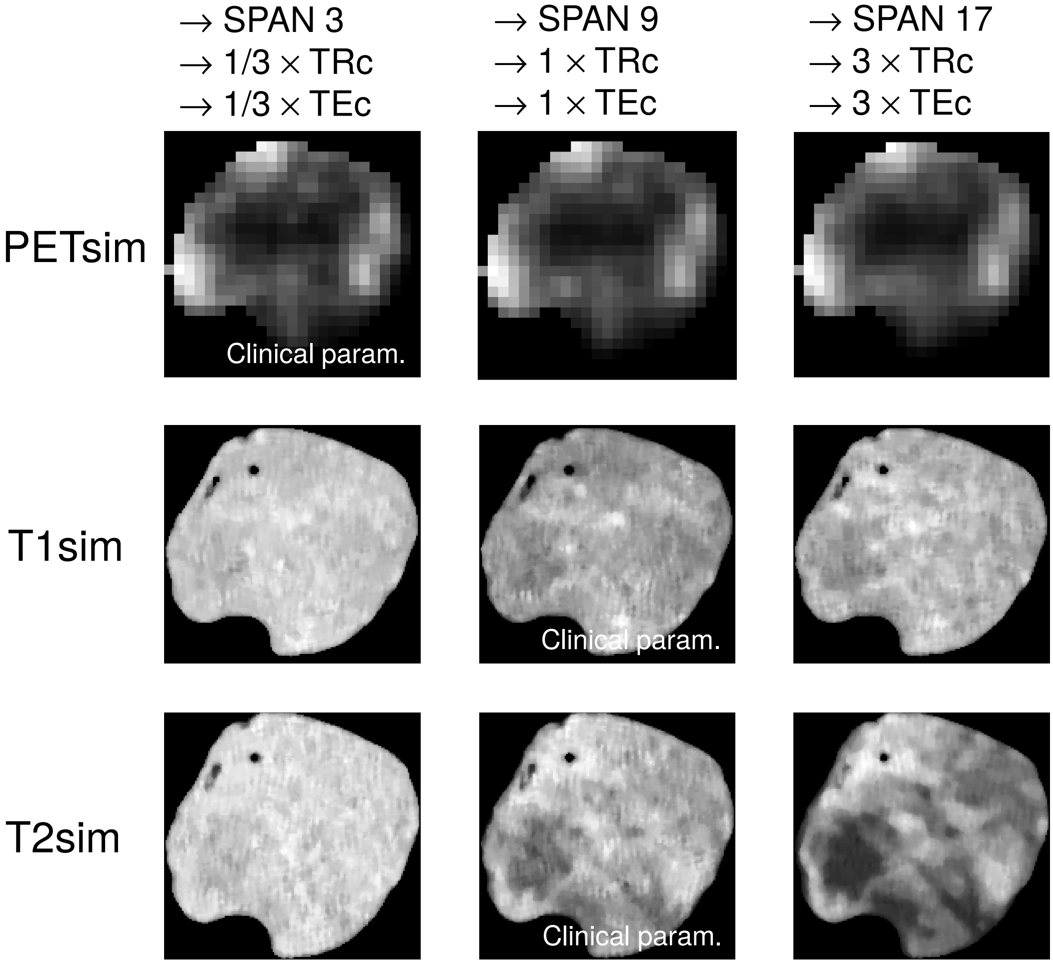


Figure 5.5: Example PET and MR simulated images acquired using different parameters. Top row: PET simulated images (PETsim) acquired with a span of 3 (left), 9 (middle) and 17 (right). Middle row: T_1-weighted simulated images (T1sim) acquired with a repetition time (TR) of $\frac{1}{3}$ (left), 1 (middle) and 3 (right) times the repetition time used in clinical acquisitions for each patient (TRc). Bottom row: T_2-weighted simulated images (T2sim) acquired with an echo time (TE) of $\frac{1}{3}$ (left), 1 (middle) and 3 (right) times the echo time used in clinical acquisitions for each patient (TEc). Simulated images acquired using clinical parameters are identified by "Clinical param." (span 3, $1 \times$ TRc, $1 \times$ TEc).

For T_1-weighted simulated image acquisitions, changes in TR in the range tested in this work do not appear to considerably affect the images. Small contrast increase and better definition of small tumour sub-regions can be observed with increasing TR, but the effect is not conclusive.

For T_2-weighted simulated image acquisitions, it can be observed that increasing TE has for effect to considerably increase the contrast between the different tumour sub-regions in the image. High-intensity and low-intensity regions are better defined and may thus improve intratumoural heterogeneity characterization (e.g., $LRLGE_{\mathrm{GLRLM}}$).

Texture measurements

We investigated the overall changes in the $\mathsf{PETsim-HGZE}_{GLSZM}$, $\mathsf{T1sim-ZSV}_{GLSZM}$ and $\mathsf{T2sim-LRLGE}_{GLRLM}$ features when extracted from simulated images acquired with different acquisition parameters. The different features extracted for each patient from simulated images acquired using the whole set of acquisition parameters were compared against the features extracted from simulated images acquired using the following *clinical* parameters: a span of 3 for PETsim, a repetition time equal to the one used in clinical acquisitions (TRc) for T1sim, and an echo time equal to the one used in clinical acquisitions (TEc) for T2sim. Percentage differences of each feature relative to textures extracted from simulated images acquired using clinical parameters were computed for all patients. As the overall goal of this study is to find a set of acquisition parameters from which extracted textures best discriminate between aggressive and non-aggressive tumours, we performed our analysis for two separate groups of patients: I) patients that developed lung metastases ("*Lung Mets*" patients); and II) patients that did not develop lung metastases ("*No Lung Mets*" patients). Results are summarized in Figure 5.6a using box plots.

For PET acquisitions, the general trend observed is that an increasing number of span increases the value of the $\mathsf{PETsim-HGZE}_{GLSZM}$ feature. This is consistent with the assessment made in the previous sub-section, where we observed that a higher span increases image smoothing and results in better defined tumour sub-regions. Also, it can be seen that the overall variations with different numbers of span seem to be more pronounced for *No Lung Mets* patients (higher interquartile range), but that *Lung Mets* patients experience a strict higher increase in the $\mathsf{PETsim-HGZE}_{GLSZM}$ feature with increasing span (median of % difference). Overall, the mean percentage change over all numbers of span as compared to span 3 is 0.3[-7,10] % and the *absolute* mean percentage change is 1 % for *Lung Mets* patients, whereas the mean percentage change is -0.2[-11,13] % and the *absolute* mean percentage change is 3 % for *No Lung Mets* patients.

For T_1-weighted MRI acquisitions, the general trend observed is that an increasing TR leads to an increase in the value of the $\mathsf{T1sim} - \mathsf{ZSV}_{\mathsf{\mathit{GLSZM}}}$ feature. This trend seems more pronounced for *Lung Mets* patients than for *No Lung Mets* patients. Overall, the mean percentage change over all different repetition times as compared to TRc is 2[-47,78] % and the *absolute* mean percentage change is 16 % for *Lung Mets* patients, whereas the mean percentage change is 2[-99,190] % and the *absolute* mean percentage change is 24 % for *No Lung Mets* patients. The ZSV_{GLSZM} feature thus experiences high variations when T_1-weighted images are acquired with different repetition times, an effect which could be beneficial for MRI sequence optimization.

For T_2-weighted MRI acquisitions, the general trend observed is that an increasing TE leads to an increase in the value of the $\mathsf{T2sim} - \mathsf{LRLGE}_{\mathsf{\mathit{GLRLM}}}$ feature. This trend seems again more pronounced for *Lung Mets* patients than for *No Lung Mets* patients. Overall, the mean percentage change over all different echo times as compared to TEc is -0.5[-25,25] % and the *absolute* mean percentage change is 9 % for *Lung Mets* patients, whereas the mean percentage change is 21[-38,1343] % and the *absolute* mean percentage change is 27 % for *No Lung Mets* patients. The $LRLGE_{\mathrm{GLRLM}}$ feature thus experiences high variations when T_2-weighted images are acquired with different echo times, another effect which could be beneficial for MRI sequence optimization.

Associations with clinical endpoint

In terms of texture optimization for the enhancement of a texture-based prediction model using a specific set of image acquisition parameters as compared to clinical ones, an ideal situation would entail that a given texture feature varies in one direction (e.g., increase) for all patients of a given group (e.g., *Lung Mets*) at the same time as the same feature varies in the opposite direction (e.g., decrease) for the other group (e.g., *No Lung Mets*). In reality, it is more likely that the textures of the majority of patients will vary in one direction, regardless of the patient group. In that case, it could be possible to enhance a given texture-based model if the absolute change in one overall group of patients is higher as compared to the other group. In this section, we verified if this effect can be observed in the current patient cohort using univariate analysis. We investigated the associations of the textures with

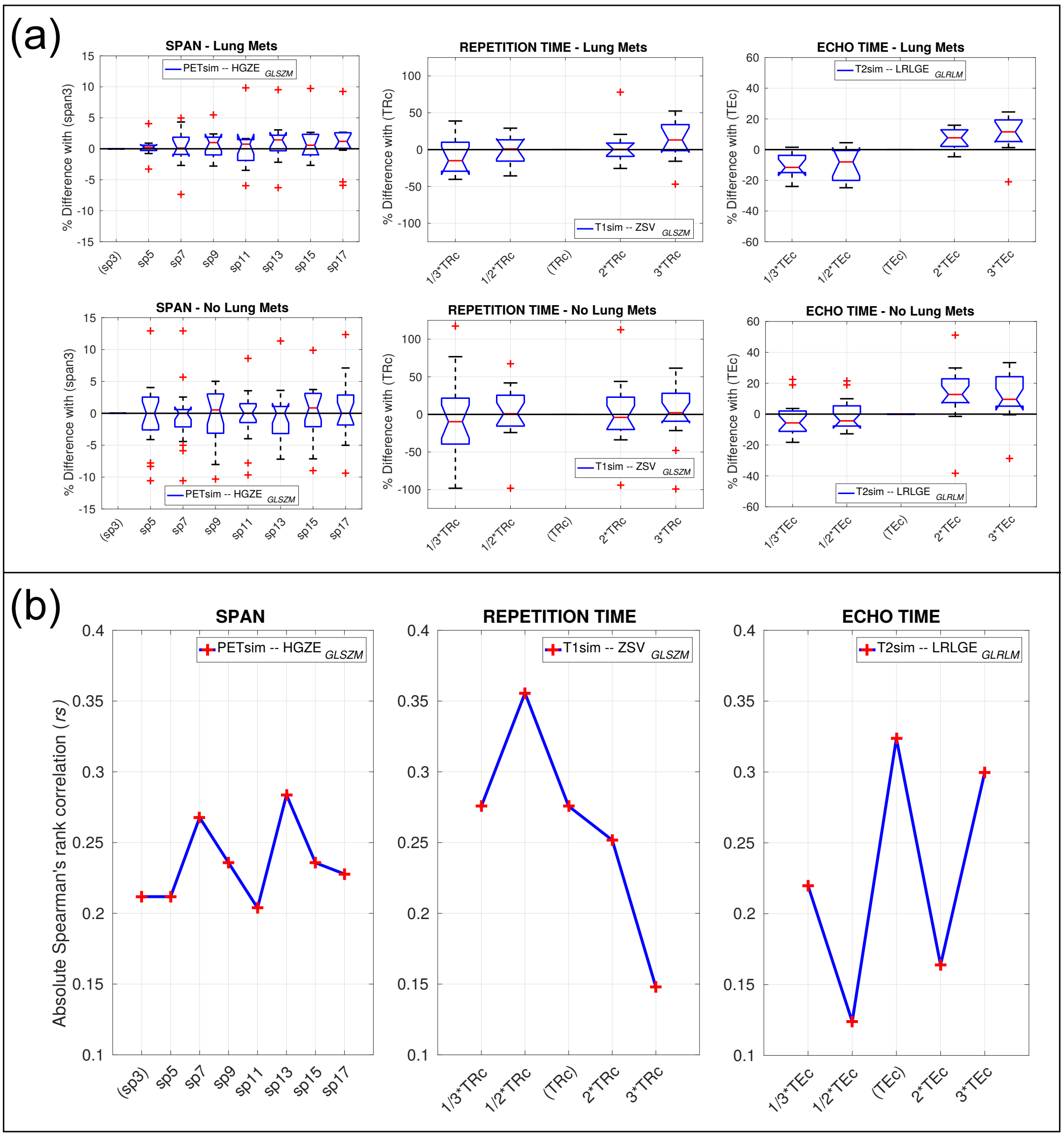


Figure 5.6: Texture variations with different acquisition parameters and resulting effects on associations with lung metastases in soft-tissue sarcomas. (a) Left: The $HGZE_{\text{GLSZM}}$ feature extracted from PET simulated images (PETsim) acquired with different numbers of span (sp) was compared against the $HGZE_{\text{GLSZM}}$ feature extracted from PET simulated images acquired with the span used in clinical acquisitions (sp3); Middle: The ZSV_{GLSZM} feature extracted from T_1-weighted simulated images (T1sim) acquired with different repetition times (TR) was compared against the ZSV_{GLSZM} feature extracted from T_1-weighted simulated images acquired with the repetition time used in clinical acquisitions for each patient (TRc); and Right: The $LRLGE_{\text{GLRLM}}$ feature extracted from T_2-weighted simulated images (T2sim) acquired with different echo times (TE) was compared against the $LRLGE_{\text{GLRLM}}$ feature extracted from T_2-weighted simulated images acquired with the echo time used in clinical acquisitions for each patient (TEc). Percentage differences relative to textures extracted from simulated images acquired with clinical parameters were computed for all patients, and results are summarized using box plots. The group of patients that developed lung metastases is denoted as "Lung Mets" and the group of patients that did not develop lung metastases is denoted as "No Lung Mets". (b). Associations of textures extracted from simulated images with lung metastases in soft-tissue sarcomas. A higher absolute Spearman's rank correlation is indicative of a stronger association.

the clinical endpoint of interest in this study, i.e., the development of lung metastases. Spearman's rank correlation coefficients (r_s) were calculated between imaging feature (j) vectors $\mathbf{x}_j = \{x_{ij} \in \mathbb{R} : i = 1, 2, \ldots, N\}$ extracted from simulated images acquired using the whole set of different parameters, and the outcome vector $\mathbf{y} = \{y_i \in \{0 : No\ Lung\ Mets, 1 : Lung\ Mets\} : i = 1, 2, \ldots, N\}$. Results are summarized in Figure 5.6b. In part due to the small number of patients used in this study (30), only one significant association was found and was obtained for T_1-weighted MRI acquisitions using a TR equal to $\frac{1}{2} \times \mathrm{TRc}$. Nonetheless, the results obtained here are informative about the possible optimization extent of each feature.

For PET acquisitions, the general trend observed is that an increasing span as compared to a number of span used in clinical acquisitions (span 3) has for effect to increase the predictive power of the $HGZE_{\mathrm{GLSZM}}$ texture. The highest absolute correlation between $\mathsf{PETsim} - \mathsf{HGZE}_{\mathsf{GLSZM}}$ and lung metastases was found at span 13, with $r_s = 0.28, p = 0.13$. In comparison, the correlation found using PET simulated images acquired with span 3 was $r_s = 0.21$, $p = 0.26$.

For T_1-weighted MRI acquisitions, the general trend observed is that a TR lower than TRc increases the predictive power of the ZSV_{GLSZM} texture, and a TR higher than TRc decreases the predictive power of the ZSV_{GLSZM} texture. The highest absolute correlation between $\mathsf{T1sim} - \mathsf{ZSV}_{\mathsf{GLSZM}}$ and lung metastases was found for a TR equal to $\frac{1}{2} \times \mathrm{TRc}$, with $r_s = -0.36, p = 0.05$. In comparison, the correlation found using T_1-weighted simulated images acquired with TRc was $r_s = -0.28, p = 0.14$.

For T_2-weighted MRI acquisitions, no general trend is observed. The highest absolute correlation between $\mathsf{T2sim} - \mathsf{LRLGE}_{\mathsf{GLRLM}}$ and lung metastases was found for a TE equal to TEc, with $r_s = 0.32, p = 0.08$. The only comparable correlation was obtained for a TE equal to $3 \times \mathrm{TEc}$, with $r_s = 0.30$, $p = 0.11$.

5.5.2 Evaluation on a texture-based model

Bootstrapping optimization

The results for the optimization of the texture-based prediction model as described in section 5.4.6 are presented in Figure 5.7. In both Figure 5.7a and Figure 5.7b, the average AUC values obtained in bootstrap testing samples for each of the 200 experiments performed in this work are shown. The total of 200 bootstrapping experiments results from all possible combinations of different PET, T_1-weighted and T_2-weighted acquisition parameters tested in this work: 8 numbers of span × 5 repetition times × 5 echo times. In Figure 5.7a, results are presented for varying numbers of span and echo times with fixed repetition times, whereas in Figure 5.7b results are presented for fixed echo times.

Similarly to the assessment made in section 5.5.1, it can be seen that an increasing span generally improves the prediction performance of the texture-based model as compared to span 3 used in clinical acquisitions. In terms of T_2-weighted MRI acquisitions, the prediction performance seems to be considerably higher when the texture-based model is constructed using $\frac{1}{2} \times \mathrm{TEc}$ as seen by the green line in Figure 5.7a. In terms of T_1-weighted MRI acquisitions, the prediction performance seems to be slightly higher for $\frac{1}{3} \times \mathrm{TRc}$ and $3 \times \mathrm{TRc}$ as seen by the blue and black lines in Figure 5.7b, respectively. Other types of trends are difficult to infer, as the overall response of the model seems to be significantly influenced by small perturbations with varying acquisition parameters.

Overall, the highest estimation of the prediction performance of the texture-based model under varying PET and MR image acquisition parameters was reached with: I) $\mathsf{PETsim-HGZE}_{\mathsf{GLSZM}}$ feature obtained with a span of 15; II) $\mathsf{T1sim-ZSV}_{\mathsf{GLSZM}}$ feature obtained with a TR equal to $\frac{1}{3} \times \mathrm{TRc}$; and III) $\mathsf{T2sim-LRLGE}_{\mathsf{GLRLM}}$ feature obtained with a TE equal to $\frac{1}{2} \times \mathrm{TEc}$. This particular model constructed by combining SUVpeak with textures extracted from simulated images acquired with optimal acquisition parameters reached an average AUC of 0.89 ± 0.01 in bootstrap testing experiments. In comparison, the model constructed with textures extracted from simulated images acquired with clinical acquisition parameters (span3, TRc, TEc) reached an average AUC of 0.84 ± 0.01 in bootstrap testing experiments.

Model response improvement

From the whole set of bootstrapping optimization experiments performed in the last sub-section for varying PET and MR image acquisition parameters, a single multivariable model combining the $\{\mathsf{SUVpeak}, \mathsf{PETsim-HGZE}_{\mathsf{GLSZM}}, \mathsf{T1sim-ZSV}_{\mathsf{GLSZM}}, \mathsf{T2sim-LRLGE}_{\mathsf{GLRLM}}\}$ features was estimated to possess the highest predictive properties for lung metastases in STSs (span 15, $\frac{1}{3} \times \mathrm{TRc}$, $\frac{1}{2} \times \mathrm{TEc}$). From this model, final logistic regression coefficients and bootstrap confidence intervals (95 %) were computed in the same manner as described in section 5.4.2 (1000 bootstrap samples), and the final model response for each patient of this cohort was subsequently calculated. The same process

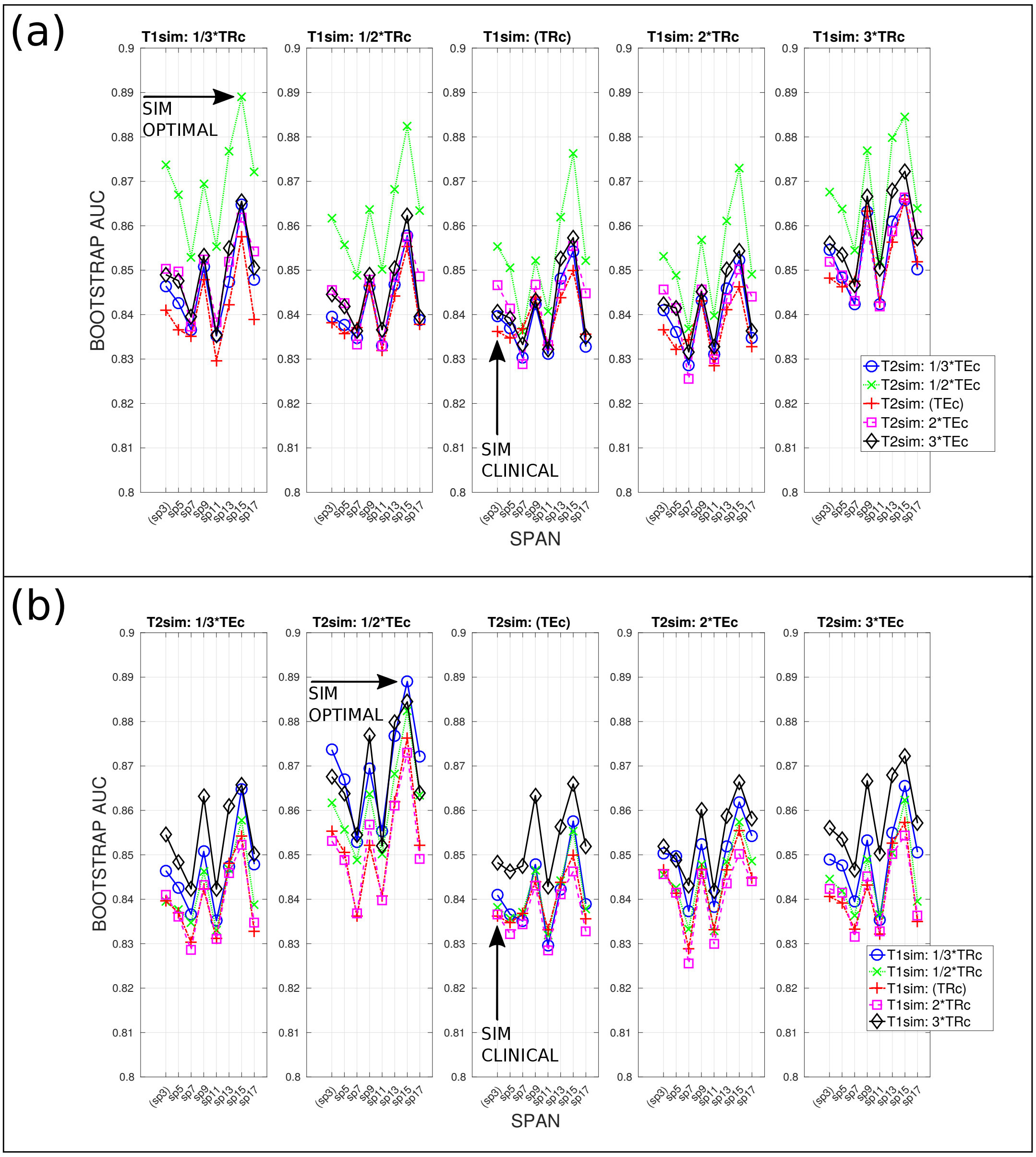


Figure 5.7: Optimization of a texture-based prediction model with respect to PET and MR simulated image acquisition parameters using bootstrapping experiments. In this work, the $HGZE_{\mathrm{GLSZM}}$ texture feature was extracted from PET simulated images (PETsim) acquired with different numbers of span (sp). The ZSV_{GLSZM} texture feature was extracted from T_1-weighted simulated images (T1sim) acquired with different multiples of the repetition time (TR) used in clinical acquisitions for each patient (TRc). The $LRLGE_{\mathrm{GLRLM}}$ texture feature was extracted from T_2-weighted simulated images (T2sim) acquired with different multiples of the echo time (TE) used in clinical acquisitions for each patient (TEc). The multivariable model combines the SUVpeak feature extracted from clinical images (sp3) with the three texture features extracted from simulated images using logistic regression. Different models were trained in 1000 bootstrap training samples for all possible combinations of the tested acquisition parameters of simulated images, and the trained models were tested in the corresponding 1000 bootstrap testing samples. The average AUC computed over all bootstrap testing samples was then calculated ("bootstrap AUC"). "SIM OPTIMAL" and "SIM CLINICAL" point to the results of the models constructed using the optimal (span 15, $\frac{1}{3} \times$ TRc, $\frac{1}{2} \times$ TEc) and clinical (span3, TRc, TEc) sets of PET and MR image acquisition parameters, respectively. (a) Plots of bootstrap AUC as a function of varying numbers of span and TE, with TR fixed. (b) Plots of bootstrap AUC as a function of varying numbers of span and TR, with TE fixed.

was also repeated for the multivariable model combining SUVpeak and the texture features extracted from simulated images acquired using the set of clinical acquisition parameters (span 3, TRc, TEc). In this section, we evaluate how the response of these two models vary, and how the response of the model constructed using optimal PET/MR acquisition parameters improves in comparison to the response of the model constructed using standard clinical acquisition parameters. The responses of these two models for each of the 30 patients of the cohort along with associated bootstrap confidence intervals are presented in Figure 5.8.

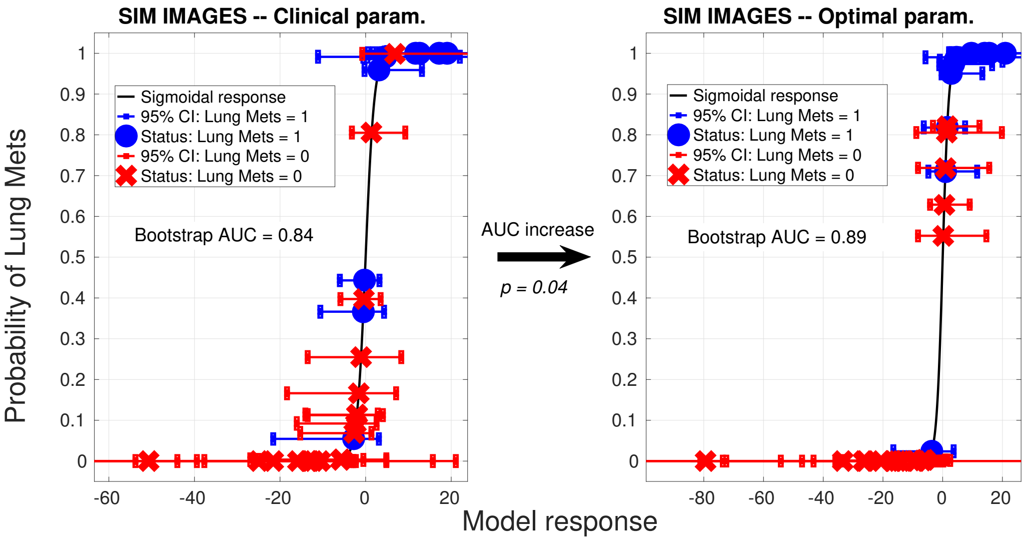


Figure 5.8: Texture-based prediction model response enhancement using an optimal set of PET and MR image acquisition parameters. Final model responses for the prediction of lung metastases in soft-tissue sarcomas and associated confidence intervals (95 %) are constructed using bootstrapping. Left: Probability of developing lung metastases as a function of the response of the final model constructed by combining SUVpeak with textures (PETsim – HGZE_{GLSZM}, T1sim – ZSV_{GLSZM}, T2sim – LRLGE_{GLRLM}) extracted from simulated images acquired with *clinical* acquisition parameters (span 3, TRc, TEc). Right: Probability of developing lung metastases as a function of the response of the final model constructed by combining SUVpeak with textures (PETsim – HGZE_{GLSZM}, T1sim – ZSV_{GLSZM}, T2sim – LRLGE_{GLRLM}) extracted from simulated images acquired with *optimal* acquisition parameters (span 15, $\frac{1}{3}$ × TRc, $\frac{1}{2}$ × TEc). The increase in AUC of the model responses is significant, with $p = 0.04$.

Overall, results presented in Figure 5.8 demonstrate the possibility of enhancing a texture-based predictive model by optimizing PET and MR image acquisition parameters. It was verified that the increase in AUC of the model responses obtained from simulated images acquired with clinical parameters to simulated images acquired with optimal parameters is significant under the Delong test [59], with $p = 0.04$. In Figure 5.8, two false negatives (*Lung Mets* patients with a probability < 50 %) using clinical parameters become true positives (*Lung Mets* patients with a probability > 50 %) using optimal parameters, as it was verified by inspecting model response values of each patient. On the other hand, some true negatives (*No Lung Mets* patients with a probability < 50 %) also become false positives (*No Lung Mets* patients with a probability > 50 %). Nonetheless, the significant AUC increase implies that the overall separation between *Lung Mets* and *No Lung Mets* patients under optimal image acquisition parameters is better than under clinical parameters, and thus that the predictive properties of the final model response are enhanced.

5.6 Discussion

Radiomics analysis is envisioned to make a significant impact in current routine clinical practice supporting more personalized cancer treatment management. In particular, the analysis of PET, CT and MR images with textural metrics have the potential to comprehensively characterize intratumoural heterogeneity and to provide crucial information about tumour aggressiveness. Tumour outcome prediction models combining textural metrics would be in turn constructed to improve prognostic assessment. However, it is recognized that routine clinical use would demand such radiomics models to be highly robust and predictive. Therefore, in the past few years, many studies have investigated the stability of textural features under varying imaging conditions in attempt to identify the most robust features, but not arguably the most optimal ones. In this work, we pursued an alternative but complementary approach that aims to identify the conditions and acquisition settings for these features to provide optimal predictive value. A proof of concept was carried out using computerized simulations of PET and MR image acquisitions, a type of framework which would provide an effective and controlled environment to study the effects of different acquisition parameters on many different types of textural measurements of intratumoural heterogeneity. We demonstrated the feasibility of enhancing a texture-based predictive model by optimizing targeted image acquisition parameters. Such identified parameters could thereafter be standardized and possibly become part of new protocols in future prospective studies designed to use radiomic models for tumour response assessment.

A multivariable model for the prediction of lung metastases in soft-tissue sarcomas (STSs) was first constructed from PET and MR clinical images. This model is composed of four features: 1) the *SUVpeak* value extracted from

PET; II) the $HGZE_{\text{GLSZM}}$ texture extracted from PET; III) the ZSV_{GLSZM} texture extracted from T_1-weighted images; and IV) the $LRLGE_{\text{GLRLM}}$ texture extracted from T_2-weighted fat-saturated images. In our work, we attempted to optimize the computation of these three texture features that are part of the model by varying different image acquisition parameters, but the *SUVpeak* feature was not optimized and was only extracted from clinical images. While the actual value of the *SUVpeak* feature would also change with varying PET acquisition parameters, the potential extent of optimization for that feature is likely less than for textures. Also, for practicality reasons in real life situations, it is possible that only a single additional set of images meeting the optimization requirements of textures would be acquired, whereas other more conventional prognostic factors would be extracted from images acquired using standard clinical acquisition protocols meeting the demands and requirements of radiologists.

In this work, simulated PET images were acquired by varying the number of span, or the number of neighboring ring detectors that are allowed to be in coincidence in the image acquisition, a procedure that increases slice sensitivity at the expense of resolution loss. Overall, we observed that an increasing number of span generally resulted in an increase of the $HGZE_{\text{GLSZM}}$ texture, a metric that quantifies the dominance or the emphasis of high-intensity subregions within the analyzed ROI. We also noted that this effect seemed to be more pronounced for *Lung Mets* patients than for *No Lung Mets* patients. In principle, increasing the number of span increases the image smoothing in the whole image and would thus result in a better definition of both the high- and low-intensity tumour sub-regions. However, as high-intensity tumour sub-regions in PET are likely more dominant for aggressive tumours with high metabolism, a higher increase of the $HGZE_{\text{GLSZM}}$ texture is more likely for metastatic STSs. To our knowledge, no other work has investigated the effect of the number of span on textures. Galavis *et al.* [30] studied the effect of different acquisition modes and reconstruction parameters on GLRLM textures (1D run length counterpart of the 2D or 3D GLSZM zone size computations), and concluded that those textures display intermediate variability. In this study, the variability of the $HGZE_{\text{GLSZM}}$ texture with varying numbers of span allowed for intermediate variations in the association of this metric with lung metastases in STSs.

In terms of MRI acquisitions, T_1-weighted simulated images were acquired with a standard spin-echo sequence using different repetition times (TR), and T_2-weighted simulated images were acquired using different echo times (TE). Overall, we observed that an increasing TR generally resulted in an increase of the ZSV_{GLSZM} texture, a metric quantifying the variance in the size of the different sub-regions within the tumour. An increasing TE also resulted in an increase of the $LRLGE_{\mathrm{GLRLM}}$ texture, a metric quantifying the dominance or emphasis of continuous long runs of low-intensities. These effects could be explained by the increased contrast resulting between low- and high-intensity tumour sub-regions in simulated images with increasing TR and TE, although the effect is less conclusive for T_1-weighted images. However, we also observed that small perturbations in T_1-weighted and T_2-weighted simulated images with different TR and TE could provok considerably high variations in the ZSV_{GLSZM} and the $LRLGE_{\mathrm{GLRLM}}$ metrics, respectively. Mayerhoefer *et al.* [33] and Waugh *et al.* [34] suggested that spatial resolution (i.e., voxel size) would nonetheless have higher impact than varying imaging sequence parameters on the outcome of texture analysis, but here our results obtained on realistic and heterogeneous tumour models advise about the need to find an effective trade-off between optimization and robustness of features in such experiments.

We demonstrated in this work that the enhancement of a multivariable texture-based predictive model via the optimization of PET and MR image acquisition protocols is feasible, as the increase in the estimation of the prediction performance by extracting texture features from images acquired using an optimal set of acquisition parameters was significant. Specifically, the optimization process led to a direct increase in the prediction of true positives (i.e., higher sensitivity) as seen in Figure 5.8. As the most important variable in the model (*SUVpeak*) was not optimized, we believe that degrees of enhancement higher than those observed in this work may be achievable if a given multivariable prediction model is texture-based only. However, we also observed a lot of statistical fluctuations in bootstrap AUC results in the optimization process of Figure 5.7, which limits our ability to decipher optimization patterns. Features weigh differently in different multivariable experiments, and as a consequence, it may be difficult to predict the global response pattern of a model under different acquisition parameters. Furthermore, it is important to recognize that the optimal set of PET and MR image acquisition parameters identified in multivariable analysis (span 15, $\frac{1}{3} \times \mathrm{TRc}$ and $\frac{1}{2} \times \mathrm{TEc}$ in Figure 5.7) is not exactly the same as the one identified in univariate analysis (span 13, $\frac{1}{2} \times \mathrm{TRc}$ and $1 \times \mathrm{TEc}$ in figure 5.6b). This attests to the complexity of machine learning problems where variables that are considered "less informative" by themselves can generate valuable predictions when combined together [60].

Despite the promising demonstration of our investigation, there are several limitations in this proof-of-concept study that we want to point out. First, in order to draw valid interpretations about texture variations in clinical settings from simulations, it is necessary to establish similarities between textures extracted from clinical and simulated images. Investigations on this issue were also performed and are presented Supplementary Material section 5.9.4. Our results show that the $HGZE_{\mathrm{GLSZM}}$ texture values are similar between clinical and simulated image acquisitions, but that the ZSV_{GLSZM} and $LRLGE_{\mathrm{GLRLM}}$ textures can considerably differ for some patients. For MRI, this effect may be explained by the high sensitivity of those textures to varying TR and TE as seen in Figure 5.6a. Although the differences between the individual textures of the clinical images and of the simulated images acquired using clinical parameters are not significantly different under the two-sided Wilcoxon signed rank sum test (PET: $p = 0.12$, T_1-weighted: $p = 0.17$, T_2-weighted: $p = 0.08$), our MRI simulation framework could be further refined in future work to achieve better results. The validity of tumour models prior to simulations is fundamental, and wrong assumptions in the creation of those models could have affected our results, especially in the case of MR image acquisitions. Furthermore, full construction of prediction models including feature selection should ideally be performed for every set of simulated images acquired with different parameters. For example, as increasing the number of span affects the intrinsic resolution of the image, optimal textures for the prediction of lung metastases in STSs may be found at different resolutions than the one identified for the $HGZE_{\mathrm{GLSZM}}$ texture extracted from clinical images in Equation 5.3 (*scale* of 5 mm). Finally, bootstrapping experiments only provides an estimation of the predictive properties of our texture-based model, and further validation of the results obtained in this study on independent external datasets is needed.

Overall, our work is only a first step towards the enhancement of texture-based prediction models via the optimization of image acquisition parameters. The type of simulation framework developed here could be useful to investigate how a wider range of radiomic features vary in different acquisition settings, to then identify which features are stable enough to further undergo imaging acquisition optimization. Furthermore, this type of simulation framework could also be effective to assess inter-scanner texture variability, and consequently to examine how texture-based prediction models may need to be optimized differently for different scanners. Ultimately, specific "radiomics" acquisition protocols optimized to generate superior texture measurements for a given clinical problem would be proposed and thereafter validated in prospective studies using clinical scanners.

5.7 Conclusion

In the past few years, many studies have examined the impact of different imaging settings on textural measurements, with the aim of identifying the most stable features under varying conditions. In this work, we pursued an alternative but complementary study paradigm by evaluating the feasibility of optimizing textures extracted from images acquired using different acquisition protocols for better prediction of a given clinical endpoint. As a proof of concept, we developed a workflow based on computerized simulations of PET and MR image acquisitions to test if a texture-based model constructed for the prediction of lung metastases in soft-tissue sarcomas could be enhanced by optimizing targeted image acquisition parameters. Results obtained in bootstrapping experiments suggest that it is possible to enhance texture-based prediction models by extracting features from images acquired using an optimal set of acquisition parameters. However, further validation on independent datasets is required, and optimal trade-offs should be attained between stability and optimization of texture features when different image acquisition parameters are varied. In this context, simulations of image acquisitions using realistic digital tumour models would constitute and effective framework to evaluate the extent of texture variations under different acquisition settings and the resulting impact on tumour outcome prediction models for prospective radiomics studies.

5.8 Acknowledgments

This work was supported by the Canadian Institutes of Health Research (CIHR) under grants MOP-114910 and MOP-136774, and the Natural Sciences and Engineering Research Council (NSERC) of Canada under the grant RGPIN 397711-11. Special thanks to Ives R. Levesque for his help in understanding the many subtleties of MRI clinical acquisitions.

5.9 Supplementary Material

5.9.1 Description of 3D radiomic features

In this thesis, please see Appendix A.

5.9.2 Construction of radiomic models

Feature set reduction

From the initial whole set of radiomic features (5 shape and 10 intensity features, 40 textures × 40 extraction parameters) extracted for each of the scans used in this work (PET, T1, T2FS), feature set reduction was performed via a stepwise forward feature selection scheme. This process aims to create a reduced feature set containing a total of 25 features balanced between predictive power and non-redundancy. This procedure is carried out using the following *Gain* equation [37]:

$$\begin{aligned}\widehat{\mathrm{Gain}}_j &= \gamma \cdot |\widehat{r}_s(\mathbf{x}_j, \mathbf{y})| \\ &\quad + \delta_a \cdot \left[\sum_{k=1}^{f} \left(\frac{2(f-k+1)}{f(f+1)} \right) \widehat{\mathrm{PIC}}(\mathbf{x}_k, \mathbf{x}_j) \right] \\ &\quad + \delta_b \cdot \left[\frac{1}{F} \sum_{l=1}^{F} \widehat{\mathrm{PIC}}(\mathbf{x}_l, \mathbf{x}_j) \right], \\ \text{where} \quad \widehat{r}_s(\mathbf{x}_j, \mathbf{y}) &= \frac{1}{B} \sum_{b=1}^{B} r_s(\mathbf{x}_j^{*b}, \mathbf{y}), \\ \text{and} \quad \widehat{\mathrm{PIC}}(\mathbf{x}_k, \mathbf{x}_j) &= \frac{1}{B} \sum_{b=1}^{B} \mathrm{PIC}(\mathbf{x}_k^{*b}, \mathbf{x}_j^{*b}). \end{aligned} \tag{5.8}$$

In Equation 5.8, $r_s(\mathbf{x}_j, \mathbf{y})$ is the Spearman's rank correlation computed between a given feature vector $\mathbf{x}_j$ and an outcome vector $\mathbf{y}$. $\mathrm{PIC}(\mathbf{x}_k, \mathbf{x}_j)$ is the *potential information coefficient* defined as $\mathrm{PIC}(\mathbf{x}_k, \mathbf{x}_j) = 1 - \mathrm{MIC}(\mathbf{x}_k, \mathbf{x}_j)$, where $\mathrm{MIC}(\mathbf{x}_k, \mathbf{x}_j)$ is the *maximal information coefficient* [61] between feature k and j. The sum over k is a sum over all f features that have already been chosen to be part of the reduced feature set (employed in forward selection schemes), whereas the sum over l is a sum over all F features that have not yet been removed from a larger initial set (employed in backward selection schemes). The sum over the k features is always done in order of appearance of the different features in the reduced set in order to favour the features from the larger initial set with the least dependence with the features chosen first in the reduced set. In this work, γ was set to 0.5, δ_a to 0.5 and δ_b to 0. Every time a new feature was chosen in the reduced set, a new *Gain* was calculated for all remaining features in the larger initial set using a different set of 100 bootstrap samples ($*b$, with $b = 1, \ldots, B$). Also, once a given texture extracted from a given scan with specific extraction parameters was chosen (scale, algo, Ng), all the other variants of that texture feature for that scan using other extraction parameters were deleted from the initial larger set of radiomic features. Note that Equation 5.8 allows to rank specific scan-texture-parameter features, as part 1 of the *Gain* equation uses Spearman's rank correlations varying over all variants of texture extraction parameters. However, to speed up calculations, average scan-texture features over all texture extraction parameters were used in part 2 (and 3 if needed) of the *Gain* equation.

Feature selection

The feature selection step was first divided into 25 experiments. In each of these experiments, a different feature from the reduced set was used as a different "starting feature". For a given starting feature, all possible logistic regression models of order 2 (i.e., combination of 2 variables) were created by combining that feature with each of the remaining features in the reduced feature set still available for that particular experiment. Bootstrap resampling (100 samples) was performed for each of these models in order to calculate the 0.632+ bootstrap AUC [41, 42], a process in which logistic regression models are trained in bootstrap training samples and tested in corresponding bootstrap testing samples. Then, the single remaining feature that maximized the 0.632+ bootstrap AUC when combined with the starting feature was selected, and the process was repeated up to model order 10 for each experiment. Finally, for each model order, the experiment that yielded the highest 0.632+ bootstrap AUC was identified, and combinations of features were thereby chosen for model orders of 1 to 10. Figure 5.9 illustrates the feature selection process.

5.9.3 Software integration for texture analysis of PET and MR simulated images

Figure 5.10, Figure 5.11 and Figure 5.12 present screenshots of the three main GUIs used in STAMP. The first complete version of STAMP is currently a

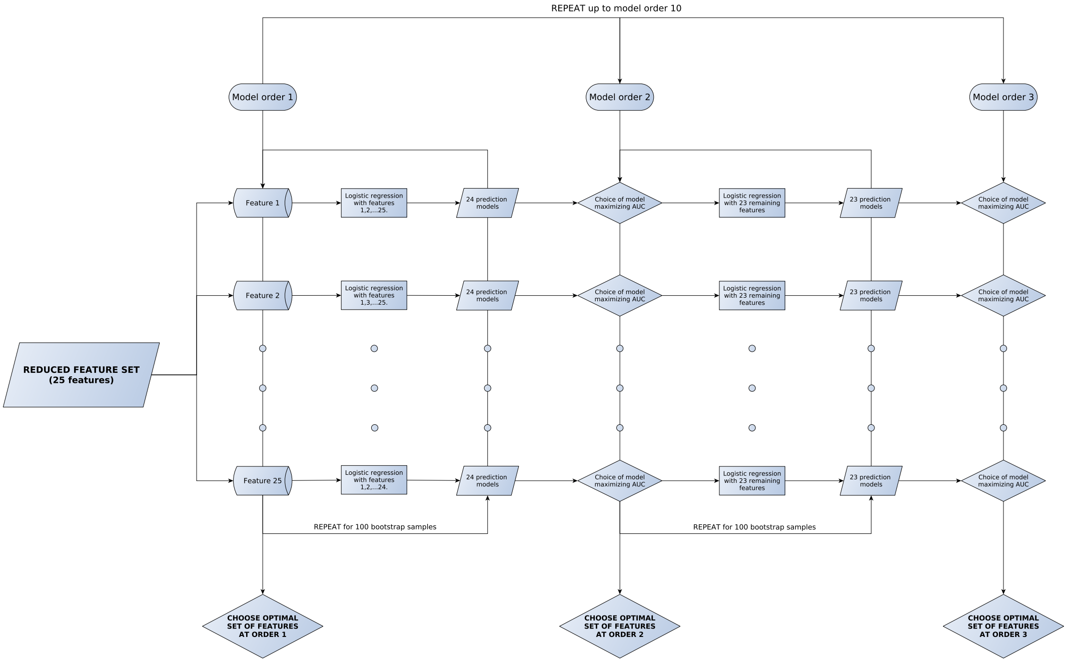
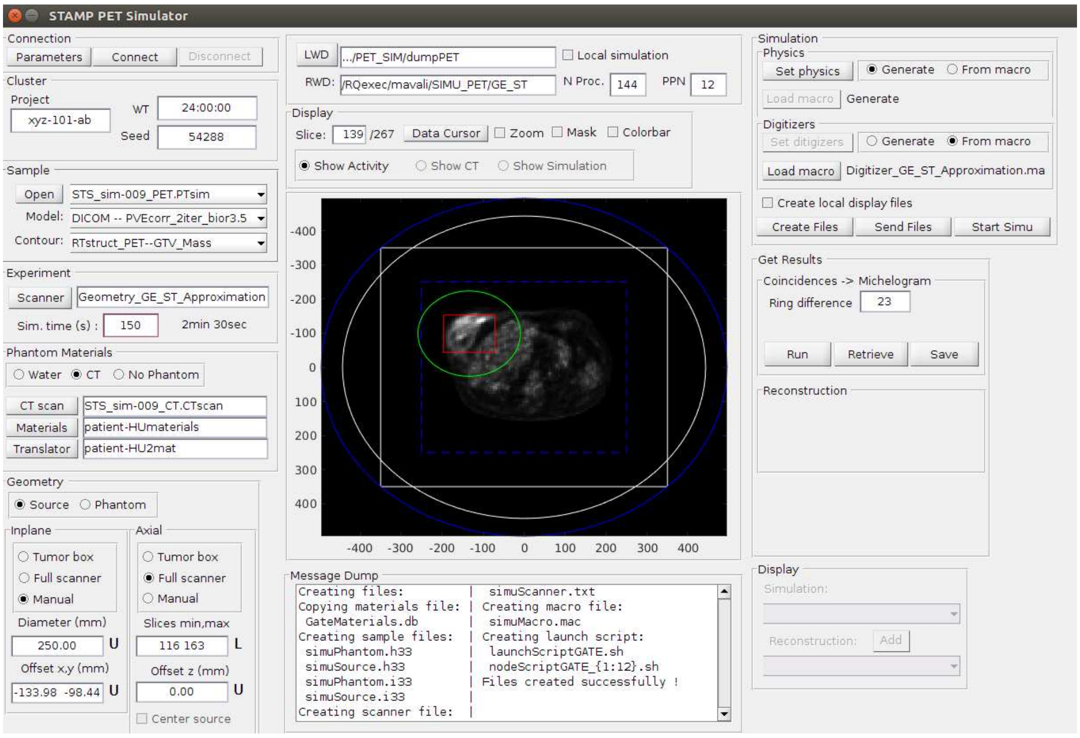

REPEAT up to model order 10
Model order 1
Model order 2
Model order 3
REDUCED FEATURE SET (25 features)
Feature 1
Feature 2
Feature 25
Logistic regression with features 1,2,...25.
Logistic regression with features 1,3,...25.
Logistic regression with features 1,2,...24.
24 prediction models
Choice of model maximizing AUC
Logistic regression with 23 remaining features
23 prediction models
REPEAT for 100 bootstrap samples
CHOOSE OPTIMAL SET OF FEATURES AT ORDER 1
CHOOSE OPTIMAL SET OF FEATURES AT ORDER 2
CHOOSE OPTIMAL SET OF FEATURES AT ORDER 3

Figure 5.9: Radiomic feature selection.

work-in-progress, and our plan for the future is to provide a public release of the software to the radiomics community. All researchers interested in the software or interested in joining our development team are most welcome, and may contact Martin Vallières at mart.vallieres@gmail.com.



Project: xyz-101-ab

WT: 24:00:00

Seed: 54288

LWD: .../PET_SIM/dumpPET

☐ Local simulation

RWD: /RQexec/mavali/SIMU_PET/GE_ST

N Proc.: 144

PPN: 12

Slice: 139 /267

☐ Zoom ☐ Mask ☐ Colorbar

● Show Activity ○ Show CT ○ Show Simulation

Physics: ● Generate ○ From macro

Load macro Generate

Digitizers: ○ Generate ● From macro

Load macro Digitizer_GE_ST_Approximation.ma

☐ Create local display files

Ring difference: 23

Open: STS_sim-009_PET.PTsim

Model: DICOM -- PVEcorr_2iter_bior3.5

Contour: RTstruct_PET--GTV_Mass

Scanner: Geometry_GE_ST_Approximation

Sim. time (s) : 150 2min 30sec

○ Water ● CT ○ No Phantom

CT scan: STS_sim-009_CT.CTscan

Materials: patient-HUmaterials

Translator: patient-HU2mat

● Source ○ Phantom

Inplane: ○ Tumor box ○ Full scanner ● Manual

Diameter (mm): 250.00 U

Offset x,y (mm): -133.98 -98.44 U

Axial: ○ Tumor box ● Full scanner ○ Manual

Slices min,max: 116 163 L

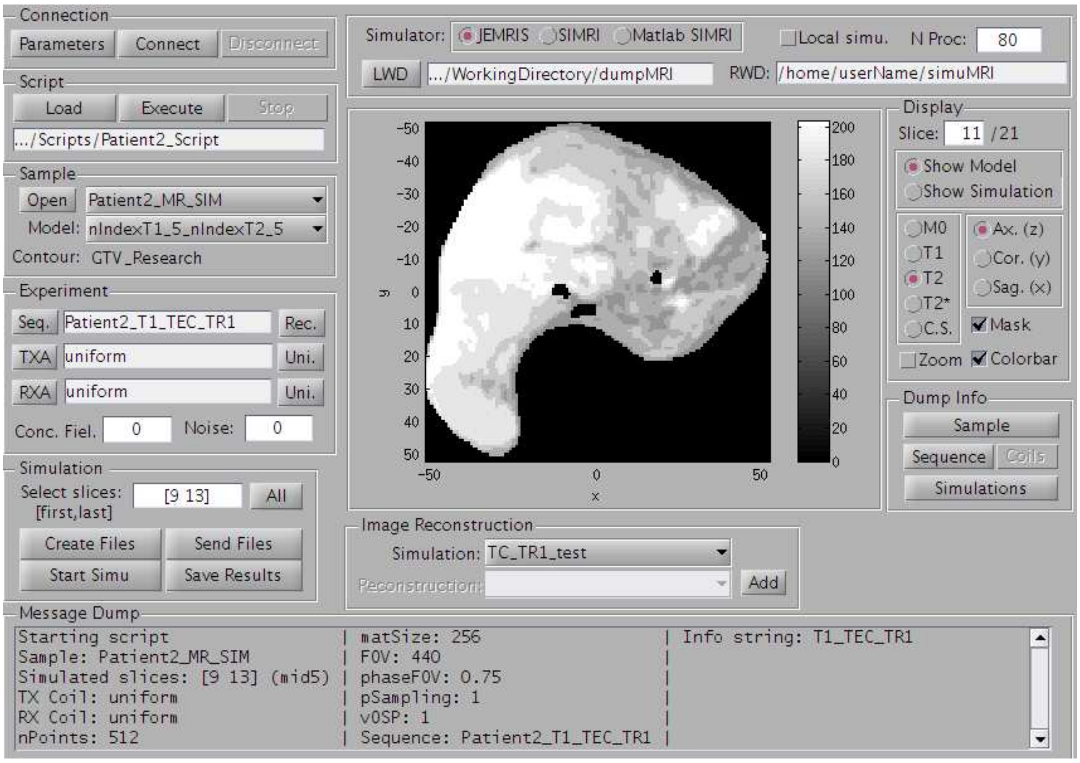
Offset z (mm): 0.00 U

☐ Center source

Message Dump

```
Creating files:          |  simuScanner.txt
Copying materials file:  | Creating macro file:
 GateMaterials.db        |  simuMacro.mac
Creating sample files:   | Creating launch script:
 simuPhantom.h33         |  launchScriptGATE.sh
 simuSource.h33          |  nodeScriptGATE_{1:12}.sh
 simuPhantom.i33         | Files created successfully !
 simuSource.i33          |
Creating scanner file:   |
```

Figure 5.10: Example image of the PET simulation GUI of STAMP.



Simulator: ● JEMRIS ○ SIMRI ○ Matlab SIMRI

□ Local simu. N Proc: 80

.../WorkingDirectory/dumpMRI RWD: /home/userName/simuMRI

.../Scripts/Patient2_Script

Patient2_MR_SIM

Model: nIndexT1_5_nIndexT2_5

Contour: GTV_Research

Seq.: Patient2_T1_TEC_TR1

TXA: uniform

RXA: uniform

Conc. Fiel. 0 Noise: 0

Select slices: [first,last] [9 13]

Slice: 11 /21

● Show Model
○ Show Simulation

○ M0
○ T1
● T2
○ T2*
○ C.S.

● Ax. (z)
○ Cor. (y)
○ Sag. (x)

☑ Mask

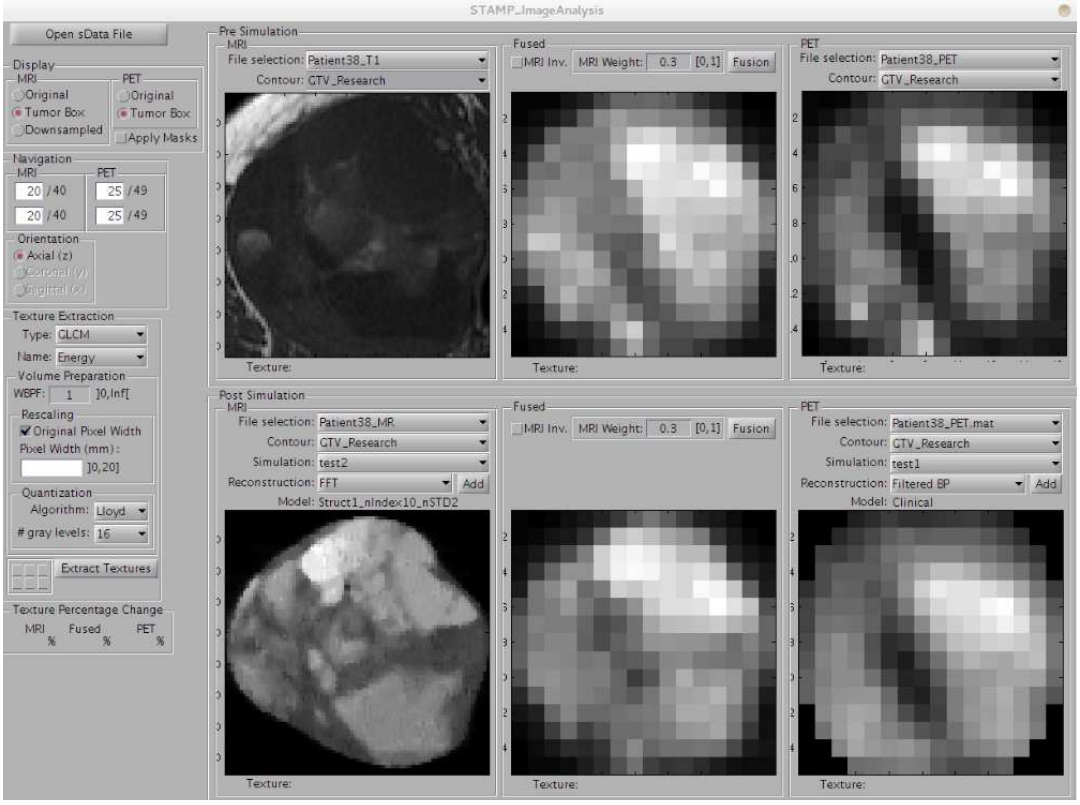
□ Zoom ☑ Colorbar

Simulation: TC_TR1_test

Reconstruction:

```
Starting script                   | matSize: 256               | Info string: T1_TEC_TR1
Sample: Patient2_MR_SIM           | FOV: 440                   |
Simulated slices: [9 13] (mid5)   | phaseFOV: 0.75             |
TX Coil: uniform                  | pSampling: 1               |
RX Coil: uniform                  | vOSP: 1                    |
nPoints: 512                      | Sequence: Patient2_T1_TEC_TR1 |
```

Figure 5.11: Example image of the MRI simulation GUI of STAMP.



MRI: ○ Original ● Tumor Box ○ Downsampled

PET: ○ Original ● Tumor Box □ Apply Masks

MRI: 20 /40, 20 /40

PET: 25 /49, 25 /49

Orientation: ● Axial (z) ○ Coronal (y) ○ Sagittal (x)

Type: GLCM

Name: Energy

WBPF: 1]0,Inf[

☑ Original Pixel Width

Pixel Width (mm):]0,20]

Algorithm: Lloyd

gray levels: 16

Pre Simulation

MRI: File selection: Patient38_T1 Contour: GTV_Research

Fused: □ MRI Inv. MRI Weight: 0.3 [0,1]

PET: File selection: Patient38_PET Contour: GTV_Research

Post Simulation

MRI: File selection: Patient38_MR Contour: GTV_Research Simulation: test2 Reconstruction: FFT Model: Struct1_nIndex10_nSTD2

Fused: □ MRI Inv. MRI Weight: 0.3 [0,1]

PET: File selection: Patient38_PET.mat Contour: GTV_Research Simulation: test1 Reconstruction: Filtered BP Model: Clinical

Figure 5.12: Example image of the image analysis GUI of STAMP.

5.9.4 Comparison of clinical and simulated textures

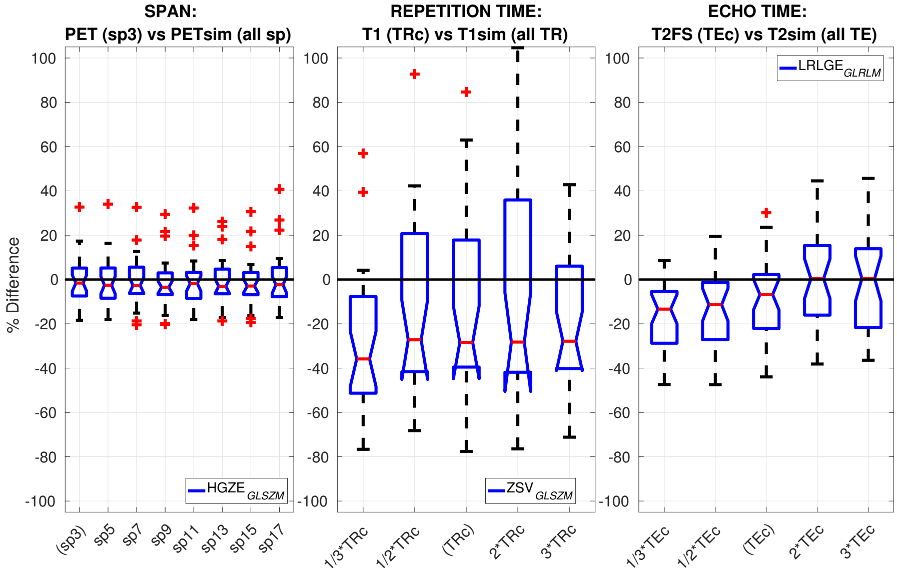


Figure 5.13: Comparison between textures extracted from clinical and simulated images. Left: The PETsim – HGZE$_{GLSZM}$ feature extracted from PET simulated images (PETsim) acquired with different numbers of span (sp) was compared against the PET – HGZE$_{GLSZM}$ feature extracted from PET clinical images acquired with span 3; Middle: The T1sim – ZSV$_{GLSZM}$ feature extracted from T_1-weighted simulated images (T1sim) with different repetition times (TR) was compared against the T1 – ZSV$_{GLSZM}$ feature extracted from T_1-weighted clinical images (T1) acquired with clinical repetition time TRc; and Right: The T2sim – LRLGE$_{GLRLM}$ feature extracted from T_2-weighted simulated images (T2sim) with different echo times (TE) was compared against the T2FS – LRLGE$_{GLRLM}$ feature extracted from T_2-weighted fat-saturated clinical images (T2FS) acquired with clinical echo time TEc. Percentage differences relative to textures extracted from clinical scans were computed for all patients, and results are summarized using box plots.

Overall, it can be seen that the HGZE$_{GLSZM}$ texture extracted from PET simulated scans is similar to the same texture extracted from PET clinical scans. This suggests that our PET simulation framework could be sufficiently accurate to approximate textures computed from images acquired on clinical scanners. Also, the smallest percentage differences between clinical and simulated textures seem to be obtained for span 3. This is consistent with the number of span used in PET clinical acquisitions.

For MR comparisons, differences between simulated and clinical textures are much larger than for PET. Furthermore, the percentage differences between clinical and simulated textures for the LRLGE$_{GLRLM}$ feature seem to increase with increasing TE, with the smallest difference (median of distribution) obtained at 3 × TEc. This suggests that our MRI simulation framework may require further refinement, from MRI tumour modeling to MRI simulations. In future work, we will perform a comprehensive assessement of the origins of the differences between clinical and simulated images.

5.10 References

1. El Naqa, I., Li, R. & Murphy, M. J. *Machine Learning in Radiation Oncology: Theory and Applications* 1st ed. 336 pp. (Springer International Publishing, Cham, Switzerland, 2015).
2. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278,** 563–577 (2016).
3. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25,** 675–680 (2007).
4. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105,** 5213–5218 (2008).
5. El Naqa, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* **42,** 1162–1171 (2009).
6. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin. Radiol.* **65,** 517–521 (2010).
7. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30,** 1234–1248 (2012).
8. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48,** 441–446 (2012).
9. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).
10. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).
11. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).
12. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

13. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

14. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).

15. Fidler, I. J. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res.* **50,** 6130–6138 (1990).

16. Yokota, J. Tumor progression and metastasis. *Carcinogenesis* **21,** 497–503 (2000).

17. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).

18. Lambin, P. *et al.* Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat. Rev. Clin. Oncol.* **10,** 27–40 (2013).

19. Hatt, M. *et al.* Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur. J. Nucl. Med. Mol. Imaging,* 1–15 (2016).

20. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61,** R150–R166 (2016).

21. Bogowicz, M. *et al.* Stability of radiomic features in CT perfusion maps. *Phys. Med. Biol.* **61,** 8736 (2016).

22. Molina, D. *et al.* Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images. *Computers in Biology and Medicine* **78,** 49–57 (2016).

23. Hatt, M., Tixier, F., Rest, C. C. L., Pradier, O. & Visvikis, D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur. J. Nucl. Med. Mol. Imaging* **40,** 1662–1671 (2013).

24. Leijenaar, R. T. H. *et al.* Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncologica* **52,** 1391–1397 (2013).

25. Parmar, C. *et al.* Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One* **9,** e102107 (2014).

26. Van Velden, F. H. P. *et al.* Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol. Imaging Biol.* **18,** 788–795 (2016).

27. Orlhac, F. *et al.* Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J. Nucl. Med.* **55,** 414–422 (2014).

28. Tixier, F. *et al.* Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *J. Nucl. Med.* **53,** 693–700 (2012).

29. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6,** 23428 (2016).

30. Galavis, P. E., Hollensen, C., Jallow, N., Paliwal, B. & Jeraj, R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* **49,** 1012–1016 (2010).

31. Nyflot, M. J. *et al.* Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J. Med. Imaging* **2,** 041002 (2015).

32. Yan, J. *et al.* Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J. Nucl. Med.* **56,** 1667–1673 (2015).

33. Mayerhoefer, M. E., Szomolanyi, P., Jirak, D., Materka, A. & Trattnig, S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study. *Med. Phys.* **36,** 1236–1243 (2009).

34. Waugh, S. A., Lerski, R. A., Bidaut, L. & Thompson, A. M. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. *Med. Phys.* **38,** 5058–5066 (2011).

35. Zhao, B., Tan, Y., Tsai, W. Y., Schwartz, L. H. & Lu, L. Exploring variability in CT characterization of tumors: a preliminary phantom study. *Translational Oncology* **7,** 88–93 (2014).

36. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features: *Investigative Radiology* **50,** 757–765 (2015).

37. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60,** 5471–5496 (2015).

38. Fahey, F. H. Data acquisition in PET imaging. *J. Nucl. Med. Technol.* **30,** 39–49 (2002).

39. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26,** 1045–1057 (2013).

40. El Naqa, I. *et al.* Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys. Med. Biol.* **51,** 5719–5735 (2006).

41. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560 (1997).

42. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.* **35,** 1559–1570 (2008).

43. Papadimitroulas, P. *et al.* Investigation of realistic PET simulations incorporating tumor patient's specificity using anthropomorphic models: Creation of an oncology database. *Med. Phys.* **40,** 112506 (2013).

44. Boussion, N., Rest, C. C. L., Hatt, M. & Visvikis, D. Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body PET imaging. *Eur. J. Nucl. Med. Mol. Imaging* **36,** 1064–1075 (2009).

45. Richardson, W. H. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* **62,** 55–59 (1972).

46. Lucy, L. B. An iterative technique for the rectification of observed distributions. *The Astronomical Journal* **79,** 745 (1974).

47. Chang, S. G., Yu, B. & Vetterli, M. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Process.* **9,** 1532–1546 (2000).

48. Jan, S. *et al.* GATE: a simulation toolkit for PET and SPECT. *Phys. Med. Biol.* **49,** 4543–4561 (2004).

49. Jan, S. *et al.* GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy. *Phys. Med. Biol.* **56,** 881–901 (2011).

50. Bettinardi, V. *et al.* Performance evaluation of the new whole-body PET/CT scanner: Discovery ST. *Eur. J. Nucl. Med. Mol. Imaging* **31,** 867–881 (2004).

51. Thielemans, K. *et al.* STIR: software for tomographic image reconstruction release 2. *Phys. Med. Biol.* **57,** 867 (2012).

52. Collewet, G., Strzelecki, M. & Mariette, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn. Reson. Imaging* **22,** 81–91 (2004).

53. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28,** 129–137 (1982).

54. Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory* **6,** 7–12 (1960).

55. Aisen, A. M. *et al.* MRI and CT evaluation of primary bone and soft-tissue tumors. *Am. J. Roentgenol.* **146,** 749–756 (1986).

56. Kroeker, R. M., Mcveigh, E. R., Hardy, P., Bronskill, M. J. & Henkelman, R. M. In vivo measurements of NMR relaxation times. *Magn. Reson. Med.* **2,** 1–13 (1985).

57. Ling, G. N. & Tucker, M. Nuclear magnetic resonance relaxation and water contents in normal mouse and rat tissues and in cancer cells. *J. Natl. Cancer Inst.* **64,** 1199–1207 (1980).

58. Stöcker, T., Vahedipour, K., Pflugfelder, D. & Shah, N. J. High-performance computing MRI simulations. *Magn. Reson. Med.* **64,** 186–193 (2010).

59. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44,** 837–845 (1988).

60. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3,** 1157–1182 (2003).

61. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334,** 1518–1524 (2011).

Chapter 6

Integration of radiomic-based prediction models with clinical prognostic factors

6.1 Foreword 201
6.2 Abstract 201
6.3 Introduction 202
6.4 Results 204
6.4.1 Summary of presentation of results 204
6.4.2 Association of variables with tumour outcomes 204
6.4.3 Construction of prediction models 207
6.4.4 Performance of prediction models 208
6.4.5 Comparison with other prognostic factors 209
6.4.6 Risk assessment of tumour outcomes 213
6.5 Discussion 214
6.6 Methods 220
6.6.1 Data sets availability 220
6.6.2 Sample size and division of cohorts 222
6.6.3 Extraction of radiomic features 222
6.6.4 Construction of radiomic models 224
6.6.5 Combination of radiomic and clinical variables 225
6.6.6 Imbalance-adjustment strategy 225
6.6.7 Random forest training 226
6.6.8 Calculation of performance metrics 227
6.6.9 Code and models availability 228
6.7 Acknowledgments 228
6.8 Supplementary Material 229
6.8.1 Supplementary results 229
6.8.2 Supplementary methods 235
6.9 References 252

6.1 Foreword

This Chapter presents a study submitted as the following paper: Martin Vallières, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo J. W. L. Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, Jan Seuntjens & Issam El Naqa. "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer". *Sci. Rep.* [submitted March 16, 2017].

In this study, radiomic-based models for the prediction of locoregional recurrences and distant metastases in head-and-neck cancer were constructed using the methods developed in Chapter 3 and a novel imbalance-adjustment strategy. The models were tested onto independent patient cohorts. Furthermore, we used a random forest algorithm to combine radiomics data with patient clinical information. Finally, please note that a particularity of the *Scientific Reports* journal is that the Results section comes before Methods.

6.2 Abstract

Quantitative extraction of high-dimensional mineable data from medical images is a process known as radiomics. Radiomics is foreseen as an essential prognostic tool for cancer risk assessment and the quantification of intratumoural heterogeneity. In this work, 1615 radiomic features (quantifying tumour image intensity, shape, texture) extracted from pre-treatment FDG-PET and CT images of 300 patients from four different cohorts were analyzed for the risk assessment of locoregional recurrences (LR) and distant metastases (DM) in head-and-neck cancer. Prediction models combining radiomic and clinical variables were constructed via random forests and imbalance-adjustment strategies using two of the four cohorts. Independent validation of the prediction and prognostic performance of the models was carried out on the other two cohorts (LR: $\mathrm{AUC} = 0.69$ and $\mathrm{CI} = 0.67$; DM: $\mathrm{AUC} = 0.86$ and $\mathrm{CI} = 0.88$). Furthermore, the results obtained via Kaplan-Meier analysis demonstrated the potential of radiomics for assessing the risk of specific tumour outcomes using multiple stratification groups. This could have important clinical impact, notably by allowing for a better personalization of chemo-radiation treatments for head-and-neck cancer patients from different risk groups.

6.3 Introduction

Precision oncology promises to tailor the full spectrum of cancer care to an individual patient, notably in terms of personalization of cancer prevention, screening, risk stratification, therapy and response assessment. With sufficient infrastructure support and concerted efforts from the different stakeholders, it is possible to foresee that personalized therapy would become the standard of care in oncology [1]. Cancer mechanisms are increasingly elucidated as functions of different biomarkers or tumour genetic mutations, thereby changing the way we design clinical trials to achieve better cancer management efficacy in specific patient sub-populations [2]. On the other hand, “rapid learning paradigms” (i.e., knowledge-driven healthcare) consisting of reusing routine clinical data to develop knowledge in the form of models that can predict treatment outcomes for a larger portion of the population have also gained popularity in the oncology community [3, 4]. Although most research approaches to precision oncology are centered on genomics technologies [5, 6], it is thought that only the integration of multiple-omics, i.e., panomics data (genomics, transcriptomics, proteomics, metabolomics, etc.) could efficiently unravel biological mechanisms [7, 8].

The importance of panomics integration for cancer risk assessment emerges from the tremendous extent of heterogeneous characteristics expressed at multiple levels of tumours. Genes, proteins, cellular microenvironments, tissues and anatomical landmarks within tumours exhibit considerable spatial and temporal variations that could potentially yield valuable information about tumour aggressiveness. Tumours are generally composed of multiple clonal sub-populations of cancer cells forming complex dynamic systems that exhibit rapid evolution as a result of their interaction with their microenvironment and therapy perturbations [9]. Differing properties can be attributed to the different sub-populations in terms of growth rate, expression of biomarkers, ability to metastasize, and immunological characteristics [10]. These properties could be described by differences in metabolic activity, cell proliferation, oxygenation levels, pH, blood vasculature and necrotic areas observed within tumours. Such intratumoural differences are related to the concept of tumour heterogeneity, a characteristic that can be observed with significantly different extents even amongst tumours of the same histopathological type. Tumours exhibiting such heterogeneous characteristics are thought to be associated with high risk of resistance to treatment, progression, metastasis or recurrence [11–13].

Nowadays, medical imaging plays a central role in the investigation of intratumoural heterogeneity, as radiological images are acquired as routine practice for almost every patient with cancer. Medical images such as 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) positron emission tomography (PET) and X-ray computed tomograph (CT) are minimally invasive and they carry an immense source of potential data for decoding the tumour phenotype [14]. The quantitative extraction of high-dimensional mineable data from all types of medical images and whose subsequent analysis aims at supporting clinical decision-making is a process coined with the term "radiomics" [15–18]. The demonstration that gene-expression signatures and clinical phenotypes could be inferred from tumour imaging features [19–21] has led to an exponential growth of this field in the past few years [22, 23]. The underlying hypothesis of radiomics is that the genomic heterogeneity of aggressive tumours could translate into heterogeneous tumour metabolism and anatomy, thereby envisioning the quantitative analysis of diagnostic medical images as an essential prognostic tool for cancer risk assessment and as an integral part of panomic tumour signature profiling.

The translation of radiomics analysis into standard cancer care to support treatment decision-making involves the development of prediction models integrating clinical information that can assess the risk of specific tumour outcomes [24] (Figure 6.1). In this work, our main objective is to construct prediction models using advanced machine learning to evaluate the risk of locoregional recurrences and distant metastases prior to chemo-radiation of head-and-neck (H&N) cancer, a group of biologically similar neoplasms originating from the squamous cells that line the mucosal surfaces in the oral cavity, paranasal sinuses, pharynx or larynx. The locoregional control of H&N cancer is usually good, but this is, however, not matched by improvements in survival, as the development of distant metastases and second primary cancers are the leading causes of treatment failure and death [25, 26]. In order to improve patient survival and outcomes, the importance of identifying relevant prognostic factors that can better assess the aggressiveness of tumours at the moment of diagnosis is crucial.

We hypothesize that radiomic features are important prognostic factors for the risk assessment of specific H&N cancer outcomes [27]. The machine learning strategy employed in this work involves the extraction of 1615 different radiomic features from a total of 300 patients from four different institutions. Two cohorts are used to construct the prediction models by combining radiomics (intensity, shape, textures) and clinical attributes (patient age, H&N type, tumour stage) via random forests classifiers and imbalance adjustments of training samples, and the remaining two cohorts are reserved to evaluate the prediction (binary assessment of outcome) and prognostic (time-to-event assessment) performance of the corresponding models (Figure 6.2). Throughout this study, results obtained for locoregional recurrences and distant metastases are also compared against prediction models constructed for the general risk assessment of overall survival in H&N cancer. A comprehensive comparison of the prediction/prognostic performance of radiomics versus clinical models and volumetric variables is also performed. Our results suggest that the integration of radiomic features into clinical prediction models has considerable potential for assessing the risk of specific outcomes prior to treatment of H&N cancer. Accurate stratification of locoregional recurrence and distant metastasis risks could eventually provide a rationale for adapting the radiation doses and chemotherapy regimens that the patients receive. Overall, combining quantitative imaging information with other categories of prognostic factors via advanced machine learning could have a profound impact on the characterization of tumour phenotypes and would increase the possibility of translation of outcome prediction models into the clinical environment as a means to personalize treatments.

6.4 Results

6.4.1 Summary of presentation of results

To ease reading and the understanding of this study, a summary of how results are presented in the text is provided in Supplementary Figure 6.5.

6.4.2 Association of variables with tumour outcomes

In order to assess the value of quantitative pre-treatment imaging to predict specific cancer outcomes in H&N cancer, we performed a comprehensive univariate analysis of the association of radiomic features with locoregional recurrences ("LR" or "Locoregional"), distant metastases development ("DM" or "Distant") and overall survival ("OS" or "Survival" or "Death"). A total of 1615 radiomic features (Figure 6.1b and Supplementary Material section 6.8.2 under "Description of 3D radiomic features" for complete description) were first extracted from the gross tumour volume ($\text{GTV}_{\text{primary}} + \text{GTV}_{\text{lymph nodes}}$) of the FDG-PET and CT images (Figure 6.1a), for all 300 patients from the four

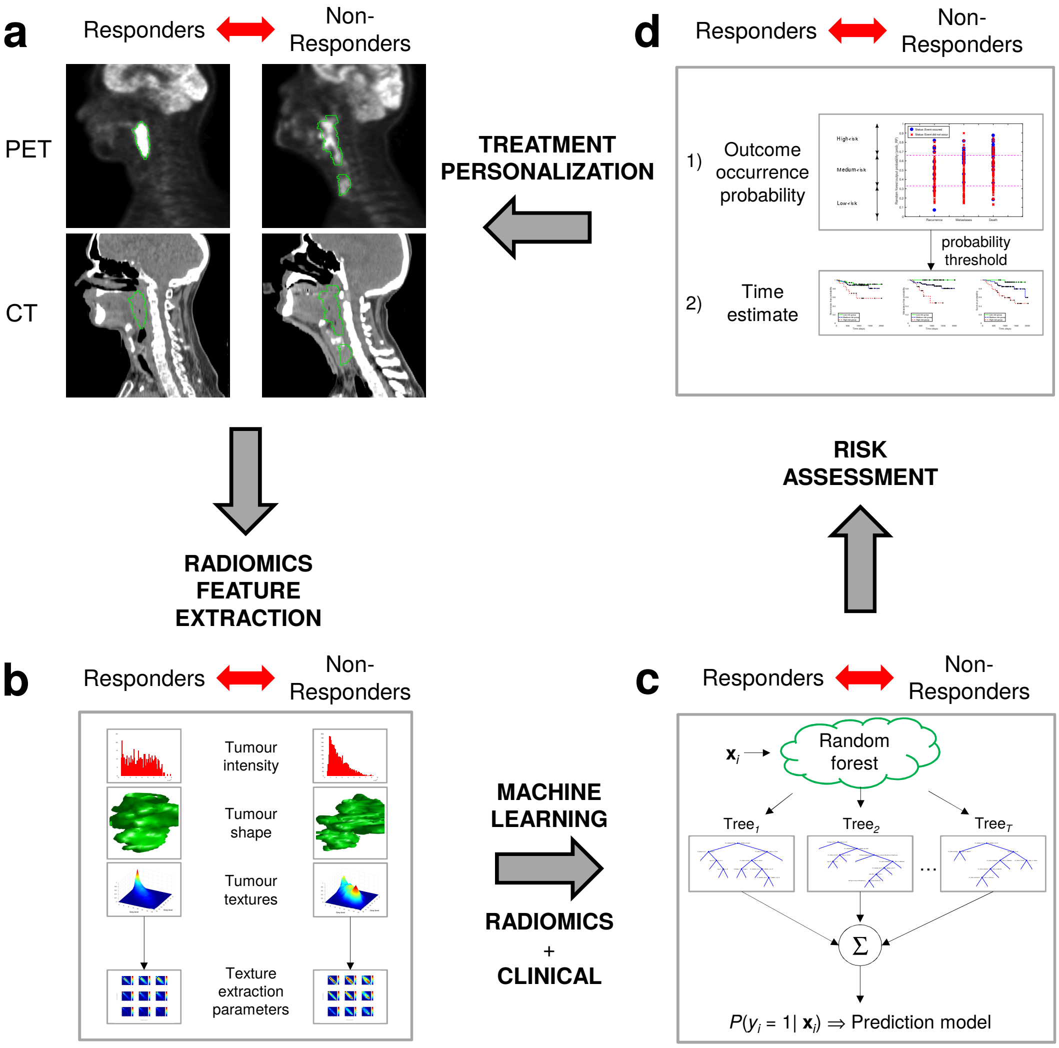


Figure 6.1: From radiomics analysis to treatment personalization. **(a)** Example of diagnostic FDG-PET and CT images of two head-and-neck cancer patients with tumour contours. The patient that did not respond well to treatment (right) has a more heterogeneous intratumoural intensity distribution in both FDG-PET and CT images than the patient that responded well to treatment (left). **(b)** The radiomics analysis strategy involves the extraction of features differentiating responders from non-responders to treatment. Features are extracted from the FDG-PET and CT tumour contours and quantify tumour shape, intensity, and texture. **(c)** Advanced machine learning combines radiomics features and patient clinical information via a random forest algorithm. The classifier is trained to differentiate between responders and non-responders to treatment (prediction model). **(d)** The output probability of the random forest classifier computed on new patients can be used to assess the risk of non-response to treatment via probabilities of occurrence of outcome events and time estimates. Eventually, accurate risk assessment of specific tumour outcomes via radiomics analysis could help to better personalize cancer treatments.

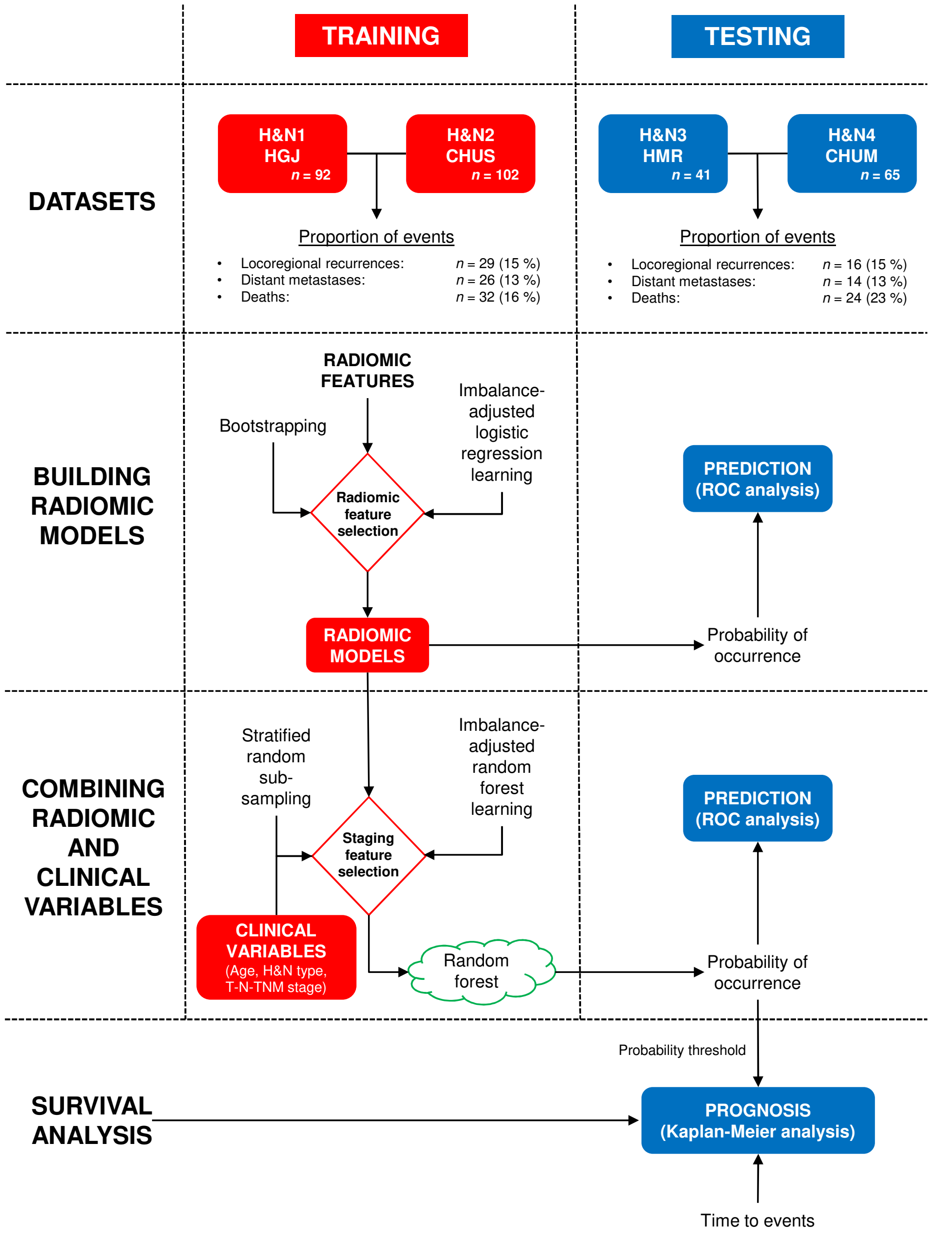

TRAINING
TESTING
DATASETS
H&N1 HGJ n = 92
H&N2 CHUS n = 102
H&N3 HMR n = 41
H&N4 CHUM n = 65
Proportion of events
• Locoregional recurrences: n = 29 (15 %)
• Distant metastases: n = 26 (13 %)
• Deaths: n = 32 (16 %)
Proportion of events
• Locoregional recurrences: n = 16 (15 %)
• Distant metastases: n = 14 (13 %)
• Deaths: n = 24 (23 %)
BUILDING RADIOMIC MODELS
RADIOMIC FEATURES
Bootstrapping
Imbalance-adjusted logistic regression learning
Radiomic feature selection
RADIOMIC MODELS
PREDICTION (ROC analysis)
Probability of occurrence
COMBINING RADIOMIC AND CLINICAL VARIABLES
Stratified random sub-sampling
Imbalance-adjusted random forest learning
Staging feature selection
CLINICAL VARIABLES (Age, H&N type, T-N-TNM stage)
Random forest
PREDICTION (ROC analysis)
Probability of occurrence
Probability threshold
SURVIVAL ANALYSIS
PROGNOSIS (Kaplan-Meier analysis)
Time to events

Figure 6.2: Models construction strategy and analysis workflow. Four different cohorts were used to demonstrate the utility of radiomics analysis for the pre-treatment assessment of the risk of locoregional recurrence and distant metastases in head-and-neck cancer. The H&N1 and H&N2 cohorts were combined and used as a single training set ($n = 194$), whereas the H&N3 and H&N4 cohorts were combined and used as a single testing set ($n = 106$). The best combinations of radiomics features were selected in the training set using imbalance-adjusted logistic regression learning and bootstrapping validations. These radiomics features were combined with selected clinical variables in the training set using imbalance-adjusted random forest learning and stratified random sub-sampling validations. Independent prediction analysis was performed in the testing set for all classifiers fully constructed in the training set. Independent prognosis analysis and Kaplan-Meier risk stratification was carried out in the testing set using the output probability of occurrence of events of random forests fully constructed in the training set.

H&N cancer cohorts (Figure 6.2): I) 10 first-order statistics features (intensity); II) 5 morphological features (shape); and III) 40 texture features each extracted using 40 different combinations of parameters. We also compared these results to the predictive power of the tumour volume ("*Volume*") and of the following clinical variables: *Age*, *T-Stage*, *N-Stage*, *TNM-Stage* and human papillomavirus status (*HPV status*), where *HPV status* was available for 120 of the 300 patients (Supplementary Tables 6.6, 6.7, 6.8 and 6.9). The association of the different variables with the different H&N cancer outcomes (binary endpoints) was then analyzed using Spearman's rank correlations (r_s) computed on all patients, and significance was assessed by applying multiple testing corrections using the Benjamini-Hochberg procedure [28] with a false discovery rate of 10 %.

Overall, we found that 0 %, 63 % and 12 % of the total radiomic features extracted from PET scans, and that 0 %, 61 % and 34 % of the total radiomic features extracted from CT scans were significantly associated with LR, DM and OS, respectively (after multiple testing corrections). The radiomic features (PET or CT) with the highest associations with LR, DM and OS were $LZHGE_{\text{GLSZM}}$ from CT scans ($r_s = -0.15$, $p = 0.007$), ZSN_{GLSZM} from CT scans ($r_s = -0.29$, $p = 2 \times 10^{-7}$) and GLV_{GLRLM} from CT scans ($r_s = 0.24$, $p = 4 \times 10^{-5}$), respectively (Supplementary Table 6.2). Tumour volume was not found to be significantly associated with LR ($r_s = -0.04$, $p = 0.48$), but was significantly associated with DM ($r_s = 0.24$, $p = 3 \times 10^{-5}$) and OS ($r_s = -0.18, p = 2 \times 10^{-3}$). Finally, we found that {*Age*, *T-Stage*, *N-Stage*, *HPV status*}, {*N-Stage*} and {*Age*, *T-Stage*, *HPV status*} were significantly associated with LR, DM and OS, respectively. The clinical variables with the highest associations with LR, DM and OS were $HPV-$ ($r_s = 0.39, p = 8 \times 10^{-6}$), higher *N-Stage* ($r_s = 0.18, p = 1 \times 10^{-3}$) and higher *T-Stage* ($r_s = 0.21, p = 3 \times 10^{-4}$), respectively (Supplementary Table 6.3).

6.4.3 Construction of prediction models

The construction of prediction models for LR, DM and OS was carried out using a training set consisting of the combination of 194 patients from the H&N1 and H&N2 cohorts (Figure 6.2). Three initial radiomic feature sets were considered: I) the 1615 radiomic features extracted from PET scans ("*PET*" feature set); II) the 1615 radiomic features extracted from CT scans ("*CT*" feature set); and III) a combined set containing all PET and CT radiomic features used in feature sets I and II ("*PETCT*" feature set).

Prediction models consisting of radiomic information only were first constructed for each of the three H&N outcomes and the three initial radiomic feature sets. *Feature set reduction*, *feature selection*, *prediction performance estimation*, *choice of model complexity* (Supplementary Figure 6.6) and *final model computation* processes were carried out using logistic regression and bootstrap resampling, similarly to the methodology developed in the study of Vallières *et al.* [29]. To account for the disproportion of occurrence of events and non-occurrence of events in the training set (15 % LR, 13 % DM, 16 % deaths), an imbalance-adjustment strategy adapted from the study of Schiller *et al.* [30] was also applied during the training process. Overall for the *PET*, *CT* and *PETCT* feature sets, the number of variables forming the final radiomic models for each outcome were, respectively: I) 8, 3 and 3 radiomic variables for the LR outcome; II) 6, 3 and 3 radiomic variables for the DM outcome; and III) 4, 3 and 6 radiomic variables for the OS outcome.

The construction of prediction models combining radiomic and clinical variables was then carried out for the nine identified radiomic models (3 feature sets × 3 outcomes). By estimating prediction performance via stratified random sub-sampling in the training set, the following group of clinical variables were first selected for each outcome: I) {*Age*, *H&N type*, *T-Stage*, *N-Stage*} for LR prediction; II) {*Age*, *H&N type*, *N-Stage*} for DM prediction; and III) {*Age*, *H&N type*, *T-Stage*, *N-Stage*} for OS prediction. Final prediction models were ultimately constructed for each radiomic feature set and H&N outcome by combining the selected radiomic and clinical variables via random forests and imbalance adjustments.

6.4.4 Performance of prediction models

The performance of the radiomic prediction models constructed using logistic regression and of the prediction models constructed by combining radiomic and clinical variables via random forests was validated in a testing set consisting of the combination of 106 patients from the H&N3 and H&N4 cohorts (Figure 6.2) using receiver operating characteristic (ROC) metrics (binary endpoints).

Figure 6.3 presents the performance results (AUC: area under the ROC curve) obtained in the testing set for the *radiomics* and *radiomics + clinical* models, where the significance of the increase in AUC when combining clinical to radiomic variables is assessed using the method of DeLong *et al.* [31].

Sensitivity, specificity and accuracy of predictions are also presented in Supplementary Figure 6.7. Overall, it can be observed that there is a general increase in prediction performance for most of the different categories of models that we constructed in this work. For LR prediction, the increase in AUC is significant for prediction models from the *PET* ($p = 0.03$) and the *CT* ($p = 0.01$) radiomic feature sets. For DM prediction, none of the radiomic models show a significant AUC increase when combined with clinical variables. For OS (death) prediction, the increase in AUC is significant for prediction models from the *PET* ($p = 0.01$) and the *PETCT* ($p = 0.006$) radiomic feature sets. Furthermore, we verified that the increase in performance is not explained by the use of a more complex and potentially more predictive learning algorithm: random forests classifiers constructed with radiomic variables alone preserved the predictive properties obtained by logistic regression models constructed with the same variables, but without improving them (Supplementary Table 6.4). These results point to the potential of random forests in successfully combining the complementary value of different categories of prognostic factors such as radiomic and clinical variables.

In Figure 6.3, the highest performance for LR prediction was obtained using the model combining the *PETCT* radiomic and clinical variables, with an AUC of 0.69. For DM prediction, the highest performance was obtained using the *CT* radiomic model, with an AUC of 0.86. These results demonstrate that different radiomic-based models could successfully be used to predict specific outcomes such as locoregional recurrences and distant metastases in H&N cancer. Finally, the highest performance for OS (death) prediction was obtained using the model combining the *PET* radiomic and clinical variables, with an AUC of 0.74. For subsequent analysis in the next section, only the prediction models (*radiomics* and *radiomics* + *clinical*) constructed from these radiomic features sets (*PETCT* for LR, *CT* for DM, *PET* for OS) are used. The complete description of these identified radiomic models (specific features, texture extraction parameters, logistic regression coefficients) is given in Supplementary Material section 6.8.1 under “Complete description of radiomic models”.

6.4.5 Comparison with other prognostic factors

The performance of the best radiomic prediction models and the best prediction models combining radiomic and clinical variables identified in this study (shown with arrows in Figure 6.3) were further compared against other

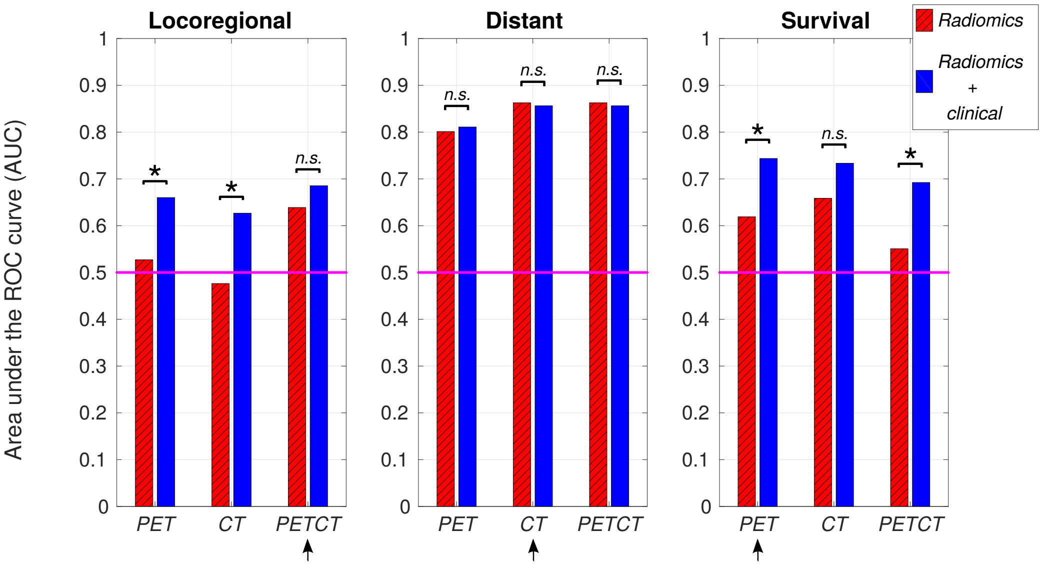


Figure 6.3: Prediction performance of selected models. All prediction models were selected and built using the training set (H&N1 and H&N2; $n = 194$) for three initial radiomic feature sets: I) PET radiomic features (*PET*); II) CT radiomic features (*CT*); and III) PET and CT radiomic features (*PETCT*). The prediction performance is evaluated here in terms of the area under the receiver operating characteristic curve (AUC) for patients of the testing set (H&N3 and H&N4; $n = 106$), for two types of prediction models: I) Radiomic models constructed using logistic regression (*Radiomics*); and II) Radiomic models combined with clinical variables via random forests (*Radiomics + clinical*). Significant increase in AUC from *Radiomics* to *Radiomics + clinical* models is identified with an asterisk (*), and non-significant increase is identified by "*n.s.*". The radiomic feature sets providing the prediction models with highest performance in this study are identified with an arrow for each outcome.

prognostic factors: I) *Volume*; II) "clinical-only" models; III) combination of *Volume* and clinical variables; and IV) a validated radiomic signature developed for the prognosis assessment of overall survival [21, 32]. In addition to the prediction performance evaluated using ROC metrics, the prognostic performance of the models was also assessed using: I) the concordance index (CI) [33] between the output probability of occurrence of an event (LR, DM, death) of prediction models and the time elapsed before an event occurred ("time-to-event"); and II) the *p*-value obtained from Kaplan-Meier analysis using the log-rank test between two risk groups. The models consisting of only radiomic or the *Volume* variables were optimized using logistic or cox regression, and all models involving clinical variables were optimized using random forest classifiers, still using the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$). Fully independent results are then presented in Table 6.1 for models evaluated in the testing set (H&N3 and H&N4 cohorts; $n = 106$).

For locoregional recurrences, we found that the model combining the *PETCT* radiomic and clinical variables provided the best performance in terms of predictive/prognostic power and balance of classification of occurrence of events and non-occurrence of events, notably with an AUC of 0.69, a sensitivity of 0.63, a specificity of 0.68, an accuracy of 0.67, a CI of 0.67 and a Kaplan-Meier *p*-value of 0.03. Using random permutation tests, each variable was calculated to be approximately of equal importance in the random forest model (Supplementary Table 6.5). Similarly to univariate analysis, *Volume* was not found to be a significant prognostic factor for LR. On the other hand, clinical variables alone had high performance with an AUC of 0.72 and a CI of 0.69, but this type of modeling did not provide sufficient balance between the prediction of occurrence and non-occurrence of events (sensitivity of 0.50, specificity of 0.76).

For distant metastases, we found that the model combining the *CT* radiomic and clinical variables provided the best overall performance, notably with an AUC of 0.86, a sensitivity of 0.86, a specificity of 0.76, an accuracy of 0.77, a CI of 0.88 and a Kaplan-Meier *p*-value of 3×10^{-6}. However, radiomic variables were found to be of much higher importance than the clinical variables in the random forest model (Supplementary Table 6.5). In fact, the model composed of clinical variables alone did not perform well. *Volume* was again found to be a significant prognostic factor for DM, but radiomic variables outperformed it.

Table 6.1: Comparison of prediction/prognostic performance of models constructed in this work with other variable combinations. Performance is shown for models constructed in the training set (H&N1 and H&N2; $n = 194$) and independently evaluated in the testing set (H&N3 and H&N4; $n = 106$).

Outcome	Variables	Prediction				Prognosis	
		AUC[a]	Sensitivity[a]	Specificity[a]	Accuracy[a]	CI[b]	p-value[c]
Locoregional	Radiomics$_{\text{PETCT}}$	0.64	0.56	0.67	0.65	0.63	0.28
	Volume	0.43	0.31	0.58	0.54	0.40	0.80
	Clinical	0.72	0.50	0.76	0.72	0.69	0.05
	$Radiomics_{PETCT} + Clinical$	*0.69*	*0.63*	*0.68*	*0.67*	*0.67*	*0.03*
	Volume + Clinical	0.71	0.50	0.76	0.72	0.68	0.06
Distant	Radiomics$_{\text{CT}}$	0.86	0.79	0.77	0.77	0.88	0.0001
	Volume	0.80	0.86	0.65	0.68	0.83	0.10
	Clinical	0.55	0.64	0.46	0.48	0.60	0.61
	$Radiomics_{CT} + Clinical$	*0.86*	*0.86*	*0.76*	*0.77*	*0.88*	*0.000003*
	Volume + Clinical	0.78	1	0.50	0.57	0.80	0.0004
Survival	Radiomics$_{\text{PET}}$	0.62	0.58	0.66	0.64	0.60	0.03
	Volume	0.68	0.67	0.57	0.59	0.67	0.29
	Clinical	*0.78*	*0.92*	*0.57*	*0.65*	*0.76*	*0.00003*
	Radiomics$_{\text{PET}}$ + Clinical	0.74	0.79	0.57	0.62	0.71	0.002
	Volume + Clinical	0.79	0.88	0.52	0.60	0.76	0.0006
Survival[d]	Radiomics$_{\text{CTcompleteSign}}$[e]	–	–	–	–	0.66	0.70
	Radiomics$_{\text{CTsign}}$[f]	0.68	0.71	0.50	0.55	0.66	0.05
	Radiomics$_{\text{CTsign}}$[g] + Clinical	0.80	0.96	0.38	0.51	0.75	0.001

→ Models involving *Radiomic* variables only or the *Volume* variable only were optimized using logistic/cox regression. All models involving *Clinical* variables were optimized using random forests.

→ The best predictive/prognostic and balanced models for each outcome (final models) are identified in italic and are fully described in Supplementary Table 6.5.

[a] Binary prediction of outcome using logistic regression/random forest output responses.

[b] Concordance-index between cox regression/random forest output responses and time to events.

[c] Log-rank test from Kaplan-Meier curves with a risk stratification into two groups (thresholds: median hazard ratio for cox regression, output probability of 0.5 for random forests).

[d] Radiomic signature variables as defined in Aerts *et al.* [21].

[e] Using the original definition of the radiomic signature variables, and the original cox regression coefficients and median hazard ratio trained from the Lung1 cohort in the study of Aerts *et al.* [21].

[f] Using a revised version of the radiomic signature variables (Supplementary Material section 6.8.2 under "Revised version of the radiomic signature") and new cox/logistic regression coefficients trained using the current training set of this work.

[g] Using a revised version of the radiomic signature variables (Supplementary Material section 6.8.2 under "Revised version of the radiomic signature") and a random forest classifer trained using the current training set of this work.

For overall survival, we found that the model composed of clinical variables alone provided the best overall performance, notably with an AUC of 0.78, a sensitivity of 0.92, a specificity of 0.57, an accuracy of 0.65, a CI of 0.76 and a Kaplan-Meier *p*-value of 3×10^{-5}. Furthermore, the *H&N type* variable had the highest and *N-Stage* the lowest importance in the model (Supplementary Table 6.5). Another important finding was that *Volume* alone provided similar or better prognosis assessment of OS than any of the following radiomic-based models: I) the best radiomic model for OS constructed in this work; II) the original radiomic signature using the cox regression coefficients employed in the work of Aerts *et al.* [21]; and III) a revised version of the radiomic signature computation (Supplementary Material section 6.8.2 under "Revised version of the radiomic signature") using new sets of regression coefficients trained with the current training set of this work.

6.4.6 Risk assessment of tumour outcomes

The work performed in this study leads to the identification of three prediction models based on three final random forest classifiers, one for each of the outcome studied here (identified with italic fonts in Table 6.1): I) {*PET-GLN*$_{\mathrm{GLSZM}}$, *CT-Correlation*$_{\mathrm{GLCM}}$, *CT-LGZE*$_{\mathrm{GLSZM}}$, *age*, *H&N type*, *T-Stage*, *N-Stage*} for LR; II) {*CT-LRHGE*$_{\mathrm{GLRLM}}$, *CT-ZSV*$_{\mathrm{GLSZM}}$, *CT-ZSN*$_{\mathrm{GLSZM}}$, *age*, *H&N type*, *N-Stage*} for DM; and III) {*age*, *H&N type*, *T-Stage*, *N-Stage*} for OS. A property of a random forest is that the binary prediction of each of its decision tree can be averaged to serve as an output probability of occurrence of a given event (*prob*$_{\mathrm{RF}}$). This output probability, similarly to other machine learning algorithms, can constitute one of the tools to be used for the risk assessment of specific tumour outcomes. For example, the final random forest classifiers constructed in the training set (H&N1 and H&N2 cohorts; $n = 194$) can be used to stratify the risk of occurrence of the outcome events for each patient of the testing set (H&N3 and H&N4 cohorts; $n = 106$) into three groups (Figure 6.4a): I) low-risk group → $0 \leq prob_{\mathrm{RF}} < \frac{1}{3}$; II) medium-risk group → $\frac{1}{3} \leq prob_{\mathrm{RF}} < \frac{2}{3}$; and III) high-risk group → $\frac{2}{3} \leq prob_{\mathrm{RF}} < 1$. Thereafter, this stratification scheme can be used to evaluate the probability of non-occurrence of the events after a given time for the different risk groups via Kaplan-Meier analysis. Standard Kaplan-Meier analysis using two risk groups ($prob_{\mathrm{RF}} \leq 0.5$, $prob_{\mathrm{RF}} > 0.5$) is first shown in Figure 6.4b for all patients of the testing set. These curves demonstrate the possibility of prognostic risk assessment of specific outcomes in H&N cancer such as locoregional recurrences ($p = 0.03$) and distant metastases ($p = 3 \times 10^{-6}$) using specific prediction models combining different radiomic and clinical variables, but also of the general outcome of overall survival ($p = 3 \times 10^{-5}$) using a prediction model composed of clinical variables only. More accurate prognostic risk assessment can then be further performed using Kaplan-Meier analysis with three risk groups (as defined above: low-risk, medium-risk, high-risk) as shown in Figure 6.4c for all patients of the testing set. For the risk assessment of LR, the developed prediction model is, however, not powerful enough to significantly separate the patients between the high/medium ($p = 0.62$) and medium/low ($p = 0.10$) risk groups. In the case of DM, the developed prediction model allows to significantly separate the patients between the high/medium ($p = 0.05$) and medium/low ($p = 0.03$) risk groups. For OS, the developed prediction model does not significantly separate the patients between the high/medium risk groups ($p = 0.07$), but it does significantly separate the patients between the medium/low risk groups ($p = 0.02$).

6.5 Discussion

Increasing evidence suggests that the genomic heterogeneity of aggressive tumours could translate into intratumoural spatial heterogeneity exhibited at the anatomical and functional scales [19–21]. This constitutes the central idea of the emerging field of “radiomics”, in which large amounts of information via advanced quantitative analysis of medical images are used as non-invasive means to characterize intratumoural heterogeneity and to reveal important prognostic information about the cancer [15–18]. Ultimately, the objective is to narrow down this extensive quantity of information into simple prediction models that can aid in the identification of specific tumour phenotypes for improved treatment management. In this study, we were able via advanced machine learning to develop two prediction models combining PET/CT radiomics and clinical information for the early assessment of the risk of locoregional recurrences and distant metastases in head-and-neck cancers.

First, we extracted a total of 1615 radiomic features from PET and CT pretreatment images of 300 patients with head-and-neck cancer from four different cohorts. These features are composed of 10 intensity features, 5 shape features and 40 textures computed using 40 different combinations of extraction parameters (five isotropic voxel sizes, two quantization algorithms and four numbers of gray levels). In general, different texture features will better

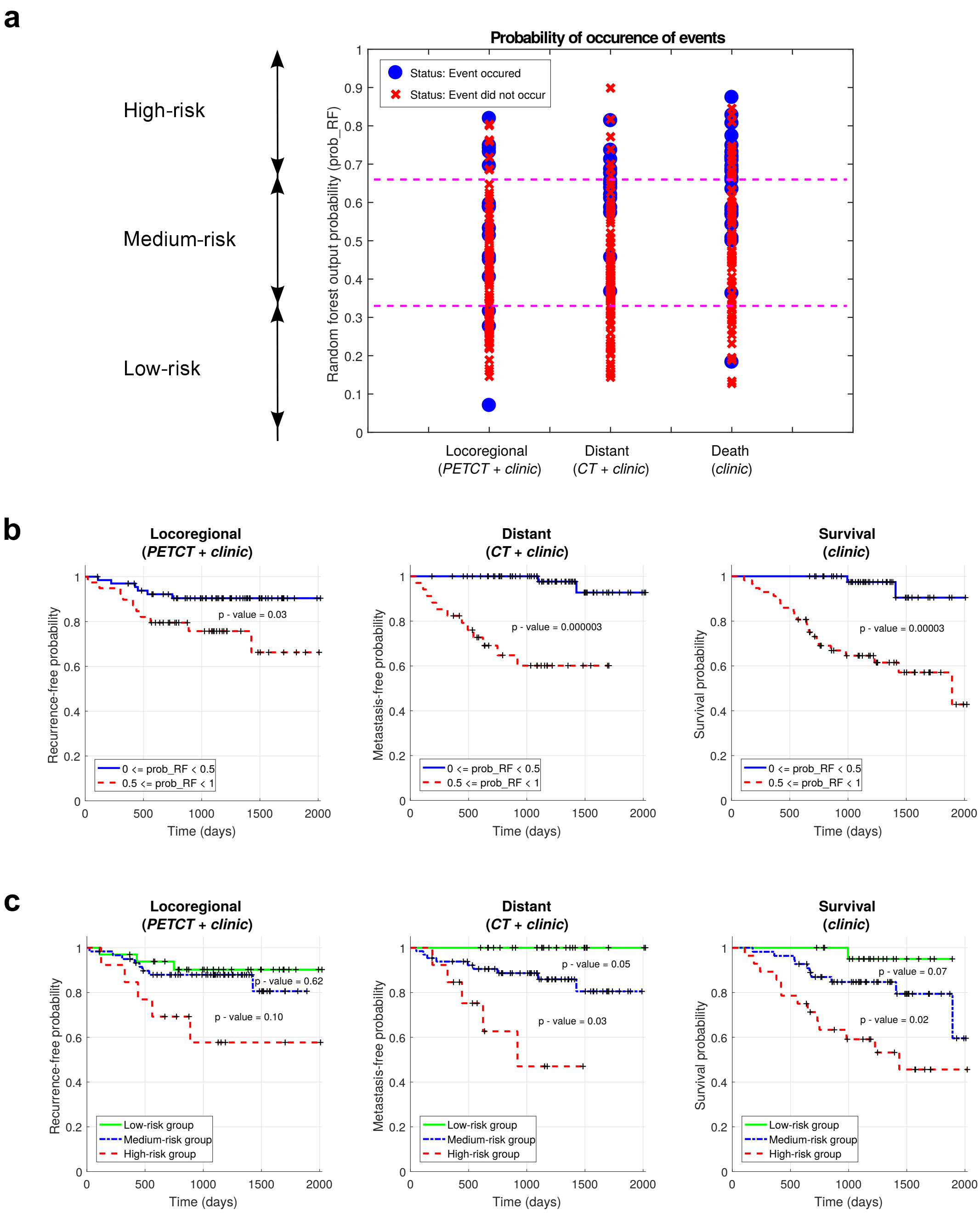


Figure 6.4: Risk assessment of tumour outcomes. **(a)** Probability of occurrence of events (locoregional recurrence, distant metastases, death) for each patient of the testing set (H&N3 and H&N4; $n = 106$) as determined by the random forest classifiers built using the training set (H&N1 and H&N2; $n = 194$). The output probability of occurrence of events of random forests allows for risk stratification; for example, three risk groups can be defined (low, medium, high) using probability thresholds of $\frac{1}{3}$ and $\frac{2}{3}$. **(b)** Kaplan-Meier curves of the testing set using a risk stratification into two groups as defined by a random forest output probability threshold of 0.5. All curves have significant prognostic performance, thus demonstrating the possibility of outcome-specific risk assessment in head-and-neck cancer. **(c)** Kaplan-Meier curves of the testing set using a risk stratification in three groups as defined by random forest output probability thresholds of $\frac{1}{3}$ and $\frac{2}{3}$. Some pair of curves have significant prognostic performance, thus demonstrating the possibility of risk stratification into multiple groups for treatment escalation/personalization in head- and-neck cancer.

represent the underlying tumour biology using different extraction parameters, and the optimal set of parameters to use is application-specific and depends on many factors such as the clinical endpoint studied and the imaging modalities employed. Texture optimization has the potential to enhance the predictive value of the extracted features as Vallières *et al.* [29] have previously shown, and we suggest to incorporate this step in the texture extraction workflow of future similar studies.

Univariate analysis showed that the majority of the features extracted from both PET and CT images are significantly associated with the development of distant metastases, suggesting that the metastatic phenotype of tumours can be captured via quantitative image analysis. On the other hand, none of the radiomic features were significantly associated with locoregional recurrences after multiple testing corrections with a FDR of 10 %. Although combinations of these metrics still proved useful for prognostic risk assessment, it does reveal the need of using other types of metrics such as radiation dose characteristics to enhance the predictive properties of the models constructed for locoregional recurrences. In addition to radiomic features, we also investigated the association of clinical variables with the different head-and-neck cancer outcomes studied in this work. The most significant association was found between *HPV status* and locoregional recurrences, a currently known result that agrees with other studies [34, 35]. However, this result was obtained with only 120 of the 300 patients with available *HPV status*, and this variable could not be used in the subsequent multivariable analysis.

Next, we constructed multivariable prediction models from radiomic information alone by using the methodology developed by Vallières *et al.* [29]. All models were entirely produced from the defined training set of this work combining two head-and-neck cancer cohorts (H&N1 and H&N2; $n = 194$). The best radiomic model for locoregional recurrences (Table 6.1) was found to possess good predictive properties in the defined testing set of this work combining two head-and-neck cancer cohorts (H&N3 and H&N4; $n = 106$). This model is composed of one metric extracted from PET images (GLN_{GLSZM}: *gray-level nonuniformity*$_{\text{GLSZM}}$) and two metrics extracted from CT images (*correlation*$_{\text{GLCM}}$ and $LGZE_{\text{GLSZM}}$: *low gray-level zone emphasis*$_{\text{GLSZM}}$). The best radiomic model for distant metastases (Table 6.1) was found to possess high predictive properties in the testing set, and is composed of three metrics extracted from CT images ($LRHGE_{\text{GLRLM}}$: *long run high gray level emphasis*$_{\text{GLRLM}}$, ZSV_{GLSZM}: *zone size variance*$_{\text{GLSZM}}$ and ZSN_{GLSZM}: *zone size nonuifomity*$_{\text{GLSZM}}$). These results suggest that radiomic models can be specific enough to assess the risk of different outcomes in head-and-neck cancer. The models we developed for locoregional recurrences and distant metastases are in fact overall different and they capture specific tumour phenotypes. It is also noteworthy that all the selected radiomic features of these two models are textural, attesting to the high potential of textures to characterize the complexity of spatial patterns within tumours. As mentioned earlier, aggressive tumours tend to show increased intratumoural heterogeneity [11–13], notably in terms of the heterogeneity in size and intensity characteristics of the different tumour sub-regions in PET and CT images. This effect may be captured by the $\textit{PET-GLN}_{\mathrm{GLSZM}}$, $\textit{CT-LRHGE}_{\mathrm{GLRLM}}$, $\textit{CT-ZSV}_{\mathrm{GLSZM}}$ and $\textit{CT-ZSN}_{\mathrm{GLSZM}}$ texture features in our radiomic models, a result in agreement with a previous study describing the importance of zone-size nonuniformities for the prognostic assessment of head-and-neck tumours [36]. From our experience, we have observed that aggressive tumours also frequently contain large inactive or necrotic regions of uniform intensities, suggesting that these tumours could be rapidly increasing in size and that they could be more at risk to metastasize, for example [37–39]. Here, this effect may be captured by the $\textit{CT-Correlation}_{\mathrm{GLCM}}$ and $\textit{CT-LGZE}_{\mathrm{GLSZM}}$ texture features. Overall, these results suggest that radiomic features could be useful to improve our understanding of the underlying biology of tumours.

We also attempted in this study to improve the predictive power of our prediction models by combining radiomic variables with clinical data. The first step of our method is based on a fast mining of radiomic variables using logistic regression. Then, random forests [40] are used as a means to combine radiomic and clinical information into a single classifier. It would also be feasible to only use random forests to mine the radiomic variables, but our method is advantageous in terms of computation speed. Our results showed that the combination of clinical variables with the optimal radiomic variables via random forests had a positive impact on the prediction and prognostic assessment of locoregional recurrences and distant metastases, although with minimal impact in the latter case (Figure 6.3, Table 6.1). As seen in Supplementary Table 6.5, this can be explained from the fact that the identified radiomic features are the strong and dominant variables in the model for distant metastases predictions. Nonetheless, we believe that random forests is one effective algorithm well-suited to combine variables of different types (categorical and continuous inputs) such as clinical and panomic tumour profile information. In general, the ongoing optimization of machine learning techniques in radiomic applications [41–43] is a step forward to improve clinical predictions.

In this work, we also performed a comprehensive comparison of the prediction/prognostic performance of radiomics versus clinical models and volumetric variables (Table 6.1). Metabolic tumour volume has already been shown to be an independent predictor of outcomes in head-and-neck cancer [44], but it was also suggested by Hatt *et al.* [45] that heterogeneity quantification via texture analysis may provide valuable complementary information to the tumour volume variable for volumes above 10 cm^3. In this study, 85 % of the patients had a gross tumour volume greater than 10 cm^3 and we consequently found that radiomic models performed considerably better than tumour volume alone for the prediction of locoregional recurrences and distant metastases. On the other hand, clinical variables alone did not perform well on their own for distant metastases, but they had good performance for locoregional recurrences by outperforming radiomic models, thus suggesting that our radiomic models need to be improved to better model locoregional recurrences.

In terms of overall survival assessment, our results indicate that the tumour volume variable matched or outperformed all radiomic models thus far we developed or tested in this work, including a previously validated radiomic signature [21, 32]. For one, it is unsurprising that the original radiomic signature [21] did not perform better than tumour volume, as it can be verified that all its feature components are very strongly correlated with tumour volume: the Pearson linear coefficients between tumour volume and the four features of the signature [21] were calculated to be 0.62 (*Energy*), 0.80 (*Compactness*), 0.99 (GLN_{GLRLM}) and 0.94 (GLN_{GLRLM}_HLH) using the whole set of 300 patients of this study, all with $p \ll 0.001$. On the other hand, all the features forming the other radiomic models developed in this work showed potential complementarity value to tumour volume (but the models still did not perform better than tumour volume alone for overall survival assessment): all the features of the revised version of the radiomic signature (Supplementary Material section 6.8.2 under "Revised version of the radiomic signature") had a Pearson linear coefficient lower than 0.5 except one (*Energy*), and all the variables forming the final radiomic models constructed in this work (italic fonts in Table 6.1, including those for locoregional recurrences and distant metastases) had linear coefficients lower than 0.40. This suggests that overall survival may be harder to model than specific tumour outcomes due to a larger number of confounding factors being involved, and it may thus be more prone to overfitting during training. As a consequence, tumour volume may currently be a more robust and reproducible metric than imaging features for modeling this outcome. In the end, the best global performance for overall survival was however obtained with clinical variables alone. This would emphasize that clinical data remains the important source of information to consider for the evaluation of the likelihood of occurrence of a general outcome with many confounding factors such as overall survival, and that more work is required to understand how to adequately model overall survival using radiomic features.

The optimal results in terms of predictive/prognostic performance and balance of prediction between the occurrence and non-occurrence of locoregional recurrences and distant metastases were found in this work by constructing models combining radiomic and clinical variables via random forests (full description of the models in Supplementary Material section 6.8.1 under "Final random forest models and variable importance"). Compared to the general assessment of overall survival as in previous studies [21, 32], our results demonstrate the possibility of decoding specific tumour phenotypes for the risk assessment of specific outcomes in head-and-neck cancer. The final results obtained for distant metastases were considerably higher than those obtained for overall survival, but those obtained for locoregional recurrences were lower albeit clinically significant (Table 6.1). Also, as seen in Figure 6.4 with patients of the testing set, the output probability of occurrence of events of our prediction models allow to significantly separate patients into two locoregional recurrence risk groups and into three distant metastases risk groups. The clinical impact of our results and of the risk assessment of specific outcomes in head-and-neck cancer could be substantial, as it could allow for a better personalization of treatments. For example, higher radiation doses could be considered for patients at higher risks of locoregional recurrences. For distant metastases, the chemotherapy regimens could be strengthened for patients in the high risk group to reduce potential metastatic invasion, and lessened for patients in the low risk group to improve quality of life. These are hypothetical scenarios that, at the moment, are not ready to be implemented in the clinical environment, as our models first need to be constructed and validated on larger patient cohorts, and robust clinical trials are required to validate their benefits on patient survival. Furthermore, the heterogeneity of the patient cohorts used in this work including varying image acquisition parameters may undermine the power of the developed models. However, it may also improve their generalizability, and the results presented in this study could now be useful for the generation of new hypotheses driving future prospective studies.

Overall, we showed in this study that radiomics provide important prognostic information for the risk assessment of locoregional recurrences and distant metastases in head-and-neck cancer. In general, the combination of panomics data into clinically-integrated prediction models should allow to more comprehensively assess cancer risks and could improve how we adapt treatments for each patient. As the standardization efforts of radiomics analysis continue to rapidly progress [46, 47], we can envision the clinical implementation of radiomic-based decision-support systems in the future. Full transparency on data and methods is the key for the progression of the field, and our research efforts needs to include large-scale collaborations and reproducibility practices to increase the possibility of translation of radiomics into the clinical environment [48].

6.6 Methods

6.6.1 Data sets availability

Our analysis was conducted on imaging and clinical data of a total of 300 H&N cancer patients from four different institutions who received radiation alone ($n = 48$, 16 %) or chemo-radiation ($n = 252$, 84 %) with curative intent as part of treatment management. The median follow-up period of all patients was 43 months (range: 6-112). The Institutional Review Boards of all participating institutions approved the study. Retrospective analyses were performed in accordance with the relevant guidelines and regulations as approved by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50).

- The H&N1 data set consists of 92 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Hôpital général juif (HGJ) de Montréal, QC, Canada. During the follow-up period, 12 patients developed a locoregional recurrence (13 %), 16 patients developed distant metastases (17 %) and 14 patients died (15 %). This data set was used as part of the *training set* of this work.

- The H&N2 data set consists of 102 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Centre hospitalier universitaire de

Sherbrooke (CHUS), QC, Canada. During the follow-up period, 17 patients developed a locoregional recurrence (17 %), 10 patients developed distant metastases (10 %) and 18 patients died (18 %). This data set was used as part of the *training set* of this work.

- The H&N3 data set consists of 41 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Hôpital Maisonneuve-Rosemont (HMR) de Montréal, QC, Canada. During the follow-up period, 9 patients developed a locoregional recurrence (22 %), 11 patients developed distant metastases (27 %) and 19 patients died (46 %). This data set was used as part of the *testing set* of this work.
- The H&N4 data set consists of 65 head-and-neck squamous cell carcinoma (HNSCC) patients treated at Centre hospitalier de l'Université de Montréal (CHUM), QC, Canada. During the follow-up period, 7 patients developed a locoregional recurrence (11 %), 3 patients developed distant metastases (5 %) and 5 patients died (8 %). This data set was used as part of the *testing set* of this work.

All patients underwent FDG-PET/CT imaging scans within a median of 18 days (range: 6-66) before treatment. For 93 of the 300 patients (31 %), the radiotherapy contours were directly drawn on the CT of the PET/CT scan by expert radiation oncologists and thereafter used for treatment planning. For 207 of the 300 patients (69 %), the radiotherapy contours were drawn on a different CT scan dedicated to treatment planning and were propagated/resampled to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM® (MIM software Inc., Cleveland, OH).

Further information specific to each patient cohort (e.g., treatment details) is presented in Supplementary Material section 6.8.2 under "Patient datasets" and Supplementary Tables 6.6, 6.7, 6.8 and 6.9. Pre-treatment FDG-PET/CT imaging data, clinical data, radiotherapy contours (*RTstruct*) and MATLAB®routines allowing to read imaging data and their associated region-of-interest (ROI) are made available for all patients on The Cancer Imaging Archive (TCIA) [49]: `http://doi.org/10.7937/K9/TCIA.2017.8oje5q00`. The Research Ethics Committee of McGill University Health Center approved online publishing of clinical and imaging data following patient anonymisation.

6.6.2 Sample size and division of cohorts

Patients with recurrent H&N cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The four patient cohorts were then divided into two groups to create one combined training set (H&N1 and H&N2; $n = 194$) and one combined testing set (H&N3 and H&N4; $n = 106$). Bootstrap resampling and stratified random sub-sampling were always performed with patients from the training set to estimate the relevant performance metrics of interest and to construct the final prediction models, and fully independent validation results were computed with patients from the testing set. This precise type of division of patient cohorts allowed to: I) Train on a combined set of different cohorts to allow the models to take into account some institutional variability; II) Reduce the number of testing results reported; III) Create a training set size to testing set size ratio of approximately 2:1; and IV) Conduct partition sampling such that the proportion of occurrence of events (locoregional recurrences, distant metastases) are approximately the same in the training and testing sets.

6.6.3 Extraction of radiomic features

Starting from the original FDG-PET/CT imaging data and associated radiotherapy contours in DICOM format, the complete set of data was read and transferred into MATLAB® (MathWorks, Natick, MA) format using in-house routines. PET images were converted to standard uptake value (SUV) maps and CT images were kept in raw Hounsfield Unit (HU) format. In this work, we then extracted a total of 1615 radiomic features for both the PET and CT images from the tumour region defined by the "GTV$_{\text{primary}}$ + GTV$_{\text{lymph nodes}}$" contours as delineated by the radiation oncologists of each institution. These features can be divided into three different groups: I) 10 first-order statistics features (intensity); II) 5 morphological features (shape); and III) 40 texture features each computed using 40 different combinations of extraction parameters.

Intensity features are computed from histograms ($n_{bins} = 100$) of the intensity distribution of the ROI. The features extracted in this work were the *variance*, the *skewness*, the *kurtosis*, *SUVmax*, *SUVpeak*, *SUVmean*, the *area under the curve of the cumulative SUV-volume histogram* [50], the *total lesion glycolysis*, the *percentage of inactive volume* and the *generalized effective total uptake* [51]. Shape feature describe geometrical aspects of the ROI. The features extracted in this work were the *volume*, the *size* (maximum tumour diameter), the *solidity*, the *eccentricity* and the *compactness*.

Texture features measure intratumoural heterogeneity by quantitatively describing the spatial distributions of the different intensities within the ROI. In this work, 9 features from the Gray-Level Co-occurrence Matrix (GLCM) [52], 13 features from the Gray-Level Run-Length Matrix (GLRLM) [53–55], 13 features from the Gray-Level Size Zone Matrix (GLSZM) [53–56] and 5 features from the Neighbourhood Gray-Tone Difference Matrix (NGTDM) [57] were computed. All texture matrices were constructed using 3D analysis/26-voxel connectivity of the tumour region resampled to a defined isotropic voxel size. For each of the four texture types, only one matrix was computed per scan by simultaneously taking into account the neighbouring properties of voxels in the 13 directions of 3D space. However, the 6 voxels at a distance of 1 voxel, the 12 voxels at a distance of $\sqrt{2}$ voxels, and the 8 voxels at a distance of $\sqrt{3}$ voxels around center voxels were treated differently in the calculation of the matrices to take into account discretization length differences.

All 40 texture features from the ROI of both PET and CT volumes were extracted using all possible combinations (40) of the following parameters:

- Isotropic voxel size (5): Voxel sizes of 1 mm, 2 mm, 3 mm, 4 mm and 5 mm.
- Quantization algorithm (2): *Equal-probability* (equalization of intensity histogram) and *Uniform* (uniform division of intensity range) quantization algorithms with fixed number of gray levels.
- Number of gray levels (4): Fixed number of gray levels of 8, 16, 32 and 64 in the quantized ROI.

Detailed description with supplementary references and methodology used to extract all radiomic features is further provided in Supplementary Material section 6.8.2 under “Description of 3D radiomic features”.

6.6.4 Construction of radiomic models

The construction of prediction models from the total set of radiomic features for each of the three initial feature sets (I: PET features; II: CT features; and III: PET and CT features) and three H&N cancer outcomes was performed from the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$) using the methodology developed in the work of Vallières *et al.* [29] The process of combining radiomic features into a multivariable model was achieved using the logistic regression utilities of the software DREES [58]. The general workflow is presented in Supplementary Figure 6.9.

First, feature set reduction was performed for each of the initial feature sets via a stepwise forward feature selection scheme in order to create reduced feature sets containing 25 different features balanced between predictive power (Spearman's rank correlation) and non-redundancy (maximal information coefficient[59]). This procedure was carried out using the *Gain* equation [29], which is detailed in Supplementary Material section 6.8.2 under "Feature set reduction".

From the reduced feature sets, stepwise forward feature selection was then carried out by maximizing the 0.632+ bootstrap AUC [60, 61]. For a given model order (number of combined variables) and a given reduced feature set, the feature selection step was divided into 25 experiments. In each of these experiments, all the different features from the reduced set were used as different "starters". For a given starting feature, 100 logistic regression models or order 2 were first created using bootstrap resampling (100 samples) for each of the remaining features in the reduced feature set. Then, the single remaining feature that maximized the 0.632+ bootstrap AUC of the 100 models was chosen, and the process was repeated up to model order 10. Finally, for each model order of each feature set, the experiment that yielded the highest 0.632+ bootstrap AUC was identified, and combinations of features were chosen for model orders of 1 to 10. The whole feature selection process is pictured in Supplementary Material section 6.8.2 under "Feature selection".

Once optimal combinations of features were identified for model orders of 1 to 10 for all feature sets, prediction performance was estimated using the 0.632+ bootstrap AUC (100 samples). By inspecting the prediction estimates shown in Supplementary Figure 6.6, a single combination of features (i.e., model order) potentially possessing the best parsimonious properties was then chosen for each feature set and each outcome (identified as circles in Supplementary Fig. S2). The final logistic regression coefficients of these selected radiomic prediction models (3 feature sets × 3 outcomes) were then found by averaging all coefficients computed from another set of 100 bootstrap samples. These prediction models in their final form were thereafter directly tested in the defined testing set of this work (H&N3 and H&N4 cohorts; $n = 106$).

6.6.5 Combination of radiomic and clinical variables

The construction of prediction models combining radiomic and clinical variables was also carried out using the training set consisting of the combination of 194 patients from the H&N1 and H&N2 cohorts (Figure 6.2). First, random forest classifiers [40] containing only the following clinical variables were constructed for the LR, DM and OS outcomes: I) *Age*; II) *H&N type* (oropharynx, hypopharynx, nasopharynx or larynx); and III) Tumour stage. The selection of the following best groups of tumour stage variables to be incorporated into the "clinical-only" random forest classifiers was performed: I) *T-Stage*; II) *N-Stage*; III) *T-Stage* and *N-Stage*; and IV) *TNM-Stage*. Estimation of prediction performance for feature selection and subsequent random forest training was performed in the training set using stratified random subsampling and imbalance adjustments to account for the disproportion between the occurrence and non-occurrence of events. Overall, the following staging variables were estimated to possess the highest prediction performance in the training set when combined into random forest classifiers with *Age* and *H&N type*, and were thereafter used for the rest of the work accordingly for each outcome: I) *T-Stage* and *N-Stage* for LR prediction; II) *N-Stage* for DM prediction; and III) *T-Stage* and *N-Stage* for OS prediction. Finally, the variables of the previously identified radiomic prediction models (3 feature sets × 3 outcomes) were incorporated with the corresponding clinical variables identified for each outcome via the separate construction of final random forests classifiers.

6.6.6 Imbalance-adjustment strategy

To obtain models with predictive power equally balanced between the prediction of occurrence of events and non-occurrence of events, an imbalance-adjustment strategy adapted from the work of Schiller *et al.* [30] was used in this work (Supplementary Figure 6.8). Imbalance adjustments become an essential part of the training process when the proportion of instances (e.g., patients) of a given class (e.g., occurrence of an event) is much lower than the proportion of instances of the other class (e.g., non-occurrence of an event). This is the case in this work for the proportion of locoregional recurrences, distant metastases and death events in the training and testing sets (Figure 6.2).

In this work, every time a different bootstrap sample was drawn from the training set to construct a logistic regression or a random forest classifier, a different ensemble of multiple balanced classifiers was used in the training process instead of using only one unbalanced classifier. The ensemble classifier is composed of a number of $P = [N^{-}/N^{+}]$ partitions, where N^{-} is the number of instances from the majority class and N^{+} the number of instances from the minority class in a particular bootstrap sample. The N^{+} instances are copied and used in every partition, and the N^{-} instances are randomly sampled without replacement in the P partitions such that the number of instances of the majority class is either $\lfloor N^{-}/P \rfloor$ or $\lceil N^{-}/P \rceil$ in each partition. For example, for $N^{-} = 168$ and $N^{+} = 32$, five partitions would be created: two would contain 33 instances from the majority class, three would contain 34 instances from the majority class, and all would contain the 32 instances from the minority class.

For the logistic regression training process, a different classifier (i.e., different coefficients) is then trained for each of the created partitions, and the final ensemble classifier consists in the average of the corresponding coefficients from each partition. For random forest training, each partition is used to create a decision tree to be appended to a final forest instead of creating only one tree per bootstrap sample.

6.6.7 Random forest training

The process of random forest training inherently uses bootstrapping in order to train the multiple decision trees of the forest. Conventionally, one different decision tree is trained for each bootstrap sample. In this work, we used 100 bootstrap samples to train each random forest constructed from the training set (H&N1 and H&N2 cohorts; $n = 194$). For each bootstrap sample, the imbalance-adjustment strategy detailed above was used such that each bootstrap sample produced multiple decision trees (one per partition) to be appended to a random forest. Therefore, the final number of decision trees per random forest was dependent on the actual proportion of events in each bootstrap sample for each outcome studied. The three final random forest models developed in this work (italic fonts in Table 6.1, Supplementary Table 6.5) were constructed using 582, 661 and 518 decision trees for LR, DM and OS, respectively.

In addition to the imbalance-adjustment strategy adopted in this work, under/oversampling of the instances in each partition of an ensemble was used to further correct for data imbalance in the random forest training process. Under/oversampling weights of the minority class of 0.5 to 2 with increments of 0.1 were tested in this work. Stratified random sub-sampling was used to estimate the optimal weight for a given training process (and also to estimate the optimal clinical staging variables to be used) in terms of the maximal average AUC, a process randomly separating the training set of this work into multiple sub-training and sub-testing sets ($n = 10$) with 2:1 size ratio and equal proportion of events. The final random forest models developed in this work (italic fonts in Table 6.1, Supplementary Table 6.5) used oversampling weights of 1.4, 1.6 and 1.7 (in conjunction with the previously described imbalance-adjustment strategy) to train the decision trees of the forests for LR, DM and OS, respectively. The overall random forest training process is pictured in Supplementary Figure 6.10.

6.6.8 Calculation of performance metrics

In this work, all prediction models were fully trained in the defined training set of this work (H&N1 and H&N2 cohorts; $n = 194$). Models were then independently tested in the defined testing set of this work (H&N3 and H&N4 cohorts; $n = 106$). Prediction performance was assessed using ROC metrics in terms of the AUC, sensitivity, specificity and accuracy of classification of binary clinical endpoints (locoregional recurrences, distant metastases, deaths). Prognostic performance in terms of time estimates of clinical endpoints was assessed using the concordance-index (CI) [33] and the p-value obtained from Kaplan-Meier analysis using the log-rank test between risk groups.

For prediction performance, the output of the linear combination of features of logistic regression models was directly used to calculate the AUC with binary outcome data. The multivariable response was then transformed into the posterior probability of occurrence of an event using a logit transform to calculate the sensitivity, specificity and accuracy of prediction using a probability threshold of 0.5. Similarly, the output probability of occurrence of an event of random forest models was directly used to calculate the AUC with binary outcome data, and an output probability of 0.5 was also used to calculate the remaining metrics.

For prognostic performance, the output of the linear combination of features of cox proportional hazard regression models was directly used to calculate the CI with time-to-event data (time elapsed between the date the treatment ended and the date when an event occurred or the date of last-follow-up). The median of the output of the cox regression models found in the training set was used to separate the patients of the testing set into two risk groups for Kaplan-Meier analysis. For random forests, the output probability of occurrence of an event was directly used to calculate the CI with time-to-event data, and a probability threshold of 0.5 was used to separate the patients of the testing set into two risk groups (or $\frac{1}{3}$ and $\frac{2}{3}$ for three risk groups) for Kaplan-Meier analysis.

6.6.9 Code and models availability

All software code used to produce the results presented in this work is freely shared under the GNU General Public License on the GitHub website at: `https://github.com/mvallieres/radiomics`. Notably, a single organized script allowing to run all the experiments performed in this work is available.

6.7 Acknowledgments

Martin Vallières acknowledges partial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (scholarship CGSD3-426742-2012), and from the CREATE Medical Physics Research Training Network of NSERC (Grant number: 432290). We would also like to thank François Deblois from HGJ, Vincent Hubert-Tremblay, Luc Ouellet and Isabelle Gauthier from CHUS, and Jean-François Carrier, Chantal Boudreau and Karim Zerouali from CHUM for their help in retrieving the data at their respective institutions. Special thanks to Aditya Kumar for his help in the initiation of the project.

6.8 Supplementary Material

6.8.1 Supplementary results

Summary of presentation of results

To ease reading and the understanding of this study, a summary of how results are presented in the main text is provided in Supplementary Figure 6.5.

Choice of complexity of radiomic models

From the defined training set of this work (H&N1 and H&N2; $n = 194$) and similarly to the methodology developed in the study of Vallières *et al.* [29], all initial radiomic feature sets (*PET*, *CT* and *PETCT*) first underwent: I) feature set reduction; and II) feature selection of models combining 1 to 10 variables via logistic regression. Prediction performance was then estimated in the training set in terms of the AUC_{632+} metric using the 0.632+ bootstrap resampling technique [60, 61], for all the 10 different logistic regression models computed on each of the initial feature sets (Supplementary Figure 6.6).

One radiomic model was then chosen for each outcome and feature set, by identifying the lowest number of variables in each model before the prediction performance started reaching a plateau or decreasing (i.e., best parsimonious models). These choices of radiomic model complexity are shown as circles in Supplementary Figure 6.6. The logistic regression coefficients forming the final prediction models for these 9 different choices of radiomic models (3 feature sets $\times$ 3 outcomes) were ultimately fitted using the whole training set.

Univariate analysis

Supplementary Table 6.2 shows the Spearman's rank correlation coefficients (r_s) between the best PET/CT radiomics variables and the binary outcome vectors for all patients of the four cohorts (H&N1, H&N2, H&N3 and H&N4; $n = 300$). Supplementary Table 6.3 shows the Spearman's rank correlation coefficients (r_s) between the clinical variables and the binary outcome vectors for all patients of the four cohorts.

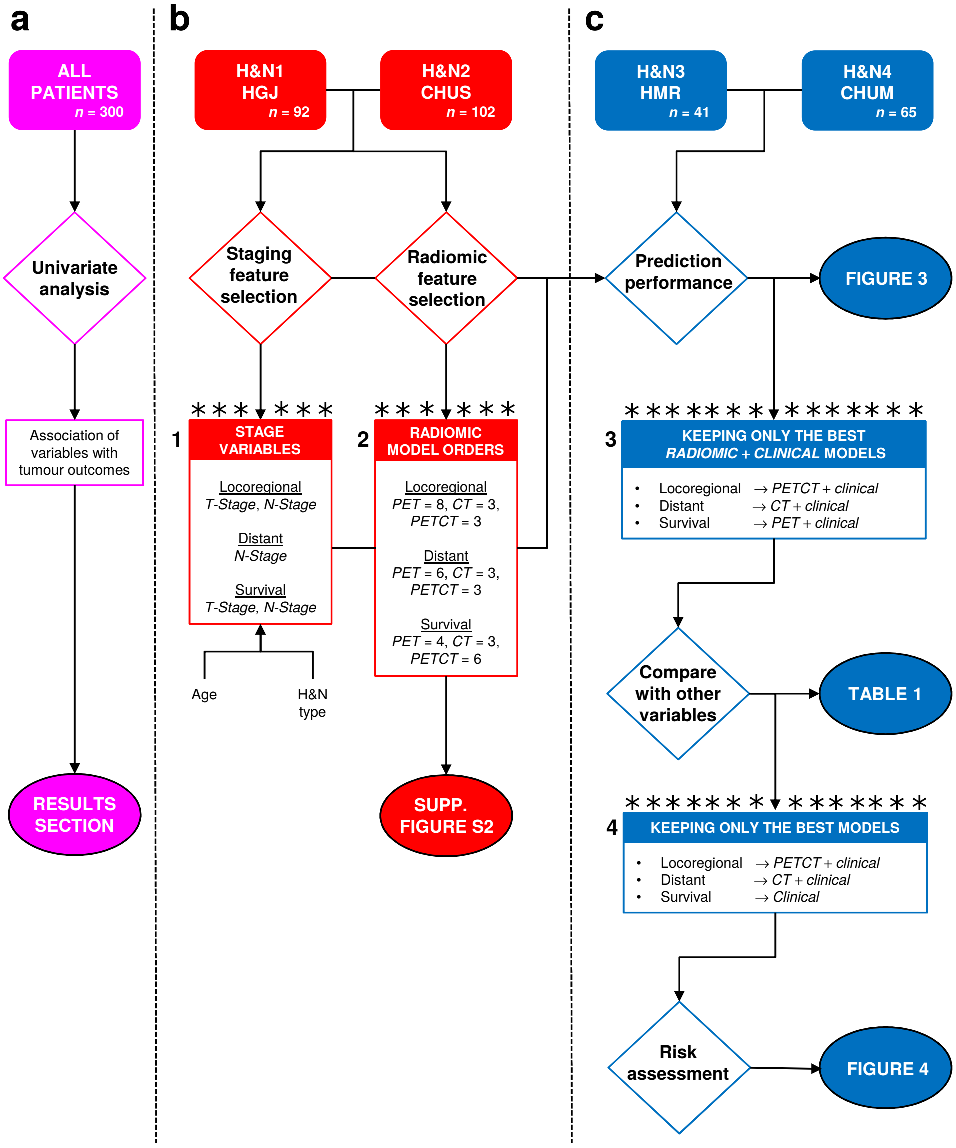


Figure 6.5: Summary of presentation of results. The boxes identified by asterisks represent study checkpoints where only a subset of variables are retained for the remainder of the study. **(a)** Univariate analysis results are computed using the four patient cohorts (H&N1, H&N2, H&N3 and H&N4; $n = 300$) and are presented in the *Results* section of the main text. **(b)** Clinical staging and radiomic feature selection processes are performed using the patient cohorts forming the training set (H&N1 and H&N2; $n = 194$). The clinical staging variables selected for the construction of prediction models are shown in box (1) for each tumour outcome. Radiomic prediction models were selected and built for three initial feature sets: I) PET radiomic features (*PET*); II) CT radiomic features (*CT*); and III) PET and CT radiomic features (*PETCT*). Box (2) shows the radiomic models orders (number of combined variables) chosen in Supplementary Figure 6.6 for each feature set and outcome. **(c)** Performance of prediction models, comparison with other prognostic factors and risk assessment processes are carried out using the patient cohorts forming the testing set (H&N3 and H&N4; $n = 106$). Prediction performance of all models selected and constructed in the training stage is displayed in Figure 6.3 of the main text, and the radiomic feature sets with best prediction performance when combined with clinical variables are shown in box (3) for each outcome. These $radiomic + clinical$ models are further compared against other prognostic factors (e.g. tumour volume, clinical variables alone, etc.) in Table 6.1. The final three models with best overall prediction/prognostic performance for each outcome are shown in box (4), and only these three models are used to perform outcome risk assessment in Figure 6.4 of the main text.

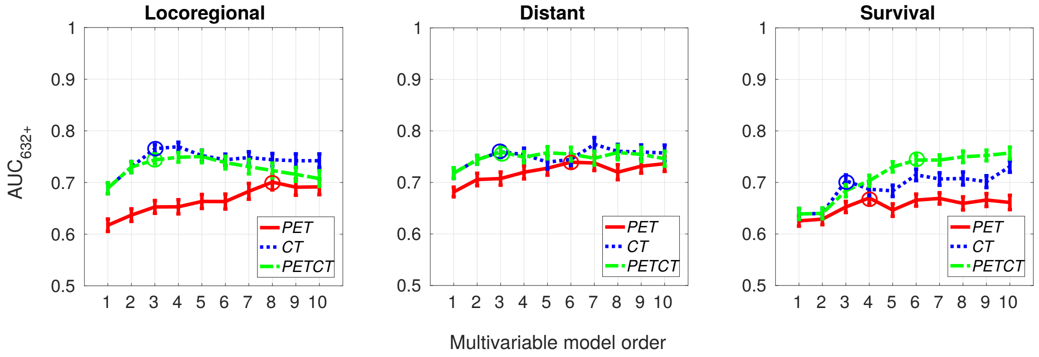


Figure 6.6: Choice of complexity of radiomic models. Choice of the lowest model order (number of combined variables) providing the combination of radiomic variables with the best predictive properties (shown as circles) for each tumour outcome and each of the three initial radiomic feature sets: I) PET radiomic features (*PET*); II) CT radiomic features (*CT*); and III) PET and CT radiomic features (*PETCT*). Prediction performance is estimated in the training set (H&N1 and H&N2; $n = 194$) in terms of the AUC_{632+} metric using bootstrap resampling. Error bars represent the standard error of the mean over 100 bootstrap samples.

Table 6.2: Univariate analysis of radiomics variables.

Metric	Locoregional	Distant	Survival
Best PET	(a) $r_s = -0.14, p = 0.02$	(c) $r_s = 0.28, p = 5.8e-07$*	(e) $r_s = 0.20, p = 3.6e-04$*
Best CT	(b) $r_s = -0.15, p = 7.3e-03$	(d) $r_s = -0.29, p = 2.4e-07$*	(f) $r_s = 0.24, p = 3.7e-05$*

* Significant associations after multiple testing corrections with a FDR of 10 %.

(a) PET-GLN_{GLSZM}: Scale = 2 mm, Quant. algo = *Uniform*, Ng = 16.

(b) CT-LZHGE_{GLSZM}: Scale = 5 mm, Quant. algo = *Equal*, Ng = 64.

(c) PET-Busyness_{NGTDM}: Scale = 3 mm, Quant. algo = *Uniform*, Ng = 64.

(d) CT-ZSN_{GLSZM}: Scale = 1 mm, Quant. algo = *Uniform*, Ng = 16.

(e) PET-$\text{Coarseness}_{NGTDM}$: Scale = 5 mm, Quant. algo = *Uniform*, Ng = 16.

(f) CT-GLV_{GLRLM}: Scale = 1 mm, Quant. algo = *Uniform*, Ng = 16.

Table 6.3: Univariate analysis of clinical variables.

Metric	Locoregional	Distant	Survival
Age	$r_s = 0.15, p = 7.5e-03$*	$r_s = -0.03, p = 0.59$	$r_s = -0.14, p = 0.01$*
T-Stage	$r_s = 0.11, p = 0.07$*	$r_s = 0.10, p = 0.09$	$r_s = -0.21, p = 3.0e-04$*
N-Stage	$r_s = -0.10, p = 0.08$*	$r_s = 0.18, p = 1.4e-03$*	$r_s = -0.07, p = 0.20$
TNM-Stage	$r_s = -0.09, p = 0.13$	$r_s = 0.09, p = 0.14$	$r_s = -0.08, p = 0.15$
HPV status	$r_s = -0.39, p = 8.0e-06$*	$r_s = -0.12, p = 0.19$	$r_s = 0.23, p = 0.01$*

* Significant associations after multiple testing corrections with a FDR of 10 %.

Performance of prediction models

Complete results: AUC, sensitivity, specificity, accuracy. Supplementary Figure 6.7a presents the prediction results obtained in the testing set (H&N3 and H&N4; $n = 106$) using the *radiomics* models, and Supplementary Figure 6.7b presents the prediction results obtained in the testing set using the models formed from the combination of radiomic and clinical variables (*radiomics + clinical*).

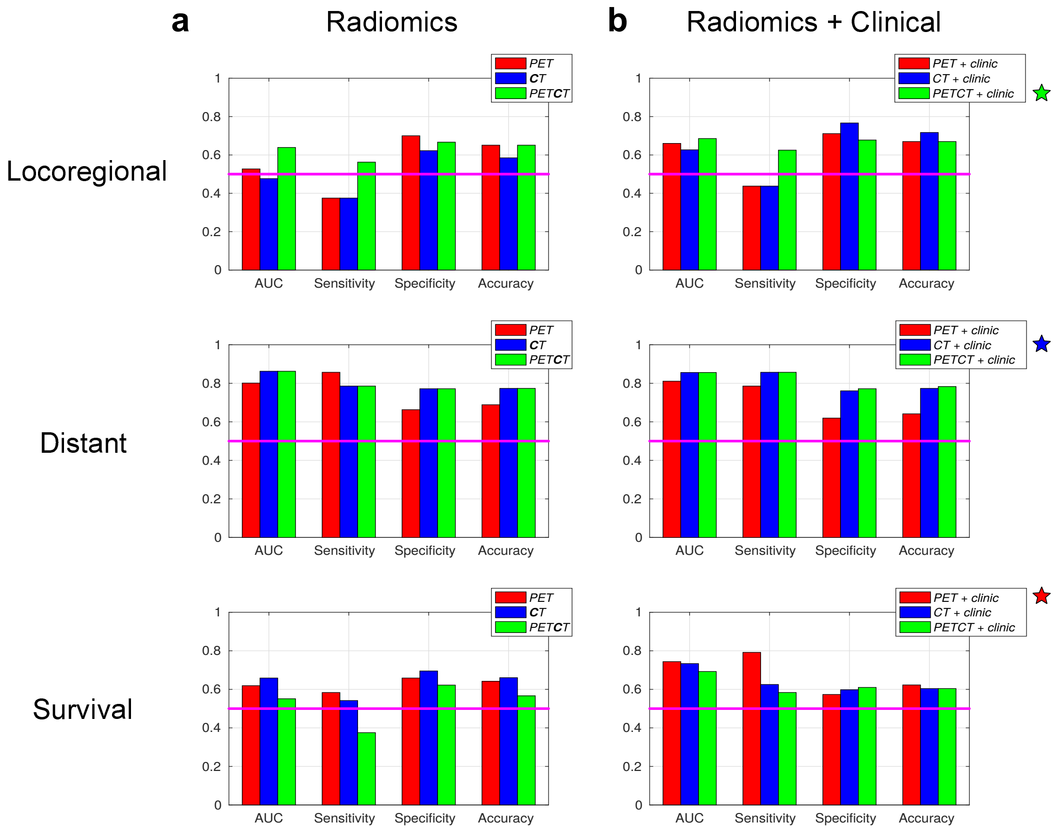


Figure 6.7: Prediction performance of selected models – complete results. All prediction models were selected and built using the training set (H&N1 and H&N2; $n = 194$) for three initial radiomic feature sets: I) PET radiomic features (*PET*); II) CT radiomic features (*CT*); and III) PET and CT radiomic features (*PETCT*). The prediction performance is evaluated here for patients of the testing set (H&N3 and H&N4; $n = 106$). **(a)** Prediction performance of radiomic models constructed using logistic regression. **(b)** Prediction performance of radiomic models combined with clinical variables via random forests. The models providing the best overall performance in terms of predictive power and balance of classification of occurence of events and non-occurrence of events are identified with stars.

For locoregional prediction, the model composed of three variables from the *PETCT* radiomic feature set obtained the best overall performance in terms of predictive power and balance of classification of occurrence of events and non-occurrence of events, with an AUC of 0.64, a sensitivity of 0.56, a specificity of 0.67 and an accuracy of 0.65. The addition of the clinical variables {*Age*, *H&N type*, *T-Stage*, *N-Stage*} to this radiomic model via random forests reached an AUC of 0.69, a sensitivity of 0.63, a specificity of 0.68 and an accuracy of 0.67.

For distant metastases prediction, the best overall performance was obtained with the model composed of three variables from the *CT* radiomic feature set, with an AUC of 0.86, a sensitivity of 0.79, a specificity of 0.77 and an accuracy of 0.77. The addition of the clinical variables {*Age*, *H&N type*, *N-Stage*} to this radiomic model reached and AUC of 0.86, a sensitivity of 0.86, a specificity of 0.76 and an accuracy of 0.77.

For overall survival prediction (death), the best overall performance was obtained with the model composed of four variables from the *PET* radiomic feature set, with an AUC of 0.62, a sensitivity of 0.58, a specificity of 0.66 and an accuracy of 0.64. The addition of the clinical variables {*Age*, *H&N type*, *T-Stage*, *N-Stage*} to this radiomic model reached and AUC of 0.74, a sensitivity of 0.79, a specificity of 0.57 and an accuracy of 0.62.

Complete description of radiomic models. This section provides the complete description (specific features, texture extraction parameters, logistic regression coefficients) of the three best radiomics models of this work, one for each outcome. Significance of the variables in the logistic regression models constructed from the training set (H&N1 and H&N2; $n = 194$) was assessed using the Wald's test implemented in the software DREES [58].

→ Locoregional recurrence

- PET-GLN$_{GLSZM}$: Scale = 2 mm, Quant. algo = *Uniform*, Ng = 64
- CT-Correlation$_{GLCM}$: Scale = 1 mm, Quant. algo = *Uniform*, Ng = 16
- CT-LGZE$_{GLSZM}$: Scale = 1 mm, Quant. algo = *Equal*, Ng = 8
- Significance of variables: $p = 0.04$, $p = 0.004$, $p = 0.02$
- Complete multivariable model response:

$$g(\mathbf{x}_i) = -350.1 \times \text{PET-GLN}_{GLSZM} + 7.42 \times \text{CT-Correlation}_{GLCM} + 21.14 \times \text{CT-LGZE}_{GLSZM} - 0.635$$

→ Distant metastases

- CT-LRHGE$_{GLRLM}$: Scale = 1 mm, Quant. algo = *Equal*, Ng = 8

- CT-ZSV$_{GLSZM}$: Scale = 5 mm, Quant. algo = *Equal*, Ng = 8
- CT-ZSN$_{GLSZM}$: Scale = 1 mm, Quant. algo = *Uniform*, Ng = 16
- Significance of variables: $p = 0.03$, $p = 0.03$, $p = 0.03$
- Complete multivariable model response:

$$g(\mathbf{x}_i) = 0.0233 \times \text{CT-LRHGE}_{GLRLM} - 226.7 \times \text{CT-ZSV}_{GLSZM} - 14.9 \times \text{CT-ZSN}_{GLSZM} + 1.21$$

→ Overall survival (death)

- PET-LGRE$_{GLRLM}$: Scale = 4 mm, Quant. algo = *Equal*, Ng = 64
- PET-SZE$_{GLSZM}$: Scale = 3 mm, Quant. algo = *Uniform*, Ng = 16
- PET-HGZE$_{GLSZM}$: Scale = 1 mm, Quant. algo = *Uniform*, Ng = 64
- PET-ZSN$_{GLSZM}$: Scale = 1 mm, Quant. algo = *Equal*, Ng = 8
- Significance of variables: $p = 0.2$, $p = 0.009$, $p = 0.04$, $p = 0.2$
- Complete multivariable model response:

$$g(\mathbf{x}_i) = -136.8 \times \text{PET-LGRE}_{GLRLM} + 11.49 \times \text{PET-SZE}_{GLSZM} - 0.0035 \times \text{PET-HGZE}_{GLSZM} - 25.91 \times \text{PET-ZSN}_{GLSZM} + 3.921$$

Random forests: radiomic variables only

Results for random forests constructed using radiomic variables only are presented in Supplementary Table 6.4.

Final random forest models and variable importance

In Supplementary Table 6.5, the features of the three final random forest models developed in this work are listed by order of importance in the models. To assess the importance of each feature in each model, an approach combining random permutations and bootstrap resampling was used. First, 100 bootstrap samples were drawn from the testing set (H&N3 and H&N4 cohorts; $n = 106$). For each bootstrap sample, the feature values of all patients of the testing set were permuted once (same permutation for all features). The average percent AUC change over all permutations was then calculated by comparing random permutation AUCs of each variable separately to the true bootstrap AUCs. Significance of each variable in the model (*p*-value) was calculated by comparing the distribution of true bootstrap AUCs to the

Table 6.4: Performance of random forest classifiers constructed using *radiomic* variables only.

Outcome	Selected features[a]	Prediction				Prognosis	
		AUC[b]	Sensitivity[b]	Specificity[b]	Accuracy[b]	CI[c]	*p*-value[d]
Locoregional	PET-GLN$_{GLSZM}$ CT-Correlation$_{GLCM}$ CT-LGZE$_{GLSZM}$	0.61	0.56	0.68	0.66	0.60	0.16
Distant	CT-LRHGE$_{GLRLM}$ CT-ZSV$_{GLSZM}$ CT-ZSN$_{GLSZM}$	0.86	0.79	0.77	0.77	0.88	0.000007
Survival	PET-LGRE$_{GLRLM}$ PET-SZE$_{GLSZM}$ PET-HGZE$_{GLSZM}$ PET-ZSN$_{GLSZM}$	0.60	0.71	0.45	0.51	0.58	0.28

[a] See Supplementary Material section 6.8.1 under "Complete description of radiomic models" for the list of extraction parameters of texture features.

[b] Binary prediction of outcome using random forest probability output.

[c] Concordance-index between random forest probability output and time to event.

[d] Log-rank test from Kaplan-Meier curves with a risk stratification into two groups (probability threshold of 0.5).

distribution of permuted AUCs via the Wilcoxon right-sided test. The more the AUC decreases as a result of random permutations, the more important the variable is to the model.

6.8.2 Supplementary methods

Imbalance-adjustment strategy

In Supplementary Figure 6.8, the imbalance-adjustment strategy used in this work is detailed. In our work, this strategy is combined to uniform bootstrap resampling: every time a boostrap sample is created for prediction estimation using logistic regression or for random forest construction, an ensemble of multiple balanced classifiers is used (in contrast to using only one unbalanced classifier).

In Supplementary Figure 6.8, please note that "$[x]$" refers to a rounding operation, "$\lceil x \rceil$" refers to a ceiling operation, and "$\lfloor x \rfloor$" refers to a floor operation. For example, for $N^- = 56$ and $N^+ = 11$, 5 partitions would be created. All partitions would contain the initial 11 positive instances. The 56 negative instances would be distributed between the 5 partitions such that the first 4

Table 6.5: Best predictive/prognostic and balanced random forest models found in this work.

Outcome	Selected features[a]	AUC change[b]	p-value[c]
Locoregional	Age	-22.8%	$\ll 0.001$
	CT-LGZE$_{GLSZM}$	-16.3%	$\ll 0.001$
	PET-GLN$_{GLSZM}$	-16.1%	$\ll 0.001$
	CT-Correlation$_{GLCM}$	-14.6%	$\ll 0.001$
	H&N type	-14.2%	$\ll 0.001$
	N-Stage	-13.4%	$\ll 0.001$
	T-Stage	-12.3%	$\ll 0.001$
Distant	CT-ZSN$_{GLSZM}$	-15.9%	$\ll 0.001$
	CT-ZSV$_{GLSZM}$	-7.7%	$\ll 0.001$
	CT-LRHGE$_{GLRLM}$	-3.1%	0.00002
	H&N type	$+0.2\%$	0.40
	N-Stage	$+2.7\%$	1
	Age	$+3.5\%$	1
Survival	H&N type	-13.8%	$\ll 0.001$
	T-Stage	-9.9%	$\ll 0.001$
	Age	-9.6%	$\ll 0.001$
	N-Stage	-1.2%	0.30

[a] See Supplementary Material section 6.8.1 under “Complete description of radiomic models” for the list of extraction parameters of texture features.

[b] Average of (AUC$_{\text{perm}}$ - AUC$_{\text{true}}$)/AUC$_{\text{true}}$ over all permutations. The more negative, the more important the variable is in the model.

[c] Significance in the model via the Wilcoxon right-sided test.

partitions would contain 11 negative instances and the last one 12 negative instances.

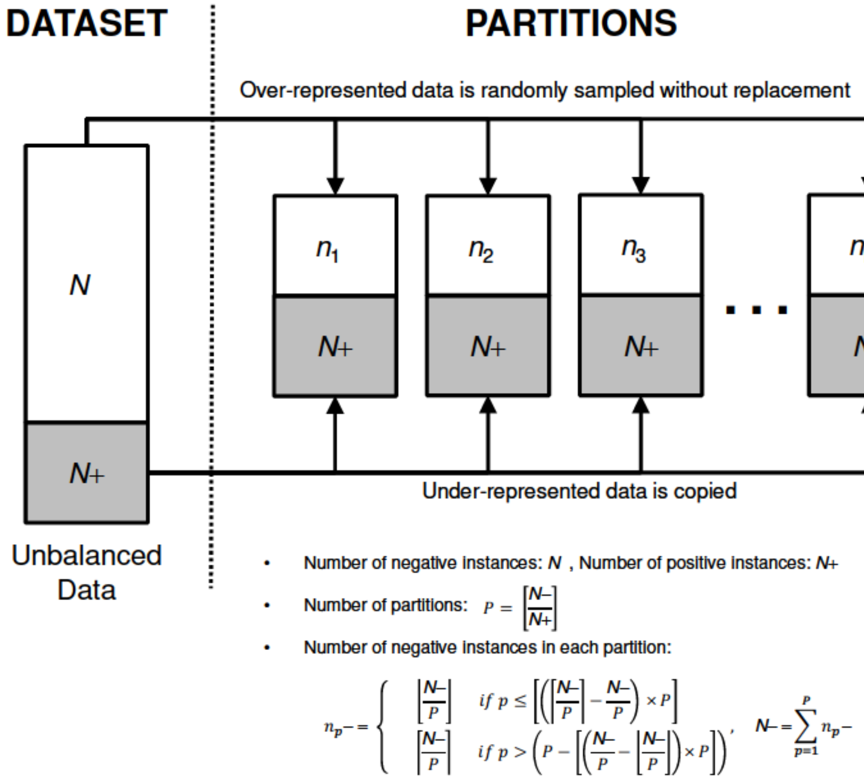

DATASET
PARTITIONS
Over-represented data is randomly sampled without replacement
N
N+
Unbalanced Data
n_1
n_2
n_3
n_P
N+
N+
N+
N+
. . .
Under-represented data is copied
Number of negative instances: N , Number of positive instances: N+
Number of partitions: $P = \left\lceil \frac{N-}{N+} \right\rceil$
Number of negative instances in each partition:
$n_p- = \begin{cases} \left\lfloor \frac{N-}{P} \right\rfloor & if\ p \leq \left[\left(\left\lceil \frac{N-}{P} \right\rceil - \frac{N-}{P}\right) \times P\right] \\ \left\lceil \frac{N-}{P} \right\rceil & if\ p > \left(P - \left[\left(\frac{N-}{P} - \left\lfloor \frac{N-}{P} \right\rfloor\right) \times P\right]\right) \end{cases}, \quad N- = \sum_{p=1}^{P} n_p-$

Figure 6.8: Imbalance-adjustment strategy. Adapted from Schiller *et al.* [30]

Construction of radiomic models

General workflow Supplementary Figure 6.9 presents the general wokflow used to construct radiomic models. For more details, please see the next two sections and the work of Vallières *et al.* [29].

Feature set reduction In this thesis, please see section 5.9.2 under "Feature set reduction".

Feature selection In this thesis, please see section 5.9.2 under "Feature selection".

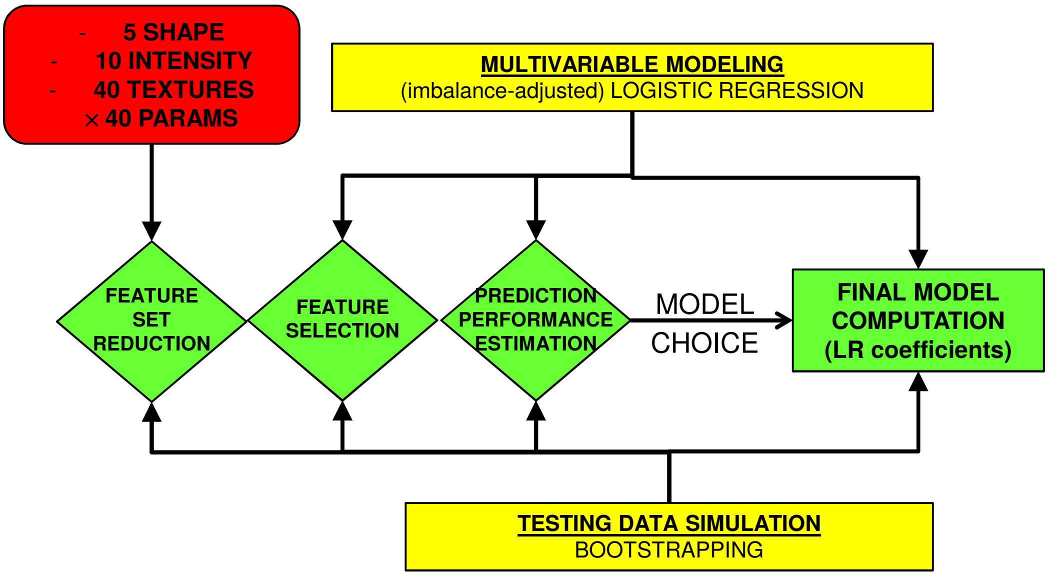


Figure 6.9: Workflow of construction of radiomic models.

Random forest training

Supplementary Figure 6.10 presents the methodology used in this work for random forest training. Stratified random sub-sampling is used to estimate the predictive properties of the random forests (e.g., estimating the best tumour staging metric addition and positive instances weight in the forests by maximizing $\widehat{\mathrm{AUC}}$). For each training sub-sample, boostrap resampling is used to grow a single random forest to be tested in the corresponding testing sub-sample. Through the imbalance-adjustment strategy, each bootstrap sample produces multiple decision trees (one decision tree per partition) to be appended to the random forest of the corresponding training sub-sample (in contrast to conventionally producing a single decision tree per bootstrap sample).

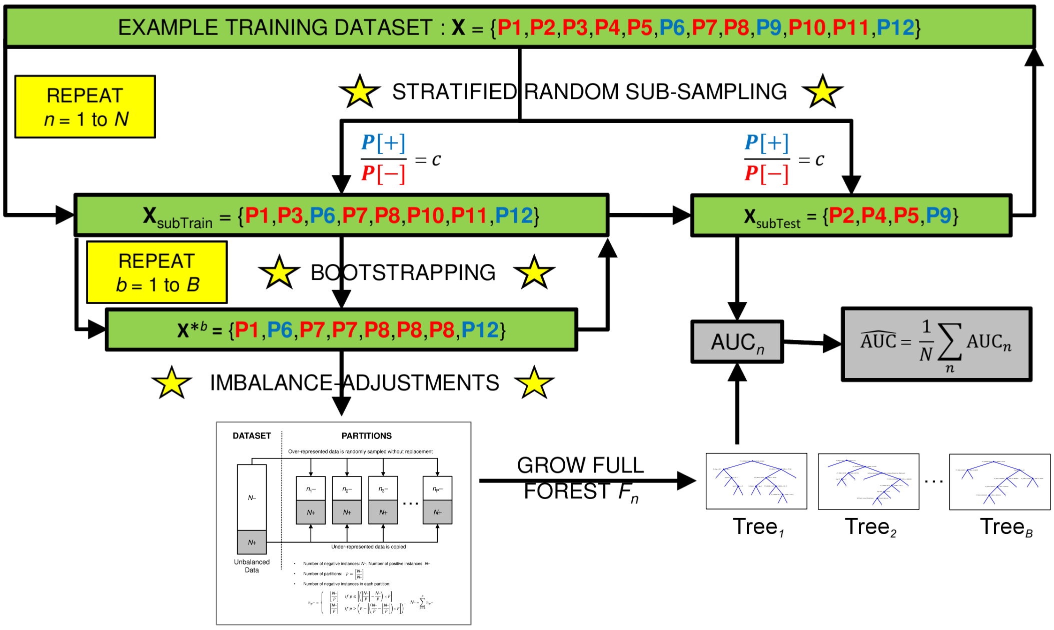

EXAMPLE TRAINING DATASET : X = {P1,P2,P3,P4,P5,P6,P7,P8,P9,P10,P11,P12}
REPEAT n = 1 to N
STRATIFIED RANDOM SUB-SAMPLING
P[+] / P[−] = c
P[+] / P[−] = c
X_subTrain = {P1,P3,P6,P7,P8,P10,P11,P12}
X_subTest = {P2,P4,P5,P9}
REPEAT b = 1 to B
BOOTSTRAPPING
X*b = {P1,P6,P7,P7,P8,P8,P8,P12}
IMBALANCE-ADJUSTMENTS
AUC_n
AUC = 1/N Σ_n AUC_n
DATASET
PARTITIONS
Unbalanced Data
GROW FULL FOREST F_n
Tree_1
Tree_2
Tree_B

Figure 6.10: Random forest training.

Patient datasets

Head and Neck 1 → Hôpital général juif, Montréal, QC

PATIENT POPULATION. This cohort is composed of 92 patients with primary squamous cell carcinoma of the head-and-neck (stage I-IVb) treated between 2006 and 2014 at Hôpital général juif, Montréal, QC. Included patients were treated with curative intent with radiation alone or with chemoradiation. Patients with recurrent head-and-neck cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. The median follow-up period of the cohort was 46 months (range: 11-112). Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The study has been approved by the institutional review board of Hôpital général juif. Detailed information about this patient cohort is provided in Supplementary Table 6.6.

TREATMENT DETAILS. Patients with stage I-II disease were treated with definitive radiotherapy alone while patients with stage III-IV disease were treated using concurrent chemo-radiation. The radiotherapy regimen was

Table 6.6 : Characteristics of H&N1 cohort – HGJ.

Characteristic	Type	No. of patients
Gender	Male	75 (82 %)
	Female	17 (18 %)
Age	Range	18-84
	Mean ± STD	61 ± 11
Tumour type	Oropharynx	56 (61 %)
	Hypopharynx	4 (4 %)
	Nasopharynx	14 (15 %)
	Larynx	14 (15 %)
	Unknown	4 (4 %)
T-Stage	T1	20 (22 %)
	T2	20 (22 %)
	T3	35 (38 %)
	T4	13 (14 %)
	Tx	4 (4 %)
N-Stage	N0	13 (14 %)
	N1	18 (20 %)
	N2	58 (63 %)
	N3	3 (3 %)
TNM-Stage	Stage I	1 (1 %)
	Stage II	5 (5 %)
	Stage III	28 (30 %)
	Stage IV	58 (63 %)
HPV status	Positive	30 (33 %)
	Negative	25 (27 %)
	N/A	37 (40 %)
Treatment	Radiation only	4 (4 %)
	Chemo-radiation	88 (96 %)
Outcome	Locoregional recurrence	12 (13 %)
	Distant metastases	16 (17 %)
	Death	14 (15 %)

planned using Volumetric Arc Modulated Radiotherapy – Rapidarc planning system (Varian Medical Systems). The radiotherapy regime consisted of hypofractionated fractionated radiotherapy with simultaneous integrated boost where the GTV was planned to receive a total of 67.5 Gy in fractions of 2.25 Gy over 6 weeks, while CTV received a total of 54-60 Gy in fractions of 1.8-2 Gy over 30 fractions. The treatment was delivered on a Linac equipped with HD120 Multileaf Collimator, with Image Guided Radiotherapy using daily kv-kv imaging and weekly Cone beam CT-scan (CBCT). Concomitant chemotherapy was given via weekly administration of Carboplatin at AUC 2-3 and Paclitaxel at dose of 40 mg/m^2.

FDG-PET/CT SCANS. All 92 eligible patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery ST, GE Healthcare) within 37 days before treatment (median: 14 days). For the PET portion of the FDG-PET/CT scan, a median of 584 MBq (range: 368-715) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 180-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a span (axial mash) of 5. The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52×3.52 mm^2 (range: 3.52-4.69). For the CT portion of the FDG-PET/CT scan, an energy of 140 kVp with an exposure of 12 mAs was used. The CT slice thickness resolution was 3.75 mm and the median in-plane resolution was 0.98×0.98 mm^2 for all patients. Contours defining the gross tumour volume (GTV) and lymph nodes were drawn by an expert radiation oncologist in a radiotherapy treatment planning system. For 2 of the 92 patients, the radiotherapy contours were directly drawn on the CT scan of the FDG-PET/CT scan. For 90 of the 92 patients, the radiotherapy contours were drawn on a different CT scan dedicated to treatment planning. In the latter case, the contours were propagated to the FDG-PET/CT scan reference frame using deformable registration with the software MIM® (MIM software Inc., Cleveland, OH).

Head and Neck 2 → Centre hospitalier universitaire de Sherbooke, Sherbrooke, QC

Table 6.7 : Characteristics of H&N2 cohort – CHUS.

Characteristic	Type	No. of patients
Gender	Male	74 (73 %)
	Female	28 (27 %)
Age	Range	34-88
	Mean ± STD	64 ± 10
Tumour type	Oropharynx	73 (72 %)
	Hypopharynx	1 (1 %)
	Nasopharynx	6 (6 %)
	Larynx	22 (22 %)
T-Stage	T1	9 (9 %)
	T2	45 (44 %)
	T3	31 (30 %)
	T4	17 (17 %)
N-Stage	N0	38 (37 %)
	N1	11 (11 %)
	N2	50 (49 %)
	N3	3 (3 %)
TNM-Stage	Stage I	3 (3 %)
	Stage II	17 (17 %)
	Stage III	22 (22 %)
	Stage IV	60 (59 %)
HPV status	Positive	26 (25 %)
	Negative	13 (13 %)
	N/A	63 (62 %)
Treatment	Radiation only	33 (32 %)
	Chemo-radiation	69 (68 %)
Outcome	Locoregional recurrence	17 (17 %)
	Distant metastases	10 (10 %)
	Death	18 (18 %)

PATIENT POPULATION. This cohort is composed of 102 patients with primary squamous cell carcinoma of the head-and-neck (stage I-IVb) treated between 2007 and 2014 at Centre hospitalier universitaire de Sherbooke, Sherbrooke, QC. Included patients were treated with curative intent with radiation alone or with chemo-radiation. Patients with recurrent head-and-neck cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. The median follow-up period of the cohort was 44 months (range: 8-93). Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The study has been approved by the institutional review board of Centre hospitalier universitaire de Sherbooke. Detailed information about this patient cohort is provided in Supplementary Table 6.7.

TREATMENT DETAILS. All patients have had a pathological confirmation of squamous cell carcinoma and imaging examination for tumor staging before all treatments. All those patients have had a treatment position PET imaging in our center. The PET images have been merged with dosimetry CT imaging, and the dosimetry plan has been performed with teraplan for 3D-conformal technique and pinnacle system for IMRT. The 3D-conformal technique has been used for all patients before 2008, and since 2008, all patients have been treated by IMRT. The treatment approaches consisted of either radiotherapy alone or radiotherapy with concurrent chemotherapy or concurrent Cetuximab. The treatment dose varied according to the tumor staging. The patients with T1 glottic laryngeal cancer have been treated mostly by 2.5 Gy daily for total dose of 50Gy, some patients have been treated with daily dose of 2.25 Gy for 63 Gy totally. All other patients with T1, T2, N0 cancers have been treated with standard fractionated radiation schedules of 60-66 Gy; for the patients with T3-4, or N+, the treatment dose varied from 68.8 Gy in 32 fractions to 70 Gy in 33 fractions. All treatments have been performed by 6 MV linear accelerator. The concurrent chemotherapy was either cisplatin 100 mg/m^2 at D1, D22 & D43, or cisplatin 40 mg/m^2, weekly. According to the consideration of the oncologist, some patients have been treated by radiotherapy associated with Cetuximab, due to the problems of kidney function, audition, elder or weak general performance status. The treatment schedule of concurrent Cetuximab was administrated according to the study of Bonner *et al.* [62].

FDG-PET/CT SCANS. All 102 eligible patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (GeminiGXL 16, Philips) within 54 days before treatment (median: 19 days). For the PET portion of the FDG-PET/CT scan, a median of 325 MBq (range: 165-517) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 150 s (range: 120-151) per bed position. Attenuation corrected images were reconstructed using a LOR-RAMLA iterative algorithm. The FDG-PET slice thickness resolution was 4 mm and the median in-plane resolution was 4×4 mm^2 for all patients. For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 12-140) with a median exposure of 210 mAs (range: 43-250) was used. The median CT slice thickness resolution was 3 mm (range: 2-5) and the median in-plane resolution was 1.17×1.17 mm^2 (range: 0.68-1.17). Contours defining the gross tumour volume (GTV) and lymph nodes were drawn by an expert radiation oncologist in a radiotherapy treatment planning system. For 91 of the 102 patients, the radiotherapy contours were directly drawn on the CT scan of the FDG-PET/CT scan. For 11 of the 102 patients, the radiotherapy contours were drawn on a different CT scan dedicated to treatment planning. In the latter case, the contours were propagated to the FDG-PET/CT scan reference frame using deformable registration with the software MIM® (MIM software Inc., Cleveland, OH).

Head and Neck 3 → Hôpital Maisonneuve-Rosemont, Montréal, QC

PATIENT POPULATION. This cohort is composed of 41 patients with primary squamous cell carcinoma of the head-and-neck (stage II-IVb) treated between 2008 and 2014 at Hôpital Maisonneuve-Rosemont, Montréal, QC. Included patients were treated with curative intent with radiation alone or with chemo-radiation. Patients with recurrent head-and-neck cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. The median follow-up period of the cohort was 38 months (range: 6-70). Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The study has been approved by the institutional review board of Hôpital Maisonneuve-Rosemont. Detailed information about this patient cohort is provided in Supplementary Table 6.8.

Table 6.8 : Characteristics of H&N3 cohort – HMR.

Characteristic	Type	No. of patients
Gender	Male	31 (76 %)
	Female	10 (24 %)
Age	Range	49-85
	Mean ± STD	67 ± 9
Tumour type	Oropharynx	19 (46 %)
	Hypopharynx	7 (17 %)
	Nasopharynx	6 (15 %)
	Larynx	9 (22 %)
T-Stage	T1	2 (5 %)
	T2	17 (41 %)
	T3	9 (22 %)
	T4	12 (29 %)
	Tx	1 (2 %)
N-Stage	N0	5 (12 %)
	N1	4 (10 %)
	N2	27 (66 %)
	N3	5 (12 %)
TNM-Stage	Stage I	0 (0 %)
	Stage II	3 (7 %)
	Stage III	5 (12 %)
	Stage IV	33 (80 %)
HPV status	Positive	2 (5 %)
	Negative	0 (0 %)
	N/A	39 (95 %)
Treatment	Radiation only	7 (17 %)
	Chemo-radiation	34 (83 %)
Outcome	Locoregional recurrence	9 (22 %)
	Distant metastases	11 (27 %)
	Death	19 (46 %)

TREATMENT DETAILS. The treatment options consisted of either definitive radiotherapy alone or concurrent chemo-radiation. All patients received continuous course of radiotherapy delivered by a 6 MV linear accelerator using 7 to 9 fields inverse planning IMRT. Only one patient was planned with 5 fields and another was treated using 6 fields forward planning IMRT to the upper neck and direct anterior field with a spinal cord block to the lower neck. For the patients receiving radiotherapy alone, 4 patients had stage II disease including a T1N1 nasopharyngeal cancer and received a dose 69.96 Gy in 33 fractions, 2 oropharyngeal and 1 hypopharyngeal cancer receiving altered fractionation with a dose of 66 to 67.5 Gy in 30 fractions. The 3 patients were offered but declined the chemotherapy and received 69.36 Gy in 33 fractions. Among patients receiving chemo-radiation, the radiation fractionation mostly used was 69.96 Gy in 33 fractions ($n = 31$) and the remaining received 70 Gy in 35 fractions ($n = 2$). The concurrent chemotherapy was in most cases cisplatin 100 mg/m2 i.v. every 3 weeks.

FDG-PET/CT SCANS. All 41 eligible patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery STE, GE Healthcare) within 60 days before treatment (median: 34 days). For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227-859) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 360 s (range: 120-360) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3-5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52×3.52 mm^2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120-140) with a median exposure of 11 mAs (range: 5-16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98×0.98 mm^2 (range: 0.98-1.37). For all 41 patients, the radiotherapy contours defining the gross tumour volume (GTV) and lymph nodes were drawn by an expert radiation oncologist on a different CT scan dedicated to treatment planning. The contours were then propagated to the FDG-PET/CT scan reference frame using deformable registration with the software MIM® (MIM software Inc., Cleveland, OH).

Head and Neck 4 → Centre hospitalier de l'Université de Montréal, Montréal, QC

Table 6.9 : Characteristics of H&N4 cohort – CHUM.

Characteristic	Type	No. of patients
Gender	Male	49 (75 %)
	Female	16 (25 %)
Age	Range	44-90
	Mean ± STD	63 ± 9
Tumour type	Oropharynx	58 (89 %)
	Hypopharynx	0 (0 %)
	Nasopharynx	2 (3 %)
	Larynx	0 (0 %)
	Unknown	5 (8 %)
T-Stage	T1	8 (12 %)
	T2	28 (43 %)
	T3	19 (29 %)
	T4	5 (8 %)
	Tx	5 (8 %)
N-Stage	N0	4 (6 %)
	N1	8 (12 %)
	N2	45 (69 %)
	N3	8 (12 %)
TNM-Stage	Stage I	0 (0 %)
	Stage II	2 (3 %)
	Stage III	7 (11 %)
	Stage IV	54 (83 %)
	N/A	2 (3 %)
HPV status	Positive	21 (32 %)
	Negative	3 (5 %)
	N/A	41 (63 %)
Treatment	Radiation only	4 (6 %)
	Chemo-radiation	61 (94 %)
Outcome	Locoregional recurrence	7 (11 %)
	Distant metastases	3 (5 %)
	Death	5 (8 %)

PATIENT POPULATION. This cohort is composed of 65 patients with primary squamous cell carcinoma of the head-and-neck (stage II-IVb) treated between 2009 and 2013 at Centre hospitalier de l'Université de Montréal, Montréal, QC. Included patients were treated with curative intent with radiation alone or with chemo-radiation. Patients with recurrent head-and-neck cancer or with metastases at presentation, and patients receiving palliative treatment were excluded from the study. The median follow-up period of the cohort was 40 months (range: 11-66). Patients that did not develop a locoregional recurrence or distant metastases during the follow-up period and that had a follow-up time smaller than 24 months were also excluded from the study. The study has been approved by the institutional review board of Centre hospitalier de l'Université de Montréal. Detailed information about this patient cohort is provided in Supplementary Table 6.9.

TREATMENT DETAILS. Most patients (94 %) underwent concurrent platinum based chemotherapy and radiotherapy. All patients received an IMRT type radiation (sliding window IMRT or tomotherapy) consisting of 70 Gy of radiation in 33 fractions. Immobilisation device included a thermoplastic mask of the head and shoulder fixed to the treatment table.

FDG-PET/CT SCANS. All 65 eligible patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery STE, GE Healthcare) within 66 days before treatment (median: 12 days). For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199-3182) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 120-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a medianspan (axial mash) of 3 (range: 3-5). The median FDG-PET slice thickness resolution was 4 mm (range: 3.27-4) and the median in-plane resolution was 4×4 mm^2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 120-140) with a median exposure of 350 mAs (range: 5-350) was used. The median CT slice thickness resolution was 1.5 mm (range: 1.5-3.75) and the median in-plane resolution was 0.98×0.98 mm^2 (range: 0.98-1.37). All patients received their FDG-PET/CT scan dedicated to the head and neck area right before their planning CT scan, in the same position with the immobilisation device. Contours defining the gross tumour volume (GTV) and lymph nodes were drawn by an expert radiation oncologist on the planning CT scan. The contours were then propagated to the FDG-PET/CT scan reference frame using deformable registration with the software MIM® (MIM software Inc., Cleveland, OH) to ensure proper coverage.

Description of 3D radiomic features

In this thesis, please see Appendix A.

Computation of the radiomic signature

Original radiomic signature This section details how the original radiomic signature proposed by Aerts *et al.* [21] was computed from CT scans in the current work. In their original work, Aerts *et al.* [21] extracted the four features of the radiomic signature on CT images with voxels of size $1 \times 1 \times 3$ mm^3. In the current work, the CT images were thus first resampled to the same voxel size of $1 \times 1 \times 3$ mm^3 using cubic interpolation. The four features of the radiomic signature were then computed from the region of interest of the tumour as defined by the "GTV$_{\text{primary}}$ + GTV$_{\text{lymph nodes}}$" contours (ROI) as follows:

1. **Energy**

   Let $\mathbf{X}$ define the vector of Hounsfield Units (HUs) from CT scans for the N voxels of the ROI. The feature *energy* is then defined as:

$$energy = \sum_{i=1}^{N} X(i)^2$$

2. **Compactness**

   Let V be the volume in mm^3 and A be the surface area in mm^2 of the ROI. The feature *compactness* is then defined as:

$$compactness = \frac{V}{\sqrt{\pi} A^{2/3}}$$

3. **GLN**

   To compute the Gray-Level Nonuniformity (GLN) texture feature similarly to the work of Aerts *et al.* [21], the ROI was first quantized to a number of gray levels N_g^p different for each patient p. For CT scans, bins of 25 HUs were created using a lower limit of 0 HU to the intensity range of the bins such that all voxels within the ROI with $-1000 \leq \text{HU} < 25$ were assigned to gray-level 1, all voxels within the ROI with $25 \leq \text{HU} < 50$ were assigned to gray-level 2, etc.

Then, let $P_\delta(i,j)$ define the directional GLRLM of the quantized ROI, where δ denotes one of the 13 directions around a center voxel in 3D space. Similarly to what is described in the previous section, $P_\delta(i,j)$ represents the number of runs of gray-level i and of length j, and L_r represents the length of the longest run (of any gray-level) in the quantized ROI **for direction** δ. The GLN$_\delta$ for direction δ is then defined as:

$$GLN_\delta = \frac{\sum_{i=1}^{N_g^p} \left(\sum_{j=1}^{L_r} P_\delta(i,j) \right)^2}{\sum_{i=1}^{N_g^p} \sum_{j=1}^{L_r} P_\delta(i,j)}$$

Finally, the GLN texture feature is calculated as:

$$GLN = \frac{1}{13} \sum_{\delta=1}^{13} GLN_\delta$$

4. **GLN_HLH**

This texture feature is obtained by computing the GLN texture feature described above (feature 3) in the HLH sub-band of the first decomposition level of the 3D undecimated discrete wavelet transform performed using the wavelet basis function "Coiflet 1".

The HLH wavelet decomposition is traditionally obtained by applying a high-pass filter in the x-direction, a low-pass filter in the y-direction and a high-pass filter in the z-direction. For medical images, standard practice is to consider the reference coordinate system (RCS) of the DICOM protocol in order to unambiguously define the filter directions. Hence, for an axial CT volume of a patient in the DICOM RCS, the HLH wavelet decomposition would be obtained by applying a high-pass filter in the lef-right direction, a low-pass filter in the anterior-posterior direction and a high-pass filter in the inferior-superior direction.

However, in their original work, Aerts *et al.* [21] considered the

MATLAB® conventions to define the directions of the filters. As a result, the HLH wavelet decomposition was obtained by applying a high-pass filter in the anterior-posterior direction, a low-pass filter in the left-right direction and a high-pass filter in the inferior-superior direction of axial CT images. The same filter directions as defined by MATLAB® conventions were thus also used for CT images in the current work.

Practically speaking, in this work, the undecimated wavelet transform was applied on the original ROI using the function *swt2* and the wavelet basis function "Coiflet 1" of MATLAB® . To achieve a 3D decomposition, the 2D undecimated discrete wavelet transform obtained with the *swt2* function was successively applied for all image planes of the ROI in the x-, y- and z-directions of the RCS, and the corresponding wavelet coefficients of all image planes were averaged. The resulting wavelet coefficients of the ROI corresponding to the HLH sub-band were then uniformly quantized to the same number of gray levels N_g^p (for a given patient p) as obtained with the computation of the standard GLN texture feature described above (feature 3). Finally, the GLN_HLH texture feature was obtained by computing the same GLN texture feature described above (feature 3) to the quantized ROI of the HLH wavelet sub-band.

COMPLETE MODEL. In one instance in our work, we directly tested in the testing set (H&N3 and H&N4; $n = 106$) a Cox regression model constructed using the original coefficients and median hazard ratio trained in the Lung1 cohort of the original work of Aerts *et al.* [21]. This complete Cox regression model $\lambda(\mathbf{x}_i)$ was applied as follows in our work:

$$\begin{aligned}\lambda(\mathbf{x}_i) = &\\ &- 2.42e\text{-}11 \times \text{CT-Energy}\\ &+ 5.38e\text{-}03 \times \text{CT-Compactness}\\ &+ 1.47e\text{-}04 \times \text{CT-GLN}_{GLRLM}\\ &- 9.39e\text{-}06 \times \text{CT-GLN\_HLH}_{GLRLM}\end{aligned}$$

with a median hazard ratio of 0.1191567. The greater $\lambda(\mathbf{x}_i)$, the worst the chanches of survival are.

Revised version of the radiomic signature This section details three modifications applied to the original radiomic signature in order to obtain a revised version (other than the following modifications, the computation remained the same as described in the previous section):

1. **CT resampling**
   In order to obtain isotropic voxel size, the CT images were resampled to a voxel size of $1 \times 1 \times 1$ mm^3 using cubic interpolation.

2. **Compactness**
   The definition of *compactness* in the original radiomic signature uses $A^{2/3}$ in the denominator. This is most likely an error in the original paper of Aerts *et al.* [21], as $A^{3/2}$ is required to create a dimensionless feature. The feature *compactness* is thus hereby defined as:

$$compactness = \frac{V}{\sqrt{\pi} A^{3/2}}$$

3. **Computation of GLN and GLN_HLH**

   - Only one GLRLM is computed per CT volume by simultaneously adding up the 13 GLRLMs of all 3D directions. The GLRLM averaging technique used for the original radiomic signature basically results in an average of limited run-length measurements.
   - A normalized version of the GLN feature is used in this work. This feature is defined in Appendix A.4 of this document under the "GLRLM" heading. The original GLN feature as defined by Galloway [53] is not properly normalized and is thus dependent on the total number of runs in a given volume.

6.9 References

1. Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J. & Mills, G. B. Building a personalized medicine infrastructure at a major cancer center. *J. Clin. Oncol.* **31,** 1849–1857 (2013).
2. Renfro, L. A., An, M.-W. & Mandrekar, S. J. Precision oncology: a new era of cancer clinical trials. *Cancer Lett.* **387,** 121–126 (2016).
3. Lambin, P. *et al.* Rapid Learning health care in oncology – an approach towards decision support systems enabling customised radiotherapy. *Radiother. Oncol.* **109,** 159–164 (2013).

4. Shrager, J. & Tenenbaum, J. M. Rapid learning for precision oncology. *Nat. Rev. Clin. Oncol.* **11,** 109–118 (2014).

5. Weitzel, J. N., Blazer, K. R., MacDonald, D. J., Culver, J. O. & Offit, K. Genetics, genomics, and cancer risk assessment. *CA Cancer J. Clin.* **61,** 327–359 (2011).

6. Garraway, L. A., Verweij, J. & Ballman, K. V. Precision oncology: an overview. *J. Clin. Oncol.* **31,** 1803–1805 (2013).

7. El Naqa, I. Biomedical informatics and panomics for evidence-based radiation therapy. *WIREs Data Mining Knowl. Discov.* **4,** 327–340 (2014).

8. Ebrahim, A. *et al.* Multi-omic data integration enables discovery of hidden biological regularities. *Nat. Commun.* **7,** 13091 (2016).

9. Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer* **108,** 479–485 (2013).

10. Heppner, G. H. & Miller, B. E. Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metastasis Rev.* **2,** 5–23 (1983).

11. Fidler, I. J. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res.* **50,** 6130–6138 (1990).

12. Yokota, J. Tumor progression and metastasis. *Carcinogenesis* **21,** 497–503 (2000).

13. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).

14. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278,** 563–577 (2016).

15. El Naqa, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* **42,** 1162–1171 (2009).

16. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin. Radiol.* **65,** 517–521 (2010).

17. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48,** 441–446 (2012).

18. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30,** 1234–1248 (2012).

19. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25,** 675–680 (2007).

20. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105,** 5213–5218 (2008).

21. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

22. Hatt, M. *et al.* Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur. J. Nucl. Med. Mol. Imaging,* 1–15 (2016).

23. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. *Phys. Med. Biol.* **61,** R150–R166 (2016).

24. Lambin, P. *et al.* Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat. Rev. Clin. Oncol.* **10,** 27–40 (2013).

25. Ferlito, A., Shaha, A. R., Silver, C. E., Rinaldo, A. & Mondin, V. Incidence and sites of distant metastases from head and neck cancer. *ORL* **63,** 202–207 (2001).

26. Baxi, S. S. *et al.* Causes of death in long-term survivors of head and neck cancer. *Cancer* **120,** 1507–1513 (2014).

27. Wong, A. J., Kanwar, A., Mohamed, A. S. & Fuller, C. D. Radiomics in head and neck cancer: from exploration to application. *Transl. Cancer Res.* **5,** 371–382 (2016).

28. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57,** 289–300 (1995).

29. Vallières, M., Freeman, C. R., Skamene, S. R. & El Naqa, I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys. Med. Biol.* **60,** 5471–5496 (2015).

30. Schiller, T. W., Chen, Y., El Naqa, I. & Deasy, J. O. Modeling radiation-induced lung injury risk with an ensemble of support vector machines. *Neurocomputing* **73,** 1861–1867 (2010).

31. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44,** 837–845 (1988).

32. Leijenaar, R. T. H. *et al.* External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol.* **54,** 1423–1429 (2015).

33. Harrell, F. E. J., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15,** 361–387 (1996).

34. Fakhry, C. *et al.* Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *J. Natl. Cancer Inst.* **100,** 261–269 (2008).

35. Ang, K. K. *et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* **363,** 24–35 (2010).

36. Cheng, N.-M. *et al.* Zone-size nonuniformity of 18F-FDG PET regional textural features predicts survival in patients with oropharyngeal cancer. *Eur. J. Nucl. Med. Mol. Imaging.* **42,** 419–428 (2014).

37. Vakkila, J. & Lotze, M. T. Inflammation and necrosis promote tumour growth. *Nat. Rev. Immunol.* **4,** 641–648 (2004).

38. Proskuryakov, S. Y. & Gabai, V. L. Mechanisms of tumor cell necrosis. *Curr. Pharm. Des.* **16,** 56–68 (2010).

39. Ahn, S.-H. *et al.* Necrotic cells influence migration and invasion of glioblastoma via NF-kB/AP-1-mediated IL-8 regulation. *Sci. Rep.* **6,** 24552 (2016).

40. Breiman, L. Random forests. *Machine Learning* **45,** 5–32 (2001).

41. Parmar, C. *et al.* Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci. Rep.* **5,** 11044 (2015).

42. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine learning methods for quantitative radiomic biomarkers. *Sci. Rep.* **5,** 13087 (2015).

43. Parmar, C. *et al.* Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front. Oncol.* **5,** 272 (2015).

44. Tang, C. *et al.* Validation that metabolic tumor volume predicts outcome in head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **83,** 1514–1520 (2012).

45. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56,** 38–44 (2015).

46. Nyflot, M. J. *et al.* Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J. Med. Imaging* **2,** 041002 (2015).

47. Zhao, B. *et al.* Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci. Rep.* **6,** 23428 (2016).

48. Ioannidis, J. P. A. How to make more published research true. *PLoS Med.* **11,** e1001747 (2014).

49. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26,** 1045–1057 (2013).

50. Van Velden, F. H. P. *et al.* Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur. J. Nucl. Med. Mol. Imaging* **38,** 1636–1647 (2011).

51. Rahmim, A. *et al.* A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction. *Phys. Med. Biol.* **61,** 227–242 (2016).

52. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).

53. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).

54. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

55. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).

56. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

57. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).

58. El Naqa, I. *et al.* Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys. Med. Biol.* **51,** 5719–5735 (2006).

59. Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334,** 1518–1524 (2011).

60. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560 (1997).

61. Sahiner, B., Chan, H.-P. & Hadjiiski, L. Classifier performance prediction for computer-aided diagnosis using a limited dataset. *Med. Phys.* **35,** 1559–1570 (2008).

62. Bonner, J. A. *et al.* Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **354,** 567–578 (2006).

Chapter 7

Conclusions

7.1 Summary and novelty of work 257
7.2 The fundamental necessity of standardization, transparency and online sharing in radiomics 264
7.3 Other contributions 265
7.4 References 266

7.1 Summary and novelty of work

“Radiomics” refers to the characterization of tumour phenotypes via the extraction of high-dimensional mineable data from all types of medical images and whose subsequent analysis aims at supporting clinical decision-making. A major aim of this work has been to develop new analysis methods in the field of radiomics, with the ultimate goal of creating highly predictive and generalizable radiomic-based multivariable models to be used in routine clinical practice to assist physicians in providing treatments more personalized to each patient. Overall, we have showed in this thesis that radiomics analyses are enabling factors towards precision medicine. As described in section 1.2.4, the field of radiomics opens the door to many interesting development possibilities in oncology research, notably in terms of medical imaging acquisition optimization, of radiomics modeling via improved feature extraction and machine learning, of the evaluation of tumour aggressiveness and underlying tumour biology, and of the personalization of cancer treatments via radiomic-based prognosis assessments. The work presented in this thesis provided strong progress in every of these different aspects of the field. We hope the methods developed here could one day be implemented in the clinical environment to better help patients afflicted by cancer to overcome this deadly disease. In this section, we provide a summary of the final results obtained in this work, as well as an highlight of the scientific novelty of each of the four manuscripts.

1. Development of a radiomic model for the early prediction of lung metastases

In this study, the main objectives were: I) Develop a robust methodology for the construction of multivariable radiomic-based prediction models that takes into account texture optimization; and II) Create fused FDG-PET/MR images with better textural predictive properties than the separate FDG-PET and MR images.

First, one of the main findings of this study is that the computation of textures with different extraction parameters (e.g., scale, quantization schemes, etc.) has a significant impact on prediction performance, an effect demonstrated here in the specific case of the prediction of lung metastases in soft-tissue sarcomas. The isotropic resolution at which textures are extracted is the parameter that has the most influence on texture definition. In general, different texture features will better represent the underlying tumour biology using different extraction parameters, and the optimal set of parameters to use is application-specific and will depend on many factors such as the clinical endpoint studied and the imaging modalities employed. To the best of our knowledge, no study in the literature has yet employed different texture extraction parameters to enhance the predictive properties of textures. Currently, the power of texture analysis for tumour outcome prediction may thus not be fully exploited by the radiomics community, and we recommend to always perform a similar texture optimization process for a given clinical application. From our experience, it seems clear for example that a lower number of gray-levels helps in better characterizing tumour sub-regions using GLSZM features (e.g., 8 or 16), whereas GLCM feature may be better modeled with a higher number of gray-levels (e.g., 32 or 64). In general, a radiomic user has to be careful of not using a too high number of gray-levels, as this could result in a higher influence of noise on the extracted textures.

The drawback of texture optimization is the exponential increase in size of the initial radiomic feature sets used in the subsequent machine learning processes. To overcome this issue, a novel feature set reduction method was developed in order to create reduced feature sets with only one variant of a given texture (i.e., computed using one specific set of extraction parameters).

The data mining process in this multivariable modeling step essentially allows to find the different texture variants that have the best predictive properties and the less redundancy with the other textures chosen to be part of the reduced feature set, thus also taking into account the intercorrelation between the different features.

Following feature set reduction, the selection and estimation of the prediction performance of different multivariable radiomic-based models using imbalance-adjusted bootstrapping accomplish two goals: I) Finding the set of features with potentially the highest predictive performance and generalizability to unseen data; and II) Construct models with a multivariable model response balanced between the sensitivity and specificity of predictions. For example, the final radiomic model identified in this study would not only have high predictive properties as estimated in bootstrapping evaluations with an AUC of 0.984 ± 0.002, but the model would also provide a balance between the prediction of true positive instances and true negative instances despite the data imbalance, with a sensitivity of prediction of 0.955 ± 0.006 and a specificity of 0.926 ± 0.004. Moreover, permutation tests demonstrated that the effect of the model that we observed in this patient cohort is estimated to be highly significant (AUC: $\widehat{p} = 0.004$) and is thus most likely present in the general soft-tissue sarcoma population. We also estimated the variation of the model response when the edema present in the vicinity of soft-tissue sarcomas is included in the definition of the region-of-interest.

In this work, we also developed a methodology for the fusion of FDG-PET and MR images that proved useful in creating new composite textures with valuable predictive properties. The final multivariable model identified in this study was created from fused FDG-PET/MR images, and it was estimated that this model possesses superior predictive properties than the models constructed from the separate imaging modalities. This type of image processing step (i.e., the fusion of different image modalities) prior to texture analysis is one of the many extraction parameters that could lead to a better characterization of intratumoural heterogeneity, and it could become an integral part of texture optimization in future studies.

Overall, the multivariable modeling steps leading to the identification of the optimal combination of radiomic features is far from being trivial and should be continuously improved. In future work, we intend to integrate a false-positive avoidance methodology recently developed by our group to our multivariable modeling workflow. The application of this methodology would help to reduce overfitting and would allow to improve the consistency in model performance obtained across the training, validation and testing sets. Other types of features will also be implemented to obtain different insights about the underlying tumour biology, including wavelet and texture map features. Furthermore, features highly correlated with tumour volume and clinical information will be discarded at the start of radiomics analyses. Finally, we will develop methods to estimate the robustness of features against different contouring and noise perturbations.

2. A strategy for treatment personalization

In this study, the main objectives were: I) Investigate the practical feasibility and clinical utility of acquiring four different types of biological images (FDG-PET, FMISO-PET, DW-MRI, DCE-MRI) at three different time points (pre-, mid-, post-radiotherapy) in the course of treatment management; II) Validate the predictive properties of the radiomic model developed in Chapter 3 (manuscript 1); and II) Verify the feasibility of double nested dose boosting to hypermetabolic and hypoxic tumour sub-regions inside the GTV in radiotherapy planning.

First, we experienced a difficulty in acquiring the number of imaging scans specified in our prospective protocol for all patients enrolled in the study. Many patients could not tolerate, for example, experiencing additional scanning visits and longer MRI anatomical scans. With the advent of MRI-linac technologies, we now hypothesize that the power of biological images for treatment response monitoring will be better exploited in the future if imaging acquisition is an integral part of the radiotherapy treatment. Notably, we observed that FMISO-PET did not bring sufficient complementarity value in comparison to FDG-PET to justify its use in soft-tissue sarcoma treatment managment. On the other hand, DCE-MRI methods have more potential to provide useful information about intratumoural evolution. However, the progression of simple percentile-based metrics could not yield a clear assessment of radiation treatment response. In future work, we will explore how more complex imaging metrics such as textures could better monitor the evolution of the microenvironment of soft-tissue sarcomas during radiotherapy.

Furthermore, we were able to validate the predictive properties of the radiomic model developed in Chapter 3. This model reached an AUC of 0.71, a sensitivity of 0.75, a specificity of 0.85 and an accuracy of 0.82 for the prediction of lung metastases in this independent patient cohort. The multivariable modeling methodology developed in Chapter 3 thus has certain potential, but further improvements are still needed to improve the prediction performance to clinical requirements. False-positive avoidance methods and the use of a standardized and exhaustive list of features developed in current ongoing work [1] should improve the prediction performance. Overall, the accurate prediction at the time of diagnosis of soft-tissue sarcoma patients more likely to develop lung metastases could allow to identify the subset of patients that would benefit the most from personalized radiotherapy with dose escalation to different GTV sub-volumes including hypoxic tumour sub-regions.

Finally, we verified the feasibility of dose painting using a prescription of 50 Gy to the PTV ($\mathrm{PTV}_{50\,\mathrm{Gy}}$) along with boost doses of 60 Gy to the FDG hypermetabolic GTV ($\mathrm{GTV}_{60\,\mathrm{Gy}}$) and of 65 Gy to the low-perfusion DCE-MRI hypoxic GTV contained within the $\mathrm{GTV}_{60\,\mathrm{Gy}}$ ($\mathrm{GTV}_{65\,\mathrm{Gy}}$) using volumetric arc therapy (VMAT). Despite the complexity of the multiple targets, adequate tumour coverage was achieved, with a homogeneity index of 1.09 for the difference volume of $\mathrm{GTV}_{60\,\mathrm{Gy}}$ minus $\mathrm{GTV}_{65\,\mathrm{Gy}}$, and 1.06 for $\mathrm{GTV}_{65\,\mathrm{Gy}}$. In future work, we intend to perform a multi-centric prospective study in order to validate these findings prior to the conduction of a formal clinical trial. We envision the use of dose painting as a useful strategy to improve tumour outcomes in soft-tissue sarcomas.

3. Enhancement of radiomic models via the optimization of imaging acquisition protocols

In this study, the main objective was to enhance the predictive properties of a texture-based model by optimizing FDG-PET and MR image acquisition protocols. A proof of concept for the prediction of lung metastases in soft-tissue sarcomas was carried out using computerized simulations of PET and MR image acquisitions.

First, we investigated how three different textures extracted from FDG-PET ($HGZE_{\mathrm{GLSZM}}$), T_1-weighted (ZSV_{GLSZM}) and T_2-weighted ($LRLGE_{\mathrm{GLRLM}}$) simulated images varied using different numbers of span, repetition times (TR) and echo times (TE) in the image acquisition processes, respectively. Overall, we observed that an increasing number of span generally resulted in an increase of the $HGZE_{\mathrm{GLSZM}}$ texture due to an increase in image smoothing, and that an increasing TR and TE generally resulted in an increase of the ZSV_{GLSZM} and the $LRLGE_{\mathrm{GLRLM}}$ textures due to changing image contrasts, respectively. In comparison to textures extracted from simulated images acquired with standard clinical parameters, we observed percentage variations for some patients as high as 15 %, 100 % and 50 % for different sets of acquisition parameters used to obtain the $HGZE_{\text{GLSZM}}$, ZSV_{GLSZM} and $LRLGE_{\text{GLRLM}}$ textures, respectively. These results confirm various assessments in the literature stating that different textures may vary when extracted from images acquired with different sets of acquisition parameters.

However, all studies investigating texture variations under different imaging settings have a common denominator: they aim at identifying the texture features that could be stable and that are presumably able to conserve predictive properties under varying imaging conditions. In this study, we hypothesized that it is also fundamental to identify the settings that would yield optimal use of texture features for a given clinical problem (similarly to texture extraction parameters in manuscript 1 → Chapter 3). To the best of our knowledge, our study was the first to explore the potential of varying image acquisition parameters to optimize the performance of texture features and consequently enhance texture-based predictive models. By combining the textures enumerated above into a multivariable model (constructed using the methods developed in Chapter 3), we demonstrated the feasibility of enhancing a texture-based predictive model by optimizing targeted image acquisition parameters. The model constructed with textures extracted from simulated images acquired with a standard clinical set of acquisition parameters reached an average AUC of 0.84 ± 0.01 in bootstrap testing experiments. In comparison, the model performance significantly increased using an optimal set of image acquisition parameters (p = 0.04), with an average AUC of 0.89 ± 0.01.

Overall, our work was only a first step towards the enhancement of texture-based prediction models via the optimization of image acquisition parameters. This part of the radiomics analysis workflow could become a sub-field in itself, as there are multiple avenues to explore. In future work, we notably intend to improve the realism of heterogeneous tumor models and image simulations in order to improve the similarities between textures extracted from clinical and simulated images. We will also increase patient dataset size, test the proof of concept in other cancer types (e.g., head-and-neck) and imaging modalities (e.g., CT) with other cancer-specific prediction models, investigate how a wider range of radiomic features vary in different acquisition settings, perform full construction of prediction models for every set of simulated image acquired with different parameters, investigate how the optimization of texture-based predictive models may vary on different scanners from different vendors (e.g., GE, Siemens, Phillips, etc.), and include independent testing cohorts in the analysis. In fact, the simulation methods developed in this work could provide an optimal framework to address such research questions, as clinical scanner settings with real patients do not provide enough flexibility. Ultimately, we envision that specific "radiomics acquisition protocols" optimized to generate superior texture measurements for a given clinical problem will be developed as complementary protocols to the clinical ones currently used.

4. Integration of radiomic models with clinical prognostic factors

In this study, the main objectives were: I) Validate the multivariable modeling methodology developed in Chapter 3 using independent external datasets; and II) Develop a complementary methodology for integrating radiomics data with clinical information for better prediction performance.

Overall in this work, we provided a rigorous methodology for integrating prognostic factors of different categories into risk-assessment models – in this case radiomics (continuous inputs) and clinical data (categorical inputs). Our methodology is first based on a fast mining of radiomic variables (including textures extracted with multiple parameters) using logistic regression and the multivariable methods developed in Chapter 3. As a second step, we integrate clinical information with radiomic variables into a random forest algorithm. The models we developed with this strategy overall performed better than when using radiomics or clinical information alone. We believe this strategy could be generalized to integrate multiple other types of panomics data: different -omics information could be mined separately with other existing tools [2–4] to then be combined altogether with random forests. In fact, one major advantage of this machine learning algorithm is its higher tolerance to overfitting compared to other more conventional algorithms such as logistic regression [5]; as long as the decision-trees of the random forest undergo a "decorrelation" process, a high number of input variables can be used. In future work, we intend to integrate a larger number of patient data information into random forests composed of uncorrelated decision-trees.

The final prediction models developed in this work reached a concordance-index (CI) of 0.67 and 0.88 in independent testing sets for the risk assessment of locoregional recurrences and distant metastases, respectively. These models were also constructed using robust imbalance-adjustment methods to take into account the high imbalance in the proportion of occurrence and non-occurence of outcome events in head-and-neck cancer. In comparison to our results, the popular study of Aerts *et al.* [6] reported a highest CI of 0.69

for the risk assessment of overall survival in head-and-neck cancer (although these results are more likely due to the high correlation of the radiomic signature with tumour volume), and a study by Coroller *et al.* [7] reported a different radiomic signature possessing a “strong” power for predicting distant metastases in lung cancer with a CI of 0.61. In our work, the prognostic power of the distant metastasis model is not just an improvement over other radiomics studies, but it practically *enables* prognostic capability in a clinical environment. For the locoregional recurrence model, the prognostic power is comparable to other studies and still shows that radiomics analysis has a role to play to decode that specific tumour phenotype. However, improvements are necessary, and we intend in future work to integrate dose metrics into random forests in order to increase prognostic power for that tumour outcome. Furthermore, to the best of our knowledge, no other radiomics studies in the literature present full prediction models able to stratify patients into multiple risks-groups and for multiple specific outcomes. Our models can stratify patients into two groups for the risk assessment of locoregional recurrences (low, high) in head-and-neck cancer, and into three groups for distant metastases (low, medium, high). This could have a major impact on the design of new clinical trials aiming at a better personalization of chemoradiation treatments in head-and-neck cancer, and one can envision different radiation and chemotherapy regimens being delivered to patients based on different assessments of risk for a particular tumour outcome.

Overall, we have showed with this last study that radiomics are enabling factors towards precision medicine. In the future, the integration of radiomics with other panomics data into comprehensive multivariable models should leverage the prognostic assessment of cancer risks.

7.2 The fundamental necessity of standardization, transparency and online sharing in radiomics

As previoulsy mentioned, the workflow of radiomics analysis is complex and involves many different computational parameters. For as faster translation of knowledge to the scientific community, for a faster translation of radiomics into the clinical environment, but above all for the sake of cancer patients, it is crucial that radiomics studies be as transparent as possible. Furthermore, the sharing of imaging data and programming code in online repositories such as GitHub and The Cancer Imaging Archive (TCIA) [8] unquestionably facilitates the reproducibility of radiomics studies. Assuredly, standardization and full transparency on data and methods is the key for the progression of the radiomics field [9].

Unfortunately, it frequently occurs in the current literature that crucial information about radiomic analyses is missing. Examples include model coefficients, tuning parameters of machine learning algorithms, image preprocessing operations, or radiomic computation details such as texture matrix construction and quantization range when using a fixed bin width algorithm. This adversly affects the reproducibility of radiomic studies and the transfer of scientific knowledge. I take advantage of this tribune to make a strong case for better transparency and sharing practices in the radiomics community, notably in terms of imaging data and programming code. The time we sometimes loose in chasing data and complete methodological details directly affects cancer patients and the rapidity with which we could provide anticancer strategies for them. Although this is not the case for the majority, there still exists too many researchers that unfortunately appear to value their work and their monopole of expertise more than the support it could bring to patients afflicted by cancer. Let us all remember why we are doing oncology research every time we take a research action.

7.3 Other contributions

The methods developed in this thesis were applied in other published works in which I am a co-author. Furthermore, I always strive to share imaging data and programming code online for other scientists to be able to easily reproduce my work, as I strongly believe that full transparency in scientific publications allows for a faster transfer of knowledge. This section briefly describe these other contributions to science.

Scientific publications

- I am a third author (equal contribution with the second author) on a study by Hatt *et al.* [10]. Various texture analysis methods developed in Chapter 3 were further investigated in this study to show that textures provide complementary information to tumour volumes above 10 cm^3 if proper textural analysis is carried out (e.g., full 3D extraction, lower number of gray-levels, etc.).

- I am a third author in this important and vast consortium work by Zwanenburg *et al.* [1] dedicated to the standardization of radiomics methods (notably the texture analysis methods developed in Chapter 3). To this date, about 50 researchers from 25 cancer institutions in the world participate to this consortium.
- I am a second author (equal contribution with the first author) on a study by Zhou *et al.* [11]. The methods developed in Chapter 3 and Chapter 6 were notably used to show how textural analyses of MR imaging data can be used to predict molecular profiles and tumour progression in lower-grade gliomas with high accuracy.

Online sharing

- Imaging data for the studies performed in Chapter 3 and Chapter 5 can be found here: LINK.
- Imaging data for the study performed in Chapter 6 can be found here: LINK.
- Programming code used in this work, including single organized scripts allowing to run all the experiments of the studies performed in Chapter 3 and Chapter 6, can be found here: LINK.
- Imaging data for the study of Zhou *et al.* [11] can be found here: LINK.

7.4 References

1. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative. *arXiv preprint*. arXiv: 1612.07003 (2016).
2. Nibbe, R. K., Koyutürk, M. & Chance, M. R. An Integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Computational Biology* **6,** e1000639 (2010).
3. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).
4. Yu, K.-H. & Snyder, M. Omics profiling in precision oncology. *Mol. Cell Proteomics* **15,** 2525–2536 (2016).
5. Breiman, L. Random forests. *Machine Learning* **45,** 5–32 (2001).
6. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

7. Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114,** 345–350 (2015).
8. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26,** 1045–1057 (2013).
9. Lambin, P. *et al.* Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat. Rev. Clin. Oncol.* **10,** 27–40 (2013).
10. Hatt, M. *et al.* 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J. Nucl. Med.* **56,** 38–44 (2015).
11. Zhou, H. *et al.* MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro. Oncol.* **19,** 862–870 (2017).

Appendix A

List of 3D radiomic features

A.1 Morphological (shape) features

The morphological features (5) computed from a ROI map $\mathbf{R}$ are defined as:

- **Volume**: Number of voxels in the tumour region multiplied by the dimension of voxels.
- **Size**: Maximum diameter of the tumour region.
- **Solidity**: Ratio of the number of voxels in the tumour region to the number of voxels in the 3D convex hull of the tumour region (smallest polyhedron containing the tumour region).
- **Eccentricity**: The ellipsoid that best fits the tumour region is first computed using the framework of Li & Griffiths [1]. The eccentricity is then given by $[1 - a \times b/c^2]^{1/2}$, where c is the longest semi-principal axes of the ellipsoid, and a and b are the second and third longest semi-principal axes of the ellipsoid.
- **Compactness**:
$$compactness = \frac{V}{\sqrt{\pi} A^{3/2}}$$
Where V denotes the volume and A the surface area of the ROI map.

Morphological features were extracted from T1 scans in Chapter 3 and Chapter 5. In Chapter 6, these features were extracted from CT scans. Morphological features are used in all the different feature sets.

A.2 Histogram-based (intensity) features

Let $\mathbf{P}$ define the first-order histogram of a ROI imaging volume V_{R} with isotropic voxel size. Each entry $P(i)$ of $\mathbf{P}$ represents the number of voxels with gray level i or within a pre-defined bin width, and N_g represents the number of gray-level bins set for $\mathbf{P}$. The i^{th} entry of the normalized histogram is then defined as:

$$p(i) = \frac{P(i)}{\sum_{i=1}^{N_g} P(i)}.$$

The first-order statistics features (10) are then defined as:

- **Variance**:
$$\sigma^2 = \sum_{i=1}^{N_g} (i - \mu)^2 \, p(i)$$
- **Skewness**:
$$s = \sigma^{-3} \sum_{i=1}^{N_g} (i - \mu)^3 \, p(i)$$
- **Kurtosis**:
$$k = \sigma^{-4} \sum_{i=1}^{N_g} \left[(i - \mu)^4 \, p(i) \right] - 3$$
- **SUVmax**: Maximum SUV of the tumour region. Extracted from FDG-PET scans only.
- **SUVpeak**: Average of the voxel with maximum SUV within the tumour region and its 26 connected neighbours. Extracted from FDG-PET scans only.
- **SUVmean**: Average SUV value of the tumour region. Extracted from FDG-PET scans only.
- **AUC-CSH**: Area under the curve of the cumulative SUV-volume histogram describing the percentage of total tumour volume above a percentage threshold of maximum SUV, as defined by van Velden *et al.* [2]. Extracted from FDG-PET scans only.
- **TLG**: Total lesion glycolysis. Defined as SUVmean × total volume of the tumour region. Extracted from FDG-PET scans only.
- **Percent Inactive**: Percentage of the tumour region that is inactive. A threshold of $0.05 \times (\mathrm{SUVmax})^2$ was used in Chapter 3 and Chapter 5, and a threshold of $0.01 \times (\mathrm{SUVmax})^2$ was used in Chapter 6. The thresholding process was then followed by closing and opening morphological operations to differentiate active and inactive tumour regions on FDG-PET scans. Extracted from FDG-PET scans only.
- **gETU**: Generalized effective total uptake, with parameter $a = 0.25$ as defined by Rahmim *et al.* [3]. Extracted from FDG-PET scans only.

A.3 GLCM features

Let $\mathbf{P}$ define the GLCM of a quantized ROI imaging volume V_{Q} with isotropic voxel size. Each entry $P(i, j)$ of $\mathbf{P}$ represents the number of times voxels of gray level i are neighbours with voxels of gray level j in V_{Q}. Also, N_g represents the pre-defined number of quantized gray levels set in V_{Q}. Only one GLCM of size $N_g \times N_g$ is computed per volume V_{Q} by simultaneously adding up the frequency of co-occurences of all voxels with their 26-connected neighbours in 3D space, with all voxels (*including* the peripheral region) considered once as a center voxel (as defined by Haralick *et al.* [4], thus always using $d = 1$). To account for discretization length differences, neighbours at a distance of $\sqrt{3}$ voxels around a center voxel increment the GLCM by a value of $\sqrt{3}$, neighbours at a distance of $\sqrt{2}$ voxels around a center voxel increment the GLCM by a value of $\sqrt{2}$, and neighbours at a distance of 1 voxel around a center voxel increment the GLCM by a value of 1. The entry (i, j) of the normalized GLCM is then defined as:

$$p(i, j) = \frac{P(i, j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j)}.$$

The following quantities are also defined:

$$\mu_i = \sum_{i=1}^{N_g} i \sum_{j=1}^{N_g} p(i, j), \qquad \mu_j = \sum_{j=1}^{N_g} j \sum_{i=1}^{N_g} p(i, j),$$

$$\sigma_i = \sum_{i=1}^{N_g} (i - \mu_i)^2 \sum_{j=1}^{N_g} p(i, j), \qquad \sigma_j = \sum_{j=1}^{N_g} (j - \mu_j)^2 \sum_{i=1}^{N_g} p(i, j).$$

The GLCM texture features (9) are then defined as:

- **Energy** [4]:

$$energy = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} [p(i, j)]^2$$

- **Contrast** [4]:

$$contrast = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - j)^2 \, p(i, j)$$

- **Correlation** (adapted from [4]):

$$correlation = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{(i - \mu_i)\,(j - \mu_j)\, p(i, j)}{\sigma_i \, \sigma_j}$$

- **Homogeneity** (adapted from [4]):

$$homogeneity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+|i-j|}$$

- **Variance** (adapted from [4]):

$$variance = \frac{1}{N_g \times N_g}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left[(i-\mu_i)^2\, p(i,j) + (j-\mu_j)^2\, p(i,j)\right]$$

- **Sum Average** (adapted from [4]):

$$sum\ average = \frac{1}{N_g \times N_g}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left[i\, p(i,j) + j\, p(i,j)\right]$$

- **Entropy** [4]:

$$entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2\big(p(i,j)\big)$$

- **Dissimilarity** [5]:

$$dissimilarity = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |i-j|\, p(i,j)$$

- **Autocorrelation** [6]:

$$autocorrelation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ij\, p(i,j)$$

A.4 GLRLM features

Let $\mathbf{P}$ define the GLRLM of a quantized ROI imaging volume V_Q with isotropic voxel size. Each entry $P(i,j)$ of $\mathbf{P}$ represents the number of runs of gray level i and of length j in V_Q. Also, N_g represents the pre-defined number of quantized gray levels set in V_Q, and L_r represents the length of the longest run (of any gray level) in V_Q. Only one GLRLM of size $N_g \times L_r$ is computed per volume V_Q by simultaneously adding up all possible longest run-lengths in the 13 directions of 3D space (one voxel can be part of multiple runs in different directions, but can be part of only one run in a given direction). A MATLAB® toolbox created by Wei [7] computes GLRLMs from 2D images, and it can be used to facilitate the computation of GLRLMs from 3D imaging volumes. To account for discretization length differences, runs constructed from voxels separated by a distance of $\sqrt{3}$ increment the GLRLM by a value of $\sqrt{3}$, runs constructed from voxels separated by a distance of $\sqrt{2}$ increment the GLRLM by a value of $\sqrt{2}$, and runs constructed from voxels separated by a distance of 1 increment the GLRLM by a value of 1. The entry (i, j) of the of the normalized GLRLM is then defined as:

$$p(i,j) = \frac{P(i,j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{L_r} P(i,j)}.$$

The following quantities are also defined:

$$\mu_i = \sum_{i=1}^{N_g} i \sum_{j=1}^{L_r} p(i,j), \qquad \mu_j = \sum_{j=1}^{L_r} j \sum_{i=1}^{N_g} p(i,j).$$

The GLRLM texture features (13) are then defined as:

- **Short Run Emphasis (SRE)** [8]:

$$SRE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \frac{p(i,j)}{j^2}$$

- **Long Run Emphasis (LRE)** [8]:

$$LRE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} j^2 \, p(i,j)$$

- **Gray-Level Nonuniformity (GLN)** (adapted from [8]):

$$GLN = \sum_{i=1}^{N_g} \left(\sum_{j=1}^{L_r} p(i,j) \right)^2$$

- **Run-Length Nonuniformity (RLN)** (adapted from [8]):

$$RLN = \sum_{j=1}^{L_r} \left(\sum_{i=1}^{N_g} p(i,j) \right)^2$$

- **Run Percentage (RP)** (adapted from [8]):

$$RP = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{L_r} p(i,j)}{\sum_{j=1}^{L_r} j \sum_{i=1}^{N_g} p(i,j)}$$

- **Low Gray-Level Run Emphasis (LGRE)** [9]:

$$LGRE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \frac{p(i,j)}{i^2}$$

- **High Gray-Level Run Emphasis (HGRE)** [9]:

$$HGRE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} i^2 \, p(i,j)$$

- **Short Run Low Gray-Level Emphasis (SRLGE)** [10]:

$$SRLGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \frac{p(i,j)}{i^2 j^2}$$

- **Short Run High Gray-Level Emphasis (SRHGE)** [10]:

$$SRHGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \frac{i^2 \, p(i,j)}{j^2}$$

- **Long Run Low Gray-Level Emphasis (LRLGE)** [10]:

$$LRLGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \frac{j^2 \, p(i,j)}{i^2}$$

- **Long Run High Gray-Level Emphasis (LRHGE)** [10]:

$$LRHGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} i^2 j^2 \, p(i,j)$$

- **Gray-Level Variance (GLV)** (adapted from [11]):

$$GLV = \frac{1}{N_g \times L_r} \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \left(i \, p(i,j) - \mu_i\right)^2$$

- **Run-Length Variance (RLV)** (adapted from [11]):

$$RLV = \frac{1}{N_g \times L_r} \sum_{i=1}^{N_g} \sum_{j=1}^{L_r} \left(j \, p(i,j) - \mu_j\right)^2$$

A.5 GLSZM features

Let **P** define the GLSZM of a quantized ROI imaging volume V_{Q} with isotropic voxel size. Each entry $P(i,j)$ of **P** represents the number of 3D zones of gray levels i and of size j in V_{Q}. Also, N_g represents the pre-defined number of quantized gray levels set in V_{Q}, and L_z represents the size of the largest zone (of any gray level) in V_{Q}. One GLSZM of size $N_g \times L_z$ is computed per volume V_{Q} by adding up all possible largest zone sizes, with zones constructed from 26-connected neighbours of the same gray level in 3D space (one voxel can be part of only one zone). The entry (i,j) of the normalized GLSZM is then defined as:

$$p(i,j) = \frac{P(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{L_z} P(i,j)}.$$

The following quantities are also defined:

$$\mu_i = \sum_{i=1}^{N_g} i \sum_{j=1}^{L_z} p(i,j), \qquad \mu_j = \sum_{j=1}^{L_z} j \sum_{i=1}^{N_g} p(i,j).$$

The GLSZM texture features (13) are then defined as:

- **Small Zone Emphasis (SZE)** [8, 11]:

$$SZE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z} \frac{p(i,j)}{j^2}$$

- **Large Zone Emphasis (LZE)** [8, 11]:

$$LZE = \sum_{i=1}^{N_g}\sum_{j=1}^{L_z} j^2\, p(i,j)$$

- **Gray-Level Nonuniformity (GLN)** (adapted from [8, 11]):

$$GLN = \sum_{i=1}^{N_g}\left(\sum_{j=1}^{L_z} p(i,j)\right)^2$$

- **Zone-Size Nonuniformity (ZSN)** (adapted from [8, 11]):

$$ZSN = \sum_{j=1}^{L_z}\left(\sum_{i=1}^{N_g} p(i,j)\right)^2$$

- **Zone Percentage (RP)** (adapted from [8, 11]):

$$ZP = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{L_z} p(i,j)}{\sum_{j=1}^{L_z} j \sum_{i=1}^{N_g} p(i,j)}$$

- **Low Gray-Level Zone Emphasis (LGZE)** [9, 11]:

$$LGZE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} \frac{p(i,j)}{i^2}$$

- **High Gray-Level Zone Emphasis (HGZE)** [9, 11]:

$$HGZE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} i^2\, p(i,j)$$

- **Small Zone Low Gray-Level Emphasis (SZLGE)** [10, 11]:

$$SZLGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} \frac{p(i,j)}{i^2 j^2}$$

- **Small Zone High Gray-Level Emphasis (SZHGE)** [10, 11]:

$$SZHGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} \frac{i^2\, p(i,j)}{j^2}$$

- **Large Zone Low Gray-Level Emphasis (LZLGE)** [10, 11]:

$$LZLGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} \frac{j^2\, p(i,j)}{i^2}$$

- **Large Zone High Gray-Level Emphasis (LZHGE)** [10, 11]:

$$LZHGE = \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} i^2 j^2\, p(i,j)$$

- **Gray-Level Variance (GLV)** (adapted from [11]):

$$GLV = \frac{1}{N_g \times L_z} \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} (i\, p(i,j) - \mu_i)^2$$

- **Zone-Size Variance (ZSV)** (adapted from [11]):

$$ZSV = \frac{1}{N_g \times L_z} \sum_{i=1}^{N_g} \sum_{j=1}^{L_z} \left(j\, p(i,j) - \mu_j\right)^2$$

A.6 NGTDM features

Let $\mathbf{P}$ define the NGTDM of a quantized volume V_Q with isotropic voxel size. Each entry $P(i,j)$ of $\mathbf{P}$ represents the summation of the gray-level differences between all voxels with gray level i and the average gray level of their 26-connected neighbours in 3D space. N_g represents the pre-defined number of quantized gray levels set in V_Q, and $(N_g)_{eff}$ is the effective number of gray levels in V_Q, with $(N_g)_{eff} < N_g$ (let the vector of gray level values in V_Q be denoted as $\mathbf{g} = g(1), g(2), \ldots, g(N_g)$; some gray levels excluding $g(1)$ and $g(N_g)$ may not appear in V_Q due to different quantization schemes). One NGTDM of size $N_g \times 1$ is computed per volume V_Q. To account for discretization length differences, all averages around a center voxel located at position (j,k,l) in V_Q are performed such that the neighbours at a distance of $\sqrt{3}$ voxels are given a weight of $1/\sqrt{3}$, the neighbours at a distance of $\sqrt{2}$ voxels are given a weight of $1/\sqrt{2}$, and the neighbours at a distance of 1 voxel are given a weight of 1. The i^th entry of the NGTDM is then defined as:

$$P(i) = \begin{cases} \sum_{\text{all voxels} \in \{N_i\}} |i - \overline{A}_i| & \text{if } N_i > 0, \\ 0 & \text{if } N_i = 0. \end{cases}$$

where $\{N_i\}$ is the set of all voxels with gray level i in V_Q (*including* the peripheral region), N_i is the number of voxels with gray level i in V_Q, and $\overline{A}_i$ is the average gray level of the 26-connected neighbours around a center voxel with gray level i and located at position (j,k,l) in V_Q such that:

$$\overline{A}_i = \overline{A}(j,k,l) = \frac{\sum_{m=-1}^{m=1} \sum_{n=-1}^{n=1} \sum_{o=-1}^{o=1} w_{m,n,o} \cdot V_\mathrm{Q}(j+m, k+n, l+o)}{\sum_{m=-1}^{m=1} \sum_{n=-1}^{n=1} \sum_{o=-1}^{o=1} w_{m,n,o}},$$

$$\text{where} \quad w_{m,n,o} = \begin{cases} 1 & \text{if } |j-m| + |k-n| + |l-o| = 1, \\ \frac{1}{\sqrt{2}} & \text{if } |j-m| + |k-n| + |l-o| = 2, \\ \frac{1}{\sqrt{3}} & \text{if } |j-m| + |k-n| + |l-o| = 3, \\ 0 & \text{if } V(j+m, k+n, l+o) \text{ is undefined.} \end{cases}$$

The following quantity is also defined:

$$n_i = \frac{N_i}{N}.$$

where N is the total number of voxels in V_{Q}.

The NGTDM texture features (5) are then defined as:

- **Coarseness** [12]:

$$coarseness = \left[\epsilon + \sum_{i=1}^{N_g} n_i\, P(i)\right]^{-1}$$

  where ϵ is a small number to prevent $coarseness$ becoming infinite.

- **Contrast** [12]:

$$contrast = \left[\frac{1}{(N_g)_{eff}\left[(N_g)_{eff} - 1\right]} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} n_i\, n_j\, (i-j)^2\right] \left[\frac{1}{N} \sum_{i=1}^{N_g} P(i)\right]$$

- **Busyness** [12]:

$$busyness = \frac{\sum_{i=1}^{N_g} n_i\, P(i)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i\, n_i - j\, n_j)}, \qquad n_i \neq 0, n_j \neq 0$$

- **Complexity** [12]:

$$complexity = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{|i-j| \left[n_i\, P(i) + n_j\, P(j)\right]}{N\,(n_i + n_j)}, \qquad n_i \neq 0, n_j \neq 0$$

- **Strength** [12]:

$$strength = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (n_i + n_j)\,(i-j)^2}{\left[\epsilon + \sum_{i=1}^{N_g} P(i)\right]}, \qquad n_i \neq 0, n_j \neq 0$$

  where ϵ is a small number to prevent $strength$ becoming infinite.

References

1. Li, Q. & Griffiths, J. G. *Least squares ellipsoid specific fitting. Proceedings of the Geometric Modeling and Processing 2004.* International Conference on Geometric Modeling and Processing (GMP 04) (Beijing, China, 2004), 335–340.

2. Van Velden, F. H. P. *et al.* Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur. J. Nucl. Med. Mol. Imaging* **38,** 1636–1647 (2011).

3. Rahmim, A. *et al.* A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction. *Phys. Med. Biol.* **61,** 227–242 (2016).

4. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3,** 610–621 (1973).

5. Thibault, G. *Indices de formes et de textures: de la 2D vers la 3D.* PhD thesis (Université AIX-Marseille, Marseille, France, 2009).

6. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5,** 4006 (2014).

7. Wei, X. *Gray Level Run Length Matrix Toolbox* version 1.0. 2007.

8. Galloway, M. M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing* **4,** 172–179 (1975).

9. Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11,** 415–419 (1990).

10. Dasarathy, B. V. & Holder, E. B. Image characterizations based on joint gray level–run length distributions. *Pattern Recognition Letters* **12,** 497–502 (1991).

11. Thibault, G. *et al. Texture indexes and gray level size zone matrix: application to cell nuclei classification. Proceedings of the Pattern Recognition and Information Processing 2009.* International Conference on Pattern Recognition and Information Processing (PRIP '09) (Minsk, Belarus, 2009), 140–145.

12. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics* **19,** 1264–1274 (1989).