Développement des approches radiomics à visées diagnostique et pronostique pour la prise en charge de patients atteints des sarcomes des tissus mous.

Amandine Crombé

université de BORDEAUX

THESE

pour l'obtention du grade de
Docteur de l'Université de Bordeaux

Ecole Doctorale de Mathématiques et d'Informatique
Spécialité : Mathématiques appliquées au calcul scientifique

présentée et soutenue par
Amandine Crombé
le 24 juin 2020

Développement des approches radiomics à visées diagnostique et pronostique pour la prise en charge de patients atteints des sarcomes des tissus mous.

sous la direction de
Olivier Saut

Jury

Jean-Emmanuel Bibault	Medecin, PHU	Examinateur
Irène Buvat	Directrice de recherche	Rapportrice
Vincent Dousset	Professeur	Président
Laure Fournier	Professeur	Rapportrice
Michèle Kind	Médecin, PH	Invitée
Nicolas Meunier	Professeur	Examinateur
Nicolas Papadakis	Directeur de recherche	Examinateur
Cécile Proust-Lima	Directrice de recherche	Examinatrice
Olivier Saut	Directeur de recherche	Directeur

TITRE

Développement des approches radiomics à visées diagnostique et pronostique pour la prise en charge de patients atteints des sarcomes des tissus mous.

RESUME

Les sarcomes des tissus mous (STM) sont des tumeurs malignes mésenchymateuses ubiquitaires hétérogènes en terme de présentations cliniques, radiologiques, histologiques, moléculaires et pronostiques. L'imagerie de référence des STM est l'IRM avec injection de produit de contraste qu'il s'agisse du bilan initial, de l'évaluation de la réponse aux traitements, de la planification préopératoire ou de la recherche de rechute locale. De plus, l'IRM permet d'accéder à la tumeur en place, *in vivo*, dans sa globalité et de manière non invasive, en complément des analyses anatomo-pathologiques et moléculaires qui nécessitent des prélèvements invasifs ne correspondant qu'à une infime fraction du volume tumoral. Cependant, aucun biomarqueur radiologique n'a été validé dans la prise en charge des STM. Parallèlement, se sont développés (i) d'autres modalités et séquences d'imagerie quantitative permettant d'aboutir à une quantification de phénomène physiopathologique intratumoraux, (ii) des techniques d'analyse d'image permettant de quantifier les phénotypes radiologiques au-delà de ce que peut voir l'œil humain à travers de multiples indicateurs de texture et de forme (: indices radiomics), et (iii) des outils d'analyses mathématiques (: algorithme de *machine-learning*) permettant d'intégrer et trier toutes ces informations dans des modèles prédictifs. Les approches radiomics correspondent au développement de modèles prédictifs basés sur ces algorithmes et ces indices radiomics. L'objectif de cette thèse est de mettre en application ces innovations et de les optimiser pour améliorer la prise en charge des patients atteints de STM. Pour cela, trois grands axes ont été développés. Dans une première partie, nous avons cherché à améliorer la prédiction du pronostic de patients atteints de certains sarcomes en combinant approches radiologiques classiques et approches radiomics sur leur IRM initiale, avec comme potentielle application de mieux identifier les patients à haut risque de rechute métastatique. Dans une deuxième

partie, nous avons construit un modèle basé sur l'évolution précoce de l'hétérogénéité intratumorale (: delta-radiomics) de patients atteints de STM traités par chimiothérapie néoadjuvante afin d'identifier les patients n'y répondant pas favorablement et qui pourraient bénéficier d'adaptations thérapeutiques anticipées. Dans une troisième et dernière partie, nous avons cherché à identifier et mieux contrôler les biais potentiels des approches radiomics afin, *in fine*, d'optimiser les modélisations prédictives basées sur les indices radiomics.

MOTS-CLEFS

Oncologie ; Sarcomes des tissus mous ; Hétérogénéité intra-tumorale ; Imagerie par résonance magnétique ; Imagerie dynamique de perfusion ; Radiomics ; Machine-learning ; Intelligence artificielle ; Modélisation prédictive ; Pronostic ; Evaluation de la réponse.

UNITE DE RECHERCHE

Equipe Modelisation in Oncology (MONC), INRIA Bordeaux-Sud-Ouest, CNRS UMR 5251, F-33405, Talence, France.

TITRE (ANGLAIS)

Development and applications of radiomics approaches to improve diagnostic and prognostic management for patients with soft-tissue sarcomas

RESUME EN ANGLAIS

Soft-tissue sarcomas (STS) are malignant ubiquitous mesenchymal tumors that are characterized by their heterogeneity at several levels, i.e. in terms of clinical presentation, radiological presentation, histology, molecular features and prognosis. Magnetic resonance imaging (MRI) with a contrast-agent injection is the imaging of reference for these tumors. MRI enables to perform the local staging, the evaluation of response to treatment, to plan the surgery and to look for local relapse. Furthermore, MRI can access non-invasively to the whole tumor *in situ* and *in vivo* which is complementary to histopathological and molecular analyses requiring invasive biopsy samples at risk of sampling bias. However, no imaging biomarker dedicated to STS has been validated so far. Meanwhile, technical innovations have been developed, namely: (i) alternative imaging modalities or MRI sequences that can quantify intratumoral physiopathological phenomenon; (ii) image analysis tools that can quantify radiological phenotypes better than human's eyes through hundreds of textural and shape quantitative features (named radiomics features); and (iii) mathematical algorithms that can integrate all these information into predictive models (: machine-learning). Radiomics approaches correspond to the development of predictive models based on machine-learning algorithms and radiomics features, eventually combined with other clinical, pathological and molecular features. The aim of this thesis was to put these innovations into practice and to optimize them in order to improve the diagnostic and therapeutic managements of patients with STS.

In the first part, we combined radiological and radiomics features extracted from the baseline structural MRIs of patients with a locally-advanced subtype of STS in order to build a radiomics signature that could help to identify patients with higher risk of metastatic relapse and may benefit from neoadjuvant treatments. In the second part, we elaborated a model based on the early changes in intratumoral heterogeneity (: delta-radiomics) on structural MRIs of patients with locally-advanced high-grade STS treated with neoadjuvant chemotherapy, in order to rapidly identify patients who do not respond to treatment and would benefit from early therapeutic adjustments. In the

last part, we tried to better identify and control potential bias in radiomics approaches in order to optimize the predictive models based on radiomics features.

KEY-WORDS

Oncology; Soft-tissue sarcomas; Intratumoral heterogeneity; Magnetic resonance imaging; Dynamic-contrast enhanced MRI; Radiomics; Machine-learning; Artificial intelligence; Predictive modelling; Prognosis; Response evaluation.

TRAVAUX ET PRESENTATIONS ISSUS DE CETTE THESE

Articles scientifiques

Publiés dans des revues à comité de lecture

(1) T2 -based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. Crombé A, Périer C, Kind M, De Senneville BD, Le Loarer F, Italiano A, Buy X, Saut O. J Magn Reson Imaging. 2019;50(2):497-510. doi: 10.1002/jmri.26589. PMID: 30569552

(2) Influence of temporal parameters of DCE-MRI on the quantification of heterogeneity in tumor vascularization. Crombé A, Saut O, Guigui J, Italiano A, Buy X, Kind M. J Magn Reson Imaging. 2019;50(6):1773-1788. doi: 10.1002/jmri.26753. PMID: 30980697

(3) High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models? Crombé A, Fadli D, Buy X, Italiano A, Saut O, Kind M. J Magn Reson Imaging. 2020 Jan 10. doi: 10.1002/jmri.27040. PMID: 31922323

(4) Can radiomics improve the prediction of metastatic relapse of myxoid/round cell liposarcomas? Crombé A, Le Loarer F, Sitbon M, Italiano A, Stoeckle E, Buy X, Kind M. Eur Radiol. 2020;30(5):2413-2424. doi: 10.1007/s00330-019-06562-5. PMID: 31953663

En cours d'évaluation dans des revues à comité de lecture

(5) Body-MRI Signal Intensity Harmonization Techniques Influence Radiomics Features and Can Enhance Radiomics-based Predictions. Crombé A, Kind M, Fadli D, Italiano A, Buy X, Saut O.

(6) Critical Review of Sarcomas Radiomics Studies: Quantifying Quality to Improve Reproducibility. Crombé A, Fadli D, Italiano A, MD, Saut O, Buy X, Kind M.

Articles satellites de la thèse acceptés/publiés dans des revues à comité de lecture

(7) Magnetic resonance imaging assessment of chemotherapy-related adipocytic maturation in myxoid/round cell liposarcomas: specificity and prognostic value. Crombe A, Sitbon M, Stoeckle E, Italiano A, Buy X, Le Loarer F, Kind M. Br J Radiol. 2020 Mar 6:20190794. doi: 10.1259/bjr.20190794. PMID: 32105502

(8) Progressive Desmoid Tumors: Comparison of Radiomics and Conventional Response Criteria For Predicting Progression During Systemic Therapy - A Multicenter Study from the French Sarcoma Group. Crombé A, Kind M, Ray-Coquart I, Isambert N, Chevreau C, et al. Accepté définitivement en mars 2020 dans American Journal of Roentgenology.

Congrès et séminaires

Journées françaises de radiologie (JFR)

- JFR 2018 : Communication orale relative à l'article (1)
- JFR 2020 (à venir) : Communication orale à la demande de la SIMS relative à l'article (4) ; communication orale scientifique relative à l'article (6)

European Congress of Radiology (ECR)

- ECR 2018 : Communication orale relative à l'article (1)

- ECR 2019 : Poster et communication orale courte relative à l'article (2) avec prix (Certificate of merit)
- ECR 2020 (à venir) : Communications orales relatives aux articles (3) et (8), poster et communication orale courte relative à l'article (7)

Oncosphère, Bordeaux :
- Séminaire de l'oncosphère (octobre 2019): « Défis de l'imagerie oncologique et apport des approches IA – exemple des sarcomes »
- Workshop recherche translationnel de l'oncosphère (janvier 2019), session « Nouvelles avancées techniques », « Approches radiomics appliquées aux sarcomes »

Table ronde des 10 ans de l'INRIA Bordeaux Sud Ouest : Simulation numérique pour la santé, de la recherche au transfert (Septembre 2018)

REMERCIEMENTS

Je tiens avant tout à remercier Monsieur ***Olivier Saut*** et Madame la docteure ***Michèle Kind*** qui ont permi la réalisation de cette thèse par des voies complémentaires.
Olivier, je te remercie encore d'avoir accepté de m'encadrer, de m'avoir fait découvrir et manipuler des techniques innovantes d'analyse des données tout en t'adaptant à mon profil médical. Chacun de nos échanges m'a permi d'élargir l'appréhension que je pouvais avoir d'une question et d'en approfondir les réponses. Les perspectives que tu m'as ouvertes nourriront durablement ma curiosité et j'ai espoir que nous poursuivrons nos collaborations.
Michèle, je vous remercie pour votre compagnonnage sans faille depuis cinq ans que nous nous connaissons. Rien de cette thèse n'aurait pu se faire sans vous : depuis mon temps de recherche que vous avez systématiquement protégé, jusqu'aux acquisitions d'images ici exploitées et que vous aviez initiées seule bien avant mon arrivée, en passant par le regard clinique et critique sur les résultats obtenus. J'espère prolonger votre vision éclairée, humble et toujours lucide de nos pratiques et de la radiologie.

Je remercie les rapportrices de cette thèse:
Madame la Professeure ***Laure Fournier*** et Madame ***Irène Buvat***, j'éprouve un grand respect pour vos enseignements et vos travaux et je suis très honorée que ayez accepté de relire cette thèse. Je vous remercie pour le temps que vous y avez consacré ainsi que pour votre relecture critique qui a contribué à en améliorer la présente version.

Je remercie Monsieur le Président du jury, Professeur ***Vincent Dousset.*** Vous avez su, dès les débuts de mon internat, identifier mon attrait pour la recherche et me donner, avec ***Thomas***, les moyens de le cultiver. C'est un honneur de vous voir présider cette étape essentielle de mon parcours scientifique.

Je remercie les membres du jury, qui ont su se montrer disponible en dépit des contraintes imposées par le contexte sanitaire et social de ce printemps 2020 : Monsieur le Docteur ***Jean-Emmanuel Bibault***, Monsieur le Professeur ***Nicolas Meunier***, Monsieur ***Nicolas Papadakis*** et Madame ***Cécile Proust-Lima***.

Je remercie enfin les membres de mon comité de thèse, le Professeur ***François Cornelis*** et Monsieur ***Pierrick Coupé***, ainsi que Mademoiselle ***Cynthia Périer*** pour ses nombreuses explications introductives au machine-learning.

Je remercie mes collègues et amis des départements de radiologie et de médecine nucléaire de l'institut Bergonié, ainsi que le groupe sarcome bordelais.
Cette thèse n'aurait évidemment pas pris cette tournure sans le soutien aimant et inconditionnel de ma famille, de mes amis proches et avant tout de ma moitié.

TABLE DES MATIERES

ABBREVIATIONS PRINCIPALES

^{18}F-FDG-PET/CT : Tomographie à émission de positon au ^{18}F-Fluorodesoxyglucose couplée au scanner (simplifié PET/CT)

95%CI: Intervalle de confiance à 95%

ACP : Analyse en composante principale

ADC : Coefficient apparent de diffusion

AIF : Arterial input function

ASPS : Alveolar soft part sarcoma

AUC : Area under the time-intensity curve

AUPRC : Area under the precision - recall curve

AUROC : Area under the receiver operating characteristics curve

C-index : Indice de concordance de Harrell

CINSARC : Complexity INdex in SARComas

CE : Contrast-enhanced

CR : Complete response

DCE-MRI : Dynamic contrast enhanced MRI

DWI : Diffusion weighted imaging

EEE : Espace extravasculaire extracellulaire

EIBALL : European imaging biomarker alliance

EORTC : European organization for research and treatment of cancer

FBN : Fixed by number

FBS : Fixed by size

FNCLCC : Fédération Nationale des Centres de Lutte Contre le Cancer

FS : Fat suppressed

Gd : Agent de contraste à base de chélates de gadolinium

GIST : Gastro-intestinal stromal tumor

GLCM : Gray level co-occurrence matrix

GLDM : Gray level dependence matrix

GLRLM : Gray level run length matrix

GLZM : Gray level zone matrix

HES : Hematoxylin and eosin stained slices

HR : Hazard ratio

iAUC : Integrated area under the time-dependent area under the ROC curve

IBSI : Image biomarker standardisation initiative

ICC : Intraclass correlation coefficient

INCA : Institut national du cancer

IRM : imagerie par résonance magnétique

ISM : Inter-site similarity matrix

IV: Intra-veineux

IRM : Imagerie par Résonance Magnétique

K^{ep} : Rate constant

KNN : k-nearest neighbor

K^{trans} : Volume transfer constant

LASSO : Least absolute shrinkage and selection operator

LOESS : Locally estimated scatterplot smoothing

LOOCV : Leave-one-out cross-validation

LPS : Liposarcome

MRS : Spectroscopie par résonance magnétique

MTV : Metabolically active tumor volume

NGTDM : Neighborhood gray tone difference matrix

NGS : Next-generation sequencing

MET : Micro-environnement tumoral

MFS : Myxofibrosarcome

M/RC-LPS : Myxoid/round cells liposarcoma

OMS : Organisation Mondiale de la Santé

PD : Progressive disease

PERCIST : PET response criteria in solid tumors

PR : Partial response

QUADAS-2 : Quality assessment of diagnostic accuracy studies

QIBA: Quantitative imaging biomarker alliance

RCP : Réunion de concertation pluridisciplinaire

RECIST : Response evaluation criteria in solid tumors

RRePS : Réseau de référence en pathologie des sarcomes

RQS : Radiomics quality score

SD : Stable disease

SI : Signal intensity

SMOTE : Synthetic minority over-sampling technique

STM : Sarcome des tissus mous

SUL : Standardized uptake value by lean body mass

SUV : Standardized uptake value

SVM : Support vector machine

TCIA : The cancer imaging atlas

TLG : Total lesion glycolysis

TLS : Secteur lymphoide tertiaire

TRIPOD : Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

t-SNE : t-distributed stochastic neighbor embedding

UH : Unité Hounsfield

UPS : Undifferentiated pleomorphic sarcoma

UICC : Union for international Cancer Control

V_e : Extravascular extracellular space volume

V_p : Plasmatic volume

VOI : Volume of interest

WHO : World Health Organization

WI : Weighted imaging

Les sarcomes sont des tumeurs malignes mésenchymateuses ubiquitaires caractérisées par leur hétérogénéité sous plusieurs points de vue, notamment radiologique et pronostic. De plus, en raison de leur rareté, le partage des connaissances fondamentales s'est structuré assez vite en base de données clinico-biologiques. L'ensemble en fait un modèle pour le développement d'applications des approches radiomics en cancérologie.

Dans une première partie introductive et théorique, nous résumerons les particularités des sarcomes d'un point de vue extra-radiologique, puis nous aborderons la place actuelle de l'imagerie telle qu'énoncée dans les référentiels, avant d'en aborder les limites et la nécessité de développer de nouvelles techniques plus à mêmes de prendre en compte les particularités structurales et fonctionnelles de ces tumeurs. Nous exposerons ensuite des pistes pour améliorer la prise en charge diagnostique et thérapeutique des patients atteints de sarcomes des tissus mous (STM) par l'imagerie médicale et introduirons les approches radiomics. Nous détaillerons les études jusque là réalisées relatives aux sarcomes et en ferons ressortir les avancées et les limites.

Dans une seconde partie pratique, nous détaillerons nos travaux selon 3 axes: (1) améliorer la prédiction du pronostic des patients atteints de STM par les approches radiomics; (2) améliorer la prédiction de la réponse aux traitements systémiques par les approches radiomics; et (3) améliorer ces approches prédictives en identifiant mieux les biais des études radiomics et en optimisant leur post-traitement.

1. INTRODUCTION

1.1. Sarcomes : généralités

1.1.1. Sarcomes : définitions

Les sarcomes correspondent à un ensemble de tumeurs malignes rares et ubiquitaires d'origine mésenchymateuse. Les STM de l'adulte, groupe concerné par ce travail, représentent le sous groupe le plus fréquent (environ 60% de tous les sarcomes), à côté des sarcomes viscéraux (notamment les « *gastro-intestinal stromal tumors* »

[GIST]) et des sarcomes osseux. Les STM touchent la paroi du tronc, les ceintures, les membres, envahissant de manière non exclusive des tissus superficiels, aponévrotiques et profonds (par rapport au fascia superficiel).

L'incidence annuelle des STM est estimée entre 3000 et 5000 nouveaux cas par an en France, soit 3.6 à 4.8 nouveaux cas pour 100 000 sujets par an, ce qui en fait une des maladies orphelines les plus fréquentes (Honoré et al, 2015). Les liposarcomes (LPS), les léiomyosarcomes et les sarcomes indifférenciés pléomorphes sont les trois histotypes prédominant avec des incidences estimées à environ 2 pour 100 000 en France (Fletcher et al, 2013 ; Penel et al, 2018).

Les STM sont caractérisés par leur hétérogénéité sous plusieurs points de vue. On peut les distinguer par leur lignage ou type de différenciation tissulaire (vasculaire, chrondrogénique, ostéogénique, fibroblastique / myofibroblastique, adipocytique, musculaire lisse ou strié, nerveux... etc.), par leur type histologique, par leur grade histopathologique, par leur profil génomique, par leur présentation clinique et radiologique, par leur évolution sous traitement et enfin par leur pronostic - toute autre variable étant égale par ailleurs (Fletcher et al, 2013).

1.1.2. STM : épidémiologie

L'âge moyen au diagnostic est de 60 ans avec une discrète prépondérance masculine (1.1 - 1.3 pour 1 femme).

La répartition des types histologiques diffèrent selon l'âge, certains sarcomes étant plutôt propres aux sujets jeunes (sarcomes alvéolaires des parties molles [ASPS], synovialosarcomes, LPS myxoïdes et à cellules ronds [M/RC-LPS]), d'autres plutôt spécifiques du sujet âgés (LPS pléomorphes, sarcomes indifférenciés pléomorphes).

En dehors d'antécédents d'irradiation, de lymphoedème chronique, d'expositions à des carcinogènes chimiques (notamment issus de l'agriculture), d'immunodéficience acquise, de neurofibromatose de type 1, de mutations des gènes p53, APC ou RB1, les facteurs de risque restent peu connus.

Les STM se présentent habituellement sous forme d'une tuméfaction lentement progressive, initialement indolente jusqu'à l'apparition de symptômes secondaires à l'effet de masse exercé sur les structures adjacentes. Néanmoins, certains sarcomes

peuvent se présenter sous forme de lésions hémorragiques à risque de confusion avec des hématomes, notamment chez le sujet âgé sous anticoagulant (Taieb et al, 2009)

Près de 10% des sarcomes sont métastatiques au diagnostic avec une médiane de survie d'environ 1 - 1.5 ans (Karavasilis et al, 2008).

Le pronostic des patients atteints de STM est conditionné par les rechutes locale et métastatique. Environ 10 à 30% des patients présenteront une rechute locale (selon la qualité de la chirurgie initiale) et entre 30 et 50% d'entre eux une rechute métastatique, majoritairement dans les 5 ans après la fin des traitements.

Ces localisations secondaires sont majoritairement situées dans le poumon, mais certains sarcomes sont connus pour présenter des dissémination métastatiques atypiques tels les ASPS (encéphale, foie, cerveau...), les leiomyosarcomes et les M/RC-LPS (séreuse, tissus mous, os...).

1.1.3. STM: Caractérisation anatomo-pathologique

Plus de 70 histotypes de STM ont été répertoriés dans la classification OMS sur la base d'analyses histologique et immuno-histochimique, divisés en 12 sous-groupes selon leur lignage cellulaire (ou du moins, sur la base de cellules saines présentant la plus forte ressemblance avec les cellule tumorale). Les 10 plus fréquents histotypes sont résumés dans la Table 1-1.

Table 1-1. Principaux types histologiques des sarcomes. Adapté de *Penel et al, 2018*.

	Histotype	Fréquence	Age moyen
1	Leimyosarcome	14.9%	59.1
2	Sarcome indifférencié pléomorphe	11.3%	62.7
3	Sarcome indifférencié, autre	9%	56
4	Tumeur lipomateuse/atypique*	8.9%	63.1
5	Liposarcome dédifférencié	8.7%	38.4
6	Synovialosarcome	7.4%	65.5
7	Liposarcome myxoide et à cellules rondes	6%	45.2
8	Myxofibrosarcome	5.6%	65.5
9	Angiosarcome	3.9%	65
10	Tumeur maligne des gaines des nerfs périphériques	3.2%	43.6

NOTE. * correspond aux LPS bien différenciés et aux tumeurs lipomateuses atypiques.

1.1.4. STM: Caractérisation moléculaire

Le développement des techniques de pathologie moléculaire, notamment l'hybridation génomique comparative sur réseau d'ADN (CGH-array), l'hybridation in situ fluorescent (FISH) ou encore le séquençage de l'ARN des tumeurs (RNA-seq) a rendu possible le profilage moléculaire des divers histotypes de sarcomes. On distingue ainsi (Figure 1-1):

- les sarcomes à translocations, par exemple: les synovialosarcomes, les rhabdomyosarcomes alvéolaires, les M/RC-LPS, les ASPS...;
- les sarcomes à mutation activatrice;
- les sarcomes à mutations inhibitrices;
- les sarcomes à amplification simple (LPS bien-différenciés et dédifférenciés, sarcomes intimaux);
- les sarcomes à génomique complexe aux caryotypes déséquilibrés, notamment: les sarcomes indifférenciés à cellules pléomorphes (UPS), les myxofibrosarcomes (MFS), les léiomyosarcomes, les rhabdomyosarcomes pléomorphes et les LPS pléomorphes.

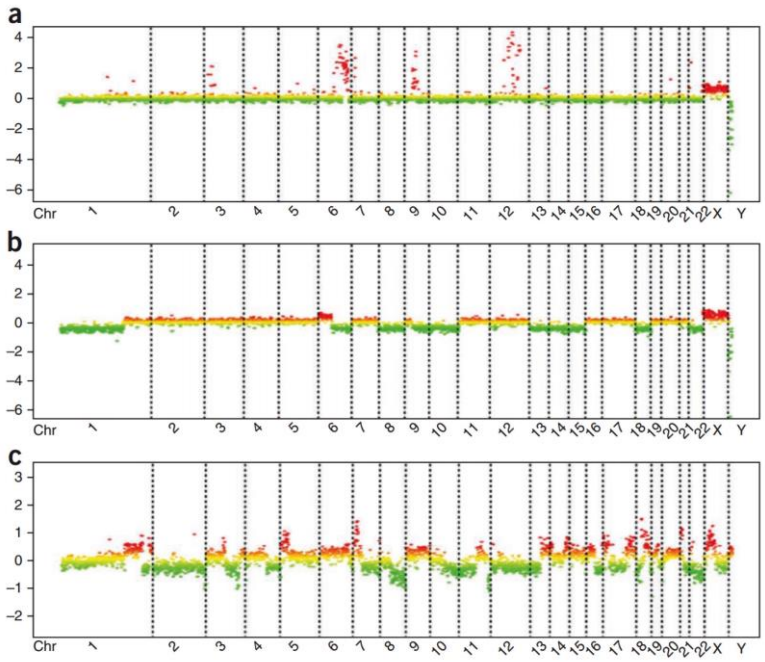
Les sarcomes présentant une mutation activatrice, ou inhibitrice ou une amplification sont habituellement regroupés sous l'appellation "sarcomes à génomique simple" et représentent près de 18% des sarcomes tandis que les sarcomes à translocation en représentent environ 25% et les sarcomes à génomique complexe environ 57% (Lucchesi et al. 2018).

Plusieurs altérations génomiques récurrentes ont pu être identifiées parmi les STM, notamment de TP53, CDK4, MDM2, RB1, CDKN2A/B, PTEN... etc. (Lucchesi et al, 2018; Groisberg et al, 2017).

Au-delà d'une meilleure connaissance du caryotype des STM, un meilleur profilage moléculaire peut permettre d'identifier des mutations ciblables par des thérapies. En effet, en s'appuyant sur 584 patients atteints de STM de la *database* AACR GENIE avec séquençage de l'ADN tumoral par NGS (pour "*next-generation sequencing*"), Lucchesi et *al.* (2018) ont montré qu'environ 41% des STM avaient au moins une mutation ciblable chez 40% des patients dans le groupe des sarcomes à génomique simple, 13% du groupe des sarcomes à translocation et 81% du groupe des sarcomes à génomique complexe. Ces mutations impliquaient les voies de signalisation des

récepteurs aux tyrosine kinases/Ras, les gènes *ALK, ARAF, ERBB2, FGFR1/2/3/4, KRAS, MET, NF1/2* et *NRAS* (Lucchesi et al, 2018).

Figure 1-1. Trois principaux profils génomiques des STM obtenus par CGH-array (d'après *Chibon et al, 2010*). L'abscisse représente les chromosomes. (**a**) STM à amplification; (**b**) STM à translocation; (**c**) STM avec multiples réarrangements à génomique complexe.



Deux points sont à noter:

(1) À ce jour, aucune étude n'a étudié d'éventuelles corrélations entre les caractéristiques radiologiques et ces 3 grandes catégories moléculaires de STM ou ces mutations d'intérêt.

(2) Pareillement, la variation de la présence de ces mutations d'une métastase à l'autre au sein d'un même patient ou leur éventuelle apparition au cours de l'évolution du cancer est méconnue.

1.1.5. STM: micro-environnement tumoral (MET)

Le MET peut être défini par tout ce qui n'est pas des cellules tumorales dans une tumeur, c'est-à-dire la matrice extracellulaire, les fibroblastes, les vaisseaux sanguins,

les cellules immunitaires (macrophages, neutrophiles, lymphocytes), ou les molécules de signalisation.

Peu d'études se sont intéressées au MET des sarcomes avant les premier essais cliniques testant l'immunothérapie, en l'occurrence le pembrolizumab, un anti-PD-1 (pour « *programmed cell death protein 1* ») (Toulmonde et al, 2018). Pour expliquer le faible taux de réponse (3 sur 50 patients) des patients atteints de STM traités par pembrolizumab et cyclophosphamide, Toulmonde et *al.* ont analysé le MET des bons et mauvais répondeurs et montré une moindre abondance de lymphocytes CD8 comparativement à d'autres cancers répondant bien à l'immunothérapie, un moindre immuno-marquage PD-1, une surreprésentations de macrophages M2 et une sur-activation de la voie IDO. Récemment, Petitprez et *al.* (2020) ont analysé de manière extensive le MET des sarcomes et proposé une classification immunologique en 5 catégories dont la dernière, dite classe E, est caractérisée par un secteur lymphoïde tertiaire (TLS) enrichi en lymphocyte B et répondant mieux aux anti-PD-1. D'éventuelles corrélations avec l'imagerie restent à évaluer.

1.1.6. STM: facteurs pronostics en situation localement avancée

A l'heure actuelle, les facteurs pronostics indépendant validés dans les dernières recommandations de l' « *European Society of Medical Oncology* » (ESMO) (Casali et al. 2018) et utilisés en routine clinique sont:

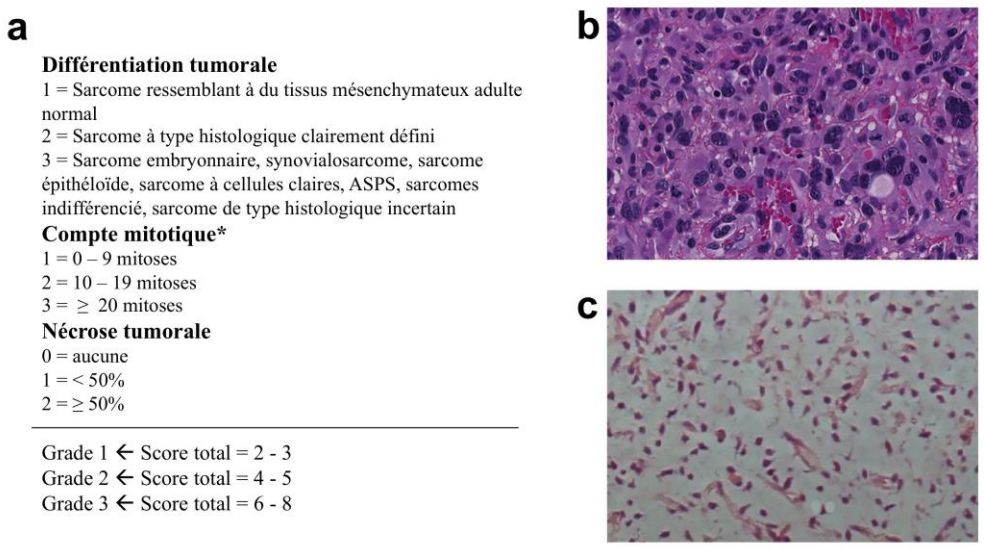
(1) la profondeur par rapport à l'aponévrose superficielle, en considérant qu'une localisation profonde et/ou atteignant l'aponévrose superficielle est de plus mauvais pronostic pour les STM localement avancés en terme de survie globale (ou « *overall survival* » [OS]), de survie sans rechute métastatique (ou « *metastatic relapse free survival* » [MFS]) et de la survie sans rechute locale (ou « *local relapse free survival* » [LFS]) (Coindre et al, 1996);

(2) le plus grand diamètre tumoral, en considérant qu'un diamètre supérieur à 5 cm est un facteur indépendant des OS, MFS et LFS (Coindre et al, 1996);

(3) le grade histopathologique selon la classification de la Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC). Ce grade a été développé en 1984 et consiste à évaluer 3 items sur des lames HES (pour « *hematoxylin and eosin stained slices* ») de la tumeur entière, à savoir: le nombre de mitoses, le pourcentage de

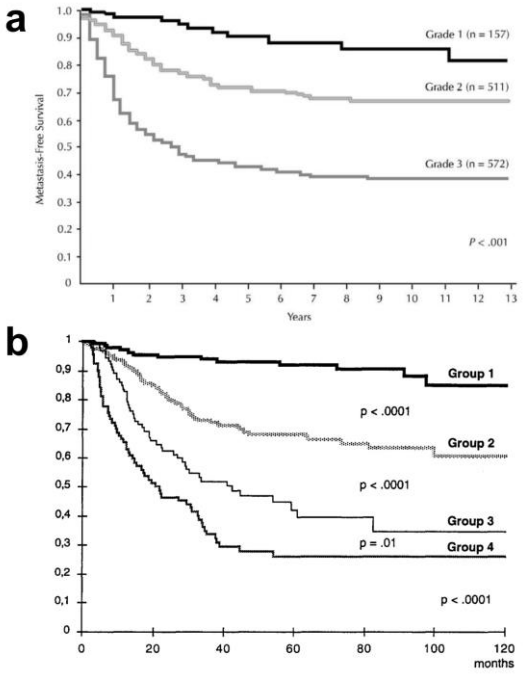
nécrose et le degré de différentiation tumorale. En sommant les points attribués à chacun des ces items, on aboutit à un score traduisant l'agressivité tumorale (Figure 1-2) (Trojani et al, 1984).

Figure 1-2. Grading histopathologique selon la FNCLCC. (**a**) Items du grade. (**b**) Lame HES d'un STM de grade 3 FNCLCC (sarcome indifférencié pléomorphe). (**c**) Lame HES d'un STM de grade II (liposarcome myxoïde – ici sans cellule ronde). *: Pour 10 champs à fort grossissement.



Coindre et al. ont montré que le grade était le principal facteur pronostic indépendant de l'OS, la MFS et la LFS, cela dans une population générale de STM, mais aussi dans les principaux sous-types histologiques (Figure 1-3) (Coindre et al, 1996 ; Le Doussal et al, 1996 ; Coindre et al, 2001).

Figure 1-3. Courbes de survie de Kaplan-Meier pour la survie sans rechute métastatique (MFS) selon le grade FNCLCC (**a**) et en prenant en compte la profondeur et la taille du sarcome (**b**) – extraits de: *Coindre et al, 1996* et *Saponara et al, 2017*. Le groupe 1 correspond aux STM de grade 1 et superficiel (n = 139), le groupe 2 aux STM de grade 2 et superficiel (n = 211), le groupe 3 aux STM de grade 3 et de moins de 10 cm (n = 96) et le groupe 4 aux STM de grade 3 et ≥ 10 cm (n = 91).



La Table 1-2 résume les pronostics des patients selon ces 3 facteurs.

Table 1-2. Principales variables pronostiques des STM. Adapté de *Coindre et al, 1996* et *Saponara et al, 2017*.

Variables pronostiques	LFS à 5 ans (%)	MFS à 5 ans (%)	OS à 5 ans (%)
Taille			
< 5 cm	76.7	80.8	83
5 - 9 cm	69.3	64.1	66.4
≥ 10 cm	60.2	48.2	55.9
Profondeur			
Superficiel	78.3	87	85.5
Profond	64.4	55.9	61.4
Grade FNCLCC			
I	76.3	91.4	94.7
II	71.6	76.4	74.4
III	60.1	38.4	47.4

D'autres facteurs pronostics ont été identifiés de manière moins constante comme un envahissement vasculaire, nerveux et/ou osseux, ou le stade AJC/UICC (pour « *Union for international Cancer Control* ») (Coindre et al, 1996).

Ces facteurs pronostics sont évaluables dès le diagnostic initial, permettent d'établir l'agressivité du STM et de décider de la stratégie thérapeutique. Néanmoins, à l'heure actuelle, aucune des innovations moléculaires ou radiologiques depuis les années 1990 n'est officiellement prise en compte en routine pour stratifier les traitements et organiser la surveillance des patients atteints de STM.
Or, Chibon et *al.* (2010) ont développé une signature moléculaire nommée CINSARC (pour « *Complexity INdex in SARComas* ») basée sur l'expression de 67 gènes impliqués dans la mitose et les réarrangements chromosomiques. CINSARC propose une classification en 2 groupes: haut risque et bas risque. Les auteurs ont montré que CINSARC prédisait mieux la MFS que le grade FNCLCC lors d'analyses univariée et multivariée pour les STM, mais aussi pour les GIST, les cancers du sein et les lymphomes.

De plus, notre groupe a aussi montré que l'imagerie par résonance magnétique (IRM) initiale avec injection de produit de contraste (« *contrast-enhanced* » [CE]) comportait une information pronostique sous la forme d'une combinaison de 3 variables (Figure 1-4):

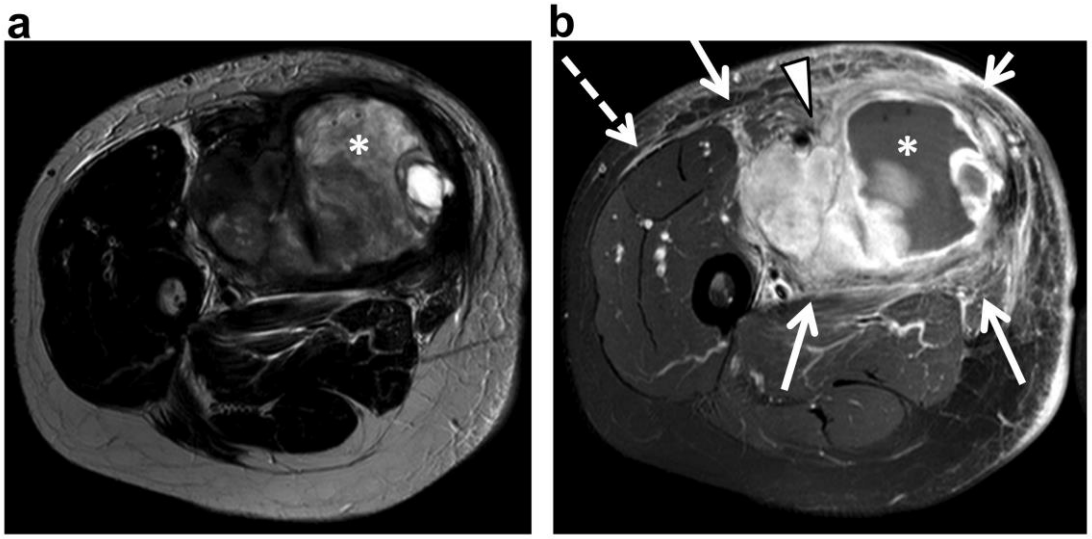
- la présence de plus de 50% de nécrose en IRM
- une hétérogénéité sur les séquences pondérées T2 avec plus de 50% du volume tumoral comportant des aires d'intensités de signal élevés, intermédiaire, et basses.
- la présence de prises de contraste péritumorales, c'est-à-dire dans les tissus environnant de la tumeur, au-delà de ses limites apparentes.

La présence d'au moins 2 parmi ces 3 variables était fortement corrélée à l'OS et à la MFS indépendamment du grade FNCLCC (Crombé et al, 2019a).

Au vue de ces travaux basés sur des approches cliniques, histologiques, immunologiques, génomiques / transcriptomiques et radiologiques, il est envisageable que se développent à moyen terme des signatures pronostiques composites synergiques capables d'intégrer l'ensemble des spécificités d'une tumeur et d'un patient afin de lui proposer une prise en charge personnalisée, lui donnant accès à des

molécules ou des combinaisons de molécules les plus susceptibles de permettre sinon une rémission, au moins une réponse partielle ou une stabilisation de la maladie. Ce type d'approche porte le nom de *médecine personnalisée*.

Figure 1-4. Indices radiologiques associés aux STM de haut grade et à valeur pronostique péjorative. Exemple d'un sarcome indifférencié pléomorphe de grade 3 FNCLCC avec IRM pré-thérapeutique comportant: (**a**) une séquence axiale T2 turbo spin echo et (**b**) une séquence axiale T1 Fat Sat turbo spin echo après injection intraveineuse de chélates de gadolinium.Cet examen montre un STM de topographie profonde de la cuisse droite, aux contours infiltrant, envahissant l'aponévrose et présentant: (i) une vaste plage de nécrose (*), (ii) d'importants rehaussements péritumoraux (flèche blanche) et (iii) une importante hétérogénéité T2 sur plus de 50% du volume tumoral. Il existe aussi des rehaussements fusant le long des aponévroses (flèche en pointillé), un envahissement de la veine fémorale externe droite, du nerf fémoral et un engainement de l'artère fémorale externe (tête de flèche).



1.2. Prise en charge des STM

1.2.1. Organisation en réseaux

Les dernières recommandations ESMO et de l' « *European Society for Skeletal Radiology* » énoncent clairement que les STM doivent être pris en charge dès la phase diagnostique par une équipe pluridisciplinaire de médecins experts de centre de référence habitués à traiter plusieurs dizaines voire centaines de patients par an (Casali et al, 2018; Noebauer-Huhmann et al, 2015).

En France, il existe 26 centres experts en sarcomes viscéraux et STM (dont 3 centres coordonnateurs: l'institut Bergonié, le centre Léon Bérard et l'institut Gustave Roussy) regroupés au sein du réseau NETSARC (https://netsarc.sarcomabcb.org), financé par l'Institut National du CAncer (INCA) et adossé un réseau de pathologistes experts en sarcome nommé RRePS (Réseau de Référence En Pathologie des Sarcomes, https://rreps.sarcomabcb.org).

NETSARC a pour objectifs d'actualiser les données épidémiologiques relatives aux sarcomes en France (incidence, prévalence), d'enregistrer les diagnostics et procédures thérapeutiques initiales ainsi que les survies et rechutes, et de favoriser la compliance aux recommandations.

Pareillement, RRePS propose une double lecture des prélèvements permettant de limiter les erreurs diagnostiques, mais aussi de colliger les nouveaux cas de sarcomes dans des bases de données, de former les médecins et de proposer de nouvelles recommandations. Près de 60 000 patients sont enregistrés dans NETSARC et plus de 50 000 dans RRePS début 2020.

L'équipe pluridisciplinaire des centres de référence sarcomes comporte des oncologues médicaux, des radiologues diagnosticiens et interventionnels, des médecins nucléaires, des anatomopathologistes, des chirurgiens et des radiothérapeutes. Cette équipe établit la stratégie pour aboutir au diagnostic via la réalisation: (i) d'examens d'imagerie adéquats, (ii) de prélèvements biopsiques (14-16G) guidés par l'imagerie, (iii) d'analyse histologiques, immuno-histochimiques et moléculaires adaptées, et (iv) de corrélations radio-anatomopathologiques. A l'issue de ce diagnostic, les différentes options thérapeutiques sont discutées et appliquées par ce groupe ainsi que l'évaluation de la réponse, les ajustements thérapeutiques et les modalités de surveillance.

En s'appuyant sur les données de NETSARC, Blay et *al.* ont montré que cette prise en charge multidisciplinaire en centre expert permettait d'améliorer significativement la LFS et la MFS en analyse univariée et multivariée, de permettre un plus grand nombre de chirurgie adaptée carcinologique, de réduire le nombre de reprises chirurgicales et de réduire les couts (Blay et al, 2017; Blay et al, 2019).

Il faut noter qu'il n'existe pas de *database* radiologique sur le modèle de ces bases de données cliniques et anatomopathologiques

1.2.2. Grands axes de la prise en charge diagnostique thérapeutique

1.2.2.1. Cas des patients localement avancés et opérables

Le traitement va dépendre du grade, de la profondeur et de la taille tumorale (Casali et al, 2018), ainsi que, pour les traitements adjuvants, de la qualité de la chirurgie (en utilisant la terminologie suivante: R0: marges saines micro- et macroscopiques, R1: marges saines macroscopiques mais envahissement microscopique et R2: berges envahies macroscopiquement).

- Un patient avec une tumeur bas grade FNCLCC (1 ou 1-2) devrait avoir une chirurgie R0 éventuellement complétée par une radiothérapie du lit opératoire si la lésion est profonde et/ou supérieure à 5 cm et/ou si la chirurgie s'est avérée R1.

- Un patient avec une tumeur de haut grade FNCLCC (3 ou 2-3) et une lésion superficielle devrait avoir une chirurgie R0 éventuellement complétee par une radiothérapie.

- Un patient avec une tumeur de haut grade FNCLCC (3 ou 2-3) et une lésion profonde peut bénéficier d'un traitement néoadjuvant (à savoir une chimiothérapie à base d'anthracycline - doxorubicine - ou une radiothérapie à 50 Gy en fractions de 1.9 - 2 Gy) suivie d'une chirurgie R0 et d'une radiothérapie adjuvante si non faite en néoadjuvant.

1.2.2.2. Cas des patients métastatiques d'emblée ou localement avancé mais non opérables

La prise en charge va consister en une chimiothérapie à base d'anthracyclines (doxorubicine) en 1ère ligne, éventuellement complétée par une radiothérapie locale pour les patients non métastatiques. Cette chimiothérapie a un double objectif : limiter la diffusion métastatique (50% des STM de haut grade) et faciliter la chirurgie en réduisant le volume tumoral et rapports vasculo-nerveux et osseux.

1.2.2.3. Controverse autour de la chimiothérapie néoadjuvante (NAC)

Les anthracyclines sont donc la chimiothérapie cytotoxique de référence des sarcomes en première ligne, éventuellement en association avec l'ifosfamide pour 3 à 6 cures. Les anthracyclines sont des agents s'intercalant entre les paires azotées de l'ADN pour

inhiber l'action de la topoisomérase II (enzyme contrôlant la structure de l'ADN). Les anthracyclines ont été utilisées dès les années 1970 en situation adjuvante et avaient montré une réduction significative du taux de rechute (Benjamin et al, 1974). La controverse autour de la NAC provient de l'absence de groupe contrôle (i.e. sans NAC) dans certaines études néoadjuvantes, ou l'absence d'étude construite pour affirmer la supériorité de la NAC, ou l'inclusion de patients aux pronostics très différents (Gronchi et al, 2017; Gortzak et al, 2001; Saponara et al, 2017; Issels et al 2010). Il apparait néanmoins dans la plus récente de ses études (consistant à évaluer le bénéfice d'une chimiothérapie déterminée selon le type histologique de STM de haut grade) qu'une NAC standard par anthracycline et ifosfamide améliorait significativement la survie globale et la survie sans rechute avec des courbes de survie meilleure que les courbes de survie usuelles de patients à haut risque métastatique (Gronchi et al, 2017).

1.2.2.3. Thérapies en seconde ligne

Si le patient est opérable, une chirurgie curatrice suivie d'une radiothérapie sera proposée. En cas d'échec de la première ligne chez un patient non opérable, les options suivantes sont possibles (Ray-Coquard et al, 2018):

- si le patient est atteint de M/RC-LPS ou léiomyosarcome, la trabectedine peut être proposée mais elle n'est pas remboursée en France;
- si le patient est atteint d'angiosarcome, une combinaison paclitaxel - gemcitabine peut être proposée;
- Si le patient ne présente aucun de ces types histologiques, ou rechute après trabectedine ou paclitaxel - gemcitabine, et si son état clinique le permet, un screening moléculaire lui sera proposé dans le cadre d'un essai thérapeutique - toujours en centre expert. En l'absence d'essai clinique, le pazopanib (un inhibiteur de tyrosine kinase) peut être prescrit puisqu'il améliore, modérément mais significativement la survie sans rechute et la survie globale (Van Der Graaf et al, 2012).

1.3. Place de l'imagerie dans la prise en charge des STM

L'imagerie occupe une place essentielle à chaque étape de la prise en charge diagnostique et thérapeutique. Toutes les modalités d'imagerie sont utilisées, qu'il s'agisse de l'échographie-doppler, du scanner, de la tomographie à émission de positon au ^{18}F-Fluorodesoxyglucose couplée au scanner (^{18}F-FDG-PET/CT) et de l'IRM.

1.3.1. Etape diagnostique: Tri par l'échographie

Le premier examen à réaliser afin de faire le tri devant une masse des parties molles reste une échographie couplée au doppler (Noebauer-Huhmann et al, 2015; Lakkaraju et al, 2009). Elle est l'occasion de recueillir des informations cliniques qui pourront orienter les hypothèses comme: l'âge, d'éventuels facteurs favorisants (neurofibromatose, antécédents d'irradiation, lymphoedème chronique), des antécédents de cancer, la prise d'anticoagulant, la cinétique d'évolution, un contexte traumatique...

L'échographie permet d'éliminer des causes de pseudo-tumeurs ou de tumeurs bénignes (bursite, kyste arthrosynovial, anévrisme artériel ou veineux, névrome de Morton, granulome à corps étranger, lipome superficiel, hernie digestive...).

Dans le cas où un diagnostic alternatif évident au STM n'est pas possible et (1) si l'échographie est bien réalisée par un expert et (2) si la lésion reste superficielle et (3) si (3) la lésion est intégralement vue et (4) si elle mesure moins de 5 cm (voire 10 cm pour les tumeurs intégralement adipeuses homogènes), l'IRM n'est pas obligatoire mais le patient doit dans tous les cas être adressé en centre de référence.

S'il n'y a pas de diagnostic alternatif évident au STM et si une seule des 4 conditions ci-dessus n'est pas remplie, l'IRM avec injection IV de chélates de Gadolinium doit être réalisée et le patient adressé en centre de référence.

Dans ce dernier cas de figure, l'échographie et l'IRM auront aussi pour fonction de guider les prélèvements biopsiques en radiologie interventionnelle.

1.3.2. Staging initial

1.3.2.1. Bilan local par IRM

L'IRM peut aussi servir à écarter certains diagnostics de bénignité (hématome, synvovite villo-nodulaire, élastofibrome, chondrocalcinose...).

Dans le cas d'une masse des parties molles sans diagnostic alternatif de bénignité évident, seule l'IRM permet de recueillir les éléments suivants, descripteurs essentiels de la décision thérapeutique:

- Les descripteurs anatomiques :

(1) ses rapports anatomiques;

(2) ses dimensions dans les 3 plans;

(3) sa profondeur par rapport à l'aponévrose superficielle;

(4) son caractère uni- ou multifocal;

- Les descripteurs de l'architecture tumorale

(5) son signal principal sur les séquences pondérées T1 (ou « T1-*weighted imaging* » [T1-WI]), T1-WI avec suppression du signal de la graisse (ou « *Fat Suppressed* » [FS] - en sachant qu'il existe plusieurs techniques de suppression, chacune avec ses avantages et inconvénients) et T2-WI;

(6) la présence d'un rehaussement intra-tumoral après injection IV de chélates de Gadolinium sur la séquence FS-T1-WI, son intensité et sa répartition;

(7) la matrice tumorale, consistant en l'identification de contingents adipeux, myxoïdes, kystiques, nécrotiques, hémorragiques, fibreux en combinant l'analyse du signal sur les différentes séquences (T1-WI, FS-T1-WI, T2-WI, FS-T2-WI, FS-CE-T1-WI) (Table 1-3). Identifier la présence d'un ou plusieurs de ces contingents au sein d'une tumeur peut aider au diagnostic en restreignant les hypothèses puisqu'il n'existe qu'un certain nombre de STM avec un contingent adipeux (M/RC-LPS, LPS bien différencié et dédifférencié, et LPS pléomorphe) ou avec un contingent myxoïde (M/RC-LPS, MFS, sarcome fibromyxoïde de bas grade; chondrosarcome myxoïde extra-squelettique) ou encore un contingent fibreux (fibrosarcome, MFS, sarcome fibromyxoïde bas grade, sarcome indifférencié pléomorphe ou UPS) (Van Vliet et al, 2009; Fletcher et al, 2013; Crombé et al, 2016)

Table 1-3. Caractérisation de la matrice tumorale des STM par IRM conventionnelle.

Séquence / Signal	Nécrotique	Hémorragique	Kystique	Graisseux	Myxoïde	Fibreux
T1 pré-contraste	-	+	-	+	-	-
T1 Fat Sat pré-contraste	-	+	-	-	-	-
T2	+	+	++	+	++	-
T2 STIR	+	+	++	-	++	-
T1 Fat Sat après injection	-	+	-	-	+	+/-
Soustraction	-	-	-	-	+	+/-

(8) l'importance de l'hétérogénéité sur les séquences T1-WI, T2-WI et FS-CE-T1-WI (cotée : nulle, modérée ou importante) - en particulier, l'hétérogénéité en T2 puisque celle-ci, évaluée de manière semi-quantitative, apparait corrélée au grade FNCLCC, à l'OS et à la MFS (Zhao et al, 2014; Crombé et al, 2019a);

- Les descripteurs de la périphérie tumorale

(9) les berges tumorales, en particulier, si elles sont bien limitées (autrement dit "*pushing*"), focalement mal limitées (< 25% de la circonférence tumorale) ou diffusément mal limitées (≥ 25 de la circonférence tumorale) - en effet, plus les berges sont mal limitées plus cela est à plus haut risque de rechute locale et à distance impactant ainsi l'OS (Nakamura et al, 2017) (Figure 1-5.a);

(10) la périphérie tumorale - 3 caractéristiques sont à rechercher: la présence de rehaussements péritumoraux, la présence d'un oedème péritumoral (c'est-à-dire à distance des bordures apparentes de la tumeur) et la présence de « *tail sign* » ou de rehaussements aponévrotiques d'autant plus s'ils ont une épaisseur ≥ 2 mm. Ces éléments sont associées à de plus forts risques de rechute locale à distance et à la survie globale, et au grade FNCLCC (de manière constante en univarié, plus inconstante en multivarié) (Lefkowitz et al, 2013; Yoo et al, 2014; Zhao et al, 2014; Kikuta et al, 2015; Crombé et al, 2019a) (Figure 1-5.b-d);

(11) la présence d'envahissements vasculo-nerveux (en précisant sur combien de degrés ces structures sont engainées - voire s'il existe un bourgeonnement tumoral dans la lumière vasculaire) ou osseux (en distinguant un contact périosté d'un envahissement cortico-médullaire de continuité) car ceux ci impactent péjorativement la chirurgie et augmentent les risques de rechute locale et métastatique (Panicek et al, 1997; Elias et al, 2003; Holzapfel et al, 2015) (Figure 1-6).

Figure 1-5. Anomalies radiologiques de la périphérie tumorale détectables en IRM corrélées au grade histopathologique et à la survie des patients atteints de STM. **(a)** *MRI growth-pattern* sur des séquences T1 Fat Sat après injection intraveineuse de chélates de gadolinium: *pushing type* (**a.1** - chez un patient atteint d'un liposarcome myxoïde et à cellule ronde de la loge postérieure du mollet gauche), *focal infiltrating* (**a.2** – chez une patiente atteinte de sarcome indifférencié pléomorphe de la loge antérieure de cuisse droite, flèche blanche) et *diffuse infiltrating* (**a.3** – chez un patient atteint d'un sarcome épithéloïde de l'avant bras gauche). (**b**) Œdème péritumoral s'étendant dans les fibres musculaires quadricipitales sus et sous-jacentes d'un sarcome indifférencié pléomorphe (flèches en pointillé). (**c**) Rehaussements péri-tumoraux sus et sous jacents de cette même tumeur (têtes de flèche blanche). (**d**) Rehaussements aponévrotiques épais tumoraux ou *tail sign* (têtes de flèche noires) chez un patient atteint d'un myxofibrosarcome superficiel et aponévrotique de la cuisse droite.

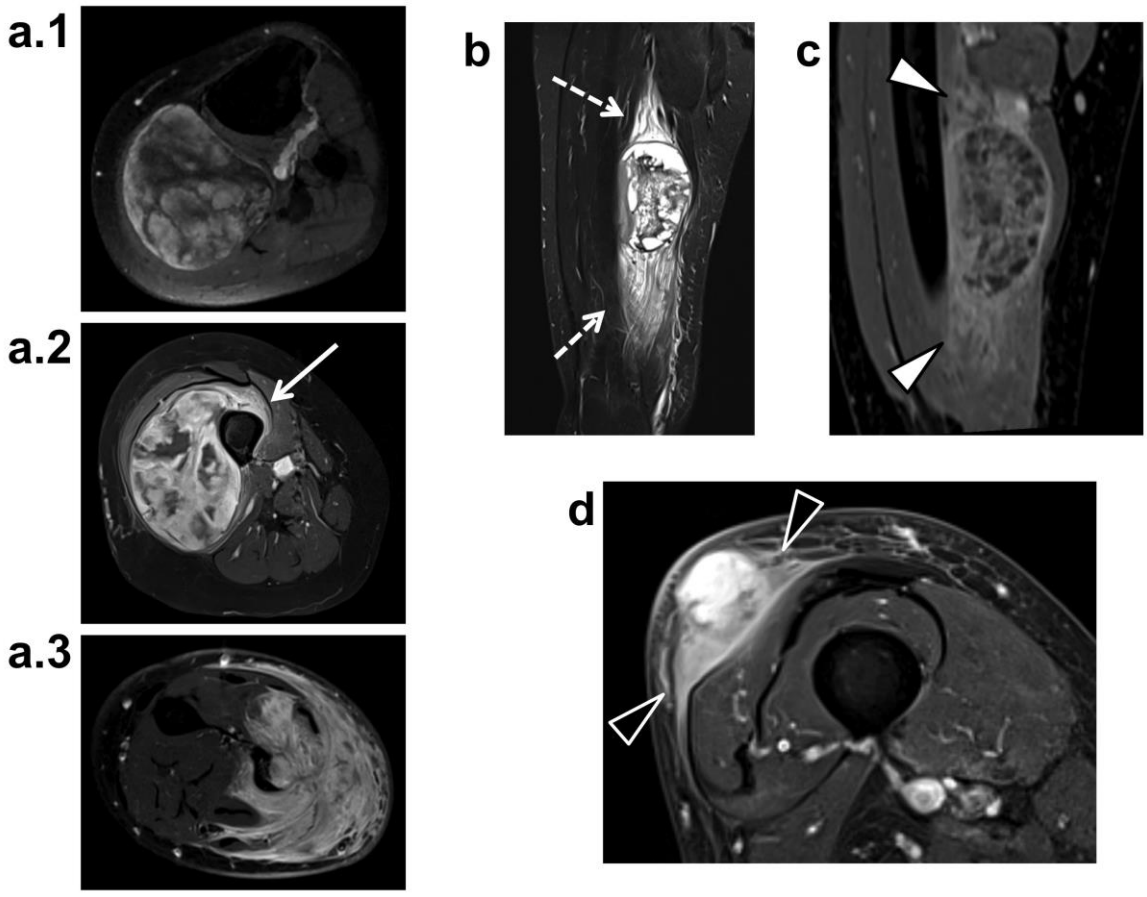


Figure 1-6. Envahissements péjoratifs des structures vasculaires, nerveuses et osseuses évaluables en IRM et impactant la stratégie thérapeutique sur des séquences axiales T1 Fat Sat après injection intraveineuse d'agents de contraste. (**a**) Envahissement du nerf sciatique (rehaussements circonférentiels) chez une patiente atteinte de sarcome indifférenciée pléomorphe de la loge postérieure de cuisse droite (flèche blanche). (**b**) Envahissement endoluminal de la veine fémorale commune chez un patient atteint de synovialosarcome de la région inguinale gauche (flèche blanche en pointillés). (**c**) Envahissement cortico médullaire de la diaphyse fémorale gauche par un sarcome indifférencié pléomorphe de la cuisse gauche (tête de flèche blanche)

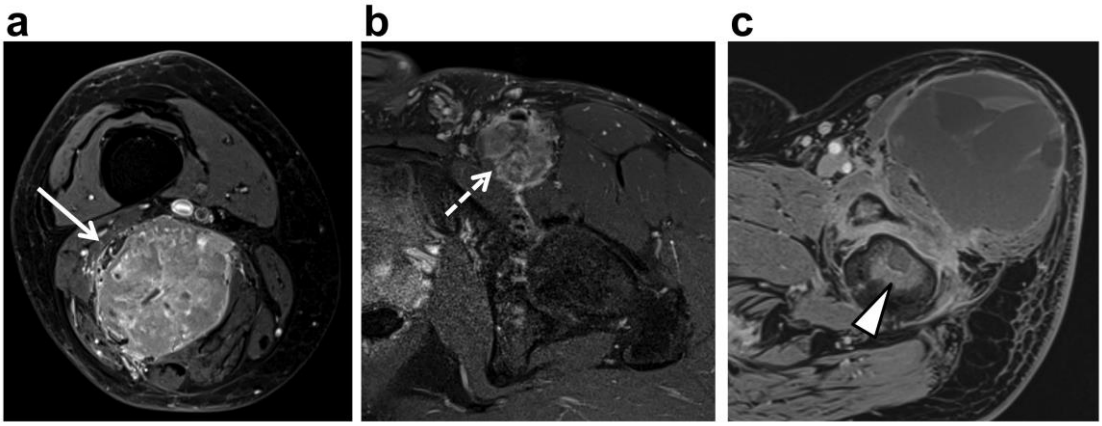


Table 1-4. Protocole standardisé IRM dédié STM à l'institut Bergonié

Séquences	**Remarques**
Coronal ou sagittal T2 STIR TSE	grand FOV sur segment entier de membre
Axial T1 TSE* °	puis FOV centré sur tumeur:
Axial T1 Fat Sat TSE* °	
Axial T2 TSE*	
Axial Diffusion Fat Sat	b0, b400, b800
… Cartographie ADC	
Axial T2 mapping	Multi echo
Axial T1 mapping	Variable flip angles 2° et 15°
Axiale T1 DCE-MRI	TWIST VIBE Dixon dt = 4s, durée 5mn, ldb x 5
… Carte K^{trans}	Alternative T1 Fat Sat GRE rapide 0s, 30s, 70s, 2mn
… Carte AUC_{90s}	
Axiale T1 Fat Sat TSE post-injection* °	
… Soustraction pré-contraste	
+/- 3D T1 GRE DIXON isotropique	optionnel pré-chirurgicale
… Reformations sagittale et coronale	

NOTE. * : les volumes d'acquisition et résolutions spatiales sont identiques pour permettre des corrélations voxel-à-voxel.
° : peut être remplacé par une séquence T1 DIXON TSE sans et avec injection et soustraction des phases WATER (ou IN).
En italique : séquences quantitatives.
En gras : séquences minimales indispensables pour l'interprétation de l'architecture tumorale.

Afin d'analyser l'ensemble de ces points, nous avons développé à l'institut un protocole standardisé pour le *staging* des STM, réalisé en routine depuis novembre 2017 et détaillé dans la Table 1-4.

Ce bilan local exhaustif sert à orienter les prélèvements biopsiques écho-guidés. Mis en parallèle avec les résultats de la biopsie, il sert aux corrélations radio-anatomopathologiques à l'occasion de la réunion de concertation pluridisciplinaire (RCP) sarcomes.

Bien que cela ne figure pas explicitement dans les recommandations ESMO et n'ait pas été validé prospectivement, l'IRM peut servir à corriger les sous-estimations du grade micro-biopsique comme cela a été proposé par notre groupe (Crombé et al, 2019a). A titre d'exemple, si une tumeur présente un grade FNCLCC 2 sur microbiopsie avec un score de nécrose coté à 0 alors que l'IRM met en évidence un volumineux contingent nécrotique, il semble licite de corriger le grade vers un grade 3.

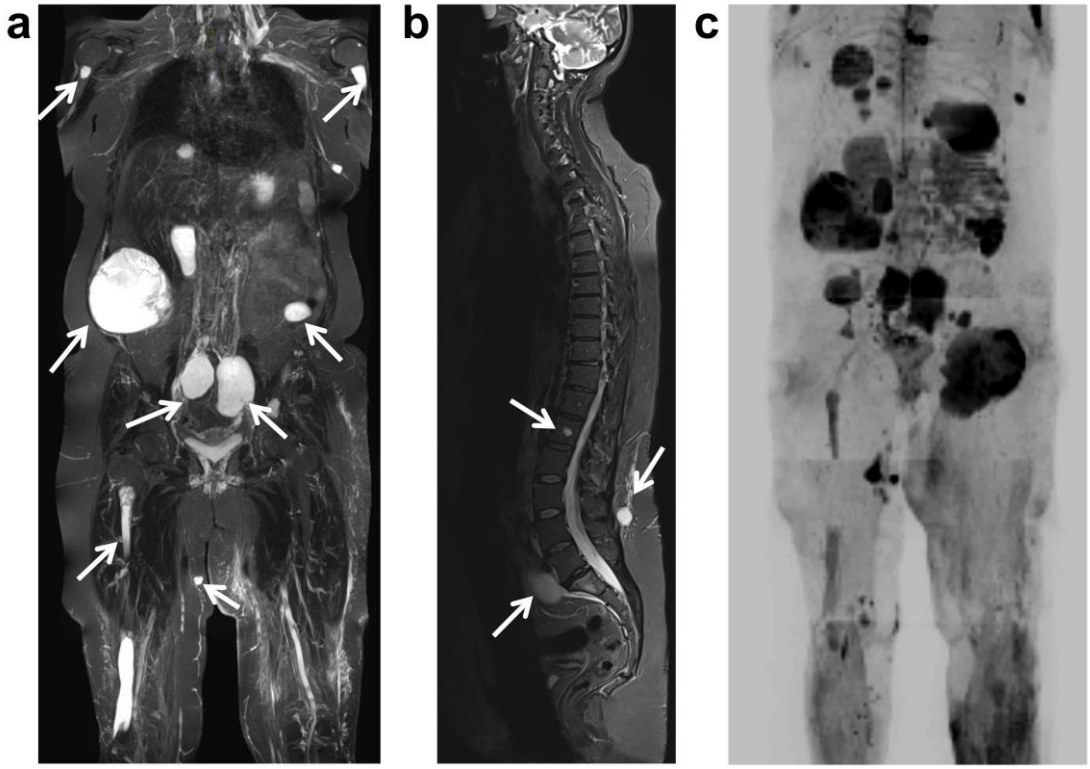
Au terme de ce bilan IRM, les différentes options thérapeutiques seront discutées. Ce *staging* local est indispensable à la planification chirurgicale (De la Hoz Polo et al, 2017). Enfin, si la tumeur est localement avancée et nécessite une chirurgie trop délabrante pour être carcinologique, à risque de séquelles fonctionnelles, il pourra être décidé d'effectuer une NAC pour réduire son extension loco-régionale et faciliter son opérabilité (Italiano et Stoeckle, 2018). Enfin, cette IRM servira de référence (ou *baseline*) si la NAC est prescrite.

1.3.2.2. Bilan régional et à distance

L'objectif de ce bilan est de détecter une maladie métastatique synchrone du diagnostic de la tumeur primitive. La dissémination métastatique principale étant pulmonaire, les dernières recommandations ESMO indiquent de réaliser au minimum un scanner thoracique non injecté (Casali et al, 2018), avec certaines variations pour des STM atypiques par leur pattern de dissémination métastatique. Les M/RC-LPS sont ainsi préférentiellement évalués par IRM corps entier puisque les métastases de signal myxoïde sont facilement identifiable en séquences T2 avec suppression de signal de la graisse et par IRM de diffusion, permettant une détection plus précoce de rechute asymptomatique (Figure 1-7) (Stevenson et al 2016; Gorelik et al 2018; Gouin et al, 2019). Le bilan à distance des ASPS peut inclure une imagerie

encéphalique en raison de la fréquence de métastases cérébrales au diagnostic (Portera et al, 2001; Crombé et al, 2019b).

Figure 1-7. Bilan d'extension et d'évaluation de la réponse aux traitements par IRM corps entier chez une patiente atteinte de liposarcome myxoïde et à cellule ronde. (**a**) Séquence coronale T2 STIR turbo spin echo sur le corps et les cuisses. (**b**) Séquence sagittale T2 STIR turbos spin echo sur le rachis entier. (**c**) Reconstruction volumique de la TRACE de la diffusion corps entier. Les métastases de liposarcome myxoïde et à cellule ronde sont facilement identifiables en diffusion et en T2 STIR (flèches blanches), alors qu'elles ont un faible contraste au scanner comparativement aux tissus mous sains environnant. Elles ont une propension à se localiser dans les tissus mous, séreuses, ou l'os.



En pratique, un scanner thoraco-abdomino-pelvien réalisé d'emblée après injection IV de produit de contraste iodé avec acquisition au temps portal voire un PET/CT injecté est de plus en plus souvent réalisé.

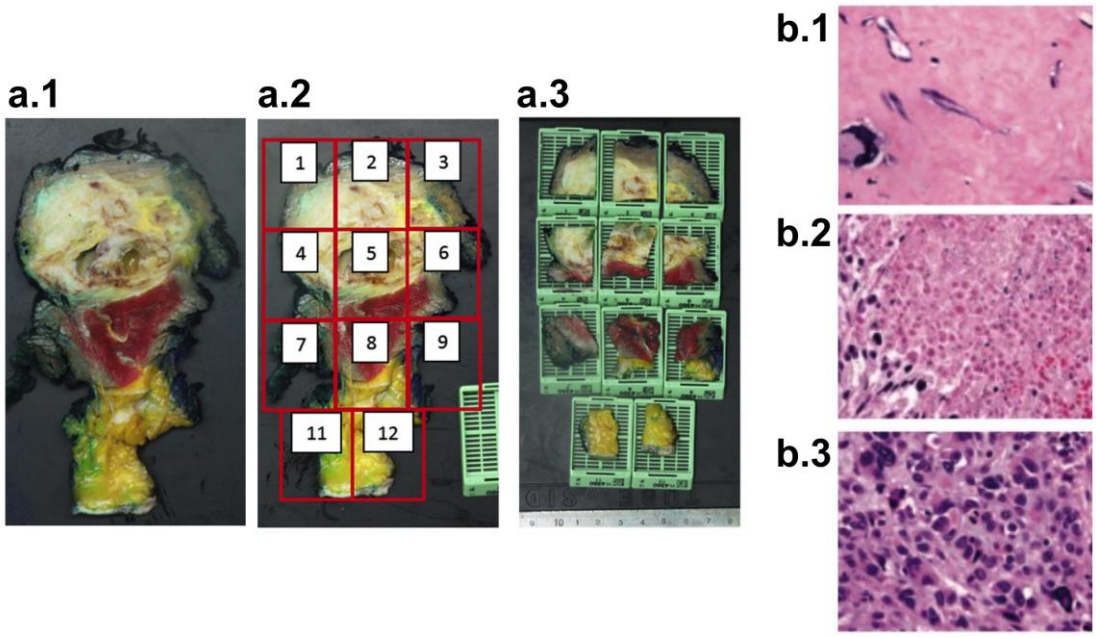
1.3.3. Evaluation de la réponse au traitement néoadjuvant

1.3.3.1. Gold standard de la réponse pour les STM?

La référence pour valider des critères de réponse radiologique peut être la survie spécifique à la maladie, ou la survie sans apparition de nouvelles lésions.

Une autre référence "intermédiaire" peut être utilisée: la réponse histologique dans la situation néoadjuvante avec chirurgie curatrice (Waldermann et al, 2016). Cette réponse histologique est évaluée sur la pièce opératoire entière après la NAC. L'anatomopathologiste va échantillonner régulièrement la tumeur (environ 1 bloc emparaffiné par cm^3) et estimer sur une lame HES issue de chacun de ces bloc les pourcentages de cellularité résiduelle, de fibrose (+/- cellules histiocytaires) et de nécrose (la somme des 3 pourcentages faisant 100%), permettant d'en inférer un pourcentage globale de ces 3 contingents à l'échelle de la tumeur entière (Figure 1-8).

Figure 1-8. Principes de l'évaluation de la réponse histologique de STM aux traitements néo-adjuvants. La pièce opératoire est orientée par un code couleur et des fils (**a.1**) puis découpée selon un schéma clairement établi par l'anatomopathologiste qui quadrille régulièrement toute la tumeur (**a.2**). Chaque cassette comprend ainsi une portion de tumeur représentant environ $1cm^3$ de tumeur et une lame HES sera réalisée par bloc (**a.3**). Pour chacune de ces lames, l'anatomopathologiste estime les pourcentages de fibrose (**b.1**), de nécrose (**b.2**) et de cellules tumorales colorées, viables (**b.3**), permettant d'estimer ces pourcentages globaux, à l'échelle de la tumeur entière.



Sur la base d'une cohorte de 150 patients, Cousin et al ont montré que la réponse histologique (avec un seuil de 10% de cellularité résiduelle pour définir bon vs.

mauvais répondeur) était effectivement corrélée à la survie (Cousin et al, 2017). Une actualisation en décembre 2019 de cette base de donnée (désormais de 176 patients) a montré qu'un seuil de 5% permettait d'identifier les patients qui ne rechuteront pas.
La réponse histologique peut donc être considérée comme un critère d'évaluation intermédiaire de la réussite de la chimiothérapie, certes imparfait mais néanmoins informatif.

1.3.3.2. Critères 'classiques': WHO et RECIST

Dans le cas où un patient présenterait un STM localement avancé ou métastatique avec indication d'un traitement systémique, il faut être en mesure de suivre l'efficacité de ce traitement de la manière la moins invasive possible afin de valider sa poursuite ou de l'ajuster précocement, voire de changer de ligne. L'évaluation de la réponse au traitement peut se faire cliniquement, et/ou via des biomarqueurs biologiques, et/ou par l'imagerie.
Les premiers critères d'évaluation radiologique de la réponse thérapeutiques ont été proposés en 1979 par l'Organisation Mondiale de la Santé (OMS, ou « *World Health Organization* » [WHO]). Ces critères WHO proposaient de sommer le produit des diamètres de la plus grande surface tumorale d'un nombre non limité de cibles tumorales et de suivre les variations relatives de cette somme. Selon ce pourcentage de variation, quatre groupes ont été définis et les noms de ces groupes se sont maintenus lors des développements ultérieurs de critères de réponse, à savoir:

- Réponse complète (ou « *complete response* » [CR]) quand toutes les localisations secondaires ont totalement disparu en imagerie
- Réponse Partielle (ou « *partial response* » [PR]) quand les métastases ont régressé jusqu'à un certain seuil sans disparaitre
- Progression (ou "*progressive disease*", PD) quand les localisations secondaires ont augmenté au-dessus d'un certain seuil ou quand une ou des nouvelle(s) métastase(s) sont apparue(s)
- Maladie stable (ou « *stable disease* » [SD]) quand les seuils de PR et PD n'ont pas été atteints, en l'absence de nouvelle lésion.

Les limites des critères WHO telles l'absence de définition d'un nombre maximal de cibles ou le fait qu'un produit des diamètres tend à produire de forts pourcentages de variations, ont conduit à développer d'autres critères notamment RECIST (pour « *Response Evaluation Criteria In Solid Tumors* ») en 2000, actualisés en 2009 (v1.1) (Miller et al, 1981; Therasse et al, 2000; Eisenhauer et al, 2009).

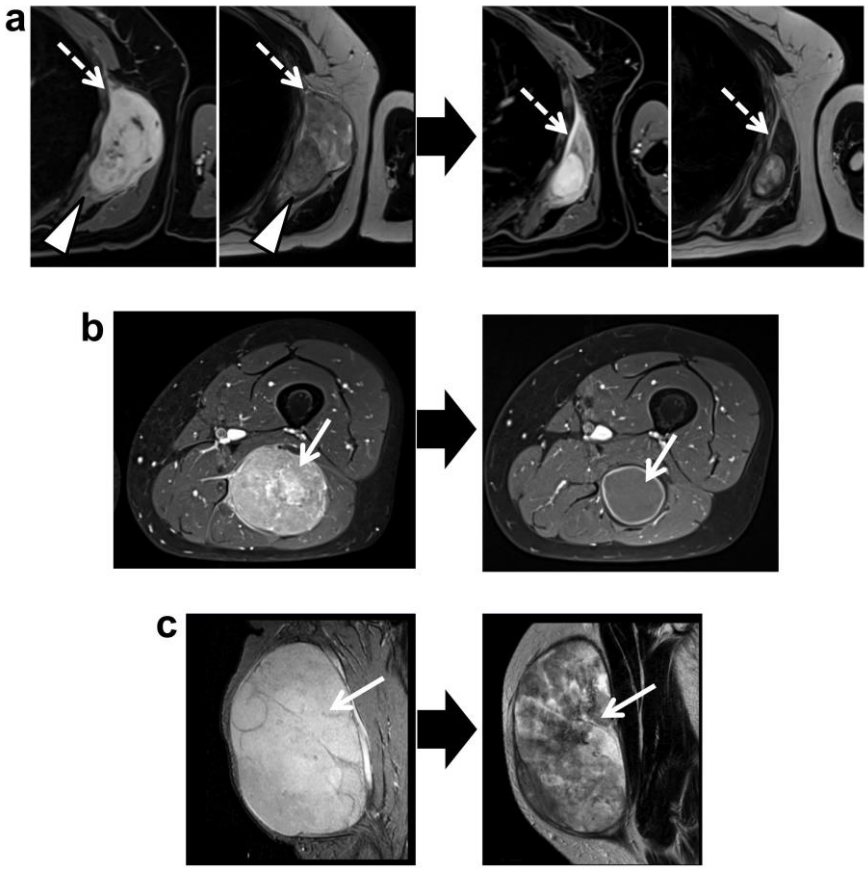
Dans les essais et la pratique cliniques, les critères de référence sont RECIST v1.1 qui définissent 3 catégories de lésions: cibles mesurables (au nombre de 5 au maximum avec 2 lésions par organe au maximum), non cibles mais mesurables et non cibles non mesurables. Ces lésions sont définies sur l'examen baseline et ré-analysées à chaque évaluation. Au lieu d'un produit de diamètre, RECIST se base sur une mesure unidimensionnelle (: le plus grand diamètre). Les seuils pour définir une PR ou une PD (-30% et +20%, respectivement) ont été définis par l'EORTC sur un premier groupe de 569 patients et confirmés sur plus de 6500 patients, issus de plusieurs essais cliniques.

Les critères RECIST v1.1 présentent les avantages d'être relativement simples, évaluables dans un délai acceptable cliniquement, de proposer un langage commun aux oncologues et radiologues, et d'apporter une évaluation relativement objective de l'efficacité des traitements dans le cadre des essais thérapeutiques et de comparer les résultats des publications scientifiques.

1.3.3.2. Limites des critères RECIST v1.1

Cependant, les critères RECIST v1.1 sont des critères généraux, difficiles à employer pour des cancers sans cible mesurable (part exemple, carcinose péritonéale, pleurésie, méningite, infiltration osseuse diffuse...), sans modulations selon les traitements et les types histologiques tumoraux et ne prenant pas en compte les modifications de l'architecture tumorale. Ainsi, sous certains traitements (par exemple anti-angiogéniques), nous pouvons assister à une forte réduction du rehaussement sur les acquisitions post-injection des tumeurs, traduisant une chute de la vascularisation tumorale à dimensions constantes. Dans le cas des STM traités par NAC et suivis par IRM, nous observons des cas d'évolutions fibrosante et/ou nécrosante extensive, sans cellules tumorales résiduelles sur la pièce opératoire finale - traduisant une excellente réponse, tandis que les dimensions restent stables. A l'inverse, nous observons aussi des cas de réponses dissociées intra-tumorales de STM hétérogènes où un contingent involue tandis qu'un autre contingent résiste au traitement, entrainant cependant une variation négative des dimensions totales de la tumeur (Figure 1-9).

Figure 1-9. Patterns de réponse des STM à la chimiothérapie néo-adjuvante susceptibles d'induire en erreur RECIST v1.1. (**a**) Evolution dissociée intra-tumorale chez une patiente atteinte de sarcome indifférencié pléomorphe de la région scapulaire gauche (séquences axiales T2 et T1 Fat Sat injectée). La portion antérieure (flèches blanches en pointillés) devient fibreuse et se réduit tandis que la portion postérieure se modifie peu sous chimiothérapie (têtes de flèche blanches). (**b**) Evolution nécrotique complète chez une patiente atteinte de synovialosarcome de la loge postérieure de cuisse gauche – avec régression de l'envahissement du nerf sciatique. (**c**) Evolution fibrosante extensive chez un patient atteint de liposarcome myxoïde et à cellules rondes de la cuisse (flèches blanches). Le cas (a) peut conduire à tort à un statut stable selon RECIST alors qu'il persiste de nombreuses cellules tumorales viables sur la pièce opératoire finale post-traitement.



1.3.3.1. Autres critères

Pour ces raisons, des alternatives ont été proposées à l'utilisation de RECIST v1.1 en prenant en compte les variations de rehaussement et de taille (produit des diamètres) de manière conjointe (critères Choi, critères Choi modifiés), ou les variations uni- ou bidimensionnelles de la portion se rehaussant après injection (critères mRECIST [pour « *modified RECIST* »] et EASL [pour *« European Association for the Study of Liver* »], respectivement) (Figure 1-10 et Table 1-5) (Bruix et al, 2001; Choi et al, 2007; Benjamin et al, 2007; Lencioni et al, 2010; Tirkes et al, 2013)

Figure 1-10. Mesures nécessaires pour l'évaluation de la réponse selon les critères de réponse radiologiques conventionnels. RECIST se base sur des variations de plus grand diamètre total; Cheson/WHO sur des variation de produit entre le plus grand diamètre et plus grand diamètre transverse; Choi prend en compte les variations des densités moyennes de la plus grande surface et du plus grand diamètre transverse; mRECIST se base sur les variations du plus grand diamètre de la portion rehaussée après injection; EASL sur les variations du produit entre le plus grand diamètre et le plus grand diamètre transverse de la portion rehaussée (voir Table 1-5)

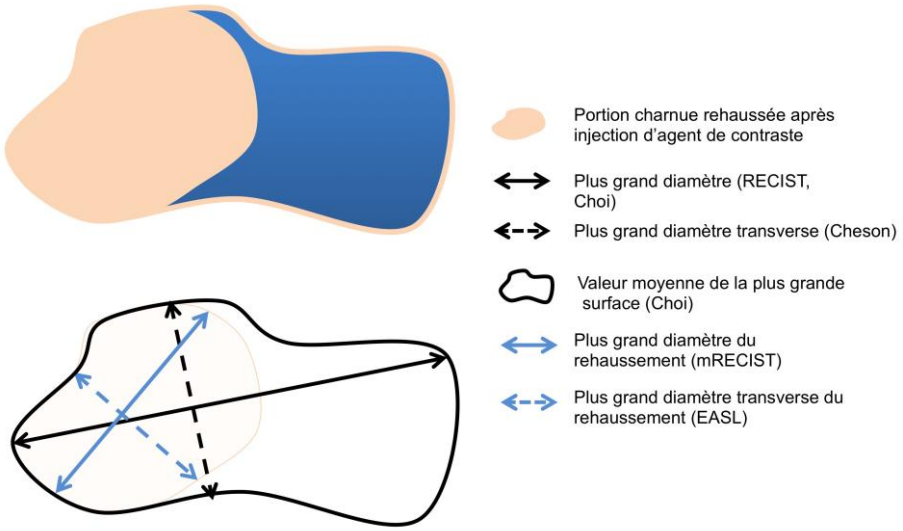


Table 1-5. Définitions des principaux critères radiologiques d'évaluation de la réponse au traitement.

	Complete response	Partial response	Stable disease	Progressive disease
RECIST 1.1	Disappearance of all target lesions	≥ 30% decrease in the sum of the diameters of target lesions	Neither partial response nor progressive disease	◆ ≥20% size increase ◆ or new disease
Modified RECIST	Disappearance of contrast-enhancement all lesions	≥ 30% decrease in the sum of the diameters of contrast-enhancement of target lesions	Neither partial response nor progressive disease	◆ ≥20% size increase in contrast-enhancement of the target lesions ◆ or new disease
EASL	Disappearance of contrast-enhancement all lesions	≥ 50% decrease in the sum of the bidimensional surface of contrast-enhancement of target lesions	Neither partial response nor progressive disease	◆ ≥ 25% increase in the sum of the bidimensional surface of contrast-enhancement of target lesions ◆ or new disease
Cheson / WHO	Disappearance of all target lesions	≥ 50% decrease in the sum of the bidimensional surface of target lesions	Neither partial response nor progressive disease	◆≥ 25% increase in the sum of the bidimensional surface of target lesions ◆ or new disease
Choi	Disappearance of all target lesions	◆ ≥10% decrease in the sum of the diameters of target lesions ◆ or ≥15% decrease in the tumor contrast-enhancement	Neither partial response nor progressive disease	◆ ≥10% increase in tumor size and does not meet the criteria of PR in tumor density, ◆ or new disease
Revised Choi	Disappearance of all target lesions	◆ ≥10% decrease in the sum of diameters of target lesions and ≥15% decrease in the tumor contrast-enhancement ◆ or in patients with no lesions suitable for density analysis, ≥30% decrease in the sum of diameters of target lesions	Neither partial response nor progressive disease	◆ ≥10% increase in tumor size ◆ or new disease

1.3.3.2. Cas des STM

Selon notre expérience, les critères RECIST v1.1 sont limités pour détecter les évolutions précoces des STM puisque l'essentiel des patients en situation néoadjuvante vont être classés en maladie stable (80%) lors des évaluations à 2 cures de chimiothérapie cytotoxique (Crombé et al, 2019c).

En reprenant les évaluations selon RECIST après 2, 4 et 6 cures de chimiothérapie cytotoxique classique et les données de survie d'une cohorte EORTC de patients très majoritairement métastatiques, Grunwald et *al.* (2016) ont montré que seule l'absence de progression selon RECIST à chacune de ces évaluations était significativement corrélée à la survie globale. Cependant, pour ces 3 temps d'évaluation, les courbes de Kaplan-Meier indiquaient que la majorité des patients des groupes non-PD (: CR, PR, SD) décéderont. Cette classification des patients groupes ne permettait pas d'identifier clairement les longs survivants (environ 25% des patients à 3 ans dans cette étude). Ces résultats montrent qu'il existe bien une association entre variation dimensionnelle et efficacité de la chimiothérapie (une progression était pratiquement toujours associée au décès), mais que le groupe non progressif dimensionnellement mélange des patients aux devenirs opposés.

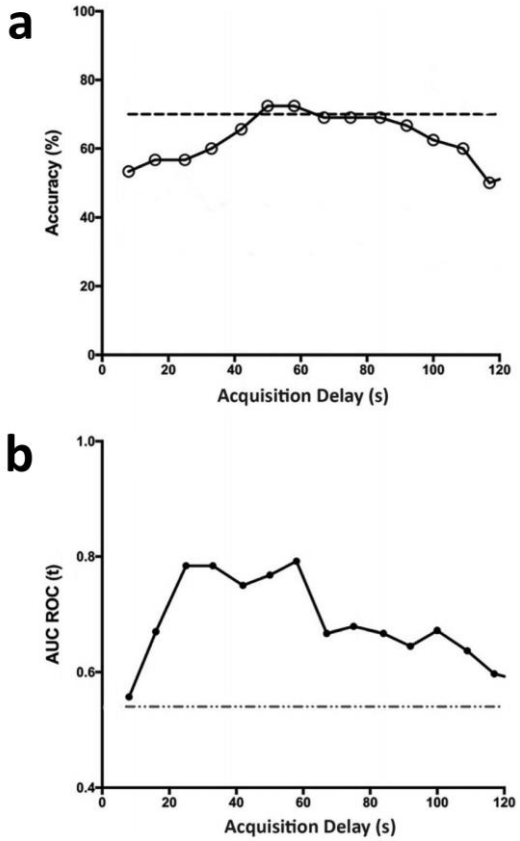
A notre connaissance, seuls les critères Choi et Choi modifiés pour l'IRM ont été utilisés pour approfondir l'évaluation de la réponse des patients atteints de STM (à l'exception d'une étude se focalisant sur la réponse de métastases hépatiques selon les critères EASL et mRECIST) (Chapiro et al, 2015).

Dans le contexte néoadjuvant, deux études sur des effectifs de moins de 30 patients ont montré un avantage des critères Choi et de la prise en compte de la variation du rehaussement tumoral pour prédire la réponse histologique, selon des seuils variables de cellularité résiduelle.

Selon Stacchiotti et *al.,* une réponse objective selon les critères Choi présenterait une plus grande sensibilité pour détecter les bons répondeurs comparativement à la réponse objective selon RECIST v1.1 (82.4% vs. 41.2% pour un seuil de 10% de cellularité résiduelle définissant la réponse histologique) (Stacchiotti et al, 2009). Une étude de notre groupe a montré que les critères Choi étaient effectivement plus performant pour identifier précocement les bons répondeurs histologiques, mais aussi que ces performances étaient accrues en utilisant un seuil de variation de rehaussement à un délai optimal post-injection de chélates de Gadolinium et propre au

moment de l'évaluation (ici 2 cures) et aux STM au lieu des seuils définis arbitrairement par Choi et *al*. - en l'occurrence -30.5% de rehaussement au temps portal (Figure 1-11) (Choi et al, 2007; Crombé et al, 2019c).

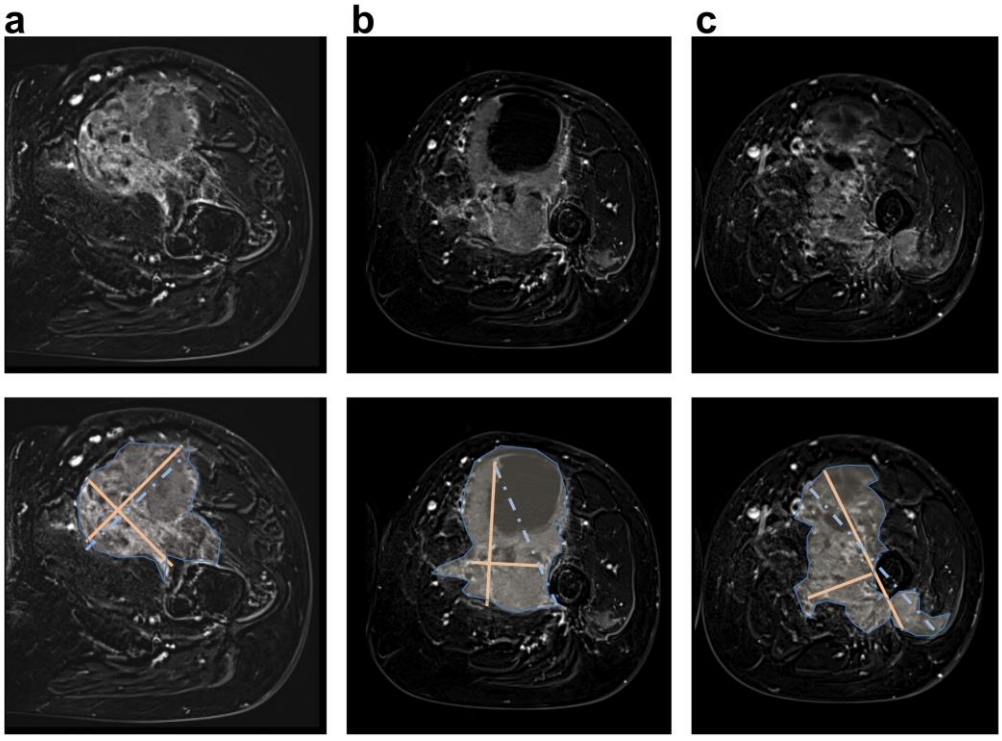
Figure 1-11. Influence du délai dt entre injection et acquisition de l'imagerie injectée sur les performances des critères Choi adaptés à l'IRM (et sur volumes entiers de STM). A chacun des délais post injection multiples de 8, les variations de rehaussement depuis l'imagerie baseline (pour un même délai) sont calculées permettant d'estimer le statut de la réponse selon Choi. Les performances de ces critères Choi(dt) pour prédire la réponse histologique définie comme < 10% de cellules viables sur la pièce opératoire finale sont évaluées selon l'*accuracy* (**a**) et l'*AUROC* (**b**). Le délai dt présentant les meilleures performances est estimé à environ 60 secondes. Extrait de *Crombé et al, 2019c*.



La seule étude évaluant les critères Choi en situation métastatique, publiée par le Groupe Sarcome Français, concernait une chimiothérapie à base de trabectedine, réservée hors AMM en seconde ligne (Taieb et al, 2015). Selon, les auteurs, l'intérêt de ces critères résiderait dans la distinction entre les fausses PD selon RECIST - phénomène qui apparait rare sous chimiothérapie à base d'anthracyclines. Cet article met surtout en avant et indirectement la plus grande difficulté à évaluer ces critères

Choi par rapport à RECIST v1.1 puisque qu'ils n'étaient pas évaluables chez 31.3% des patients. Notre expérience à l'institut Bergonié confirme la difficulté à évaluer de manière reproductible et fiable ces critères dans le cas de STM d'emblée nécrotique avec des rehaussements intra-tumoraux aux contours géographiques complexes, des difficultés à évaluer visuellement la coupe avec la plus grande surface du fait de formes tumorales elles-aussi complexes. Enfin, pour les critères mRECIST prenant en compte le plus grand axe des composantes rehaussées, il peut être difficile de choisir les limites des zones rehaussées versus non rehaussées d'autant plus sous traitement.
La Figure 1-12 illustre certains de ces cas compliqués d'évaluations de critères basés sur les rehaussements.

Figure 1-12. Exemple illustratif des difficultés limitant la reproductibilité de l'évaluation de la réponse selon les critères radiologiques conventionnels du fait des complexités de formes et d'architecture des STM. Soustraction de séquences T1 Fat Sat après - avant injection d'agent de contraste à 3 niveaux (de haut en bas) chez une patiente atteint de sarcome indifférencié pléomorphe de haut grade de la cuisse gauche. Mesures possibles selon 3 radiologues. (**a**) Montre le plus important rehaussement moyen (162 UI), (**b**) le plus important produit des surfaces (53.90 cm^2) et (**c**) le plus grand diamètre total et des portions rehaussées (110 mm). Trait bleu pointillé = plus grand diamètre (pour RECIST); trait orange = diamètres du rehaussement (pour mRECIST (le plus grand) et EASL (le produit des 2)); surface = recueil de la valeur moyenne du rehaussement (pour Choi)



1.3.4. Surveillance

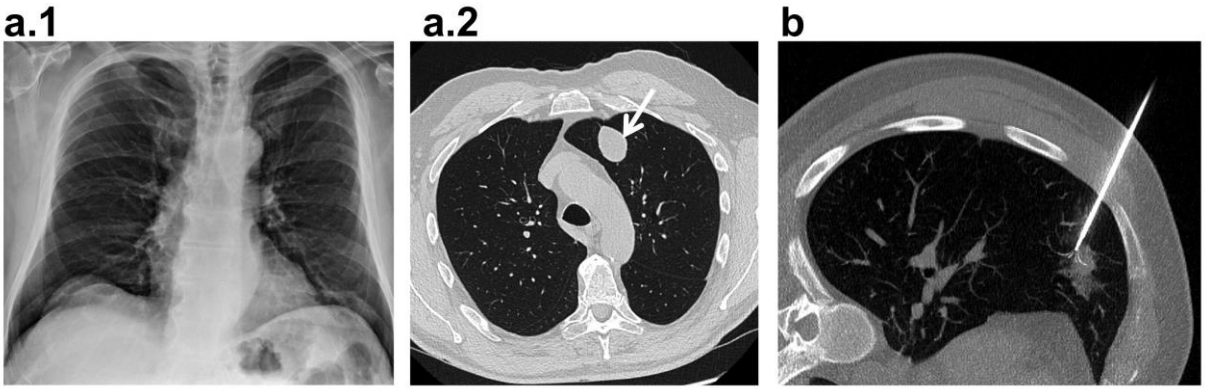
Relativement peu d'études se sont intéressées aux modalités de surveillance des STM après la fin des traitements initiaux. Les recommandations actuelles, appliquées pour les patients de l'institut Bergonié, préconisent une évaluation minimale par examen clinique spécialisé et une radiographie thoracique trimestrielle pendant 3 ans puis bisannuelle jusqu'à 5 ans puis annuelle (Casali et al, 2018).

Si la réalisation d'IRM de surveillance pour le bilan local et le scanner thoracique pour le bilan à distance pourraient servir à détecter les rechutes précocement, ces modalités d'imagerie en routine n'ont pas démontré leur valeur ajoutée. En effet, selon Rothermundt et *al.,* les rechutes locales sont pratiquement toujours détectées par le patient lui même ou le clinicien. Dans leur propre série de 174 patients, une seule rechute a été détectée par IRM (Rothermundt et al, 2014). Pareillement, dans une série de 124 patients suivis par 663 IRM planifiées, seules 2 des 11 rechutes locales ont été détectées par IRM - les autres l'ayant été cliniquement (Labarre et al, 2009).

Enfin, dans un essai contrôlé randomisé ayant inclus 500 patients avec sarcomes osseux et des tissus mous des membres, Puri et *al.* ont montré la non-infériorité d'une surveillance par radiographies simples par rapport au scanner thoracique en terme de survie globale et sans rechute (Puri et al, 2014).

Nous pouvons cependant argumenter que la détection à des stades précoces des métastases notamment pulmonaires ou des tissus mous, peut rendre plus facilement possibles des gestes d'ablathermie (type radiofréquence pulmonaire ou cryoablation), moins morbides et invasives que la chirurgie et qui peuvent être répéter plusieurs fois sans altérer la fonction respiratoire (Figure 1-13) (Nakamura et al, 2009; Palussière et al, 2011; Crombé et al, 2016). De plus, les études basées sur l'IRM n'incorporaient pas de séquences autres que structurales alors que les séquences de diffusion et dynamique de perfusion pourraient améliorer la distinction entre remaniements inflammatoires post-thérapeutiques et rechute locale.

Figure 1-13. Identification et prise en charge de la rechute métastatique pulmonaire des STM. (**a.1**) Radiographie thoracique de face de surveillance chez un patient atteint de sarcome indifférencié pléomorphe traité, sans anomalie suspecte. (**a.2**) Le scanner réalisé 5 semaines plus tard montre une métastase du segment ventral du lobe supérieur gauche possiblement masquée par la superposition avec la crosse de l'aorte et de plus petite taille sur la radiographie. (**b**) Traitement par radiofréquence d'une métastase de STM.



1.4. Limites actuelles de l'imagerie des sarcomes et axes d'évolution

1.4.1. Limites

Au terme de ce panorama introductif, les limites suivantes concernant l'imagerie des STM doivent être soulignées:

(1) à la différence des réseaux cliniques et anatomopathologiques, il n'existe pas de réseau ou de database d'imagerie des STM - bien que des initiatives de projets collectifs autour de l'imagerie émergent à l'échelle française au sein du Groupe Sarcome Français et que des réflexions s'organisent pour intégrer des items radiologiques dans la CONTICABASE ;

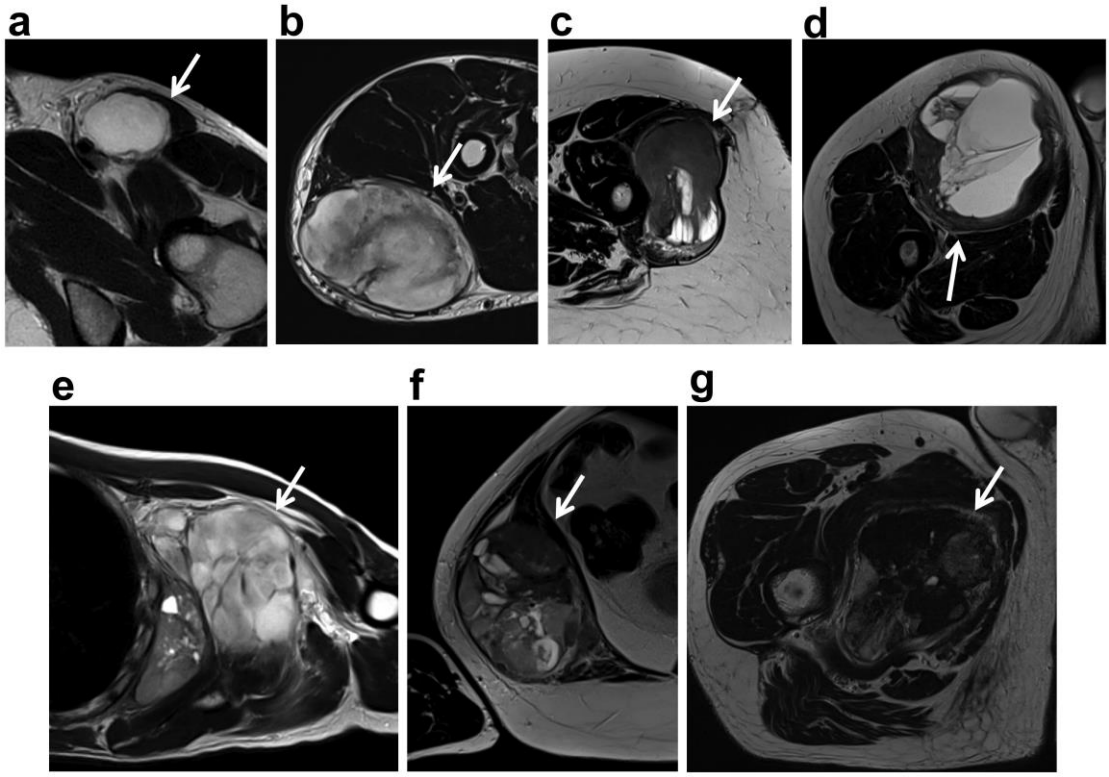
(2) les protocoles IRM ne sont pas standardisés parfois même entre radiologues d'un même centre, bien que des initiatives aient été proposées pour les uniformiser notamment dans des revues oncologiques sous l'initiative de l'« *European Organization for Research and Treatment of Cancer* » (EORTC) - malheureusement peu relayées dans la littérature radiologique (Messiou et al, 2016).

Dans les faits, la grande majorité des bilans d'imagerie initiaux sont réalisés hors centres expert et non systématiquement actualisés une fois que le patient a intégré un parcours de soin en centre de référence sarcome, en particulier si le patient n'est pas inclus dans un essai, pour des raison de couts, de disponibilité de machines, d'effectifs médicaux/paramédicaux, ou de sensibilités variables des médecins non-radiologues à l'importance de la radiologie. A titre d'exemple, sur 326 patients adultes atteints de STM traités en 2007 et 2015 par chirurgie première recensés sur l'institut Bergonié dans la CONTICABASE, 175/326 (53.7%) n'avaient pas eu d'IRM avec chirurgie hors centre, 15 sur les 151 restant (10%) avaient eu une IRM mais sans injection IV de chélates de gadolinium et 8/151 (5.3%) avaient une IRM de qualité insuffisante (tumeur non entièrement inclus dans le volume d'acquisition, bandes d'artefacts sur la tumeur... etc.) (Crombé et al, 2019a).

(3) En dehors des mesures de diamètre ou de surface, nous pouvons remarquer que la quasi-totalité des items de l'évaluation radiologique sont d'ordre qualitatif nominal ou ordinal. Cette évaluation repose donc sur la subjectivité et l'expérience des radiologues avec un risque de reproductibilité insuffisante en intra- et inter-radiologues. A titre d'exemple, l'importance de l'hétérogénéité intra-tumorale sur les séquences pondérées T2 a été corrélée au grade, à la MFS et à la PFS des STM et notamment des M/RC-LPS mais sa reproductibilité en intra et inter- observateur n'excède pas un Kappa pondéré de 0.70 (Zhao et al, 2014; Gimber et al, 2017; Crombé et al, 2019a). La Figure 1-14 illustre ces difficultés. Il apparait donc essentiel de développer des outils pour rendre plus quantitatives et objectives nos variables d'intérêt d'origine radiologique.

(4) Enfin, l'évaluation radiologique actuelle repose sur une imagerie structurale, c'est-à-dire sur des images reflétant l'anatomie du patient, les rapports et architecture tumorale et non son métabolisme ou ses propriétés de néoangiogenèse.

Figure 1-14. Difficultés à évaluer l'hétérogénéité intra-tumorale et à caractériser la texture intra-tumorale. Exemples de 8 STM classés selon l'importance de leur hétérogénéité intra-tumorale sur des séquences axiales T2 turbo spin echo (flèches blanches). Les cas (**c**) et (**d**) montrent des tumeurs avec des bourgeons charnus homogènes mais des zones kystiques/nécrotiques. Le cas (**e**) montre une tumeur avec des lobules de texture différente mais assez homogène au sein d'un même lobule.



1.4.2. Vers une imagerie plus quantitative et multimodale?

Il existe en imagerie médicale et en médecine nucléaire des techniques permettant d'estimer certaines propriétés biologiques des tumeurs et qui pourraient servir de biomarqueurs d'imagerie. Un biomarqueur au sens large correspond à un indicateur de phénomènes physiologiques biologiques, ou de processus pathologiques, ou de l'exposition à une intervention (notamment thérapeutique) et qui nécessite de remplir plusieurs conditions pour être validé, c'est-à-dire: avoir démontré lors d'études précliniques sa corrélation avec le phénomène dont il se veut le reflet, être quantifiable, reproductible, fiable, répétable et standardisé. A ce titre la « *Quantitative Imaging Biomarker Alliance* » (QIBA) et l' « *European Imaging Biomarkers Alliance* » (EIBALL) sont des organisations similaires rattachées aux sociétés nord-

américaine et européenne de radiologie afin de promouvoir de manière multidisciplinaire le développement de biomarqueurs. Nous allons ici brièvement exposer les modalités d'imagerie dites « avancées » ou « fonctionnelles » et les avancées qu'elles ont déjà pu permettre dans le cadre des STM (Ahlawat et al, 2019).

1.4.2.1. ^{18}F-FDG-PET/CT

La TEP consiste à mesurer une activité métabolique d'un organe ou d'une tumeur via les émissions produites par les positons d'un traceur radioactif. Dans le cas des STM, seul le ^{18}F-FDG a été utilisé cliniquement à notre connaissance. Le principe sous-jacent à l'utilisation de ce traceur est que les cellules cancéreuses produisent très majoritairement leur énergie via la glycolyse anaérobie puis la fermentation d'acides lactiques, nommé « effet Warburg » (Warburg et al, 1956). Ce sucre marqué va être capté, phosphorylé et s'accumuler (: « fixer ») dans les cellules cancéreuses, les rendant radioactives et détectables par la caméra TEP. Nous employons la valeur de fixation normalisée par le poids (SUV pour « *Standardized Uptake Value* ») ou la valeur de fixation normalisée par la masse maigre (SUL pour « *SUV corrected fo Lean body mass* ») pour mesurer le degré de fixation du ^{18}F-FDG dans la tumeur. Nous obtenons ainsi une information quantitative sur le métabolisme tumoral, mais dépendante de nombreux facteurs liés aux patients, aux techniques d'acquisition et à la machine, et au post-traitement des images.

La place du ^{18}F-FDG-PET/CT, éventuellement couplé à un scanner réalisé après injection IV de produit de contraste iodé (: « combitep »), par rapport au scanner simple n'est pas consensuelle ni pour le bilan local ni pour le bilan d'extension, mais plusieurs études encouragent son plus large recours.

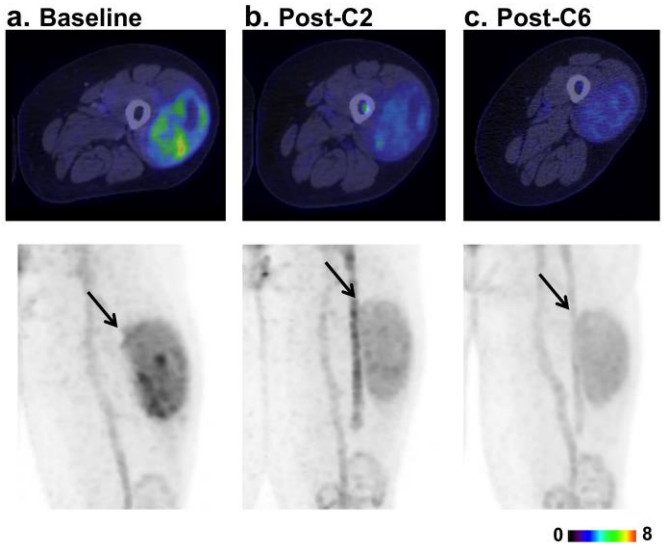
Premièrement, il semble exister une corrélation entre la valeur maximale du SUV (SUV_{max}) dans la tumeur primitive et le grade FNCLCC, avec cependant une importante superposition des SUVmax selon les grades (Tateishi et al, 2006; Benz et al, 2010; Macpherson et al, 2018).

Dans notre expérience et la littérature, le degré de fixation des STM est très variable, certains STM fixant peu (M/RC-LPS), d'autres fortement (sarcomes indifférenciés, tumeurs malignes des gaines des nerfs périphériques), avec des degrés de fixation variables pour certains mêmes types histologiques (angiosarcomes, MFS, sarcomes épithéloïdes... etc.) (Macpherson et al, 2018). Ce dernier point rend difficile l'utilisation du TEP/CT comme outil isolé d'aide au diagnostic histologique et

nécessaire de prendre le patient comme son propre référentiel afin de calculer des variations relatives si le SUV est employé pour évaluer des modifications sous certaines conditions.

Dans ce sens, plusieurs études ont montré l'intérêt des variations de SUV_{peak} (: valeur maximale du SUV dans un volume de 1 cm^3 - ou une surface de 1cm2 – centré(e) sur la région la plus métabolique de la tumeur et non la valeur maximale du SUV dans un seul voxel [SUV_{max}], non la valeur moyenne du SUV dans l'ensemble de la tumeur [SUV_{mean}]) pour prédire la réponse histologique (définie plus de 95% de nécrose sur pièce opératoire) après une cure de NAC par anthracycline et après fin de la NAC (aire sous la courbe ROC [AUROC] = 0.69 et 0.90-0.93, respectivement, avec pour seuils optimaux une diminution de -38% et de -60 - -72% du SUV_{peak}, respectivement) - cela avec de meilleurs performances que des critères unidimensionnels et volumiques (Benz et al, 2008; Benz et al, 2009; Evilevitch et al, 2008). Dans un second temps, les mêmes groupes ont montré que la variation précoce et intermédiaire du SUL_{peak} sous NAC était un prédicteur indépendant de la survie globale des patients après chirurgie curatrice (Figure 1-15) (Herrmann et al, 2012; Eary et al, 2014).

Figure 1-15. Evaluation de la réponse à la chiomiothérapie néo-adjuvante par ^{18}F-FDG-TEP/CT. Exemple d'une patiente atteinte d'un sarcome indifférencié pléomorphe de haut-grade du quadriceps gauche. (**a**) Bilan baseline : SUVmax = 7.8, SULpeak = 3.9. (**b**) Evaluation à 2 cures: SUVmax = 4 (-48.7%), SULpeak = 2.1 (-46.2%). (**c**) Evaluation à 6 cures: SUVmax = 3.9 (-50%) et SULpeak = 1.7 (-56.4%). Variation de taille > -30% au cours du suivi. Selon Benz et al. (2009), les variations précoces à partir de -35 – -38% sont prédictives d'une bonne réponse histologique (ici le cas, avec <10% de cellularité résiduelle)



D'autres indices que les SUV et SUL ont été décrits et déjà utilisés pour évaluer le métabolisme tumoral des STM, comme le « *Total Lesion Glycoloysis* » (TLG - qui correspond au produit du SUV_{mean} multiplié par le volume tumoral métaboliquement actif [MTV]), et pourraient améliorer ces performances prédictives (Andersen et al, 2015a; Andersen et al, 2015b). Il faut noter que les choix du seuil pour définir ce qui est métaboliquement actif ainsi que les algorithmes de segmentation des volumes tumoraux sont susceptibles d'influencer ces potentiels biomarqueurs et donc les prédictions basées sur eux (Stevenson et al, 2018).

Enfin, le PET/CT pourrait aider à confirmer des suspicions de rechute locale et à distance en identifiant par exemple un hypermétabolisme marqué au sein de remaniements post-thérapeutique ambigus, en complément de l'IRM (Park et al, 2016; Erfanian et al, 2018).

Pour conclure, des critères de réponse métabolique nommés EORTC et PERCIST (pour « *PET response criteria in solid tumors* », Table 1-7) sur le modèle de RECIST ont été proposés avec des seuils là-encore non spécifiques des sous-types de cancers et des traitements et dont la valeur prédictive, à notre connaissance, n'a pas été comparée aux critères de réponse radiologique préexistant dans le contexte des STM (Wahl et al, 2009).

Table 1-7. Principaux critères d'évaluation de la réponse au traitement basés sur le ^{18}F-FDG-TEP/CT.

Status	EORTC
CMR	Complete resolution of 18F-FDG uptake within all lesions, making them indistiguishable from normal surrounding tissues AND no new lesion
PMR	After 1 cycle: reduction of at least -15% in the sum of SUV-max_{lesion}
PMD	1/ Increase of at least +15% in the sum of SUV-max_{lesion} OR 2/ Appearance of new 18F-FDG avid tumoral lesions OR 3/ Visible increase in extent of 18F-FDG tumor uptake >20%
SMD	Percentage change in SUV-max_{lesion} sum between -15% and +15% Change in longest diameter of fixation < 20%

Status	PERCIST 1.0
CMR	Complete resolution of 18F-FDG uptahe within measurable target lesion, less than liver activity and undistinguishable from normal surrounding tissues AND no new lesion (AND no progressive disease according to RECIST (otherwise need confirmation))
PMR	Decrease in SUL-peak > -30% AND > -0.8 units (AND no progressive disease according to RECIST (otherwise need confirmation))
PMD	1/ Increase in SUL-peak >+30% AND > 0.8 units OR 2/ Change in TLG > +75% OR 3/ Appearance of new 18F-FDG avid tumoral lesions
SMD	Not CMR, not PMR, not PMD

NOTE. CMR: complete metabolic response, PMR: partial metabolic response, PMD: progressive metabolic disease, SMD: stable metabolic disease.

1.4.2.2. DWI

Cette technique d'IRM permet de quantifier les mouvements (ou la diffusion) des molécules d'eau libre dans les tissus (Le Bihan et al, 1986). Les séquences DWI sont typiquement des séquences de type *echo planar* utilisant une accélération en acquisition parallèle. Elles consistent en l'application de deux gradients intenses, courts et de même amplitude mais opposés, de part et d'autre d'un pulse de 180°. Un proton qui se serait déplacé aura donc subi un déphasage proportionnel à son déplacement le long de l'axe des 2 gradients. En appliquant cela dans les 3 directions, nous pouvons quantifier le mouvement dans l'espace de chacun des voxels du volume d'acquisition. Nous pouvons faire varier la pondération en diffusion (aussi appelée b-value) selon l'intensité, la rapidité de montée et la durée d'application des gradients de diffusion. A partir de l'acquisition brute, nous pouvons calculer la carte paramétrique du coefficient apparent de diffusion (ADC) ou pour chaque voxel, la valeur de l'ADC est donnée par la relation:

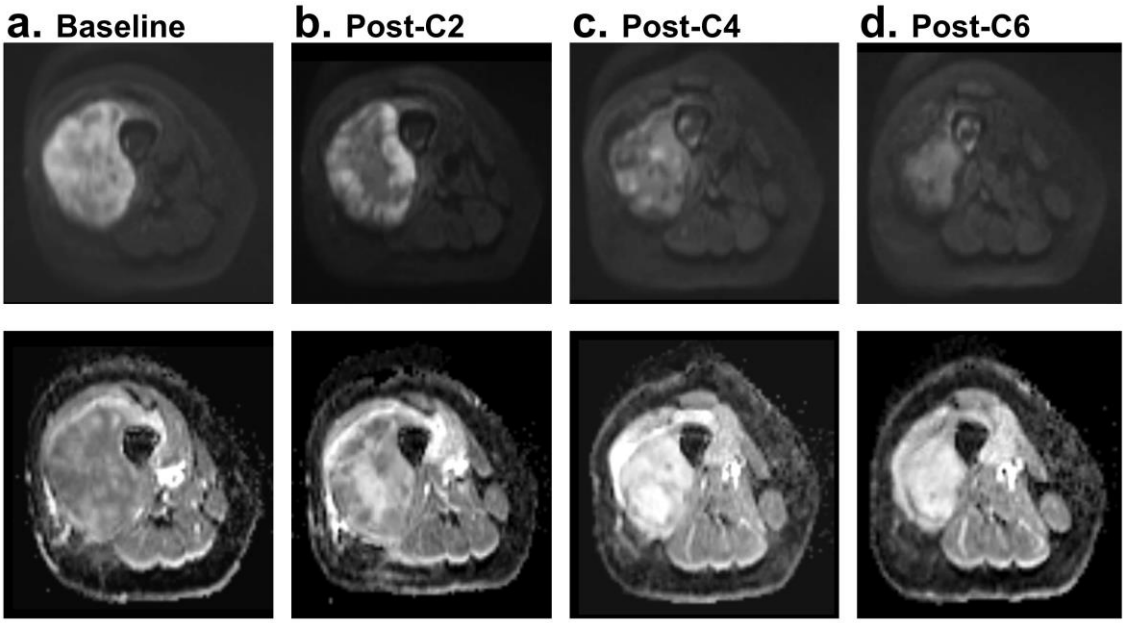
$$SI = SI_0 \times \exp[-b \times ADC]$$

Où SI signifie « *signal intensity* » et SI_0 est l'intensité de signal sans diffusion.

En cancérologie, ces molécules vont être contraintes par les membranes des cellules tumorales, des éléments du MET ou encore des débris cellulaires. Ainsi, nous supposons que plus une tumeur est cellulaire, plus la valeur de l'ADC est faible.

Cette hypothèse a été confirmée avec les STM. Une relation quasi linéaire entre la densité cellulaire sur des lames HES et l'ADC a été mise en évidence, quel le patient ait reçu ou non un traitement néoadjuvant, ainsi qu'une corrélation forte entre valeur du Ki67 et ADC (Schnappauff et al, 2009; Lee et al, 2020). L'ADC pourrait ainsi servir de marqueur de cellularité et de prolifération tumorale. Enfin, au moins deux études ont montré que les variations de l'ADC sous traitement (vers une augmentation de celui-ci, sans donner de seuil précis) étaient corrélées à la réponse dimensionnelle et à la réponse histologique (Figure 1-16) (Dudeck et al, 2008; Soldatos et al, 2016).

Figure 1-16. Evaluation de la réponse au traitement par IRM de diffusion. Exemple d'un sarcome indifférencié pléomorphe de haut-grade traité par anthracyclines. L'image du haut correspond à la TRACE à b = 800 s/mm^2, celle du bas à la cartographie ADC. Sur l'imagerie baseline (**a**), la valeur de l'ADC est de 1.50 × 10^{-3} mm^2/s (écart-type = 0.145). Après 2 cures, elle passe à 2.14 (écart-type = 0.334, +42.7%) (**b**); après 4 cures à 2.45 (écart-type = 0.173, +63.3%) (**c**) et se stabilise après 6 cures avec une moindre dispersion des valeurs (2.44, écart type = 0.092) (**d**).



1.4.2.3. DCE-MRI (ou IRM de perfusion)

1.4.2.3.1. Principe de la séquence DCE-MRI

En cancérologie périphérique, ici des tissus mous, il s'agit de séquences pondérées T1 en écho de gradient, généralement avec suppression du signal de la graisse, acquises très rapidement de manière répétée dans le temps (résolution temporelle généralement < 10s) après une ligne de base, puis l'injection IV de chélates de Gadolinium (Gd), à l'aide d'un injecteur automatique. L'agent de contraste, paramagnétique, va se lier aux

noyaux d'hydrogène et raccourcir le temps de relaxation T1 entrainant une augmentation du signal en T1 des tissus (ou rehaussement) qui est le reflet de la perfusion des tissus, de la fonction d'entrée artérielle (AIF pour « *arterial input fonction* »), de la surface capillaire, de la perméabilité capillaire et de l'espace extravasculaire - extracellulaire de diffusion de l'agent de contraste (EEE, de volume V_e ou volume interstitiel tissulaire) (O'Connor et al, 2012). Les séquences DCE-MRI permettent ainsi de visualiser et quantifier la cinétique de rehaussement intra-tumoral et d'estimer indirectement la néoangiogenèse, facteur clef de l'oncogenèse. Deux phénomène sont intriqués dans le rehaussement dynamique observé (Figure 1-17) (Cuenod et al, 2013): (i) un phénomène de perfusion tissulaire pure initiale, sans fuite interstitielle et (ii) un phénomène de fuite du produit de contraste dans l'interstitium.
La QIBA a édité des recommandations techniques pour réaliser ce type de séquences qui ont été appliquées sur l'institut Bergonié et sont de deux natures dans les travaux ici présentés: acquisition sériée de séquences DCE-MRI de type T1 « *Volume Interpolated Body Examination* » (VIBE) Fat Sat ou de type T1 « *Time-resolved angiography With Interleaved Stochastic Trajectories* » (TWIST) VIBE DIXON (avec analyse de la phase WATER) sur une durée comprise entre 2 s et 8 s, précédée d'un T1 mapping via deux angles de flip (2° et 15°) et d'une ligne de base de 5 phases à blanc, à l'aide d'un injecteur automatique.

Il existe historiquement 3 façons d'en extraire de l'information (Figure 1-17):
(1) Qualitativement, par analyse des courbes de rehaussement : en qualifiant si la tumeur se rehausse dès les phases artérielles ou veineuses (i.e. quand les artères ou les veines non pathologiques dans le volume d'acquisition commencent à s'opacifier), si elle présente un lavage (ou chute de son rehaussement) au cours de l'acquisition ou au contraire, une stabilisation voire une majoration progressive de son signal post-injection par accumulation de produit de contraste dans l'interstitium;

(2) Semi-quantitativement, sans l'intermédiaire de modèles perfusionnels, en quantifiant la courbe SI = f(temps) (ou sa transformation: [Gd] = f(temps) où [Gd] est la concentration en chélates de Gadolinium) en chaque voxel de l'image. Nous pouvons ainsi estimer la pente de rehaussement initial (« *wash-in* »), le maximum de rehaussement *(« peak-enhancement* »), le temps pour atteindre ce maximum (« *time to peak* »), l'aire sous la courbe de rehaussement à différents délais après injection,

généralement 90 s (AUC_{90s} pour « *area under the time-intensity curve at 90 sec* »), le lavage du produit de contraste après le premier passage artériel (« *wash-out* »);

(3) Quantitativement, avec utilisation de modèles pharmacocinétiques qui visent à modéliser (et donc à simplifier de manière réaliste) les échanges entre les compartiments plasmatiques et l'interstitium tumoral (ou EEE). Plusieurs modèles ont été développés, mais nous utiliserons ici le modèle de Tofts et Kety (ou Tofts modifié) (Sourbron et al, 2013).

Ce modèle considère les échanges entre 2 compartiments: le plasma (de fraction de volume total V_p avec une concentration en $[Gd]_p$) et l'EEE (de fraction de volume totale V_e avec une concentration en Gd dite $[Gd]_e$) (Figure 1-17).

Il nécessite:

- une première étape de conversion des SI en [Gd]. Cette conversion peut être faite selon plusieurs méthodes et nous avons choisi celle proposée par QIBA c'est-à-dire un T1 mapping (ou cartographie T_{10}) via l'utilisation d'angles de flip variables.
- une AIF qui permet de déterminer la concentration plasmatique artérielle et peut être mesurée directement dans une artère, ou indirectement par mesure de région de référence, ou provenir d'une AIF moyenne à l'échelle d'une population.

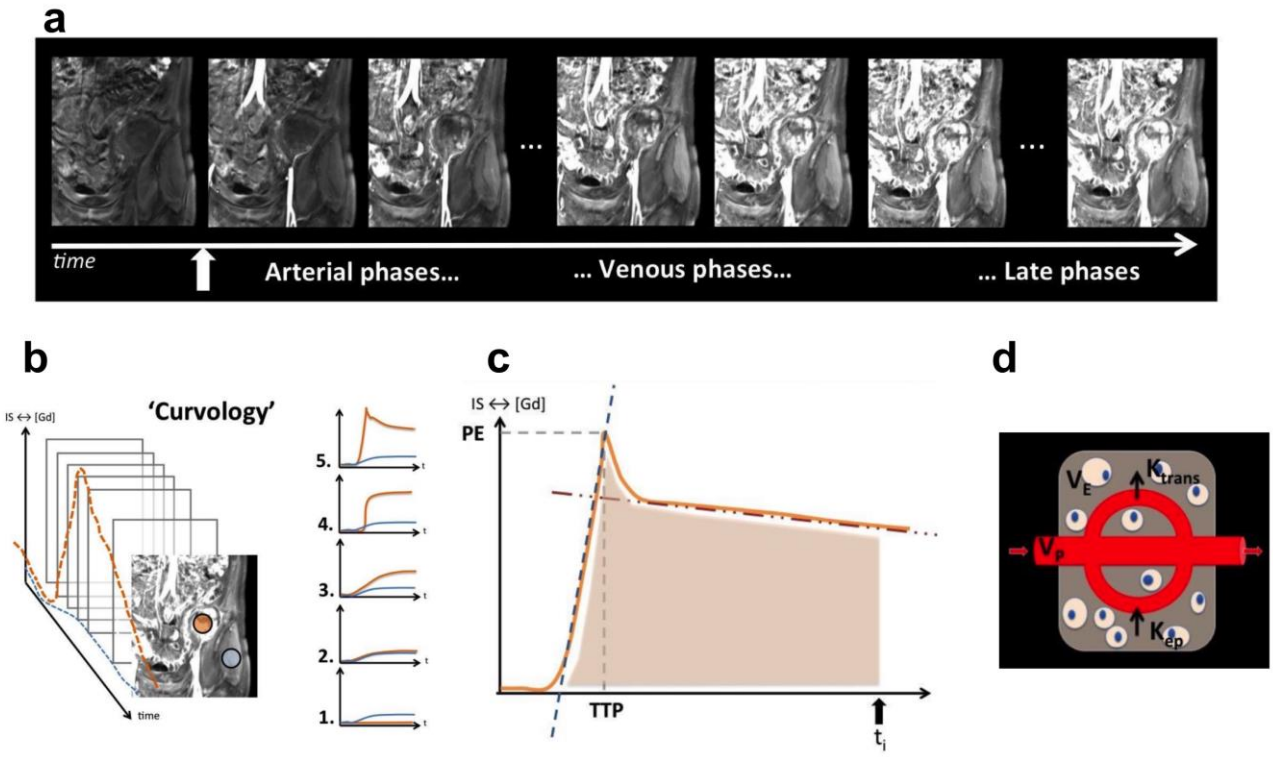
Le modèle de Tofts et Kety suppose que les échanges sont à l'équilibre entre ces 2 compartiments et peuvent être estimés par la relation suivante :

$$V_e \times \frac{d[Gd](t)_e}{dt} = K^{trans} \times ([Gd](t)_p - [Gd](t)_e)$$

Dont la solution utilise un produit de convolution et où le K^{trans} est la constante de transfert entre entre le plasma et l'interstitium tumoral (en min^{-1}). La constante K^{ep} correspond à la constante d'efflux entre l'interstitium et le plasma peut être calculée comme K^{trans} / V_e.

Ces 3 paramètres K^{trans}, K^{ep} et V_e sont estimés en chaque voxel de l'image pour obtenir des cartes dites paramétriques qui sont utilisés pour évaluer la néo-angiogenèse tumorale et ses modifications sous traitement. Il faut toutefois souligner que le K^{trans} est d'interprétation complexe et représente une combinaison variable entre le débit de perfusion tissulaire et la perméabilité surfacique (Cuenod et al, 2013).

Figure 1-17. Principe de l'analyse des séquences DCE-MRI. (**a**) Exemple chez un patient atteint d'un sarcome indifférencié pléomorphe du muscle iliaque gauche (plan coronal) consistant en une série d'images explorant un volume similaire acquises très rapidement avant et après injection intraveineuse de produit de contraste (flèche blanche verticale), et permettant de suivre le rehaussement du signal des tissus de manière 'dynamique'. (**b**) Nous pouvons tracer une région d'intérêt dans la tumeur (orange) et les tissus sains (bleu) pour suivre ce rehaussement du signaltraduisant la captation du produit de contraste dans la tumeur. Nous obtenons ainsi la courbe de rehaussement fonction du temps. Ces courbes peuvent être appréciées qualitativement, mais aussi semi-quantitativement (**c**), en estimant la pente initiale de la courbe (« *wash-in* » en pointillé bleu), la pente de décroissance tardive (« *wash-out* » en pointillés rouges), l'aire sous la courbe (orange pâle), le temps au maximum de rehaussement (TTP). Il est aussi possible de réaliser une analyse quantitative en appliquant des modèles de perfusion à un ou plusieurs compartiments et en quantifiant les constantes de transfert entre ces différents compartiments.

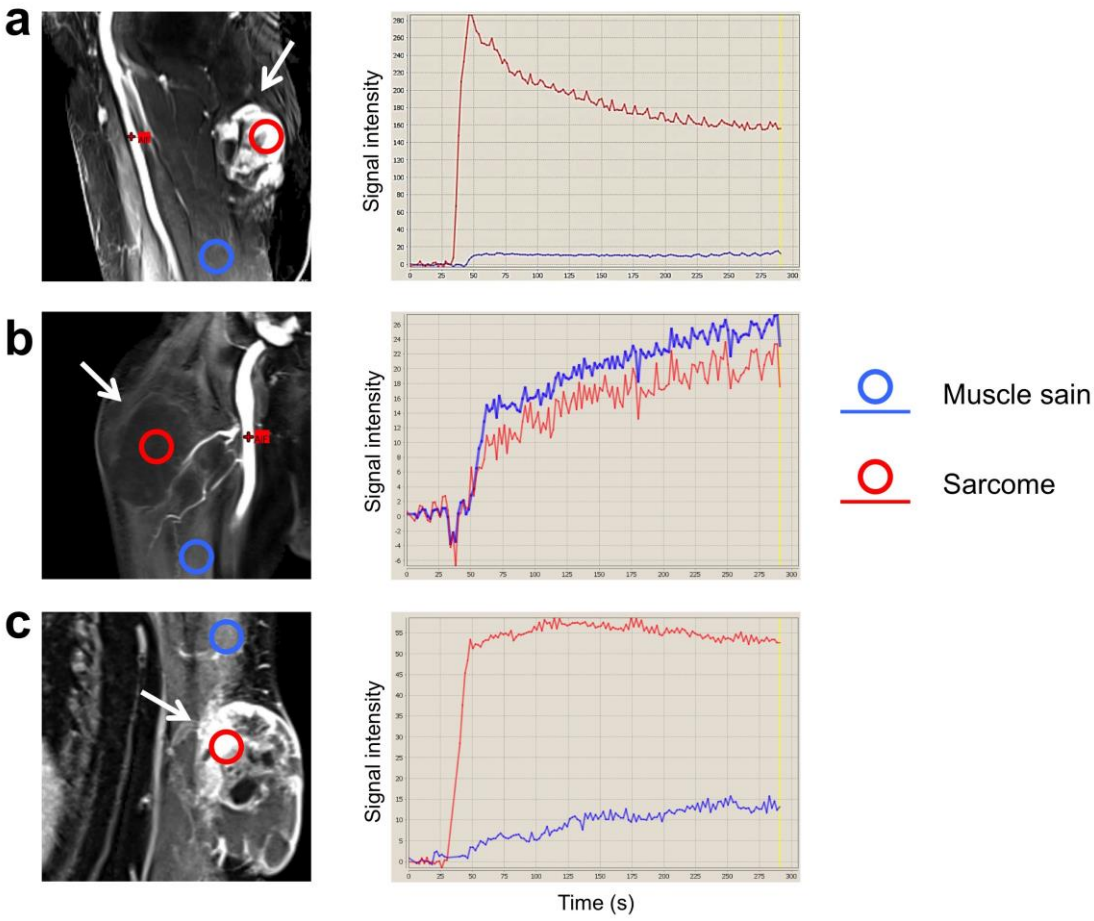


1.4.2.3.2. DCE-MRI et STM

Du fait de leur diversité histologique et moléculaire, il existe aussi une grande diversité des profils de rehaussement des STM en DCE-MRI (Figure 1-18). Si certaines études se sont intéressées à l'apport du DCE-MRI pour différencier les tumeurs des parties molles bénignes de celles malignes, aucune à notre connaissance n'a étudié les associations entre profils perfusionnels et histotypes ou profils moléculaires. Selon notre expérience, les M/RC-LPS tendent à se rehausser progressivement en nappe hétérogène pour finalement s'homogénéiser au-delà de 5

mn (courbe de type 3 de Kuhl) (Kuhl et al, 1999). A contrario, les UPS et MFS peuvent présenter des rehaussements intenses artériels suivis de *wash-out* importants cohabitant avec des rehaussements intermédiaires voire faibles.

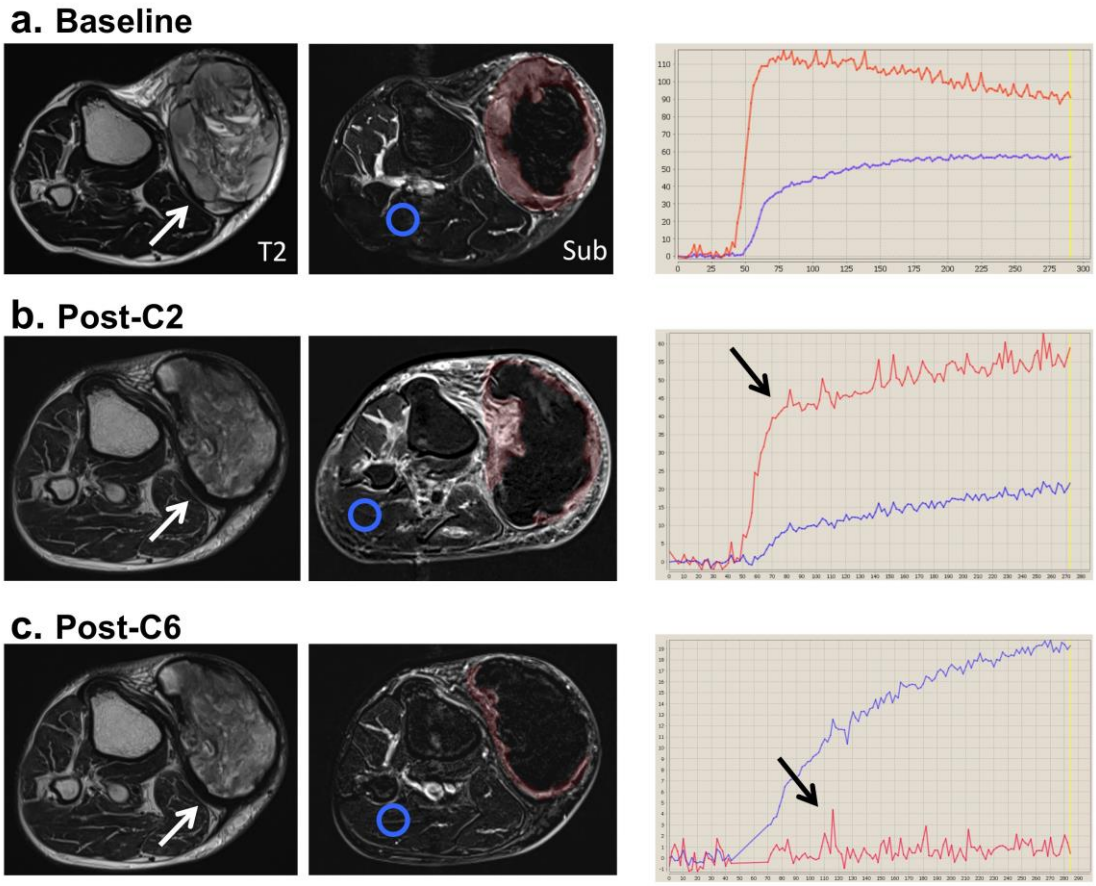
Figure 1-18. Diversité des courbes de rehaussement en fonction du temps après injection obtenues par les séquences DCE-MRI parmi les STM (flèches blanches). (**a**) Exemple d'un sarcome indifférencié pléomorphe de haut grade du grand fessier droit, avec un important *wash-in*, un pic puis un *wash-out* marqué. (b) Exemple d'un liposarcome myxoïde et à cellules rondes de la cuisse droite, se rehaussant très faiblement et très progressivement sans lavage. (**c**) Exemple d'un rhabdomyosarcome pléomorphe du bras gauche, présentant un franc pré-décalage de la courbe de rehaussement, mais sans pic et suivi d'un plateau, sans lavage. La référence correspond au muscle sain. L'acquisition pour ces 3 exemples a duré 5 minutes.



L'essentiel de la littérature relative à l'usage du DCE-MRI pour les STM s'est attachée à l'identification de corrélations entre variations ultra-précoces, précoces ou tardives de paramètres perfusionnels et la réponse histologique (définie de manière variable comme < 5% ou < 50% de cellularité résiduelle ou ≥ 95% de nécrose) après traitement par perfusion isolée de membre chez l'homme ou souris xénogreffée, ou

association radio-chimiothérapie (classique ou anti-angiogénique) néo-adjuvant chez l'homme ou le chien (où l'*outcome* était ici la survie) (Van Rijswijk et al, 2003; Viglianti et al, 2009; Alic et al, 2013; Meyer et al, 2013; Huang et al, 2016a; Soldatos et al, 2016; Xia et al, 2017). La Figure 1-19 illustre ces variations de la perfusion des STM sous NAC imagée par DCE-MRI.

Figure 1-19. Aide à l'évaluation de la réponse à la chimiothérapie néo-adjuvante des STM par les séquences DCE-MRI. Exemple d'un sarcome indifférencié de haut grade du mollet droit (flèches blanches). Pour chaque évaluation baseline (**a**), post 2 cures (**b**) et pré-chirurugicale à 6 cures (**c**), l'image de gauche correspond à une séquence axiale T2 turbo spin echo, celle du milieu une séquence de soustraction entre T1 Fat Sat turbo spin echo avec – sans injection, et celle de droite à la courbe de rehaussement (intensité de signal en fonction du délai post-injection). Au cours du traitement, nous observons successivement la disparition du pic de rehaussement et du lavage puis l'aplatissement complet de la courbe de rehaussement (flèches noires). Patient excellent répondeur sur pièce opératoire avec < 5% de cellularité tumorale résiduelle.



Malheureusement, ces études n'excédent pas la trentaine d'individus, leurs techniques d'acquisition et de post-traitement diffèrent sensiblement empêchant de leur comparaison et le mélange de leurs patients, et leurs résultats, s'ils encouragent l'utilisation des séquences DCE-MRI dans l'évaluation de la réponse des STM quelque soit le traitement systémique ou local, ne proposent pas de biomarqueurs avec des seuils exploitables. De plus, à l'exception de l'étude d'Alic et *al.* (2013), ces études ont évalué soit qualitativement les cinétiques de rehaussement des STM, soit une valeur moyenne dans une section 2D de la tumeur ou dans un volume, sans prendre en compte l'hétérogénéité perfusionnelle. Alic et *al.* (2013), les premiers, ont appliqué une méthodologie visant à quantifier les modifications perfusionnelles ultra-précoces de l'architecture tumorale sous traitement (entre une acquisition baseline puis 1h après perfusion isolée de membre par anti-TNFα et melphalan) et à identifier des corrélations avec la pièce opératoire. Si les auteurs n'ont pas utilisé d'indices radiomics tels qu'actuellement, leurs conclusions incitent à prendre en compte ces variations très précoces de l'hétérogénéité perfusionnelle accessibles par analyse radiomics des cartes paramétriques d'indices semi-quantitatifs et quantitatifs.
Enfin, nous pouvons aussi relever l'intérêt des séquences DCE-MRI comme aide au diagnostic de rechute locale des STM, sous forme de rehaussements précoces, artériels, au sein de rehaussements fibro-cicatriciels mais ces résultats restent à confirmer au vu de la nature rétrospective sur de faibles effectifs de l'étude correspondante (Del Grande et al, 2014).

1.4.2.4. Autres méthodes d'imagerie « fonctionnelle »

L'*intra-voxel incoherent motion* (IVIM) et la spectroscopie par résonance magnétique (MRS) sont deux autres méthodes susceptibles d'apporter des marqueurs radiologiques quantitatifs de phénomènes biologiques mais qui ont été très peu appliquées aux STM.

L'*IVIM* considère que la diffusion des molécules d'eau mesurée par l'ADC est due à 2 composantes: (i) la diffusion réelle des molécules d'eau dans les tissus et (ii) la perfusion du sang dans le réseau (micro) capillaire (Le Bihan, 2019). Ainsi, la relation qui relie l'atténuation globale du signal (F_{total}) à l'atténuation du signal due à la diffusion réelle (F_{diff}) et à la perfusion (F_{perf}) correspond à :

$$F_{total} = f_{IVIM} \times F_{perf} + (1 - f_{IVIM}) \times F_{diff}$$

Où f_{IVIM} est la fraction du volume sanguin soumis à l'IVIM.

Ce qui devient en appliquant l'équation du paragraphe 1.4.2.2:

$$\frac{SI}{SI_0} = f_{IVIM} \times \exp[-b \times (ADC^* + ADC_{blood})] + (1 - f_{IVIM}) \times \exp[-b \times ADC_{tissue}]$$

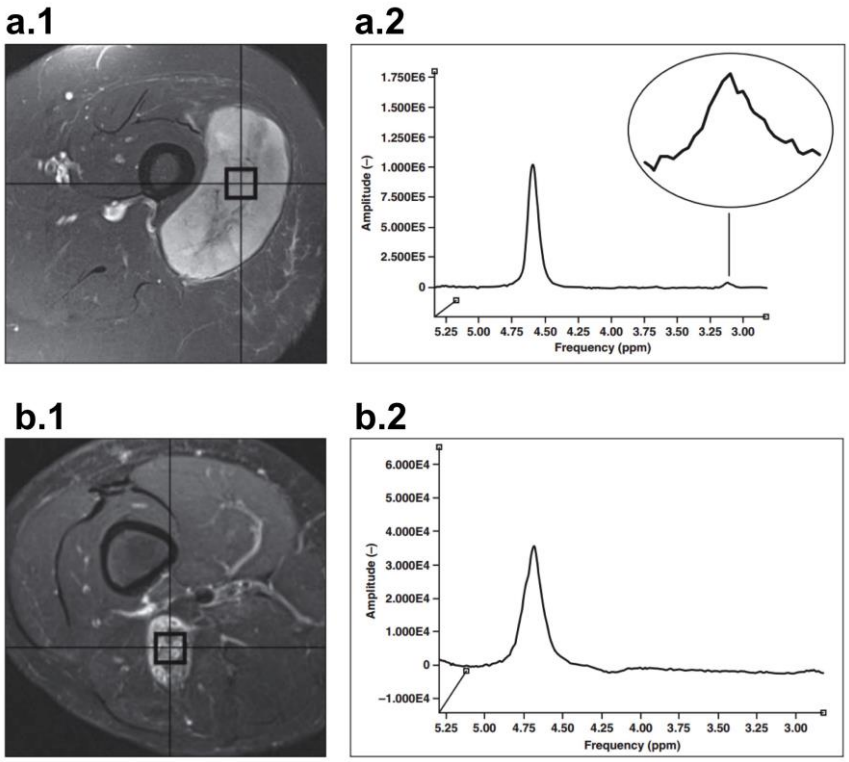
Ce phénomène est responsable du décalage de la courbe $F_{total} = S/S_0 = f$(b-value) pour les très faibles b-values (< 200 s/mm^2). La fraction perfusionnelle est estimée à l'aide de modèles bi-exponentiels ou de kurtosis en sur-échantillonnant ces b-values faibles. L'IVIM permet donc une imagerie perfusionnelle sans agent de contraste exogène. Au-delà de la recherche d'indices quantitatifs pour discriminer un STM d'une masse bénigne des parties molles dans de petites séries exploratoires, l'IVIM n'a pas été explorée dans le domaine des STM (Du et al, 2015; Wu et al, 2018).

La spectroscopie par résonance magnétique (MRS) consiste à mesurer la concentration de métabolites dans un ou plusieurs voxels. Le principe est qu'un noyau atomique placé dans un champ magnétique auquel est envoyé une impulsion de radiofréquence réémettra cette énergie sous forme d'une onde électromagnétique (la «*free induction decay*») dépendant de l'intensité du champ magnétique et de l'environnement moléculaire du noyau atomique d'intérêt (Galanaud et al, 2007). Ainsi, un proton donnera un signal différent selon qu'il est situé dans des molécules différentes. Cette différence, appelée déplacement chimique, est visualisée en MRS sous forme d'un spectre et exprimée en partie par millions (ppm). Dans le cas des STM, la MRS du proton a été principalement étudiée. En pratique, seul un faible nombre de molécules sont observées en MRS du proton - différents selon les paramètres d'acquisition tel le temps d'écho -, comme les lipides, la créatine-phosphocréatine (: marqueur du métabolisme énergétique), la choline (: marqueur de la synthèse et dégradation des membranes cellulaires) ou encore le lactate (: marqueur du métabolisme anaérobie). La Figure 1-20 illustre un spectre normal de muscle et un spectre de STM issus de la littérature.

Dans une méta-analyse, Subhawong et al (2012) ont répertorié les applications testées de la MRS du proton pour conclure sur l'intérêt d'identifier et de quantifier le pic de Choline afin de distinguer les tumeurs bénignes des parties molles des STM (à l'exception de tumeurs à cellules géantes et de tumeurs des gaines des nerfs

périphériques) en insistant sur la difficulté à obtenir une quantification reproductible. Les applications de marqueurs pronostics ou de la réponse à des traitements sont à notre connaissance inexistante.

Figure 1-20. Exemple d'application des séquences de spectroscopie par résonance magnétique du proton pour les STM: distinction bénin versus malin, extraits de *Subhawong et al, 2012*. (**a.1**) Séquence axiale T2 fat sat d'un sarcome de haut grade dans le muscle vaste latéral de cuisse gauche. (**a.2**) le spectre montre un pic de choline (: zoom). (**b.1**) Séquence axiale T2 fat sat montrant une lésion intermusculaire aux dépens du nerf sciatique correspondant à un neurofibrome. (**b.2**) Absence de pic de choline sur le spectre. Spectroscopies mono-voxel à temps d'écho long.



1.4.2.5. Résumé: quels biomarqueurs validés des STM?

Au travers de cette revue de la littérature, il apparait que:

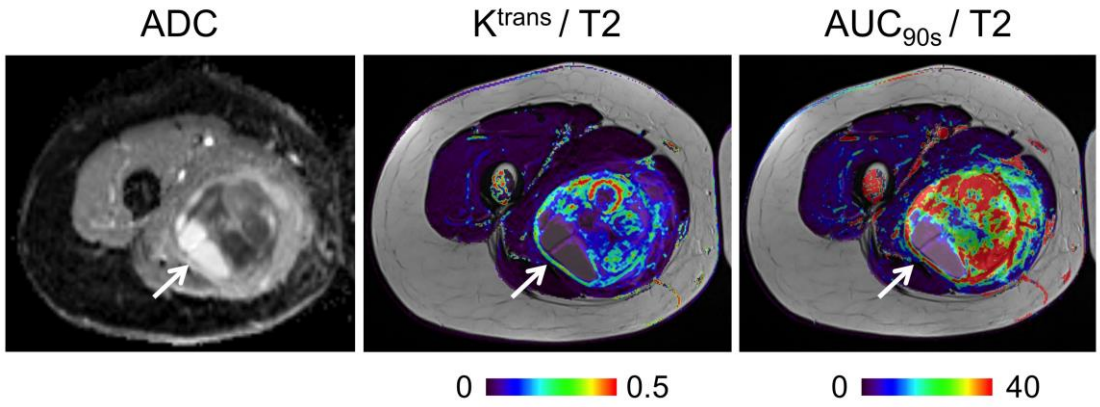
- toutes les innovations techniques de l'IRM de ces 35 dernières années (DWI, IVIM, MRS, DCE-MRI) et le TEP/CT apportent une information complémentaire et quantitative sur l'organisation de la tumeur, sa malignité, son agressivité, sa vascularisation et son métabolisme;

- cependant ces études se sont focalisées sur une modalité d'imagerie, généralement sans mise en perspective avec des indices radiologiques préexistant ou avec d'autres modalités d'imagerie « fonctionnelle ». Il est ainsi possible que certaines informations

soient redondantes ou se potentialisent, et que l'identification de biomarqueurs passe par le développement de critères composites issus d'IRM multiparamétriques (c'est-à-dire issus de combinaisons de variables issues de l'analyse radiologique classique, de mesures sur IRM structurales et de variables issues de l'analyse de séquences d'IRM « fonctionnelle »);

- en dehors des mesures TEP/CT, ces mesures quantitatives correspondent à une moyenne des voxels contenus dans une région d'intérêt (ROI pour « *region of interest* ») en 2D, soit placée dans une portion d'allure charnue, soit contourant la plus grande section tumorale dans le plan axial. Ces mesures ne donnent aucune information sur l'hétérogénéité tumorale et sont à risque de biais d'échantillonnage. Ainsi, une tumeur peu hypervascularisée et modérément cellulaire homogène peut obtenir des mêmes valeurs moyennes d'ADC et de K^{trans} qu'une tumeur très hétérogène avec de vastes plages de nécrose (aux ADC très élevé et K^{trans} nul) et des plages très cellulaires et vascularisées (aux ADC très faible et K^{trans} élevé) se compensant (Figure 1-21). Instinctivement, ce deuxième cas de figure nous apparait plus agressif. Cet exemple illustre la nécessité de mieux prendre en compte l'hétérogénéité intra-tumorale accessible en imagerie.

Figure 1-21. Hétérogénéité d'un sarcome indifférencié pléomorphe (flèches blanches) de haut grade de la loge postérieure de cuisse droite sur les séquences de diffusion (cartographie ADC) et sur les cartes paramétriques issues de la séquence DCE-MRI, K^{trans} (en s^{-1}) et AUC_{90s} (en unité arbitraire.s).



- Enfin, O'Connor *et al.* (2017) ont détaillé toutes les étapes nécessaires à l'établissement d'un biomarqueur radiologique depuis sa découverte en laboratoire jusqu'à son usage en routine avec identification de deux gaps translationnels, i.e.:

(i) la démonstration qu'il s'agisse d'une mesure fiable pour tester des hypothèses en recherche clinique (impératifs de reproductibilité, de répétabilité, de maitrise des biais, d'accessibilité, d'acceptabilité, de tolérance, de corrélation, de spécificité et de lien temporel avec le phénomène mesuré, cela pour un cout acceptable et une valeur ajoutée aux marqueurs préexistant);

(ii) la démonstration qu'il puisse être utilisée pour la prise en charge clinique quotidienne de patients dans le cadre du système de santé.

Or, bien qu'ils soient déjà mesurables en routine clinique, les potentiels biomarqueurs exposés ci-dessus n'ont pas franchi ces étapes méthodologiques.

1.5. Quantifier le phénotype des tumeurs: introduction aux approches radiomics

1.5.1. Imagerie médicale et médecine personnalisée

La médecine de précision (ou médecine personnalisée) consiste à prendre en compte de la manière la plus exhaustive et pertinente possible les caractéristiques individuelles des patients et de leurs tumeurs afin de leur proposer une prise en charge individualisée et spécifique pour une plus grande efficacité.

L'imagerie médicale est le seul moyen d'obtenir une visualisation non-invasive et en temps réel de la tumeur *in situ* et *in vivo,* et de ses variations temporelles sous traitement. Nous avons vu que de multiples variables issues de l'analyse radiologique classique (qualitative, sémantique) sur imagerie structurale et issues d'une analyse quantitative sur imagerie fonctionnelle étaient capables de subdiviser les STM en de multiples sous-groupes de tumeurs, certains de ces groupes ayant une valeur pronostique. Par exemple, un STM avec un volume de nécrose estimé > 50% en IRM, des rehaussements péri-tumoraux et une forte hétérogénéité sur les séquences pondérées T2 a significativement plus de risque de présenter une rechute métastatique qu'un STM ne présentant aucune de ces propriétés (Crombé et al, 2019a). Un STM fortement hyper-perfusé en DCE-MRI, présentant une abondante néo-vascularisation macroscopique sous forme de « *flow-voids* » intra- et péri-tumoraux a plus de chances

d'hyper-exprimer des gènes impliqués dans la néoangiogenèse et d'être potentiellement sensible aux traitements anti-angiogéniques (Ledoux en al, 2019). Ainsi, dans une optique de médecine personnalisée, intégrer ces variables issues de l'imagerie avec d'autres variables extra-radiologiques (cliniques, histologiques ou moléculaires) pourrait aider à mieux stratifier les patients et leurs tumeurs.

Nous avons cependant vu plusieurs obstacles limitant l'utilisation de ces variables radiologiques pour prendre des décisions médicales, d'ordres technique et méthodologique:

- le caractère qualitatif et subjectif de certaines d'entre elles, à risque de restreindre leur fiabilité et leur reproductibilité. En particulier, les radiologues se rendent compte de leur difficulté à caractériser verbalement le phénotype radiologique, des architectures complexes, l'hétérogénéité et la forme des tumeurs (Figure 1-13).
- l'utilisation de valeurs moyennes des variables quantitatives plutôt que des indices prenant en compte l'hétérogénéité intra-tumorale;
- l'absence de cohorte de validation rétrospective puis prospective;
- l'absence de démonstration de leur valeur ajoutée pour les patients dans des essais cliniques radiologiques avec décision médicale basée sur une stratification radiologique

Les approches dites radiomics sont un moyen de résoudre les deux premières limites techniques.

1.5.2. Principe général des approches radiomics

Les approches radiomics correspondent au processus visant (1) à extraire (semi-) automatiquement de multiples variables numériques quantifiant le phénotype d'objets d'intérêt situés dans des images médicales de toute nature (PET/CT, scanner, IRM structure, cartes paramétriques d'imageries fonctionnelles, échographie-Doppler...), (2) les post-traiter (i.e. les nettoyer, transformer, sélectionner) pour (3) les intégrer dans des modèles prédictifs (Lambin et al, 2017).

En oncologie, ces objets d'intérêt sont les tumeurs et/ou leurs tissus environnants, ou leurs métastases, voire certains compartiments dans la tumeur. Des exemples de prédiction que nous cherchons à réaliser sont: (i) le pronostic local, à distance, général; (ii) l'identification de certains sous-groupes histologiques ou moléculaires

d'intérêt car plus sensibles à telles thérapeutiques ciblées; (iii) la réponse au traitement choisi.

Le terme « radiomics » a été choisi par analogie aux autres « -omics » puisque le nombres de variables extraites peut aller de quelques dizaines à quelques milliers. Ce terme a été introduit en 2010 par Gillies et *al.* (Gillies et al, 2010) mais les méthodes permettant ces approches avaient déjà été conceptualisées et employées dès les années 1980 en echographie (Finette et al, 1983a ; Finette et al, 1983b) et les années 1990 en IRM et radiographies standard (Lynch et al, 1991; Lerski et al, 1993 ; Schad et al, 1993). Le véritable essor des approches radiomics en imagerie oncologique pourrait coïncider avec l'étude de Aerts et *al.* (2014), dans laquelle les auteurs ont construit une signature à partir de 440 indices radiomics extraits de scanners de 1019 patients atteints de cancer broncho-pulmonaires et ORL. Ils ont démontré que cette signature radiomics était corrélée au pronostic des patients et à l'expression de gènes impliqués dans l'oncogenèse aussi bien dans les cohortes d'entrainement que de validation (Aerts et al, 2014). En février 2020, nous avons répertorié plus de 1500 articles incluant le terme « radiomics » sur Pubmed et plus de 1100 dédiés à l'imagerie cancérologique.

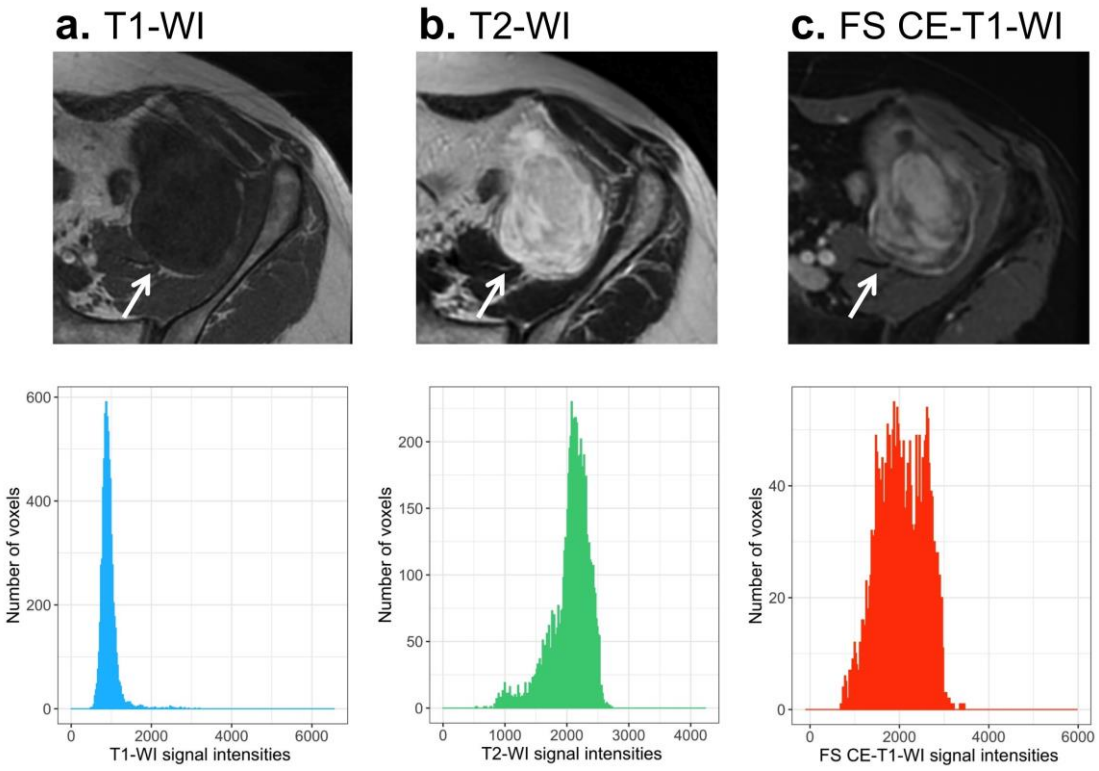
1.5.3. Principaux indices radiomics

Il existe 3 grandes catégories d'indices radiomics, les indices d'intensité, les indices de texture et les indices de forme. Ils sont calculés pour une région ou volume d'intérêt (VOI pour « *volume of interest* ») pré-déterminé comportant un nombre fini de voxels avec un nombre de valeurs de niveaux de gris aussi connu et fixé (appelé « *bins* »), selon la question posée. Afin d'homogénéiser les définitions de ces indices, l' « *Image Biomarkers Standarization Initiative* » (IBSI), une collaboration de 19 équipes issues de 8 pays, a récemment publié un document de référence décrivant les pipelines de post-traitement d'images et les formules pour les obtenir (Zwanenburg et al, 2019 ; https://ibsi.readthedocs.io/en/latest/). La liste des indices radiomics employés dans les travaux de cette thèse est donnée en Annexe 2. Plusieurs logiciels in-house, open-source et payant, ainsi que des packages dans les principaux langages de programmation permettent de calculer ces indices (Cf. infra).

1.5.3.1. Indices d'intensité

Les indices d'intensité ne prennent en compte que les valeurs, ou intensités, de gris sans considérer l'organisation spatiale des voxels dans le VOI. Ces indices sont issus de l'analyse de histogramme des intensités du VOI. La moyenne, la médiane, l'écart type, les 10^{ième} et 90^{ième} percentiles, le minimum et le maximum, l'écart interquartiles, l'écart entre minimum et maximum, le coefficient d'asymétrie (ou « *skewness* »), le coefficient d'aplatissement (ou kurtosis) et l'entropie en font partie (Figure 1-22).

Figure 1-22. Histogrammes des intensités de signal des 3 séquences structurales (ou conventionnelles) de base pour le bilan IRM des STM. Exemple d'un sarcome indifférencié pléomorphe de grade 2 du muscle ilio-psoas gauche (flèches blanches). (**a**) T1-weighted imaging (-WI), (**b**) T2-WI et (**c**) T1 Fat Sat après injection IV de chélates de gadolinium (FS CE-T1-WI). L'histogramme correspond à celui de la section (2D) affichée de la tumeur.



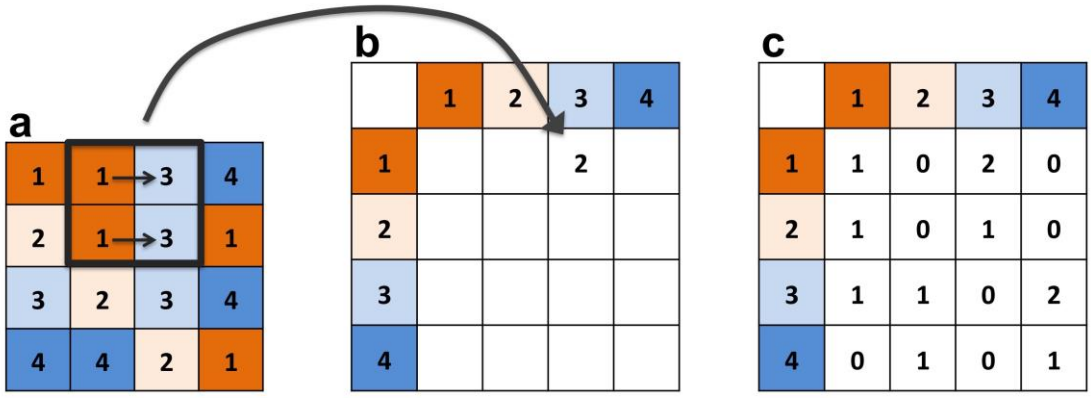
1.5.4.2. Indices de texture

En plus de l'intensité de gris, ces indices prennent en compte les réarrangements spatiaux (en 2D et en 3D) de voxels d'intensités de signal variables, afin d'identifier

des motifs récurrents susceptibles d'aider à classer des images. Plusieurs méthodes complémentaires sont décrites:

(1) Les *matrices de co-occurrence de niveaux de gris* (GLCM pour « *gray-level co-occurrence matrices* ») extraient une information sur les dimensions de zones homogènes pour chacun des niveaux de gris. Dans un premier temps, la GLCM de dimensions N par N (où N est le nombre de *bins*) d'une image est construite, en choisissant une direction θ et une distance d. La valeur de chaque point (i, j) de cette matrice correspond au nombre de fois où les voxels d'intensité i et j ont été voisins (pour θ et d) (Figure 1-23). Les indices de texture sont ensuite calculés sur cette matrice (Haralick et al, 1973).

Figure 1-23. Principe du calcul d'une matrice de coocurrence de (4) niveaux de gris (**a**) pour une distance de 1 (: 1 voxel) adjacent et un angle de 0° (: voisin direct à droite). Pour exemple, les voxels de niveau 1 sont directement adjacents à 2 reprises, à droite, de voxels de niveau 3 (**b**). Cette même opération est répétée pour chaque couple de niveaux de gris (**c**).



Dans le cas d'image en 3D, une étape clef pour calculer les indices radiomics est de déterminer la méthode d'agrégation. La valeur des indices radiomics peut être fortement influencée par la méthode choisie. L'IBSI a répertorié les méthodes suivantes (Depeursinge et al, 2017; Zwanenburg et al, 2019):

- les indices sont d'abord calculés dans chaque matrice directionnelle 2D puis moyennés pour chacune des directions en 2D et pour chacune des coupes du VOI;
- les matrices 2D directionnelles sont fusionnées pour chaque coupe du VOI, puis les indices sont calculés à chaque coupe puis moyennés sur l'ensemble des coupes;

- les matrices 2D directionnelles sont d'abord fusionnées pour chaque direction, puis les indices radiomics sont calculés et moyennés;

- les matrices 2D directionnelles sont toutes fusionnées résultant en une seule GLCM sur laquelle sont calculés les indices radiomics;

- les indices sont calculés pour chacune des matrices 3D directionnelles puis moyennés pour toutes ces directions;

- les indices sont calculés sur une seule matrice après avoir fusionné toutes les matrices 3D différant selon la direction.

Selon la méthode choisie, nous définissons les indices radiomics comme 2.5D ou 3D.

(2) Les *« gray-level run-length matices »* (GLRLM) évaluent la longueur de « *runs* » de voxels de même intensité côte à côte (i.e. consécutifs) dans une même direction. La valeur de chaque point (i, j) d'une GLRLM correspond au nombre d'occurrences de *runs* de longueur j pour un niveau de gris i. De la même manière, la technique d'agrégation devra être clairement définie pour les images 3D lors du calcul des indices radiomics (Galloway et al, 1975).

(3) Les « *gray-level size zone based matices* » (GLZM) sont calculées en comptant le nombre de groupes de voxels de même intensité de gris connectés entre eux. Ainsi, la valeur d'un point (i, j) d'une GLZM correspond au nombre d'occurrences de zones de taille j voxels et d'intensité i. Après détermination de la méthode d'agrégation pour des images 3D, les indices radiomics peuvent être calculés (Thibault et al, 2014).

(4) Les « *gray-level distance zone based matrices* » (GLDM) est une variante des GLZM qui comptabilise le nombre de zones de voxels connectés partageant une même valeur de niveau de gris et une même distance par rapport à l'extrémité du VOI (Thibault et al, 2014).

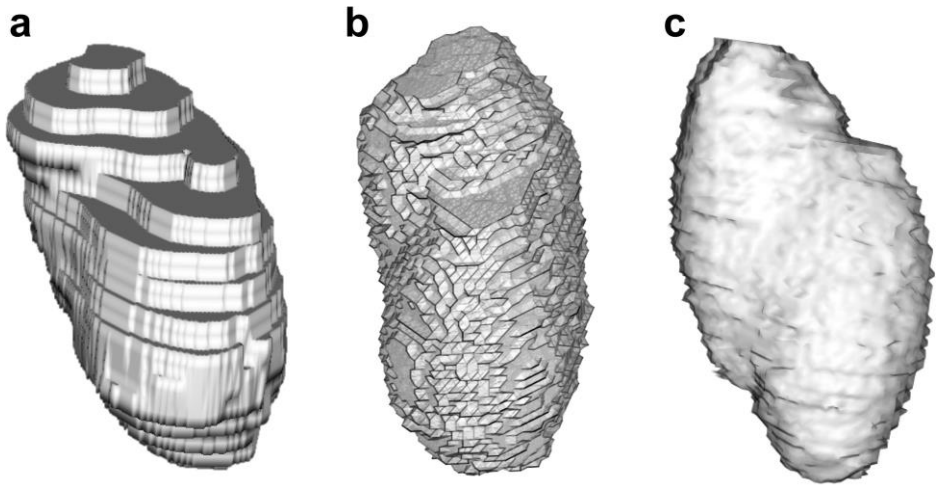
(5) Les « *neighborhood gray-level difference based matrices* » (NGTDM), qui contiennent la somme des différences de niveaux de gris par rapport à un niveau de gris donné i dans un périmètre de rayon donné. Après détermination de la méthode d'agrégation pour des images 3D, les indices radiomics peuvent être calculés (Amadasun et King, 1989).

(6) *Autres indices des texture*: d'autres indicateurs ont été employés dans la littérature radiologique mais ne sont pas répertoriés dans IBSI ou les logiciels ou implémentations que nous avons utilisés, à savoir les motifs binaires locaux, les fractales ou les ondelettes.

1.5.4.3. Indices de forme

Ces indices quantifient les propriétés géométriques du VOI. Il faut noter que le passage d'une succession de coupes 2D incluant un nombre fini de voxels (représentés sous forme de « cubes » en cas de voxels isotropiques) à une image en 3D nécessite de déterminer la méthode pour reconstruire les contours du VOI, ce qui influencera d'une part la valeur des indices de forme mais aussi de texture de 1er et 2nd ordre (Figure 1-24) (Limkin et al, 2019).

Figure 1-24. Influence sur la forme du post-traitement des volumes segmentés. Il s'agit du même exemple de sarcome indifférencié pléomorphe de grade 2 que pour la figure 1-22. (**a**) Correspond à la représentation brute de la segmentation manuelle coupe par coupe. (**b**) Correspond à la représentation après transformation de la surface du volume par maillage triangulaire (algorithme des "marching cubes"). (**c**) Correspond à (b) après application d'un filtre gaussien.



Nous pouvons choisir :

(i) de garder les volumes "brutes" ce qui donne des contours crénelés du à l'empilement des voxels cubiques avec effet de volume partiel et une surestimation du volume réel du VOI;

(ii) d'obtenir des volumes aux formes plus adoucies en reliant le centre des voxels des extrémités - ce qui est recommandé par IBSI pour des VOI avec des volumes internes importants ;

(iii) de calculer des volumes basés sur un maillage faits de triangle de la surface externe (« *mesh-based* »). La reconstruction du maillage peut se faire selon plusieurs méthodes, elles-mêmes susceptibles d'influences la valeur des indices radiomics de forme (Limkin et al, 2019).

En plus du volume, de l'aire de la surface du VOI, d'autres descripteurs de forme sont fréquemment utilisés comme l'élongation (: à quel point un objet est plus que large et étendu), la planéité (: à quel point un objet est plat comparé à sa longueur), sa sphéricité (: à quel point un objet ressemble à une sphère compacte).

Ces indicateurs de forme en situation néoadjuvante apparaissent adaptés aux STM puisque ces tumeurs ont naturellement une forme ovoïde avec un plus grand axe correspondant approximativement à l'axe principal de la loge musculaire qui les contient et elles présentent une tendance à l'aplatissement et une moindre sphéricité en cas d'efficacité des traitements (impression de « ramollissement de la tumeur » décrite pas les cliniciens)

1.5.4.4. Delta-radiomics

Les indices radiomics sont généralement utilisés à un instant « t », mais il peut aussi être pertinent de regarder leurs variations entre deux moments afin de quantifier les variations du phénotype radiologique sous une condition appliquée sur cette durée de temps. Nous pouvons calculer les delta-radiomics des façons suivantes:

- variation absolue: $\Delta X_{absolue} = X(t') - X(t)$
- variation relative: $\Delta X_{relative} = \frac{X(t')-X(t)}{X(t)}$
- variation absolue normalisée par la temps: $\Delta X_{abs-dt} = \frac{X(t')-X(t)}{t'-t}$
- variation relative normalisée le temps: $\Delta X_{rel-dt} = \frac{X(t')-X(t)}{X(t)} \times \frac{1}{t'-t}$

où t et t' sont les deux moments de l'évaluation avec t' > t.

Il n'y a pas de consensus sur la meilleure méthode pour calculer les delta-radiomics et à notre connaissance, aucune étude n'a comparé les performances de modèles prédictifs basés sur chacune des ces quatre méthodes. Cependant, nous pouvons supposer que l'utilisation d'une différence normalisée par le temps s'impose s'il existe

une forte variation des délais entre les deux examens réalisés d'un patient à l'autre (Cherezov et al, 2018).

1.5.4. Etapes des approches radiomics

Plusieurs étapes sont nécessaires à la réalisation d'études radiomics afin d'homogénéiser le jeu de données. Chacune de ces étapes est susceptible d'influencer la valeur des indices radiomics et donc la prédiction souhaitée, ce qui nécessite une méthodologie transparente et clairement énoncée, avec des particularités propres à chacune des modalités d'imagerie.

1.5.4.1. Comment définir la stabilité d'un indice radiomics?

Une première définition serait de considérer qu'un indice est stable quand il ne varie pas de manière *trop importante* selon des variations dans son processus d'obtention correspondant à des variations usuelles, du bruit, non liées aux phénomènes biologiques mesurés. Ainsi, un indices radiomics stable ne devrait pas varier s*ignificativement* selon que le radiologue qui a segmenté la tumeur, ou selon que le jour ou l'heure d'acquisition de l'imagerie dont il provient (: « *test – retest* »), ou encore selon le constructeur de la machine d'acquisition...

Une première étape de filtrage des indices radiomics est ainsi de plus en plus couramment réalisée. Elle consiste à extraire au moins deux fois les indices quantitatifs radiomics selon le facteur de variation potentiel d'intérêt et à :

(i) Soit tester si la variation n'est pas significative selon des tests de type t-test appariés ou ANOVA à mesure répétée (ou leur équivalent non-paramétrique) - selon le nombre de fois où les indices radiomics ont été mesurés.

(ii) Soit mesurer leur variabilité et exclure ceux étant sous un seuil définissant la stabilité: « *intra-class correlation coefficien*t » (ICC), rho de Spearman ou coefficient de variation. Il est ainsi fréquent de voir exclu les indices radiomics dont l'ICC est inférieur à 0.9 après 2 segmentations.

Pour limiter l'exclusion d'indices radiomics qui pourraient être discriminants pour la question oncologique posée mais qui seraient trop sensibles aux étapes d'acquisition et d'extraction, les méthodes suivantes peuvent être proposées :

(i) standardiser prospectivement en amont les acquisitions d'image ;

(ii) homogénéiser a posteriori le jeu de données par des techniques de correction ;

(iii) enfin, concernant l'influence de la machine d'acquisition ou encore de l'antenne IRM employée... etc., nous pouvons nous référer à la méthode employée par Orlhac *et al.* (2018 ; 2019) au scanner et au PET/CT, c'est-à-dire à la technique d'harmonisation ComBat. ComBat a d'abord été développé pour limiter l'effet de lot (ou « *batch effect* ») des analyses génomiques puis appliqués en neuroradiologie pour corriger les mesures de tenseur de diffusion ou d'épaisseurs corticales en IRM structurale pour des jeux de données multi-sites (Johnson et al, 2007; Fortin et al, 2017; Fortin et al. 2018). ComBat suppose qu'il existe une relation linéaire entre l'indice mesuré et toutes les variables biologiques et techniques liées à l'acquisition avec un terme d'erreur incluant un facteur multiplicatif lié au protocole d'acquisition, pouvant s'écrire ainsi:

$$y_{i,j,v} = \alpha_v + X_{Ti,j} \times \beta_v + \gamma_{j,v} + \delta_{j,v} \times \epsilon_{i,j,v}$$

Où :

- $y_{i,j,v}$: valeur d'intérêt dans la VOI v, pour le sujet i, pour la machine j
- $X_{Ti,j}$: vecteur de covariables fixes pour le sujet i sur la machine j
- α_v et β_v: sont l'interception et la pente du vecteur de covariables pour la VOI v
- $\gamma_{j,v}$: effet additif du protocole sur la machine j pour la VOI v
- $\delta_{j,v}$: effet multiplicatif du protocole sur la machine j pour la VOI v
- $\varepsilon_{i,j,v}$: terme d'erreur suivant une loi normale d'espérance nulle

L'algorithme de ComBat va estimer les paramètres $\gamma_{j,v}$ et $\delta_{j,v}$ en utilisant un modèle bayésien naïf (nommés respectivement $\boldsymbol{\gamma}^*_{j,v}$ et $\delta^*_{j,v}$). Ainsi, en nommant α^*_v et β^*_v des estimateurs de α_v et β_v, nous obtenons la correction suivant des indices radiomics « y « :

$$y_{i,j,v}^{corrigé} = [y_{i,j,v} - \alpha_v^* - X_{Ti,j} \times \beta_v^* - \gamma_{j,v}^*] / \delta_{j,v}^* + X_{Ti,j} \times \beta_v^*$$

Les scripts de ComBat sont accessibles au public en R et en python, avec des versions paramétriques et non-paramétriques.

1.5.4.2. Sélection des données: pour quelle question

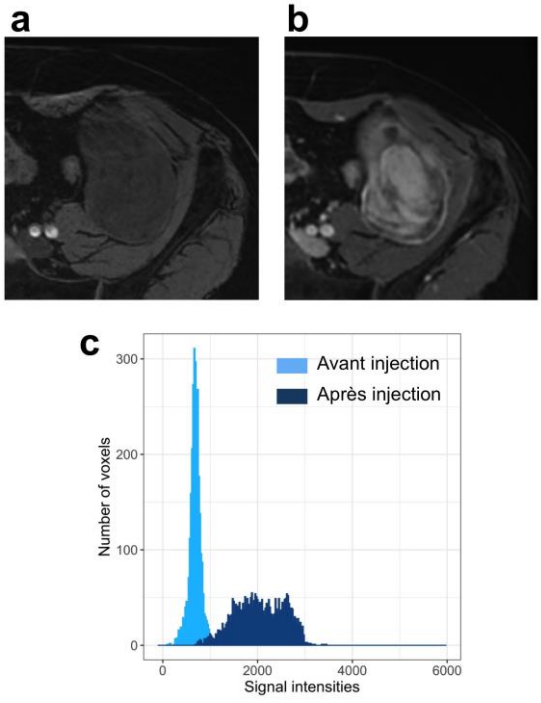
Réaliser une analyse radiomics nécessite d'avoir un jeu de donnée d'imagerie homogène et qui permet de répondre à la question posée. Chaque modalité et séquence apporte une information qui lui est propre, permettant de répondre à certaines questions mais non à d'autres

Du fait du manque de standardisation des protocoles IRM, il est fréquent de devoir exclure certains patients car les séquences les plus informatives pour la question posée n'ont pas été réalisées ou réalisées selon des techniques d'acquisition différentes.

Pour exemple, si la question est de déterminer si les variations précoces de l'hétérogénéité du rehaussement tumoral en IRM sous anti-angiogénique permettent d'anticiper le statut de la réponse selon RECIST v1.1, il est nécessaire de n'inclure que des patients ayant des séquences T1 injectées au baseline et à chacune des évaluations. Ensuite, les techniques d'acquisition d'images et notamment l'application de saturation de signal de la graisse pouvant être inconstantes, d'autres patients peuvent s'en retrouver exclus. Dans notre expérience, cette étape peut aboutir à l'exclusion de plus de 60% des patients.

Concernant le scanner, il apparait difficilement concevable d'inclure dans une même étude des acquisition sans, avec injection au temps artériel, avec injection au temps portal, tant le phénotype radiologique se trouve modifié après injection (Figure 1-25).

Figure 1-25. Influence de l'injection de produit de contraste sur les indices de texture. Exemple basé sur le même STM que les figures 1-22 et 1-23. (**a**) Séquence axiale T1 echo de gradient sans injection puis (**b**) après injection intraveineuse de chélates de gadolinium (acquisition à 2 mn). (**c**) L'histogramme retranscrit cette impression visuelle de modifications de la texture tumorale avec un étalement de l'histogramme vers de plus hautes intensités de signal. L'histogramme correspond à celui de la section (2D) affichée de la tumeur.



Le choix du matériel radiologique de base des approches radiomics est donc la résultante d'un compromis entre:

- le(s) type(s) d'imagerie(s)/séquence(s) le(s) plus pertinent(s) pour répondre à la question posée
- le plus petit commun dénominateur parmi les protocoles réalisés chez tous les patients potentiellement incluables.

Il est aussi essentiel de réaliser un contrôle qualité des examens d'imagerie, afin d'exclure ceux tronquant la tumeur (défaut de couverture, bande de saturation couvrant partiellement la tumeur...), les examens avec des bandes d'artefacts affectant la tumeur (mouvements, répétition des vaisseaux, distorsions...) ou un bruit normal (problèmes d'antenne...).

Plusieurs études ont souligné l'influence du constructeur et des paramètres d'acquisition sur les valeurs des indices radiomics issus du scanner sur fantômes et de l'IRM pour des séquences structurales et de diffusion (b-values variables) sur fantôme ou acquisitions réelles chez l'homme avec création d'imageries synthétiques (Mackin et al, 2015 ; Vallières et al, 2017 ; Buch et al, 2017 ; Buch et al, 2018 ; Berenguer et al, 2018 ; Ford et al, 2018 ; Becker et al, 2018). De plus, Vallières et *al.* (2017) ont montré que ces variations de paramètres d'acquisitions en IRM structurale pouvaient influencer la prédiction en situation clinique (ici de rechute métastatique).

Pour pallier à ce problème et puisqu'il est compliqué éthiquement, matériellement et financièrement de réitérer des acquisitions chez un patient, une proposition est d'utiliser un fantôme présentant des composantes de textures différentes pour identifier les indices radiomics les plus stables et les sélectionner dans la suite de l'analyse. Une alternative serait aussi de reconstruire synthétiquement les images selon la méthode de Vallières et *al.* (pour l'IRM et le TEP/CT) afin d'en homogénéiser l'aspect.

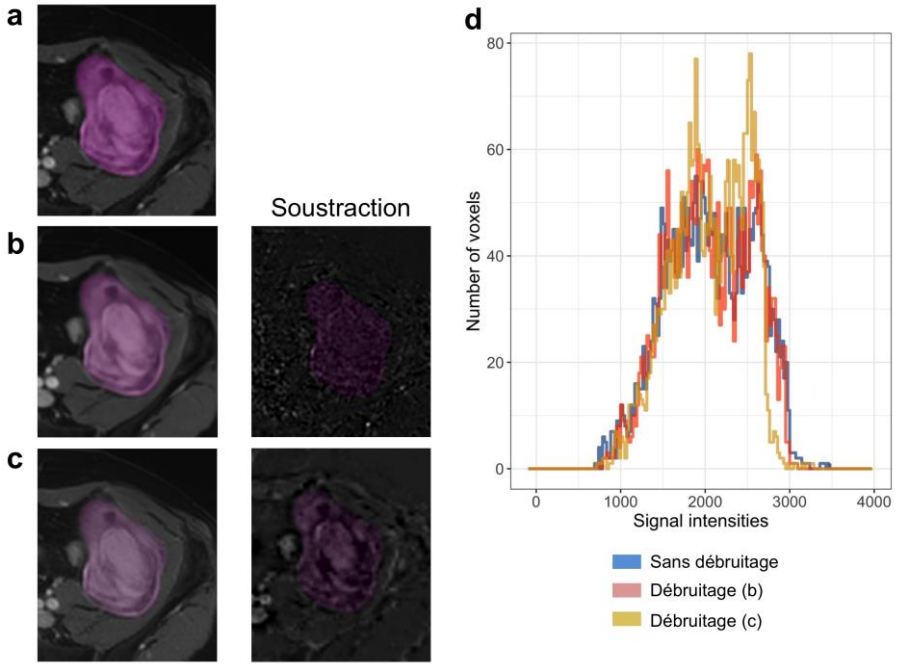
1.5.4.3. Post-traitement des images:

Le post-traitement recouvre 4 à 5 étapes selon la modalité d'imagerie choisie. Il permet d'homogénéiser les imageries du jeu de données et de permettre une extraction des indices radiomics:

1.5.4.3.1. Débruitage

Il existe un bruit gaussien (ou bruit blanc, suivant une distribution normale) et un bruit ricien (suivant une distribution de Rice) dans les images médicales susceptibles de modifier les valeurs des indices de texture au scanner, en IRM et PET/CT ainsi que les prédictions basées sur ces indices (Grootjans et al, 2016 ; Bagher-Ebadian et al, 2017 ; Bologna et al, 2019). Plusieurs algorithmes de débruitage peuvent être appliquées aux images médicales dans des librairies open source ou non (filtres gaussiens, ondelettes, moyennage des valeurs de pixel localement, semi-localement, à distance, par patch, pondéré, deep-learning... etc.). Cependant, elles risquent de rendre plus floues les images et de faire disparaitre les anomalies de texture réelles et non aléatoires que nous souhaitons mesurer par les approches radiomics (Figure 1-26).

Figure 1-26. Influence du débruitage sur la texture des tumeurs. Exemple basé sur le même STM que les figures 1-22, 1-24 et 1-25 (séquence axiale T1 fat sat après injection). (**a**) Représente l'image native, (**b**) correspond à la même image après application d'un débruitage gaussien par patch avec un rayon de patch de 1 recherché dans un rayon de 3 voxels, avec la soustraction entre l'image débruitée et l'image native. (**c**) Correspond à la même image après application d'un débruitage gaussien par patch avec un rayon de patch de 2 recherché dans un rayon de 6, avec la soustraction entre l'image débruitée et l'image native. L'algorithme utilisé est issu du package ANTsRCore et se base sur les travaux de Coupé et *al*. (2010). (**d**) Represénte les variations d'histogramme de la tumeur avant et après application de ces deux débruitages. L'histogramme correspond à celui de la section (2D) affichée de la tumeur.



Nous avons décidé de ne pas appliquer ces algorithmes dans les travaux exploratoires de cette thèse.

Depuis, Moradmand et *al.* (2019) ont montré que la reproductibilité des indices radiomics extraits d'IRM de gliobastomes était significativement influencée par les post-traitements d'images incluant, entre autres, différentes techniques de filtrage du bruit proposées par le package pyradiomics (*local binary pattern, wavelet, laplacian of gaussian, square, squareRoot*, logarithmique...).

Le recours aux algorithmes de débruitage dépend de la qualité du jeu de donnée radiologique, du volume des lésions étudiées, de la modalité d'imagerie, et n'apparaît pas inéluctable.

1.5.4.3.2. Correction N4 en IRM

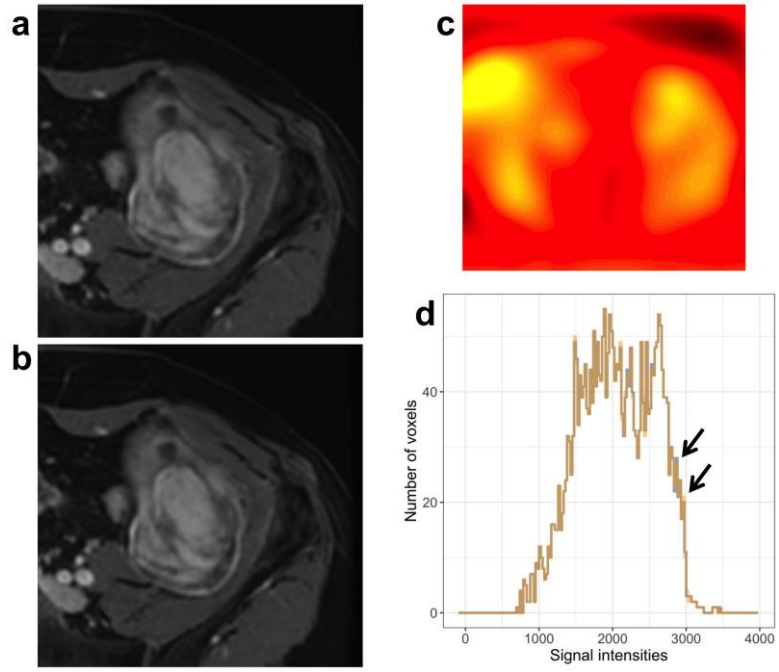
L'algorithme en question a pour objectif de corriger les défauts d'homogénéité du champ magnétique à basse fréquence. En effet, plus une région est éloignée du centre de l'aimant et de l'antenne IRM, plus il existe une dérive, progressive, de l'intensité de signal dans l'espace vers de plus faibles valeurs. Ce biais ne gêne pas/peu l'interprétation qualitative d'un radiologue mais peut influencer la valeur d'indices de texture de 1er et de 2^{ième} ordre (Figure 1-27).

Pour corriger ce biais potentiel dans nos travaux, nous avons appliqué la correction N4 (Tustison et al, 2010). Le modèle utilisé dans la correction N4 est:

$$v(x) = u(x) \times f(x) + n(x)$$

Où v est une image donnée, u est l'image idéale non biaisée, f est le biais, n est le bruit (avec pour approximation qu'il ne serait que gaussien), et x est un point dans l'image. Après transformation logarithmique de l'équation et élimination du facteur correspondant au bruit (considéré nul), la correction N4 consiste à faire une approximation itérative du biais de champ magnétique par courbes B-splines jusqu'à convergence. Quand la convergence est atteinte, l'image est retransformée dans son unité d'origine.

Figure 1-27. Influence de la correction N4 sur la texture des tumeurs. Exemple basé sur le même STM que les figures 1-22, 1-24, 1-25 et 1-27 (séquence axiale T1 fat sat après injection). (**a**) Représente l'image native, (**b**) correspond à la même image après application de la correction N4 (Tustinson et al, 2010 – package ANTsR). Les variations sont peu visibles à l'oeil nu. (**c**) Correspond au ratio entre (a) et (b) permettant d'illustrer où la correction a été la plus forte. (**d**) Représente les variations de l'histogramme des intensités de signal de la tumeur avant et après application de la correction. Les variations sont de l'ordre du décalage d'un niveau de gris pour certains voxels (flèches noires). L'histogramme correspond à celui de la section (2D) affichée de la tumeur.



1.5.4.3.3. Discrétisation des niveaux de gris

La discrétisation des niveaux de gris consiste à déterminer en combien de valeurs possibles de niveaux de gris vont être réduites l'ensemble des valeurs contenues dans le volume d'acquisition. En pratique, deux méthodes de discrétisation sont possibles, nous pouvons définir:

- soit un nombre fixe de niveaux de gris (méthode FBN pour « *fixed bin number* ») - par exemple, convertir l'ensemble des niveaux de gris possibles dans l'image en 32 valeurs possibles ;

- soit une taille de *bin* fixe (méthode FBS pour « *fixed bin size* ») - par exemple choisir qu'à une valeur de gris correspond « x » UH au scanner.

Les nombres de *bins* observés dans la littérature radiomics sont: 32, 64, 128 et 256.

En pratique, le choix de la méthode dépend de la modalité d'imagerie, de la question posée et de l'usage préalable de techniques de normalisation des niveaux de gris pour les modalités d'imagerie sans unité. Selon IBSI, les méthodes FBN et FBS sont toutes

deux envisageables si les unités sont calibrées, tandis que la méthode FBN est préconisée dès que les unités deviennent arbitraires. En effet, prenons l'exemple d'une tumeur imagée par scanner présentant deux sous-types différents à différencier par une analyse radiomics. Si dans une première approche exploratoire, il s'avère que cette tumeur présente des densités toujours comprises entre 30 et 40 unités Hounsfield (UH) quelque soit le sous-type, choisir une largeur de bin de 10 UH masquerait les différences de texture potentielles entre ces sous-types. Nous proposerions alors préférentiellement la méthode FBS avec une largeur de bin d'environ 0.1 UH (en sachant que plusieurs largeurs pourront être évaluées pour déterminer leur influence sur les prédictions).

Dans le cas de l'IRM structurale, les intensités de signal n'ont pas d'unité fixe mais une valeur dépendant des paramètres d'acquisition et de données physico-chimiques des tissus analysés. Ainsi, l'intensité de signal d'un tissu sur une séquence pondérée T2 sur une IRM donnée, un jour donné, ne sera pas la même le jour suivant ou sur une autre IRM. L'analyse du radiologue se fait plus ou moins consciemment en comparant l'intensité de signal d'un objet d'intérêt aux intensités de signal de tissus apparemment sains environnant (liquide, muscle, gras...). Les approches FBS et FBN peuvent donc se discuter, ainsi qu'un post-traitement préalable à la discrétisation visant à redonner du sens clinique aux intensités de signal brutes par normalisation des intensités de signal (Cf. infra).

Concernant l'IRM « fonctionnelle » avec analyse radiomics réalisée sur des cartes paramétriques avec une valeur théoriquement standardisée (par exemple: ADC, K^{trans}, K^{ep}... etc), nous rejoignons le cas du scanner où il semble licite de proposer une technique de discrétisation du FBS avec une largeur de *bin* fonction du tissu évalué et de la question posée.

L'influence de la méthode de discrétisation et du nombre de niveaux de gris sur la reproductibilité des indices radiomics de texture a été montrée sur des cohortes de patients atteints de cancer ORL, cérébraux, mammaires, de lésions orbitaires, ou encore sur des fantômes, cela au scanner, PET/CT et en IRM, avec, selon les séries entre 33% et 95% des indices radiomics présentant un ICC inférieur à 90% (Shafiq-Ul-Hassan et al, 2017 ; Altazi et al, 2017 ; Duron et al, 2019 ; Branchini et al, 2019 ; Goya-Outi et al, 2018). Citons à ce titre l'étude de Goya-Outi et *al.,* (2018) qui a évalué 3 méthodes (largeur de bin constant et intervalles relatifs, nombre constant de bins et intervalles relatifs, et nombre constant de bins et intervalles absolus) puis

corrélé les indices de texture extraits de gliomes pontiques ainsi obtenus à la classification de l'hétérogénéité de ces tumeurs par des neuroradiologues experts pour aboutir à la conclusion que la méthode « nombre constant de bins et intervalles absolus » devait être privilégiée.

Dans les travaux de cette thèse, les IRM structurales ont toujours bénéficié d'une transformation préalable pour normaliser les intensités de signal et les rendre comparable puis nous avons appliqué une technique FBS.

1.5.4.3.4. Standardisation des tailles des voxels

Les tailles des voxels sont souvent variables d'un patient à l'autre en raison de l'absence de *guidelines* précises et, dans le cas des STM, de différences importantes des dimensions et topographies des tumeurs. La plupart des imageries ont une résolution dans le plan entre 0.7 x 0.7 et 1.2 x 1.2 mm^2 et une épaisseur comprise entre 1 et 6 mm. Du fait de cette hétérogénéité du jeu de données, une standardisation des tailles des voxels est requise, nécessitant une interpolation des valeurs de niveaux de gris des nouveaux voxels. Cela introduit deux nouvelles variables susceptibles d'influencer les indices de texture i.e. la taille des voxels reconstruits et la méthode d'interpolation (plus proche voisin, bilinéaire, bicubique, polynomiale de degré supérieur...), et une variable susceptible de modifier les indices de forme i.e. la taille des voxels reconstruits.

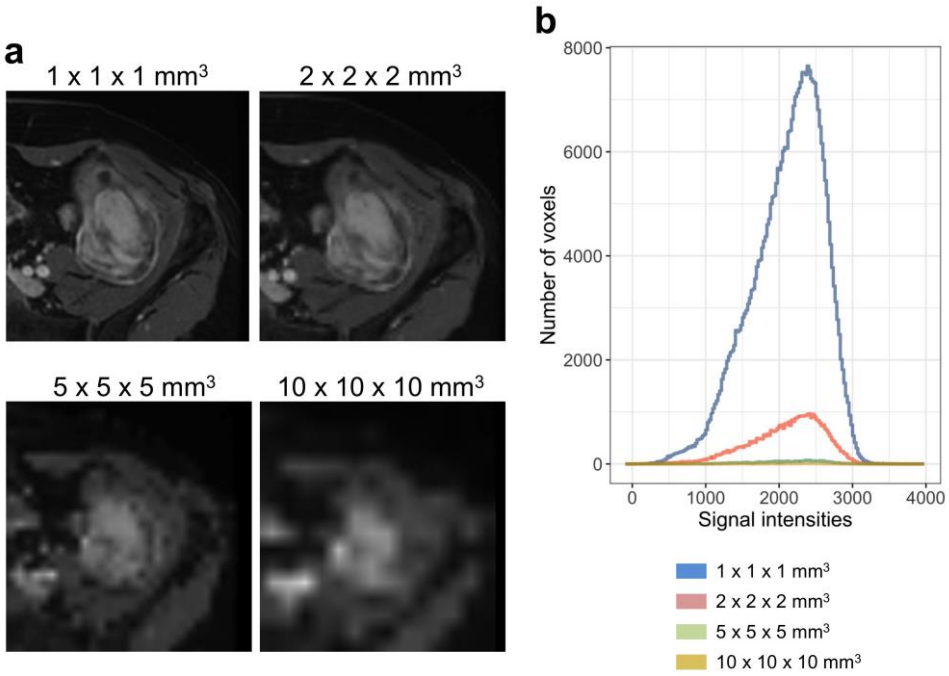
Ainsi, des études sur fantômes et sur un jeu de données de scanner de cancer pulmonaire ont montré que plusieurs indices radiomics n'étaient pas stables selon ces nouvelles variables (Shafiq-ul-Hassan et al, 2017 ; Shafiq-ul-Hassan et al, 2018 ; Mackin et al, 2017). La Figure 1-28 illustre cette influence.

L'influence de la taille des voxels a aussi été confirmée en IRM structurale (séquences T1 et T2) sur le fantôme virtuel sain de BrainWeb (Collins et al, 1998 ; Bologna et al, 2019). Les auteurs ont montré que la plupart des indices de texture de 2^{nd} ordre avaient un ICC sous la borne des 0.9, mais que réaliser une standardisation a posteriori permettait la stabilité des indices extraits des séquences T1 mais non T2. Les indices de forme et de 1er ordre cependant paraissaient d'emblée stables en l'absence de standardisation.

Citons aussi l'étude de Goya-Outi et al. (2018) basées sur une cohorte de patients atteints gliomes pontiques pédiatriques et qui montre que de nombreux indices de

texture de 1er ordre et de 2nd ordre sont fortement corrélés au volume des VOIs , ce qui suppose que ce volume soit homogénéisé dans les groupes à comparer.

Figure 1-28. Influence de la taille des voxels sur la texture des tumeurs. Exemple basé sur le même STM que les figures précédentes. (**a**) Représente la même coupe avec des reconstruction de 4 résolutions différentes (interpolation par b-splines). (**b**) Histogrammes des intensités de signal contenues dans le volume entier de la tumeur (VOI 3D) montrant un aplatissement de l'histogramme par diminution du nombre de voxels et moyennage des intensités de signal.



1.5.4.3.5. Normalisation des intensités de signal en IRM

A l'inverse du PET/CT ou du scanner, l'échelle des niveaux de gris en IRM structurale est arbitraire, c'est-à-dire que l'intensité de signal du gras ou du muscle pour un même patient peut varier selon les paramètres d'acquisition, la machine ou le jour de l'examen. Une étape supplémentaire est nécessaire consistant à normaliser les intensités de signal pour rendre comparables les IRM et s'assurer que les valeurs des organes sains environnants et des mêmes phénomènes pathologiques chez des patients différents soient similaires. Cette question a d'abord été étudiée en neuroradiologie, hors approche radiomics, afin de pouvoir regrouper des cohortes multicentriques d'IRM cérébrales et s'assurer que les mesures faites sur la substance blanche ou le cortex « apparemment sains » de patients atteints de maladies neuro-inflammatoires ou neuro-dégénératives étaient bien comparables, ou pour suivre longitudinalement

les patients. Shinohara et *al.* (2014) ont ainsi exposé 7 principes pour une normalisation idéale des IRM :

- elle doit avoir une même interprétation pour un même type de tissu quelque soit la localisation;
- elle doit être réplicable;
- elle doit préserver l'ordre des intensités de signal;
- elle doit avoir une même distribution dans un même tissu d'intérêt pour un même patient et entre des patients différents;
- elle doit être la moins sensible possible au bruit et aux artefacts;
- elle ne doit pas être influencé par l'hétérogénéité de la population;
- elle ne doit pas résulter en une perte d'information relative aux pathologies.

Plusieurs méthodes sont implémentées en langage libre, particulièrement en neuro-imagerie. Les plus simples consistent à:

(i) *standardiser* les valeurs selon:

$$SI_{corrected,v} = \frac{SI_v - mean(SIs)}{sd(SIs)}$$

Où $SI_{corrected,v}$ est l'intensité de signal du voxel « v » après correction, SI_v: Intensité de signal du voxel « v » avant correction, mean(SIs): moyenne de toutes les intensités de signal contenues dans le volume d'acquisition, sd(SIs): écart-type de toutes les intensités de signal contenues dans le volume d'acquisition ;

(ii) *normaliser* les valeurs selon:

$$SI_{corrected,v} = \frac{SI_v - \min(SIs)}{\max(SIs) - \min(SIs)}$$

Où min(SIs) est le minimum de toutes les intensités de signal contenues dans le volume d'acquisition, max(SIs) est le maximum de toutes les intensités de signal contenues dans le volume d'acquisition.

(iii) d'autres auteurs ont proposé de diviser toutes les valeurs par la valeur moyenne des intensité de signal dans un organe sain de référence que nous appellerons "*tissue-normalisation*" (selon les études, le graisse sous-cutanée, la rate, le muscle strié);

(iv) réaliser un « *histogram-matching* » des intensités de signal d'une acquisition avec celles d'une acquisition de référence (éventuellement déjà normalisée ou standardisée)

voire avec l'histogramme moyen de plusieurs acquisitions. Cette méthode a été proposée par Nyul et Udupa pour les IRM (Nyul et Udupa, 1999). Elle consiste donc à transformer de manière non linéaire l'histogramme des intensités de signal d'une imagerie pour la rendre plus proche de l'histogramme des intensités de signal d'une image de référence. Cela est réalisé en 2 étapes. La 1ère étape d'entrainement consiste à donner comme input l'acquisition de référence avec un nombre défini de percentiles qui serviront de repères réguliers le long de l'histogramme de cet input. La 2nde étape consiste en la transformation du jeu de données d'imagerie.

Soit pour une image de référence:

- p_1: intensité de signal du percentile minimal
- p_2: intensité de signal du percentile maximal
- N: le percentile N
- i: la i-ème valeur de percentile
- μ_{Ni}: intensité de signal correspondant au Ni-ème percentile
- s_1: intensité minimale de l'image d'intérêt standardisée
- s_2: intensité maximale de l'image d'intérêt standardisée

Les *landmarks* seront par exemple L = { p_1 ; p_2 ; μ_{25} ; μ_{50} ; μ_{75} }

Pour une nouvelle image « j » à matcher avec l'image de référence, la formule pour transformer x $\in [p_{1,j} ; p_{2,j}]$ en x' dans $[s_1 ; s_2]$ est donnée par la relation :

$$x' = s_1 + \frac{x - p_{1,j}}{p_{2,j} - p_{1,j}} \times (s_2 - s_1)$$

Puis, plusieurs mappings linéaires sont réalisés pour transformer chaque segment $[\mu_{Ni,j} ; \mu_{Ni+1,j}]$ en $[\mu_{Ni} ; \mu_{Ni+1}]$ (Figure 1-29). La Figure 1-30 illustre cette transformation pour plusieurs valeurs de *landmarks* possibles.

Cette méthode a été beaucoup utilisée en raison de sa facilité d'implémentation et de sa rapidité de calculs, mais elle peut produire des images d'interprétation difficile pour l'œil du radiologue, selon les références choisies ou le nombre de *landmarks*.

D'autres méthodes avancées (WhiteStripe, RAVEL, DeepHarmony) ont été développées en neuroradiologie mais elles n'ont pas d'équivalent pour les autres situations anatomiques non-cérébrales.

Figure 1-29. Principe de l'histogram-matching. (**a**) La 1ère étape consiste à aligner les histogrammes. Les bornes de l'histogramme standard/de référence sont préalablement fixées (s_1, s_2). Les histogrammes des IRM servant pour l'entrainement sont utilisées pour calculer le landmark μ sur l'échelle standard (adapté de Nyul et *al.*, 2000). (**b**) Transformation d'une IRM i par histogram-matching une fois l'échelle standard de référence construite, par alignement linéaire, morceaux par morceaux des niveaux de gris pour "l" landmarks.

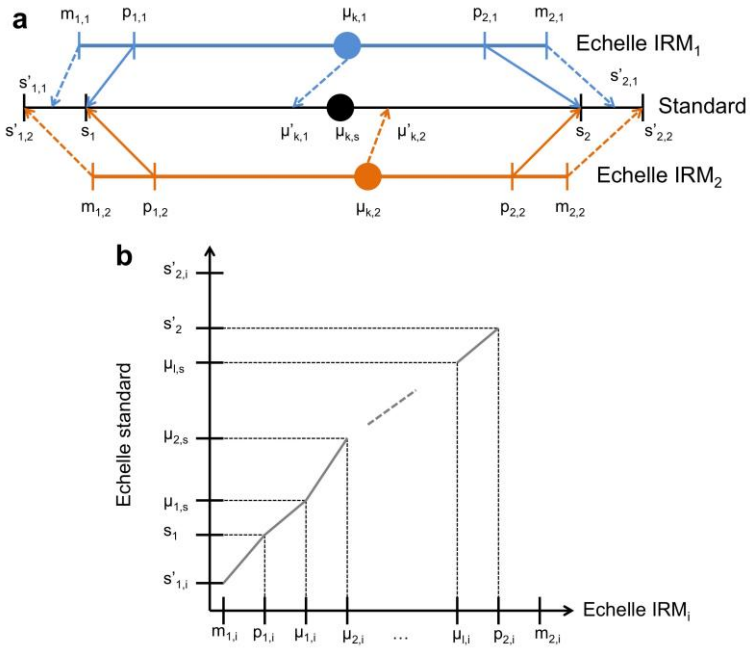
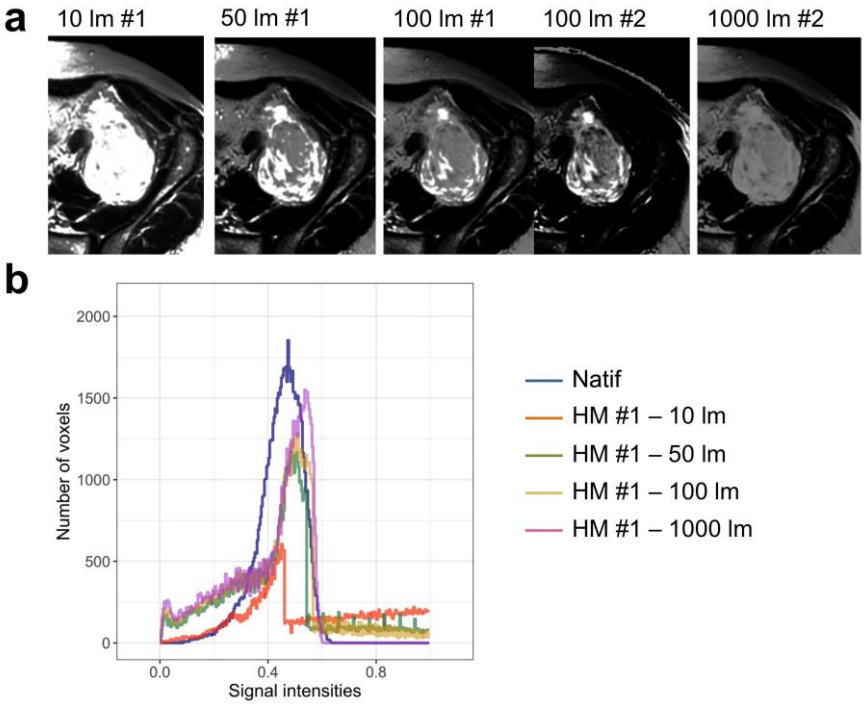


Figure 1-30. Séquences axiales pondérées T2 d'un STM reconstruites après histogram-matching avec les IRM de 2 autres patients comme références (#1 et #2) et différents landmarks, illustrant les modifications de textures des tumeurs (**a**). Histogrammes standardisés entre 0 et 1 illustrant les variations induites par le nombre de landmarks (**b**). Abbréviation: lm: landmarks. Figures réalisées avec le package R « hatch ».



Peu de travaux se sont intéressées à l'influence de méthodes de normalisation sur les approches radiomics alors que cette étape, visuellement (Figure 1-30), peut modifier l'architecture tumorale. Nous mentionnerons la publication de Collewet *et al.* (2004), un équivalent d'étude de fantôme consistant à classer des fromages par leur ancienneté selon des indices de texture de 1er et 2^{ième} ordre issus de séquence T2. Trois techniques simples pour normaliser les images avaient été utilisées: même maximum pour toutes les IRM, même moyenne pour toutes les IRM, et en ne gardant que les intensités de signal entre la moyenne des intensités de signal de la VOI ± 3 écart-types redistribuées en 64 niveaux de gris. Les auteurs en concluaient que la classification des fromages pouvait être améliorée par la méthode de normalisation des intensités de signal.

Concernant des données acquises chez l'homme, nous pouvons citer deux travaux récents concernant l'IRM prostatique et concluant à l'influence directe sur les indices radiomics. Scalco *et al.* (2020) soulignent le manque de stabilité des indices radiomics extraits d'organes sains environnant entre deux IRM acquises à deux moments différents malgré l'application des techniques de normalisation précédemment citées pour le corps, avec toutefois de meilleurs résultats obtenus par *histogram-matching* (ICC le plus élevé à 0.76). A contrario, utiliser une technique inadaptée risque de biaiser les indices radiomics (Isaksson et al, 2020). Citons aussi l'étude de Lacroix et al., (2020) utilisant une standardisation de séquences T2 d'adénocarcinomes bronchiques selon les intensités de signal d'une région saine. Les auteurs ont testé plusieurs tissus et retenu le tissu adipeux. Ils ont ensuite montré que davantage d'indices radiomics devenaient corrélés à l'outcome en combinant correction N4 et normalisation.

1.5.4.4. Segmentation des objets d'intérêt

La segmentation des volumes d'intérêt peut être faite : (i) manuellement, coupe par coupe idéalement par un radiologue expert de la pathologie, (ii) semi-automatiquement avec correction manuelle ou (iii) entièrement automatiquement.

La limite de la segmentation manuelle est son caractère chronophage (entre 10 et 45 mn pour un STM), laborieux et rapidement lassant pour le radiologue, entrainant un risque d'erreur à mesure qu'elles se répètent. La segmentation automatique apparait souhaitable pour limiter les biais de mesure, le temps passé et s'assurer d'une méthodologie similaire pour tous les patients. Cependant, à l'heure actuelle aucun

algorithme libre ne permet une segmentation satisfaisante des STM sur l'IRM structurale et le temps passé à corriger les segmentations proposées font que nous avons privilégié une segmentation manuelle. Nous pouvons toutefois noter la publication de 4 études (3 chez l'homme, 1 chez le petit animal) visant à développer une segmentation automatisée en 3D des STM soit sur la base d'imagerie multiparamétrique (IRM structurale et ^{18}F-FDG-PET/CT) ou de ^{18}F-FDG-PET/CT seule ou d'IRM structurale seule (séquences T1, T2, T1 après injection) par apprentissage profond, i.e. soit par transfert learning ou bien par réseau de neurone entièrement convolutionnel en particulier de type U-net fonctionnant notamment par patch 3D (Hermessi et al, 2018 ; Holbrook et al, 2019 ; Guo et al, 2019 ; Peng et al, 2019).

Plusieurs études se sont intéressées aux variations des indices radiomics et des prédictions selon les segmentations manuelles, ainsi qu'aux facteurs influençant la qualité de ces segmentations. En particulier, Zhang et al (2019) ont montré sur un *dataset* de cancers naso-pharyngés et de cancers du sein en IRM que plus les tumeurs étaient volumineuses, moins les indices étaient reproductibles. Si une technique de sélection des indices radiomics consiste à ne garder que ceux avec un ICC > 0.9 après plusieurs segmentations, les auteurs ont montré que seuls 15% des indices radiomics remplissant cette condition étaient conservés ensuite dans les modèles pronostics. Enfin, des variations non négligeables des AUROC de modèles prédictifs de rechute métastatique ou d'envahissement ganglionnaire étaient observées selon les VOI étaient érodées ou dilatés pour nombres variables de voxels (de 0.588 à 0.749 pour une segmentation sur du T2 par exemple). Par ailleurs, Yang et *al.* (2019) ont pu mettre une plus grande robustesse des NGTDM aux variations de contourages au PET/CT, ainsi qu'une co-influence du nombre de niveaux de gris - si la segmentation est faite après le post-traitement d'images.

1.5.4.5. Extraction des indices radiomics

Plusieurs outils permettent d'extraire les indices radiomics:

- via des software payant (TexRad, OleaSphère) ou gratuit (LIFEx, CGTIA, MaZda, IBEX) (Nioche et al, 2018 ; Zhang et al, 2015)
- via des packages en R: "RIA", "radiomics";
- via le package en python: "pyradiomics" (van Griethuysen et al, 2018);
- via le package en Matlab: "RADIOMICS" ;

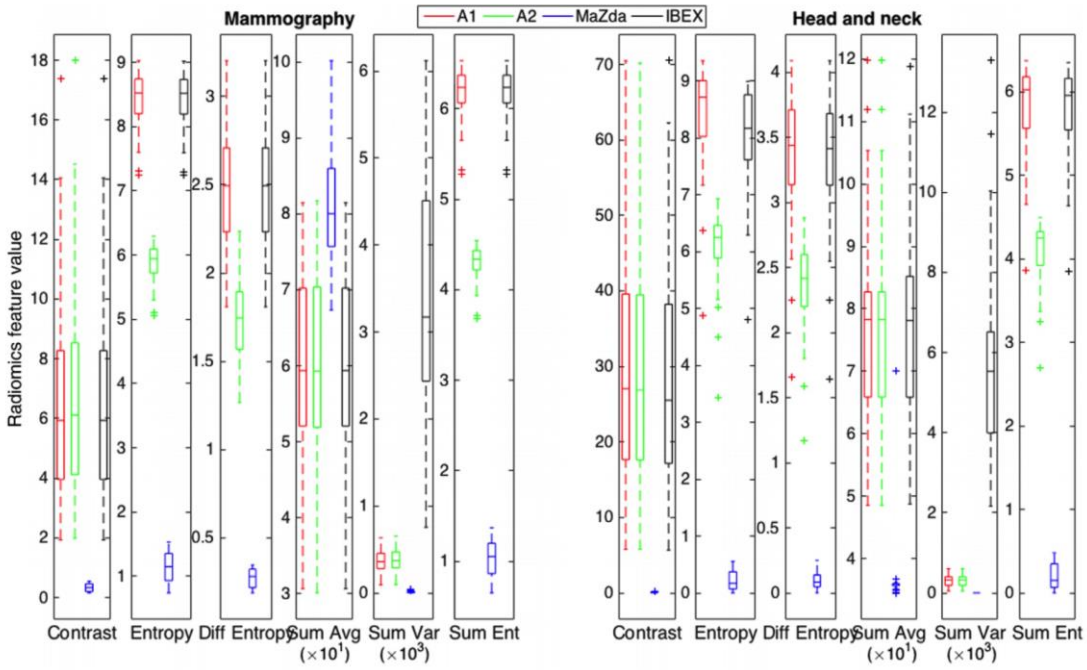
ainsi que, dans certaines études, via des codes *in-house* dans divers langages.

Si, en théorie, les indices radiomics sont calculés selon des formules issues des mêmes références bibliographiques, trois études, à notre connaissance, ont identifié des variations de ces indices selon l'outil employé (Bianchi et al, 2019 ; Foy et al, 2018 ; Fornacon-Wood et al, 2020). En particulier, Foy et *al.* ont comparé les indices radiomics obtenus selon 4 outils (MaZda, IBEX et 2 codes in-house issus de laboratoire de recherche en imagerie) sur des jeux de données d'imageries médicales cliniques (scanner injecté ORL et mammographie) et montré qu'il existait des différences significatives pour la totalité des indices radiomics de 1er et 2nd ordre issus du scanner. Pour les mammographies, seule la « *skewness* » restait stable. Les auteurs ont ensuite recherché les causes de ces différences et mis en évidence :

- l'absence de contrôle de tous les paramètres liés aux GLCM impliquant que certains soient fixés arbitrairement (notamment le nombre maximal de directions, de niveaux de gris pris en compte);
- l'import des images et leur pré-traitement. En effet, certains outils n'importent pas les voxels avec des valeurs négatives ou appliquent une transformation aux valeurs des voxels avant les calculs des indices radiomics;
- des variations dans les formules des calculs d'un indice appelé pareillement (typiquement toutes les variantes de l'entropie) ou des indices correspondant aux mêmes formules mais appelés différemment (par exemple *GLCM_energy* correspondant à *GLCM_uniformity ou GLCM_uniformity_of_energy* ou *GLCM_angular_2nd_moment* selon les outils);

La Figure 1-31 extraite de l'article de Foy et *al.* (2018) illustre ces variations.

Figure 1-31. Influence du logiciel de calcul sur les valeurs des indices radiomics – figure extraite de Foy *et al*, 2018. Les distributions de 6 indices radiomics issus des GLCM ont été affichées pour deux cohortes: mammographies et cancers ORL, selon 4 méthodes d'obtention: deux codes in house (A et B, utilisés par 2 groupes de recherche en radiomics), MaZda et IBEX.



1.5.4.6. Manipulation, transformation et analyse exploratoire des données radiomics

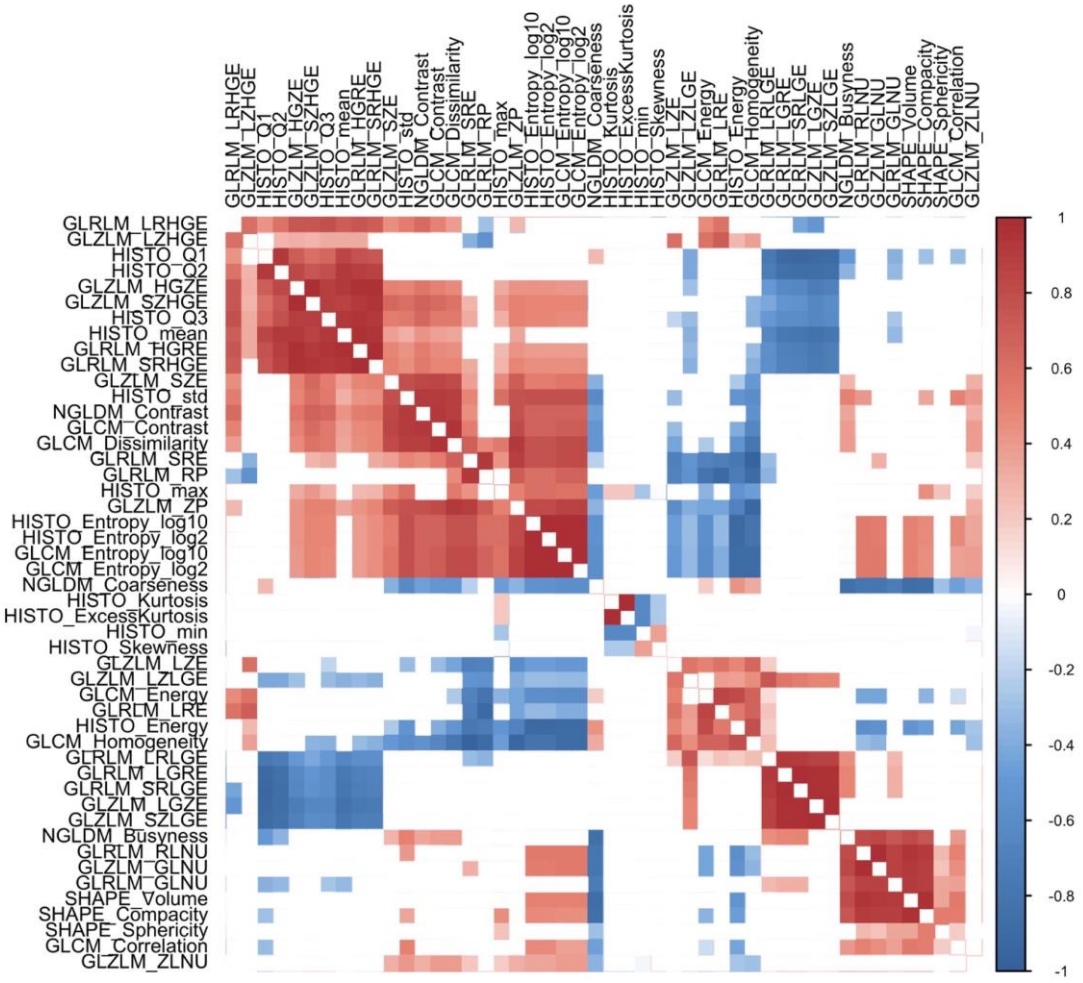
Les indices radiomics bruts issus d'une même famille sont souvent fortement corrélés entre eux (Figure 1-32) et l'échelle de leurs valeurs peut présenter de très forte variations. Les étapes initiales de l'analyse consistent donc à identifier ces corrélations, réduire les dimensions du jeu de données et remettre toutes les variables sur une même échelle.

1.5.4.6.1. Transformation des variables

Nous avons uniquement utilisé un *standard scaling* (ou *Z-score standardization*) pour que tous les indices radiomics soient centrés sur 0 avec un écart-type de 1.

Toutefois, d'autres techniques de transformation auraient pu être envisagées: la standardisation (permettant de ramener toutes les valeurs possibles de X entre 0 et 1), la transformation de Box-Cox (afin de rendre la distribution de X normale), logarithmique ou encore la transformation LOESS (pour *LOcally Estimated Scatterplot Smoothing* qui consiste à lisser les données pour en réduire le bruit)

Figure 1-32. Correlation plot des indices radiomics de 1er et 2nd ordre (issus des séquences T1 Fat Sat après injection d'agents de contraste d'une cohorte de 63 STM, post-traitée similairement). L'encodage couleur est fonction de la valeur du rho de Spearman (échelle sur la droite du graphique). Seules les corrélations significatives au seuil $p < 0.05$ sont colorées.



1.5.4.6.2. Sélection des variables

Concernant les travaux de cette thèse, nous avons employé (voire combiné) trois méthodes:

- La première méthode consiste en une *élimination récursive* des indices radiomics. Il s'agit des procédures en pas-à-pas (*stepwise*) descendante ou ascendante ou mixte. Dans la procédure pas-à-pas ascendante, le modèle initial ne comporte qu'un terme constant puis les « p » prédicteurs sont tous testés seuls, indépendamment et nous ne gardons lors du 1er pas que celui qui a obtenu le meilleur *indicateur de performance*. Au 2ième pas, la même procédure est réalisée avec les (p - 1) prédicteurs restant et nous ne gardons que celui qui a amélioré le modèle issu de la 1ère étape (composé jusqu'alors du 1er prédicteur et du terme constant)... etc., jusqu'à ce que l'indicateur de

performance n'augmente plus significativement. Dans la procédure pas-à-pas descendante, tous les « p » prédicteurs sont initialement inclus dans le modèle initial. La variable la moins importante du modèle est retiré puis nous réitérons la modélisation sur (p - 1) prédicteurs restant... etc., jusqu'à ce l'indicateur de performance du modèle chute significativement après retrait d'un prédicteur.

- La seconde méthode consiste à ne sélectionner que les indices radiomics passant un *filtre*. Ce filtre correspond généralement à un test univarié de recherche d'association avec la variable à prédire (par exemple: t-test ou régression de Cox univariée ou régression linéaire univariée selon la tache). L'indice testé serait conservé si sa p-value est inférieur à une certaines valeur fixée (selon les études, 0.2 ou 0.05).
- La dernière méthode passe par une procédure de *régularisation* / pénalisation comme LASSO (Cf. Infra).

1.5.4.6.3. Réduction des dimensions

Les données multidimensionnelles très corrélées entre elles (: multi-colinéaires) ont pour conséquence d'empêcher d'identifier des prédicteurs intéressant et d'induire des erreurs lorsque nous souhaitons réaliser des prédictions sur de nouvelles observations à partir de notre modèle. Pour limiter ce problème, une solution est de transformer le jeu de données en un jeu de données de moindre dimension tout en perdant le moins d'information possible.

Dans les travaux de cette thèse, nous avons exclusivement utilisé l'analyse en composante principale (ACP). Les variables doivent d'abord être normalisées (et la variable à prédire exclue). L'idée est de transformer des variables corrélées entre elles en de nouvelles variables décorrélées appelées les composantes principales. Ces composantes principales sont construites comme des combinaisons linéaires des variables de départ et qui expliquent le maximum de variance dans le jeu de données. La limite de l'ACP est que ces nouvelles « variables » (i.e. les composantes principales) n'ont plus de sens biologique ou clinique comme les variables initiales, ce qui peut rendre compliqué l'interprétation d'un modèle basé sur ces composantes principales.

A noter que des alternatives à l'ACP existent comme t-SNE (pour « *t-distributed stochastic neighbor encoding* ») ou l'analyse linéaire discriminante qui est une ACP supervisée.

1.5.4.6.4. Problèmatique du déséquilibre des classes

Une autre problématique, non systématique mais présente pour les STM de par leur rareté, est le déséquilibre entre les classes de la variable à prédire. Les options suivantes peuvent aider à y remédier:

- sous-échantillonner la classe majoritaire;
- sur-échantillonner la classe minoritaire: soit en répliquant aléatoirement les observations de la classe minoritaire, soit en créant de nouvelles données synthétiques à partir des donnés préexistantes. L'algorithme SMOTE (pour « *Synthetic Minority Oversampling Technique* ») est le plus couramment utilisé (Chawla et al, 2002). Il consiste à parcourir toutes les observations de la classe minoritaire, en chercher les k plus proches voisins puis à synthétiser de nouvelles données entre ces points.

Dans un des travaux de cette thèse, nous avons été confronté à cette problématique (rareté relative des bonnes réponses à la chimiothérapie comparativement aux mauvaises réponses). Nous avons adopté une autre stratégie consistant à utiliser une matrice de coût où un poids plus fort a été donné aux erreurs faites sur la classe minoritaire.

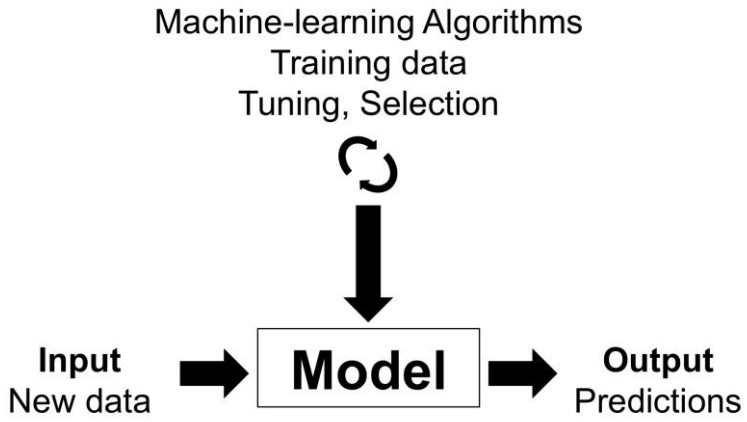
1.5.4.7. Modélisation prédictive

Suite à l'extraction et au « nettoyage » des indices radiomics, l'étape suivante consiste à construire le modèle prédictif de la variable d'intérêt (réponse au traitement, présence d'un statut mutationnel particulier, groupe de mauvais pronostic, survenue d'une rechute locale ou à distance infra-clinique / infra-radiologique...). Les jeux de données de cette thèse comportent tous une variable à prédire (ou « *outcome* ») déterminée en amont de la réalisation de l'étude et nous emploierons donc une approche dite *supervisée*. Néanmoins, une autre approche complémentaire peut être réalisée lorsque l'*outcome* n'est pas clairement identifié: les approches *non supervisées* qui cherchent à identifier des relations, des groupes (ou *clusters*) entre les observations sur la base des variables du jeu de données via des techniques de clustering (*k-means*, *hierarchical clustering*). Les deux approches (supervisées et non supervisées) emploient des algorithmes différents et nous nous concentrerons essentiellement sur les algorithmes d'analyses supervisées.

1.5.5. Algorithmes de « *machine-learning* »

Le terme *machine-learning* peut être défini comme un programme informatique permettant d'apprendre à partir d'une certaine expérience « E » pour une certaine tâche « T » et une certaine mesure de performance « P », afin d'améliorer ses performances selon « P » pour une même tache « T » pour un autre jeu de données « E' ». En pratique, nous fournissons à l'algorithme un input correspondant aux données « E » pour lequel nous connaissons le résultat réel de ce que nous souhaitons lui faire apprendre à prédire. Ces données vont lui servir à s'entrainer. Nous fournissons ensuite un second input « E' » sur lequel il fera des prédictions (« *output* ») (Figure 1-33).

Figure 1-33. Principe général de l'apprentissage statistique



La tache « T » peut être de plusieurs natures:

- soit une classification (binaire le plus souvent), quand l'output correspond à n-classes, par exemple: le grade FNCLCC (grade 1, grade 2, grade 3) ou le statut survivant à 1 an (oui / non) ou le statut mutationnel (muté BRCA1/2: oui / non);
- soit une régression, quand l'output est un nombre continu, par exemple: le pourcentage de cellules viables après traitement ou de cellules marquées par tel anticorps. Les analyses de survie (ou « *time-to-event analysis* ») y sont rattachées où une prédiction à réaliser, par exemple, est la durée avant survenue d'une rechute.

Il existe plusieurs dizaines d'algorithme de *machine-learning* accessibles en langage open source. Nous avons ici essentiellement utilisé la librairie scikit-learn dans le

langage python et le package « caret » dans le langage R (utilisant lui même des dizaines de packages R – environ 180 au début de cette thèse).

1.5.5.1. Principaux algorithmes de machine-learning employés

1.5.5.1.1. Régression logistique

La régression logistique est une régression adaptée à une classification binaire dans laquelle on souhaite prédire une variable Y (= 0 ou 1) en fonction d'un ensemble de prédicteur X = (X_1, X_2, ... X_n). Nous noterons P(Y = 1|X) = P(X). A la différence d'une régression linéaire, P(X) est bornée entre 0 et 1, la régression logistique a recours à la fonction logit (ou log-odds - mais d'autres fonctions sont possibles) où :

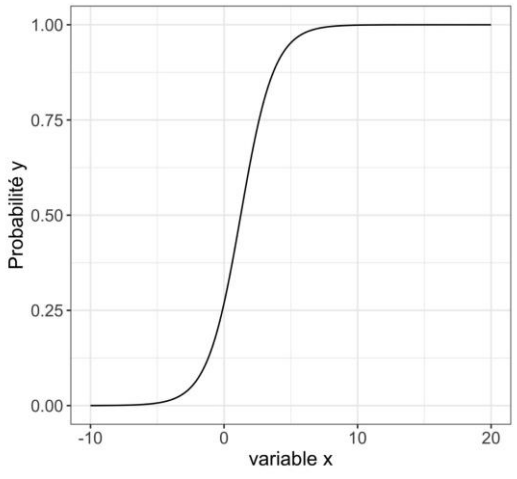
$$\log\left[\frac{P(X)}{1-P(X)}\right] = logitP(X)) = \beta_0 + \sum_{i=1}^{n}(\beta_i \times X_i)$$

Soit :

$$P(X) = \frac{\exp[\beta_0 + \sum_{i=1}^{n}(\beta_i \times X_i)]}{1 + \exp[\beta_0 + \sum_{i=1}^{n}(\beta_i \times X_i)]}$$

La Figure 1-34 représente cette fonction logistique.

Figure 1-34. Graphique de la courbe de régression logistique montrant la probabilité de l'évènement y = 1 en fonction de la valeur de la variable x.



L'étape suivante consiste à estimer les β_i de sorte que les observations appartenant à la classe Y = 1 aient effectivement une probabilité la plus proche de 1 et celles appartenant à l'autre classe aient une probabilité la plus proche de 0. Ainsi, on essaye de maximiser le produit :

$$\ell(\beta_0, \beta_1, \dots, \beta_n) = \prod_{i:Y_i=1} P(X_i) \times \prod_{j:Y_j=0} (1 - P(X_j))$$

… à l'aide de la méthode du maximum de vraisemblance qui peut être obtenue par des algorithmes numériques (en particulier l'algorithme de Newton-Raphson). Cela revient à vouloir minimiser le négatif du log vraisemblance, exprimée ainsi :

$$L = -\log[\ell(\beta_0, \beta_1, \dots, \beta_n)]$$

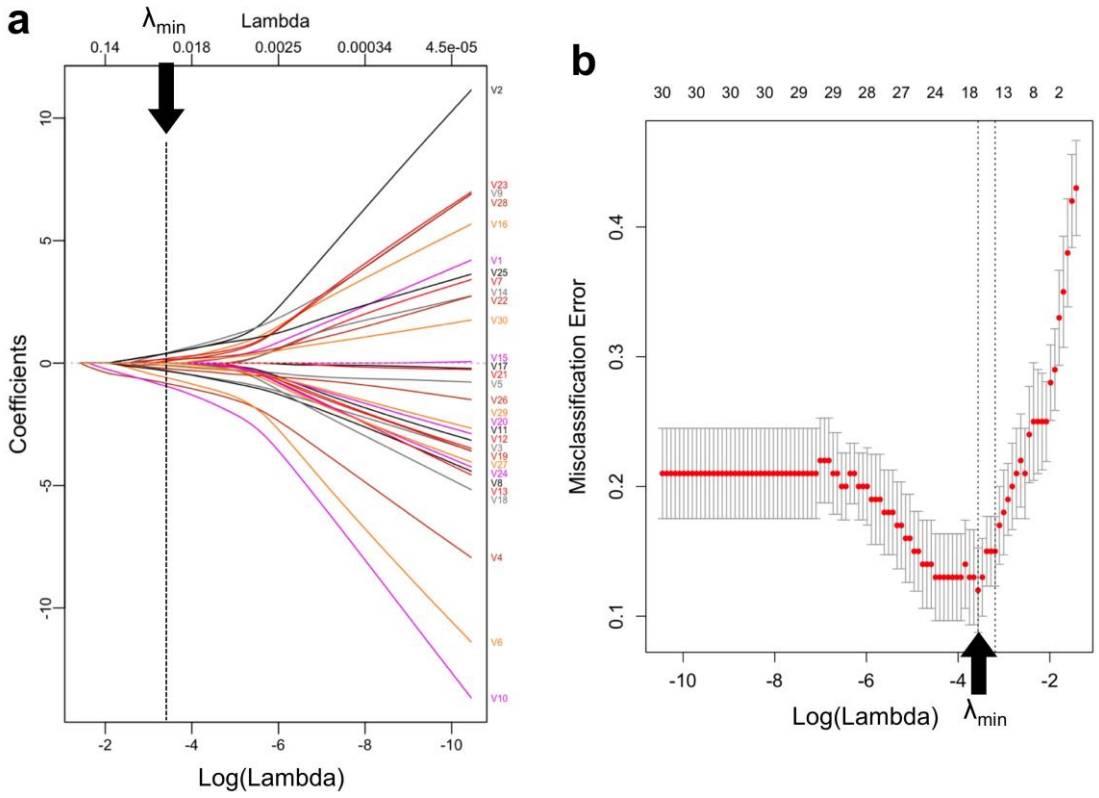
La régression logistique est le classifieur le plus simple, qui sert de référence avant d'établir des modèles plus avancés nécessitant des optimisations de leurs paramètres propres ou hyperparamètres, notion que nous reverrons ci-dessous, puisqu'il n'en possède pas. Néanmoins, la régression logistique a ses limites en cas de colinéarité des variables et lorsque le nombre de variables augmente (multidimensionnalité).

1.5.5.1.2. Régression logistique pénalisée LASSO (LASSO-LR)

La régularisation LASSO (pour « *least absolute shrinkage and selection operator* ») est une méthode de contraction des coefficients de la régression (Tibshirani et al, 1996). Elle nécessite de normaliser les données aux préalables. Le principe est le suivant : le LASSO va rajouter une pénalité (ou contrainte) à L, exprimée : $\lambda \times \sum_{i=1}^{n} |\beta_i|$ où λ est un paramètre à sélectionner pour obtenir le modèle qui minimise l'erreur de prédiction. λ est un hyperparamètre. Il va être sélectionné via un « *grid-search* » (c'est-à-dire parmi un ensemble valeur à fournir screenant les valeurs possibles) typiquement en validation croisée (Figure 1-35). Le LASSO est aussi une méthode de sélection de variable puisque si λ est suffisamment grand, la minimisation de L ne pourra se faire qu'en annulant certains coefficients β_i.

Le LASSO est implémenté en R via le package « glmnet » inclus dans « caret » (Kuhn et al, 2008; Friedman et al, 2010). Enfin, « *elasticnet* » est une autre méthode de régularisation combinant les régressions « *ridge* » et LASSO

Figure 1-35. Applications de la régression logistique avec pénalisation LASSO à un jeu de données de 30 variables (V1 à V30) pour prédire Y (0 ou 1). (**a**) Chemin de régularisation du LASSO. Chaque courbe représente une des 30 variables. Selon la valeur de l'hyperparamètre λ (en abscisse), les valeurs des 30 coefficients sont estimées par le LASSO. Pour choisir le λ, nous avons recours à la validation croisée; le graphique (**b**) représente l'erreur de classification des différents modèles (avec leur intervalle de confiance à 95%) selon les valeurs de λ. Le λ qui minimise l'erreur vaut environ 0.0283 (log = -3.565) et garde 18 des 30 variables dans le modèle final (flèches noires). Figures réalisées à partir du package R "glmnet"

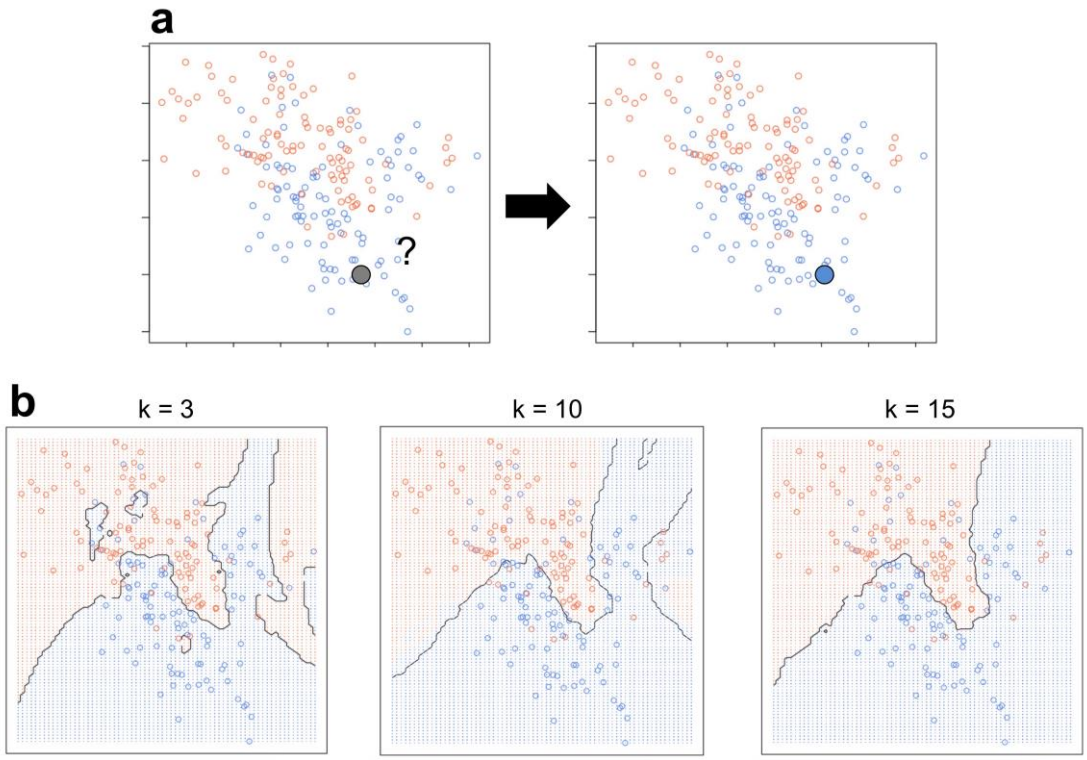


1.5.5.1.3. « K-nearest neighbors » (kNN)

Les kNN sont un algorithme de classification ou de régression qui permet de modéliser des relations non-linéaires entre plusieurs points. Pour chaque nouvelle observation, l'algorithme va rechercher les k points déjà étiquetés qui lui sont le plus proche (selon la distance euclidienne « d » pour des variables continues) (James et al, 2013). Selon leur étiquette (ou label), ces autres points vont voter pour lui assigner une classe avec une pondération « 1 / d » (Figure 1-36). Le paramètre « k » est le seul hyperparamètre pour la version de base des kNN mais d'autres hyperparamètres sont possibles comme la méthode de calcule de la distance ou l'utilisation d'une fonction noyau particulière (ou « *kernel* ») parmi plusieurs possibles qui permet de

transformer l'espace de représentations des données en un espace de plus grande dimension où les données pourraient être mieux séparables et les performances meilleures.

Figure 1-36. Méthode des k-nearest neighbours (kNN). (**a**) Principe général: soit une nouvelle observation (rond gris). L'algorithme cherche les "k" points déjà étiquetés les plus proches selon la distance euclidienne et va prédire le groupe auquel appartient la nouvelle observation (orange ou bleu). (**b**) k est l'hyperparamètre des kNN et influence donc les prédictions du modèle comme l'illustre cet exemple où les contours des prédictions se modifient selon k. Réalisé à partir des package R "ElemStatLearn" et "knn".

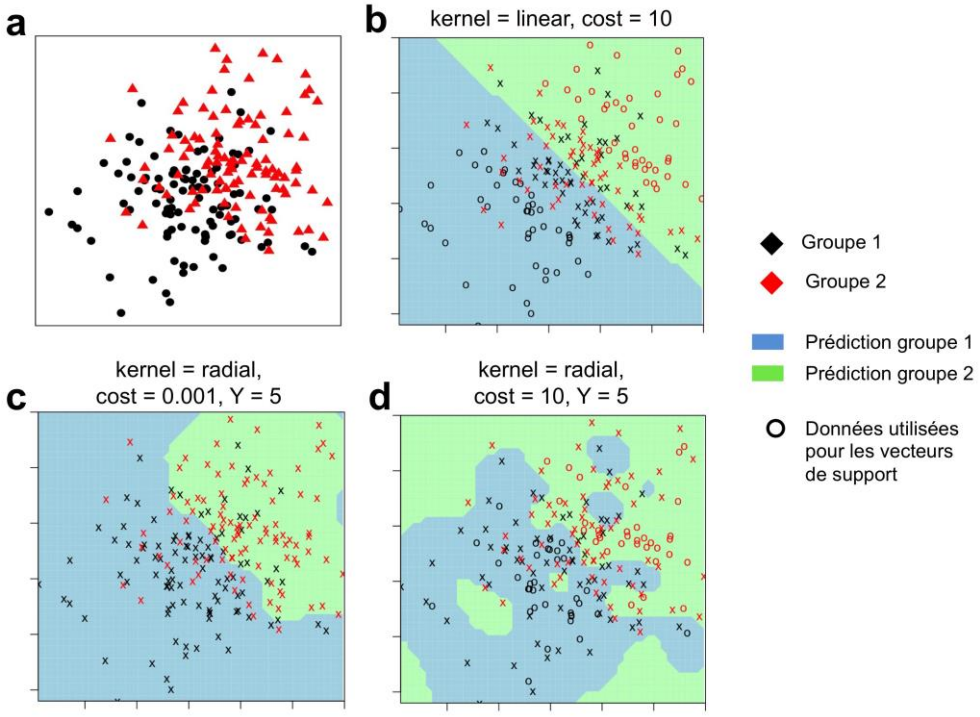


1.5.5.1.4. « Support vector machines » (SVM)

Les SVM (ou supports à vastes marges ou machines à vecteurs de support) proposent de prolonger l'espace initial des données en un espace de dimensions plus grandes afin de mieux séparer les classes. Ainsi, la séparation linéaire obtenue dans cet espace agrandi est équivalente à une séparation non-linéaire dans l'espace initial. Pour déterminer cette séparation linéaire (ou hyperplan), les SVM cherchent la séparation qui maximise la marge entre les classes, autrement dit la largeur du couloir (ou hyperplan) entre les observations des 2 classes (Figure 1-37). Plus la marge est large, plus le modèle sera facilement généralisable. Il s'y ajoute une contrainte comme pour le LASSO afin de trouver une solution plus robuste (James et al, 2013). Les SVM

utilisent aussi des *kernels* pour accéder à un espace de plus grande dimension où il est plus probable qu'une séparation linéaire existe (par exemple: linéaire, gaussien, radial, polynomial...). Les hyperparamètres varient selon le *kernel*, ils incluent le coût « C » qui pénalise les erreurs et permet de contrôler l'ajustement du modèle. Les SVM sont réputés efficaces quand le nombre de variables excède le nombre d'observations.

Figure 1-37. Exemple illustrant le principe des Support Vector Machines (SVM). (**a**) Correspond aux données initiales étiquetées (groupe 1 et 2). (**b, c, d**) Montrent les hyperplans séparateurs permettant de séparer les données en 2 groupes pour différents modèles construits selon les hyperparamètres kernel, coûts (: tolérance aux observations mal classées, équivalent de paramètre de régularisation) et gamma (: pour les kernels linéaires, détermine combien une unique observation « o » peut influencer le modèle/incurver les bords de l'hyperplan). Figures réalisées à partir du package « e1071 ».



1.5.5.1.5. « Random forests »

Les *random forests* rassemblent deux concepts: les arbres de décision et la méthode ensembliste basée sur le vote de plusieurs arbres de décision simples (Breiman et al, 2001). Ces arbres peuvent servir à des taches de régression ou de classification.

Arbres de décision. A la différence des autres approches, la frontière entre Y = 0 et Y = 1 n'est pas recherchée sous la forme d'un hyperplan mais sous la forme de segments perpendiculaires définissant des rectangles emboités (Figure 1-38). La première étape consiste à trouver une segmentation de la population en 2 rectangles permettant

au mieux de distinguer les 2 classes. Un « nœud » correspond à un rectangle qui sera redécoupé en 2 autres rectangles à l'étape suivante... etc., jusqu'à ce qu'une condition d'arrêt soit rencontrée. Une « feuille » correspond à un rectangle élémentaire. Ainsi, à chaque noeud, il faudra déterminer la variable pour découper le sous ensemble et son seuil optimal. Ces deux choix s'effectuent:

- soit en maximisant un critère d'homogénéité
- soit en minimisant un critère d'entropie ou encore d'impureté défini comme :

$$Entropie(noeud) = -\sum_i f_i \times \log(f_i)$$

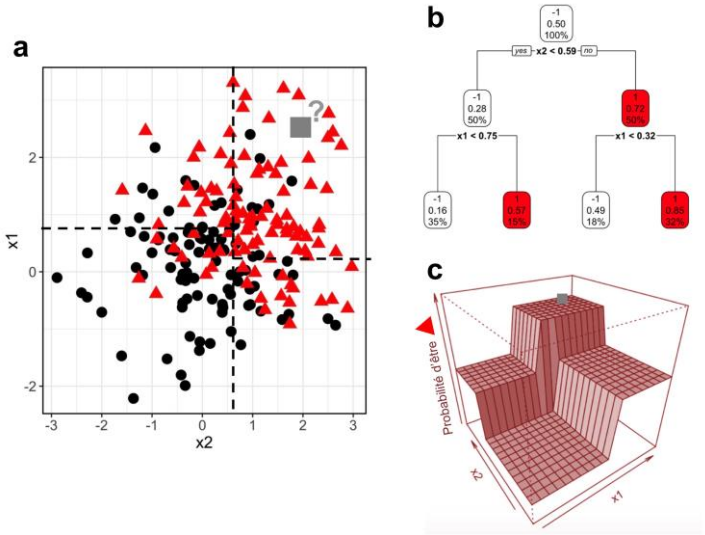
Où les f_i sont les fréquences empiriques dans le noeud des k classes de la variable Y (dans notre classification binaire, k = 2)

$$Gini(noeud) = 1 - \sum_i {f_i}^2$$

L'indice de Gini étant plus souvent utilisé car plus simple et rapide à calculer.

A mesure que la profondeur de l'arbre s'accroit, le risque de sur-apprentissage augmente. Pour limiter cela, nous devons déterminer les hyperparamètres comme la taille minimale des sous-groupes, la profondeur maximale de l'arbre, le pourcentage d'homogénéité satisfaisant...

Figure 1-38. Exemple illustrant le principe des arbres de décision. (**a**) Correspond aux données initiales étiquetées (groupe 1 et -1 – idem Figure 1-37) avec les nœuds de l'arbre selon (**b**). (**c**) Correspond à une vision 3D synthétique des probabilités d'appartenir au groupe 1 selon les valeurs des variables prédictives x1 et x2. Ainsi, une nouvelle observation (carré gris) ayant x2 >0.59 et x1 >0.35 a une probabilité de 0.85 d'appartenir au groupe 1 (triangle rouge). Figures réalisées à partir des packages R « rpart », « rpart.plot » et « plotmo ».

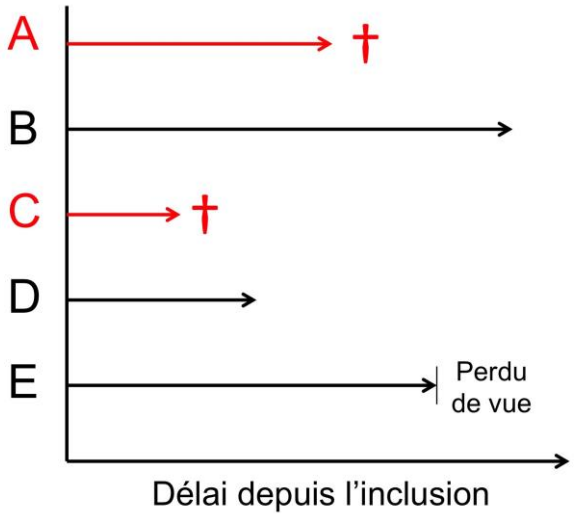


Forêts d'arbres de décision. Le principe des *random forests* est de tirer un grand nombre d'échantillons par *bootstrap*, à construire un modèle (toujours le même, type CART) sur chaque échantillon puis à agréger / faire voter les modèles (pour une classification). L'autre particularité est que seul un sous-ensemble des prédicteurs est utilisé pour chaque arbre unitaire. La classe prédite correspond à celle qui aura recueilli le plus grand nombre de votes. Nous pouvons aussi utiliser la moyenne des la probabilité d'appartenance à une classe. Les hyperparamètres à rechercher lors de l'entrainement comprennent: le nombre d'arbres agrégés (entre 500 et 1000, habituellement), le nombre de prédicteurs sélectionnés pour la scission de chaque noeud et l'effectif minimum de chaque feuille.

1.5.5.1.6. Modèle des risques proportionnels de Cox

Dans les analyses de survie, l'objectif est de déterminer le délai avant la survenue de l'évènement (i.e. une variable continue), avec la particularité que des observations peuvent être « censurées » c'est-à-dire que nous savons que les patients sont indemnes de l'événement au moment du recueil des données, mais cela ne signifie pas qu'ils le resteront toujours (l'événement pouvant survenir ultérieurement) (Figure 1-39).

Figure 1-39. Particularités des observations dans le cas des analyses de survie. Trois cas de figure peuvent se produire: survenue de l'évenement (1) après un certain délai (A, C); patients toujours suivi n'ayant pas présenté l'évènement après un certain délai (B, D); patient perdu de vue après un certain délai (E).



Soit « T » la durée de survie, nous définissons la fonction de survie « S(t) » et la fonction de risque instantané « h » (ou « *hazard function* ») qu'un événement se produise dans un intervalle de temps « δt » comme:

$$S(t) = P(T > t)\ avec : S(0) = P(T > 0) = 1\ et\ S(+\infty) = 0$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \geq 0$$

Le modèle de Cox est de loin le modèle le plus employé pour les analyses de survie en uni- ou multivarié (Cox et al, 1972). Il décrit la fonction de risque instantané « h » en fonction de « p » variables explicatives (X_1, X_2, ... X_p) potentielles pour « i » patients de la manière suivante :

$$h(t \mid X_{1,i}, X_{2,i}, \ldots, X_{p,i}) = h_0(t) \times \exp[\sum_{j=1}^{p} X_{j,i} \times \beta_j]$$

Soit :

$$\log\left[\frac{h(t \mid X_{1,i}, X_{2,i}, \ldots, X_{p,i})}{h_0(t)}\right] = \sum_{j=1}^{p} (X_{j,i} \times \beta_j) \quad (*)$$

Où les β_j correspondent aux coefficients associés à chacun des « p » prédicteurs. Nous pouvons voir que la partie gauche de (*) ne comporte aucune variable dépendant du temps. Par analogie avec la régression logistique, il faut ensuite estimer ces « p » coefficients ce qui se fait aussi par la méthode du maximum de vraisemblance. Il est courant d'utiliser l'exponentielle des β_j (ou rapport de risque ou « *hazard ratio* » [HR]) et qui sont interprétés comme suit:

- $0 \leq HR < 1$ (et l'intervalle de confiance à 95% [95%CI] n'inclut pas 1): le prédicteur X_j diminue le risque
- HR = 1 (et/ou le 95%CI inclut 1): le prédicteur X_j n'a pas d'effet sur la variable à prédire.
- HR > 1 (et le 95%CI n'inclut pas 1): le prédicteur X_j augmente le risque.

Pour que ce modèle soit applicable, il faut que les variables soient indépendantes du temps (ce qui peut être testé par les test des résidus de Schoenfeld) et que les variables ne soient ni trop nombreuses, ni trop corrélées.

1.5.5.1.7. Modèle de Cox pénalisé LASSO (LASSO-Cox)

Pour surmonter le cas où le nombre de covariables devient trop importants par rapport au nombre d'observations, la procédure de régularisation LASSO a été étendue au

modèle de Cox afin de sélectionner et mieux estimer les β_j. La procédure est accessible en *open source* dans le package R « glmnet » (Simon et al, 2010). Pareillement, un terme de pénalité λ a été ajouté à l'opposé du logarithme de la fonction de vraisemblance. Ce λ correspond à un hyper-paramètre de l'algorithme qu'il faudra optimiser (selon plusieurs critères possibles dans le package comme l'*Akaike Information Criteria* ou le *Bayesian Information Criteria*) pour obtenir la meilleure prédiction de la survie sur un échantillon d'entrainement en validation croisée (Cf. infra).

1.5.5.2. Méthodes d'entrainement des algorithmes et hyper-paramètres

A l'issue de ce panorama des algorithmes que nous emploierons, nous avons insisté sur l'existence d'hyperparamètres qui doivent être fixés avant d'appliquer le classifieur sur le jeu de données (par exemple: le λ de la régression logistique pénalisée LASSO). Ils sont différents des paramètres propres du modèle qui sont appris directement sur les données fournies (par exemple: les coefficients β des régressions logistiques).

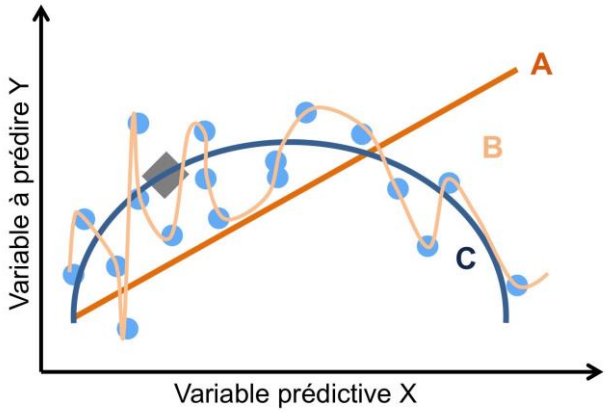
Le choix des hyperparamètres va modifier les performances du modèle final. Bien qu'il existe des valeurs par défaut dans les algorithmes proposés en R ou python, il faut donc optimiser leur valeur. Pour ce faire, nous pouvons proposer une grille (ou « *grid-search »,* manuelle ou aléatoire) contenant toutes les valeurs d'hyperparamètres à tester. Chacun des modèles possibles (selon les combinaisons d'hyperparamètres) est évalué et le plus performant est conservé.

Le processus de création d'un modèle de *machine-learning* nécessite de scinder la population totale en deux cohortes dites d'entrainement et de test:

- la cohorte d'entrainement sert à entrainer et optimiser les hyper-paramètres
- la cohorte de test n'est utilisée qu'une fois les modèles finaux (i.e. meilleur modèle basé sur les SVM, meilleur modèle basé sur les kNN...) établis afin de tester et comparer leur performance sur une population originale.

L'objectif de cette subdivision est aussi d'identifier et de limiter le sur-apprentissage, c'est-à-dire quand un modèle est extrêmement performant sur la cohorte d'entrainement mais nettement moins sur la cohorte de test (Figure 1-40).

Figure 1-40. Illustration des concepts de surapprentissage (« overfitting ») et de sous-apprentissage (« underfitting »). Trois modèles pour expliquer Y en fonction de la variable prédictive X sont proposés. Le modèle (A) correspond à un modèle trop simple de sous-apprentissage (régression linéaire) avec peu de variance et trop de biais. Le modèle (B) par contre montre un surapprentissage et ne se généralisera pas à de nouvelles données (trop de variance et peu de biais). Enfin, le modèle (C) apparait comme un meilleur compromis entre biais et variance et se généralise bien sur la nouvelle observation (losange gris).



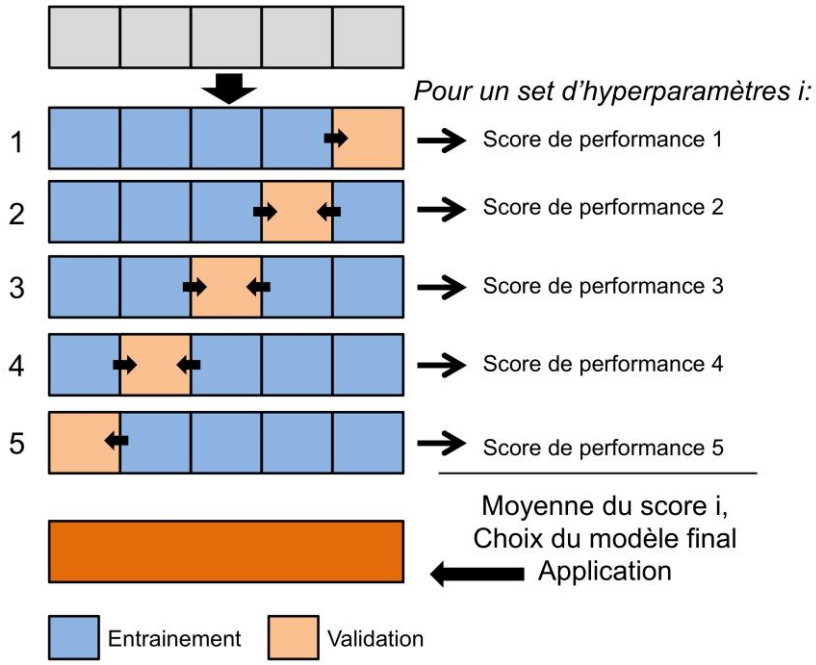
Afin de choisir le meilleur set d'hyper-paramètres et de construire le meilleur modèle avec la cohorte d'entrainement, nous aurons recours à une procédure dite de validation croisée (ou « *cross-validation »)*.

Dans la *k-fold cross-validation,* la population « n' » de la cohorte d'entrainement est divisée en « k » blocs égaux de « *round*(n'/k) » patients (avec k = 5 ou 10 le plus souvent). Ainsi, « k » cycles d'apprentissage sont réalisés au cours desquels l'algorithme apprend sur une population de « n' – *round*(n'/k) » patients et est évalué selon un indicateur de performance sur les « *round*(n'/k) » patients restant. Nous aboutissons ainsi à « k » valeurs de cet indicateur qui sont moyennées (Figure 1-41). De cette manière, chaque observation est utilisée (k - 1) fois pour construire le modèle et une unique fois pour l'évaluation du modèle.

A noter que la "*leave-one-out cross validation*" (LOOCV) est une variante de la validation croisée dans laquelle « k » est égal au nombre total d'observations de la cohorte d'entrainement « n' ». Ainsi le modèle est entrainé sur (n' - 1) observations et évalué sur l'observation restante.

La procédure de cross-validation sur la cohorte d'entrainement va être répétée pour autant de combinaisons possibles d'hyperparamètres fournis dans la grille. Le modèle final sera celui dont les paramètres et hyper-paramètres auront permis d'obtenir la meilleure valeur pour l'indicateur de performance choisi.

Figure 1-41. Principe de la validation croisée. Ici, en 5-folds consistant en la division du jeu de données en 5 groupes de taille similaire. Pour chaque fold, le modèle est entraine sur 4 des 5 sous-groupes et les performances du modèle sont évaluées sur le 5^{ième} groupe. Ces étapes permettent de choisir le meilleur modèle qui pourra ensuite être appliqué sur la cohorte en entier mais surtout sur une autre cohorte indépendante de validation (non représentée ici).



1.5.5.3. Méthodes de mesure des performances des modèles

Les indicateurs de performance sont cruciaux dans le choix du meilleur modèle, dépendent de la tache « T » et des spécificités de la question posée. De manière générale, il existe 3 grandes catégories d'indicateurs:

(i) des indicateurs de performance globale (mesurant la distance entre la prédiction et la réalité),

(ii) des indicateurs de discrimination (mesurant la capacité d'un modèle à distinguer les observations à haut risque de survenue de l'événement versus celles à bas risque),

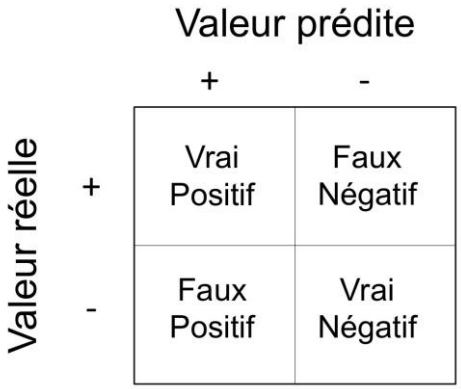
(iii) des indicateurs de calibration (mesurant la concordance entre probabilités réelles et probabilités prédites de l'évènement d'intérêt).

Nous ne détaillerons ici que les plus fréquents et utilisés dans les travaux de cette thèse pour des taches de classification et de régression – survie.

1.5.5.3.1. Classification:

Une partie des indicateurs est basée sur la matrice de confusion (Figure 1-42):

Figure 1-42. Matrice de confusion



A partir de laquelle nous calculons les indices suivant :

- l' « *accuracy* » (ou fiabilité) qui correspond à la proportion de prédictions correctes sur l'ensemble des éléments à prédire:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Il s'agit d'un indicateur de performance globale du modèle. La limite de cet indicateur est qu'il surestime les performances d'un modèle lorsque l'évènement est rare.

- La *précision* (ou valeur prédictive positive) qui définit à quel point les prédictions positives sont précises:

$$Precision = \frac{TP}{TP + FP}$$

- le « *recall* » (ou sensibilité) qui correspond à la proportion de cas prédits positifs parmi l'ensemble des cas réellement positifs:

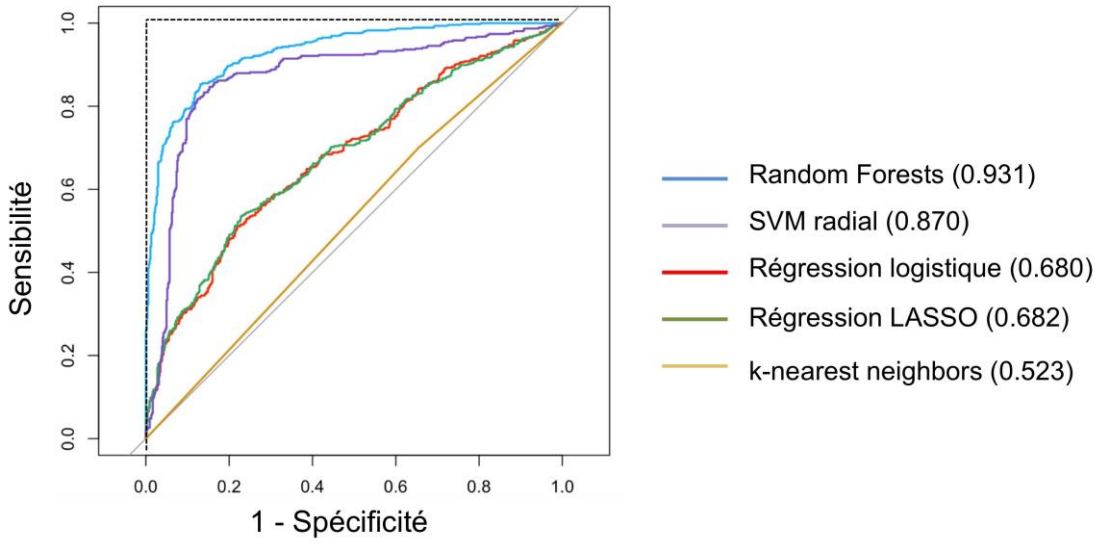
$$Recall = \frac{TP}{TP + FN}$$

- le *kappa*, qui est parfois privilégié pour les jeux de données déséquilibrés à la place de l'*accuracy*, et qui consiste à comparer la fiabilité observée du modèle ($P_{Observée}$) à la fiabilité attendue (liée au hasard, $P_{attendue}$) d'un modèle:

$$Kappa = 1 - \frac{1 - P_{obs}ervée}{1 - P_{attendue}}$$

- *La courbe ROC et l'AUROC* qui consiste à tracer la courbe avec le taux de fausse prédiction en abscisse (= 1 - spécificité) et le taux de vrai positif en ordonnée (= sensibilité) pour chaque seuil possible (Figure 1-43). Un modèle purement aléatoire aura théoriquement une aire sous la courbe ROC (AUROC) de 0.5 tandis qu'un modèle parfait aurait une AUROC de 1.

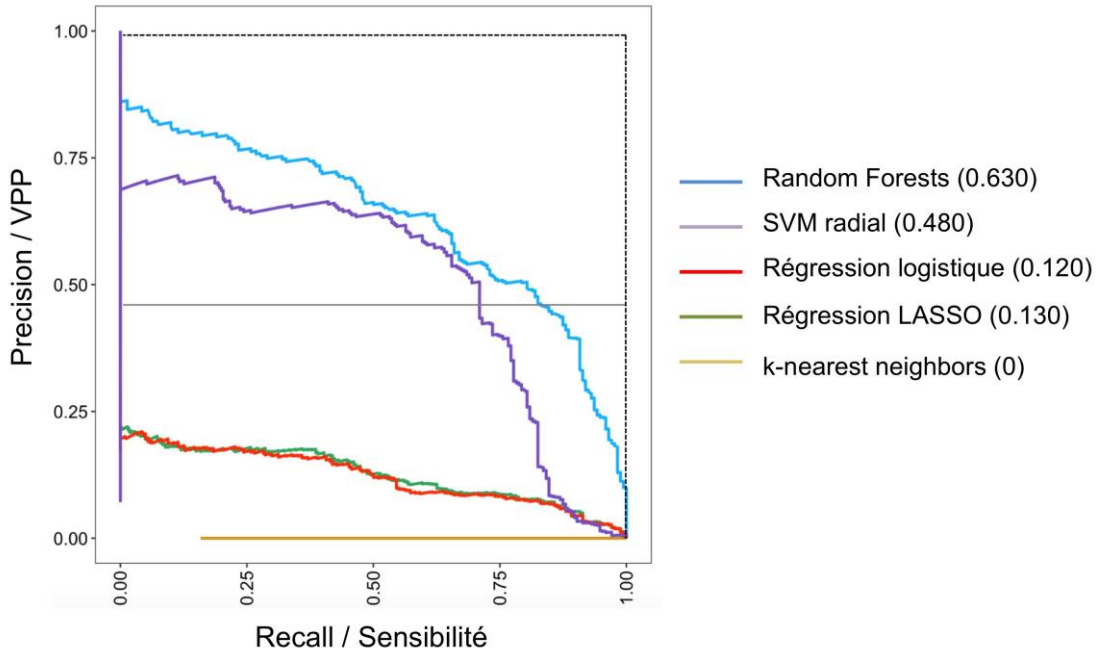
Figure 1-43. Courbes ROC permettant d'évaluer les performances de plusieurs modèles pour classer des observations en 2 classes. L'aire sous la courbe ROC est donnée entre parenthèses. Un modèle totalement aléatoire correspondrait à la diagonale grise. Un modèle parfait au tracé noir en pointillé. Le modèle le plus performant parmi les 5 proposés apparaît être les Random Forests et le moins performant les kNN. Figure réalisée à partir du package R « pROC ».



- *La courbe précision rappel:* cette courbe est plus précise que la courbe ROC pour des jeux de données déséquilibrés (Saito et al, 2015). L'abscisse correspond à la sensibilité et l'ordonnée à la précision. Tout comme pour les courbes ROC, nous pouvons calculer l'aire cous la courbe de précision rappel (AUPRC) (Figure 1-44). Un modèle totalement aléatoire aura une AUPRC = P / (P + N) où le ratio entre vrai

positif et vrai négatif est « P : N » tandis qu'un classifieur parfait aura une AUPRC = 1.

Figure 1-44. Courbe « precision-recall » pour les 5 mêmes modèles. L'aire sous la courbe de precision-recall est donnée entre parenthèse. Un modèle totalement aléatoire correspondrait à la droite grise (y = n(Class1) / n(Class1 + Class2)) et un modèle parfait à la courbe en pointillés noirs. Les random forests montrent encore les meilleures performances. Figure réalisée à partir du package R « MLeval».

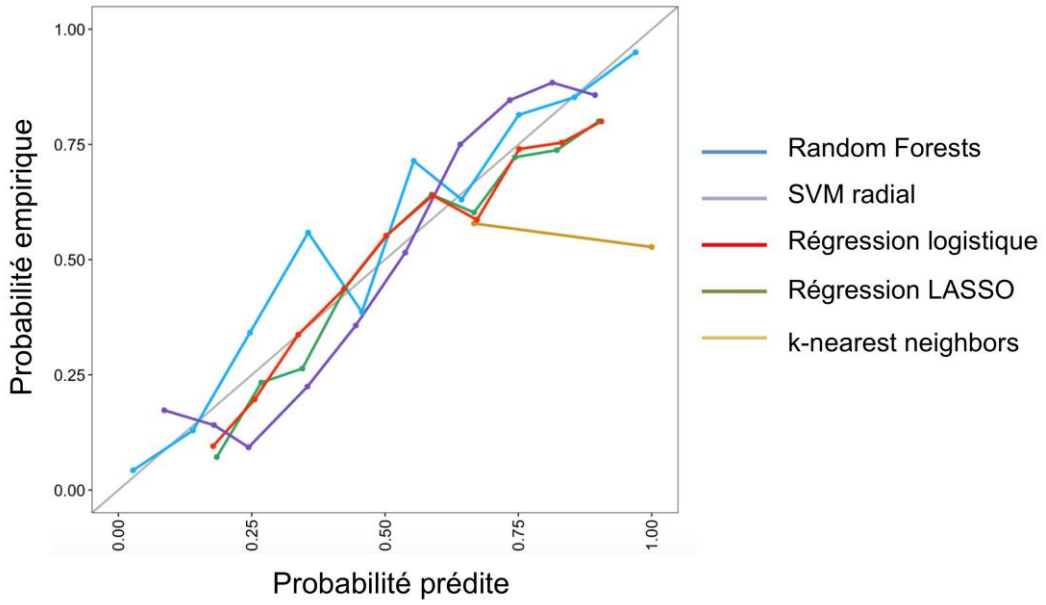


- *La courbe de calibration :* plus un classifieur est bien calibré, plus la proportion d'événements prédits doit être proche de la probabilité d'événements réels (idéalement égales). La courbe de calibration consiste à représenter en abscisse les probabilités réelles et en ordonnées les probabilités prédites (Figure 1-45). Elle montre le biais d'un classifieur mais non la qualité de ses prédictions. Pour quantifier la qualité de la calibration, nous avons recours au score de Brier (Brier et al, 1950) définit comme :

$$Brier\ score = \frac{1}{N} \times \sum_{i=1}^{N} \left(P_{prédit,i} - P_{observé,i}\right)^2$$

Où N est le nombre d'observation. Plus un score de Brier est faible, plus le modèle est bien calibré.

Figure 1-45. Courbes de calibration des 5 mêmes modèles. La courbe grise représente une calibration parfaite: pour chaque groupe ayant une probabilité prédite similaire, la proportion de ces individus dans Class1 est égale à la probabilité prédite. Ici les kNN sont mal calibrés, tandis que la régression logistique classique et la régression LASSO apparaissent les mieux calibrées. Figure réalisée à partir du package R « MLeval».



1.5.5.3.2. Survie

- *Indice de concordance de Harrell (c-index)*: il s'agit d'un estimateur de la capacité de discrimination du modèle pronostic allant de 0 à 1 (: modèle parfait), un modèle totalement aléatoire ayant un c-index = 0.5 (Harrell et al, 1996). Nous noterons « r_i » le risque qu'un patient « i » présente l'évènement à prédire. Le calcul du c-index se base sur l'hypothèse suivante. Dans un modèle pronostic performant, un patient « j » avec un délai de survenue à l'événement plus court « t_j » ($t_i < t_j$) devrait avoir un risque « r_j » tel que $r_j > r_i$: cela définit une paire (i, j) concordante. La formulation du c-index est donc :

$$c - index = P(r_i > r_j \,|\, t_j < t_i)$$

$$c - index = \frac{nb\ paires\ concordantes}{nb\ paires\ concordantes + nb\ paires\ non\ concordantes}$$

Une limite du c-index est qu'il n'évalue pas toutes les paires possibles du jeu de données mais seulement (1) celles pour lesquelles i et j auront présenté l'événement et (2) celles où i aura été censuré et j aura présenté l'événement.

- *Courbe ROC temps-dépendantes (ROC(t))*: développée par Haegerty et Zheng (2005), l'objectif est d'étendre le concept de courbes ROC aux analyses de survie. Il faut donc trouver une méthode pour estimer la sensibilité et la spécificité du modèle. Suite à la remarque précédente sur les cas censurés et en notant $D_i(t)$ le statut du patient à un instant « t » ($D_i(t) = 1$ si l'événement a eu lieu et 0 sinon) et « x » une valeur possible du risque « r », nous pouvons écrire que :

$$c - index = P\left(r_j > r_i \,|D_i(t) = 0, D_j(t) = 1\right)$$

$$Sensibilité(t,x) = P(r > x \,|D(t) = 1)$$

$$Spécificité(t,x) = P(r \leq x \,|D(t) = 0)$$

La courbe ROC au moment « t » peut donc être tracée comme l'ensemble des points:

$$\{1 - Spécificité(x); Sensibilité(x)\}, \qquad \forall x$$

Deux types de courbe ROC(t) sont définies: (i) les courbes cumulatives où nous évaluons toute la population jusqu'à l'instant « t » (notamment toutes les observations où l'évènement s'est produit bien avant « t ») et (ii) les courbes incidente (où seules sont considérées les observations qui ont survécu jusqu'à « t »). Les deux sont complémentaires, la première approche permettant d'avoir une vision globale des performances du modèle alors que la deuxième permet de mieux prendre en compte la variation de l'influence des prédicteurs au cours du temps.

Nous pouvons alors calculer pour chaque instant « t » :

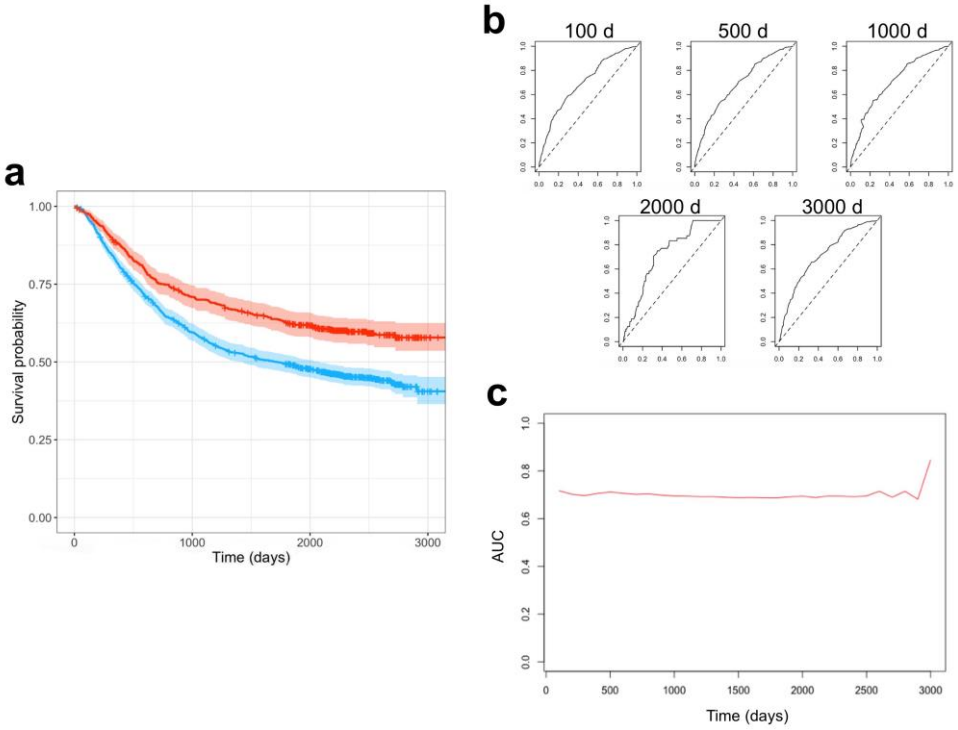
$$AUC(t) = \int_{x=0}^{1} Sensibilité(t,x) \times d[1 - Specificité(t,x)]$$

- *« Integrated AUROC » (iAUC)*: l'idée est ensuite de calculer l'AUROC des AUC(t) des ROC(t), comme suit :

$$iAUC = \int_{t=0}^{\max(t)} AUC(t) \times dt$$

La Figure 1-46 représente les relations entre ROC(t), AUC(t) et iAUC. L'iAUC est comprise entre 0 et 1 (: modèle parfait). Dans nos travaux, nous avons utilisé une méthode cumulative d'estimation de ces indicateurs (Uno et al, 2007).

Figure 1-46. Relations entre les indicateurs ROC(t), AUC(t) et iAUC pour estimer les performances de modèles pronostics. (**a**) Soit une variable X associée à la survenue d'un évènement Y au cours du temps dont est représentée la courbe de Kaplan-Meier selon chacune de ses 2 modalités. Soit un modèle de Cox incluant cette variable X. (**b**) Représente les courbes ROC(t) à 5 délais successifs. Ces courbes ROC(t) dont intégrées pour chacun de ces délais (= AUC(t)) permettant de tracer la courbe (**c**), elle même intégrée pour obtenir l'iAUC, comprise entre 0 et 1 (ici = 0.710)



1.5.5.4. Cohorte de validation

A l'issue de l'étape d'entrainement, nous avons pu identifier un certain nombre de modèles optimaux. Afin de sélectionner le meilleur modèle, la dernière étape consiste à l'appliquer sur la (ou idéalement les) cohorte(s) de validation, jusque là non utilisée(s) et à calculer les indicateurs performances. Le modèle final sera celui présentant les meilleures performances.

1.5.6. Contrôle qualité des analyses radiomics?

1.5.6.1. Résumé de tous les éléments à contrôler/préciser pour la reproductibilité d'une analyse radiomics

A l'issue de cette revue des étapes permettant de répondre à une question oncologique par une approche radiomics, il en ressort que chacune d'elle, qu'il s'agisse de

paramètres d'acquisition ou de post-traitement, est susceptible d'influencer la valeur des indices radiomics et ainsi la prédiction souhaitée:

- le choix des modalités d'imagerie, des machines d'acquisition, des séquences et leurs paramètres d'acquisition;
- l'utilisation d'une correction N4;
- l'utilisation d'un algorithme de débruitage et ses paramètres;
- la méthode de discrétisation des niveaux de gris et le nombre de niveaux;
- l'utilisation d'un algorithme de normalisation des intensités de signal en IRM;
- la méthode d'interpolation pour la standardisation de la taille des voxels et la taille des voxels;
- la méthode de segmentation et le nombre de "segmenteur";
- la technique de reconstruction des VOI en 3D;
- le logiciel choisi pour le calcul des indices radiomics;
- les distances et angle pour les calculs de certains indices de 2^{nd} ordre ;
- la méthode d'agrégation pour les VOI 3D;

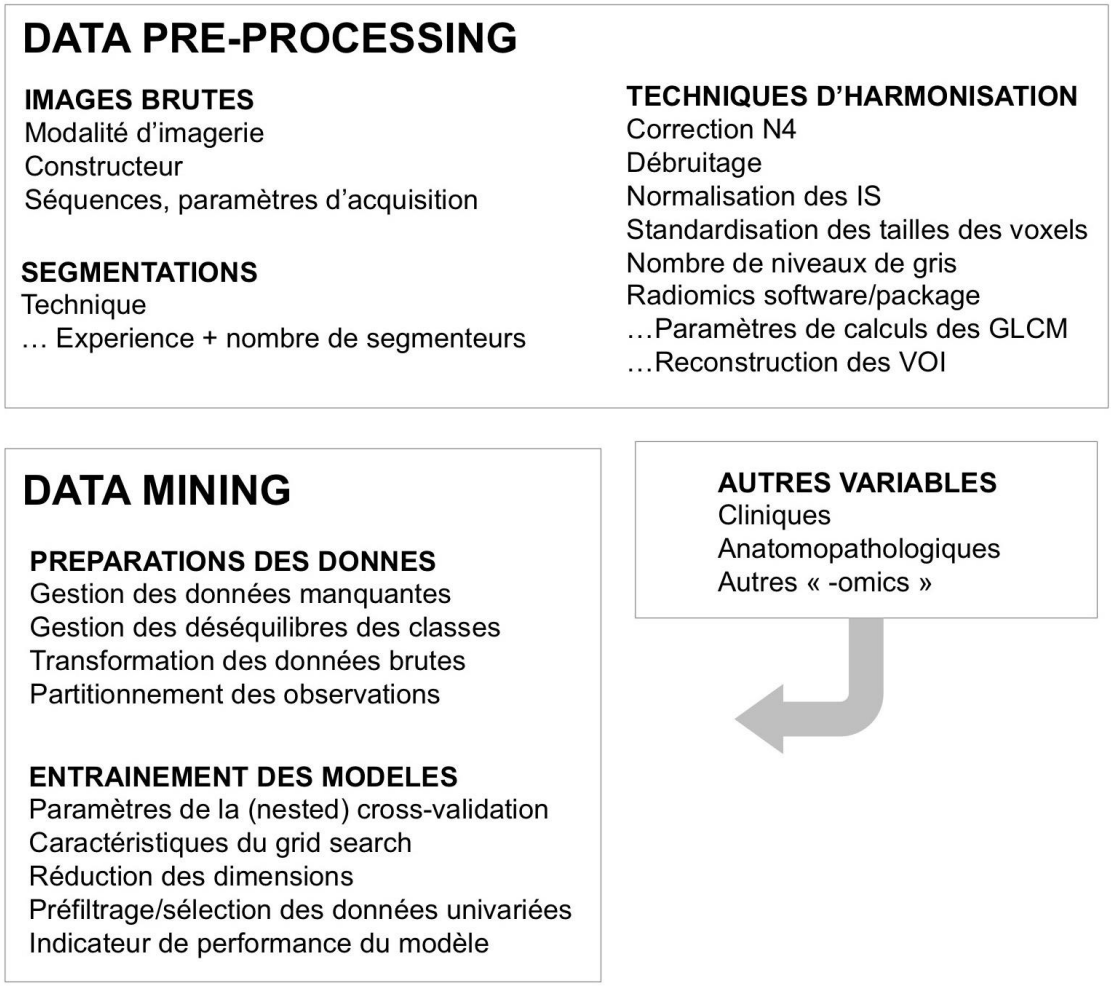
Pareillement, plusieurs aspects du *data-mining* sont eux-mêmes des sources de variation de la qualité de la prédiction, à savoir:

- devant des classes à prédire de fréquences déséquilibrée: le choix ou non d'une technique de ré-échantillonnage et (si oui) selon quelle méthode
- le choix des co-variables extra-radiologiques et radiologiques non-radiomics intégrées dans le modèles, leur encodage, leur transformation et la gestion d'éventuelles données manquantes;
- la technique de normalisation des indices radiomics et/ou la méthode de calcul des indices delta-radiomics;
- le partitionnement du jeu de données pour créer la cohorte d'entrainement et la cohorte de validation (effectifs de chaque groupe, choix des variables devant être équilibrées entre les groupes);
- l'utilisation d'une technique de ré-échantillonnage pour entrainer le modèle sur la cohorte d'entrainement et sa nature;
- l'utilisation d'une technique de réduction du nombre de dimensions du jeu de donnée et sa nature;
- l'utilisation de techniques de pré-filtrage/sélection des variables avant de les intégrer dans le modèle prédictif;

- le choix de l'algorithme de machine-learning selon la variable à prédire et l'éventuelle re-transformation du jeu de données selon l'algorithme choisi (binarisation / « *one-hot encoding* » de variables)
- les hyperparamètres de l'algorithme incluant la méthode de *tuning* de ces hyperparamètres (fixé, aléatoire, dans des valeurs prédéfinies);
- les indicateurs de performance pour comparer les modèles et choisir le plus approprié, fonction de la question posée et de la nature de la variable à prédire.

La description de chacune des ces étapes (résumées dans la Figure 1-47) dans la méthodologie des études radiomics est indispensable pour s'assurer de la reproductibilité de leurs résultats.

Figure 1-47. Etapes susceptibles d'influencer les modèles radiomics.

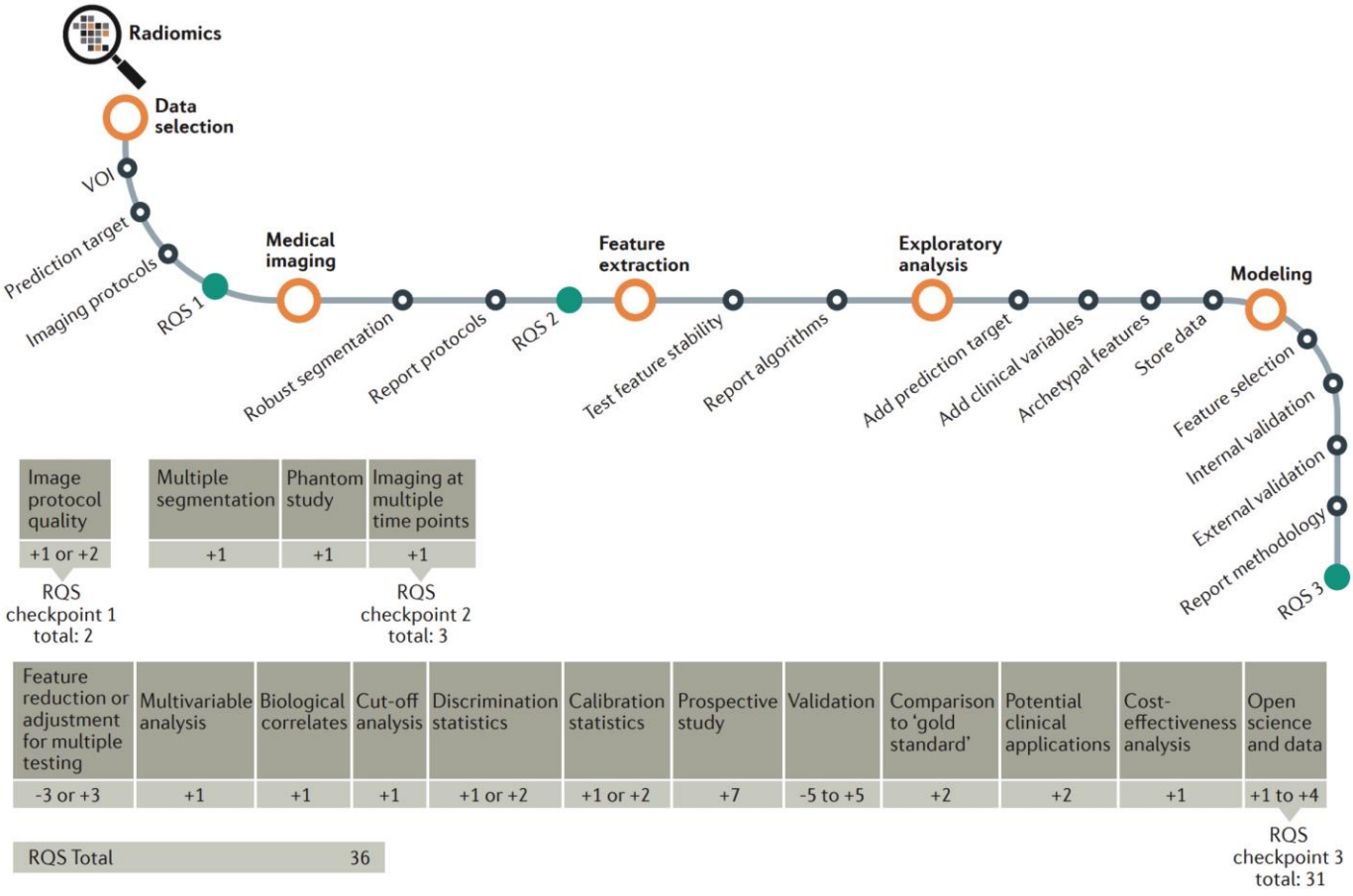


1.5.6.2. Développement d'outils de contrôle de qualité des études radiomics

Afin de juger de la qualité d'une étude radiomics en terme de méthodologie, Lambin et al ont développé un score nommé « *Radiomics Quality Score* » (RQS) (Figure 1-48) (Lambin et al, 2017). Ce RQS cote un total de 16 items répartis en 6 domaines, à savoir (Park et al, 2020):

- la qualité du protocole et la stabilité des indices
- la sélection des indices radiomics et leur validation;
- les indices de performances du modèle;
- la validation biologique et/ou clinique du modèle et son utilité;
- le niveau d'évidence de l'étude;
- l'accessibilité libre et gratuite du modèle et des données pour le construire.

Figure 1-48. Radiomics Quality Score (RQS): étapes permettant de calculer la qualité d'une étude basée sur des indices radiomics. Figure extraite de Lamblin et *al,* 2019.



Le RQS peut aller de -8 à +36. Il est complémentaire d'autres scores de qualité tels QUADAS-2 (pour *« Quality Assessment of Diagnostic Accuracy Studies 2 »*) recommandé pour évaluer les biais et l'applicabilité lors de méta-analyses d'études médicales diagnostiques; ou TRIPOD (pour « *Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis »*) qui correspond à une check-list de 22 recommandations pour améliorer la qualité des études cliniques diagnostiques et pronostiques quelque soit leur stade (développement, validation, mise à jour) (Whiting et al, 2011; Moons et al, 2015).

Devant la multiplicité des études radiomics malgré la complexité de leur réalisation, sous-estimée car vraisemblablement méconnue des radiologues, trois études (à notre connaissance) ont évalué la qualité méthodologique des articles en lien avec cette approche en incluant le RQS (Park et al, 2020a; Park et al, 2020b ; Ursprung et al, 2020). Park et *al.* (2020a) ont ainsi évalué 77 articles publié dans des revues à haut impact factor (clinique, *Radiology*, *European Radiology*) dont 91% (70/77) concernaient la cancérologie. Les auteurs ont montré que le RQS moyen étaient de 9.40 (range : [-5 ; - 21]) soit 26.1% du maximum. En détail, les 7 items du RQS les plus déficients étaient :

- la réalisation d'études de fantôme (i.e. évaluer la robustesse des indices radiomics aux changements de machine) dans aucune étude (0/77);
- la présence d'une évaluation cout-efficacité - dans aucune étude (0/77);
- l'accessibilité libre et gratuite des données, VOIs, scripts dans seulement 3.9% des études (3/77).
- le caractère prospectif de l'étude dans 3.9% des cas (3/77);
- la réalisation d'un "*test - retest*" (i.e. refaire l'acquisition d'images après un court délai afin de s'assurer de la robustesse des indices radiomics face aux variations temporelles) dans seulement 6.5% des études (5/77);
- l'identification d'une utilité clinique potentielle dans seulement 19.5% des cas (15/77);
- la recherche d'un seuil pour la signature radiomics identifiée (incluant la médiane) dans seulement 20.8% des études (16/77).

Des analyses similaires ont été reproduites spécifiquement pour le cancer du rein (57 articles) et les tumeurs gliales (51 articles) et ont abouti à des conclusions similaires (Ursprung et al, 2020 ; Park et al, 2020b).

1.5.6.3. Evolution des études radiomics?

L'absence de contrôle et de détails concernant ces points méthodologiques doivent faire questionner les résultats très majoritairement positifs jusque là obtenus.

Nous devons aussi signaler qu'une des signatures pronostiques radiomics les plus connues, celles de Aerts et *al.* (2014) s'est avérée fortement biaisée, ainsi que l'a démontré l'étude de Welch et *al.* (2019). En effet, en reprenant les données brutes (scanners de cancers ORL et broncho-pulmonaires) qui ont permi le calcul de la signature radiomics (4 indices en tout dont le volume et 3 indices de texture), Welch et al ont montré que (i) ces 3 indices de texture étaient fortement corrélés avec le volume et que (ii) les performances de la signature radiomics n'étaient pas significativement modifiées si les indices de texture étaient extraits des mêmes VOIs mais contourant des voxels randomisés. Ainsi, cette signature démontrerait plutôt un résultat déjà connu pour ces tumeurs : plus le volume tumoral est important, moins bon est le pronostic. Paradoxalement, le score de qualité des études radiomics a justement été proposé par des co-auteurs de l'étude de Aerts et al. (2014).

Il ressort l'absence quasi-totale de résultats négatifs des études radiomics (seulement 6% d'entre elles - Orlhac et Buvat, 2019). Devant la multiplicité des indices radiomics générables selon les paramètres de post-traitement et la multiplicité des analyses réalisables comparativement aux nombres d'observations des études, il ressort l'impression qu'il existera toujours une combinaison aboutissant à un résultat positif, possiblement par le jeu du hasard. Parallèlement, aucune signature radiomics, à notre connaissance, n'a été validée prospectivement. Trois hypothèses non exclusives peuvent être proposées : soit les approches radiomics sont réellement efficaces, soit les études négatives sont insuffisamment publiées, soit les résultats des études publiées sont biaisés. Comparativement aux autres approches « *-omics* », les approches radiomics sont encore jeunes mais l'engouement initial cède progressivement place à un regard plus critique, à la prise de conscience de la nécessité de mieux contrôler les biais et à l'introduction de méthodes d'identification de fausses découvertes. Orhlac et Buvat (2019) ont ainsi proposé de:

- enregistrer les études dans un registre international des études radiomics (tel clinicaltrials.gov - où 88 essais radiomics sont enregistrés au 1er mars 2020) ;
- utiliser des données dites sham obtenues, par exemple en randomisant la disposition des voxels dans une VOI, soit en utilisant des VOIs d'autres objets ou régions du corps ;

- publier les résultats négatifs sur des plateformes publiques d'archives accessibles à tous et gratuitement comme arvix.org ;
- rendre les auteurs, éditeurs et *reviewers* des revues médicales (en particulier radiologiques) plus alertes sur les particularités techniques et statistiques des études radiomics, des recommandations IBSI (Vallières et al, 2018) ; s'inspirer des méthodes statistique de contrôle des fausses découvertes des autres « -omics » ;
- rendre accessible à des agences indépendantes les analyses et résultats (négatifs ou positifs) des études radiomics financées.

1.5.7. Autres aspects des recherches en radiomics

Deux autres aspects des approches radiomics doivent être mentionnés :

1.5.7.1. Radiogenomics

L'idée est de combiner les indices radiomics avec d'autres données « -omics » (en particulier génomique et transcriptomique) afin de:

- Soit identifier des corrélations entre indices radiomics et statuts mutationnels d'intérêt ou signatures moléculaires préexistantes (tel Oncotype DX, MammaPrint ou PAM50 pour le cancer du sein, ou CLOVAR pour les carcinomes séreux ovariens de haut-grade, ou CINSARC pour les STM). En effet, au vu du cout de ces signatures (environ 3000 euros pour le test génomique Oncotype DX, non remboursé en France) et souvent de leur invasivité (nécessité d'une biopsie), il apparaitrait utile d'identifier une signature radiomics parfaitement corrélées à la signature moléculaire (et au pronostic des patients) - éventuellement en combinaison avec les données d'une prise de circulant (: biopsie liquide) - afin d'en limiter l'usage aux cas où la probabilité d'événement défavorable est intermédiaire.
- Soit de combiner ensemble données brutes radiomics et « -omics » afin de construite une signature composite pronostique améliorant les prédictions de signatures uni-modales préexistantes (Lo Gullo et al, 2020).

1.5.7.2. Quantification de l'hétérogénéité inter-sites

Les études jusqu'ici citées ont extrait les indices radiomics depuis un seul site tumoral (généralement les sites primitifs). Cette approche exclut l'information issus d'autres sites métastatiques voir de tous les patients métastatiques, les études radiomics étant

alors réalisées uniquement chez les patients en situation de primitif localement avancé. Or, si nous suivons l'hypothèse sous-jacente des approches radiomics, ces métastases pouvant être le site de nouvelles mutations (à risque d'échappement aux traitements ou de réponses dissociées à un traitement donné), cela devrait se traduire par une divergence des profils radiomics entre les métastases, et entre les métastases et le primitif au cours du temps. Il y aurait donc une information pronostique dans la quantification de l'hétérogénéité inter-sites. Vargas et al (2017) ont proposé une méthode pour permettre cette quantification sur une cohorte de 35 patientes atteintes de carcinomes séreux ovariens de haut-grade métastatiques passant par le calcul de matrices de similarité inter-sites (ISM pour « *inter-site similarity matrix* »). Ces ISM sont calculées à partir des indices radiomics de chaque site métastatique, par patient, puis des indices de texture sont eux-mêmes extraits de ces matrices à partir desquels une approche supervisée (par arbres de décision) pour prédire la survie à 5 ans et une approche non supervisée (par clustering par l'algorithme des k-means) pour identifier des clusters ensuite corrélés au pronostic et au statuts mutationnels CCNE1 et BRCA1/2 (Vargas et al, 2017; Meier et al, 2019).

1.6. Radiomics et STM

1.6.1. Revue de la littérature

En février 2020, après tri de 497 articles de *Web of Sciences* et 480 articles de *Pubmed* répondant aux mot-clefs :

["radiomics" OR "texture analysis" OR "texture" OR "histogram"]
AND ["sarcoma" OR "soft tissue tumor" OR "leiomyosarcoma" OR "chondrosarcoma" OR "osteosarcoma"]

nous avons pu identifier 47 articles scientifiques complets différents avec plus de 5 indices radiomics, en langue anglaise, sans mélange avec d'autres grandes classes de cancers.

Parmi eux, 6/47 (12.8%) concernaient les sarcomes gynécologiques, 12/47 (25.5%) les sarcomes osseux et 26/47 (57.4%) les STM (2/47 [4.2%] mélangeant STM et

sarcomes osseux). La revue de ces articles est donnée à la fin des annexes de ce manuscrit (annexe 8).

Parmi ces 27 articles, 9/27 (33.3%) étaient relatifs au diagnostic, 6/27 (22.2%) relatifs au *grading*, 8/27 (29.6%) relatifs à la prédiction pronostique (2 avec objectifs mixtes *grading* et pronostic) et 4/27 (14.8%) relatifs à la prédiction de la réponse au traitement (2/27 [7.4%] avaient d'autres objectifs, tel des corrélations au KI67, ou identifier plus précocement une rechute locale). Nous allons détailler les résultats apportés par les 23 articles publiés antérieurement à cette thèse (après exclusion des 4 articles issus de l'association Bergonié - MONC qui seront discutés dans les chapitres ultérieurs, et d'articles mixtes tumeurs osseuses et des parties molles). La Table 1-7 en résume les caractéristiques méthodologiques principales.

Table 1-8. Résumé des 23 études radiomics antérieures

Auteurs	Objectifs	Nb. patients	Modalité	Nature	
Chen et al, 2009	Distinction bénin vs. STM	114	US	retrospectif	unicentrique
Juntu et al, 2010	Distinction bénin vs. STM	135	IRM (T1)	retrospectif	unicentrique
Mayerhoefer et al, 2008	Distinction bénin vs. STM	58	IRM (T1, T2, T2 STIR)	retrospectif	unicentrique
Kim et al, 2017	Distinction tumeurs myxoides	40	IRM (T1)	retrospectif	unicentrique
Martin-Carreras et al, 2019	Distinction tumeurs myxoides	56	IRM (T1)	retrospectif	unicentrique
Malinauskaite et al, 2020	Distinction LPS vs. lipome	34	IRM	retrospectif	unicentrique
Thornhill et al, 2014	Distinction LPS vs. lipome	44	IRM (T1, T2, T2 FS)	retrospectif	unicentrique
Vos et al, 2019	Distinction LPS vs. lipome	115	IRM (T1, T2, CE-FS-T1)	retrospectif	unicentrique
Corino et al, 2017	Prédiction du grade	19	IRM (DWI)	retrospectif	unicentrique
Peeken et al, 2019a	Prédiction du grade + pronostic	225	IRM (T2 FS, CE-FS-T1)	retrospectif	Multicentrique
Peeken et al, 2019b	Prédiction du grade + pronostic	212	CT	retrospectif	Multicentrique
Wang et al, 2019	Prédiction du grade	91	IRM (T2FS)	retrospectif	unicentrique
Xiang et al, 2019	Prédiction du grade	67	IRM (T1, T2, CE-FS-T1)	retrospectif	unicentrique
Zhang et al, 2018	Prédiction du grade	34	IRM (T2 FS)	retrospectif	unicentrique
Hayano et al, 2015	Prédiction du pronostic	20	CT	retrospectif	unicentrique
Spraker et al, 2019	Prédiction du pronostic	226	IRM (T1)	retrospectif	Multicentrique
Vallières et al, 2015	Prédiction du pronostic	51	IRM + TEP	retrospectif	OpenSource
Vallières et al, 2017	Prédiction du pronostic	30	IRM + TEP	retrospectif	OpenSource
Esser et al, 2018	Prédiction de la réponse au traitement	31	CT	retrospectif	unicentrique
Esser et al, 2019	Prédiction de la réponse au traitement	33	CT	retrospectif	unicentrique
Tian et al, 2015	Prédiction de la réponse au traitement	20	CT	retrospectif	unicentrique
Meyer et al, 2020	Associations Ki67	29	IRM (T1, T2)	retrospectif	unicentrique
Tagliafico et al, 2019	Prédiction de la rechute locale	11	IRM (T1, T2, CE-FS-T1)	retrospectif	unicentrique

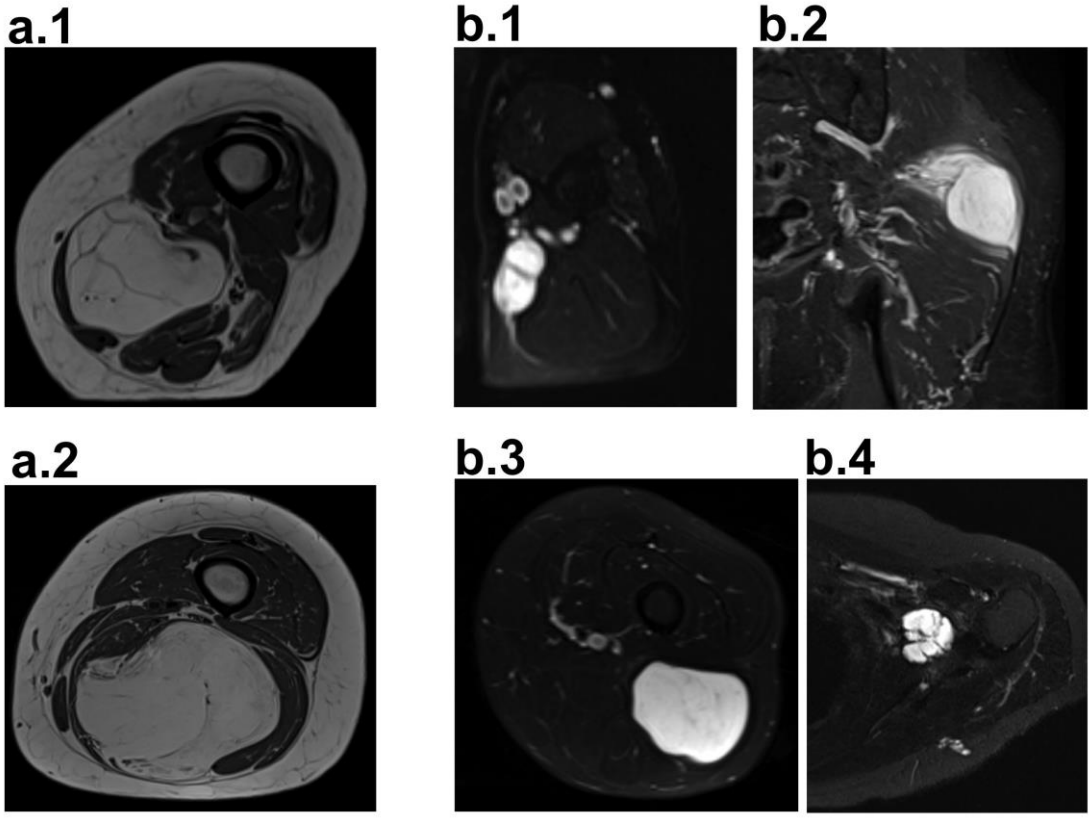
1.6.1.1. Aide à la distinction tumeur bénigne / pseudotumeur versus STM

Nous pouvons distinguer deux types d'étude concernant cette question:

(1) *la distinction entre tous les types de diagnostic bénin confondus et tous les types de STM confondus.* Bien que conceptuellement intéressante, ce type d'approche apparait éloignée de la pratique radiologique quotidienne en centre expert. En effet, le radiologue expert est généralement confronté à une tumeur dont l'imagerie est compatible avec un nombre restreint de diagnostics de tumeurs bénignes et tumeurs malignes proches sur les séquences morphologiques (Cf. infra : devant une tumeur adipeuse homogène : lipome vs. LPS bien différencié ; devant une tumeur myxoïde homogène : myxome, tumeur neurogène bénigne, chondrosarcome myxoïde extra-squelettique de petite taille, M/RC-LPS, myxome, MFS de bas grade...). De plus, au vu du nombre de types histologiques bénins et malins possibles comparés aux populations de patients inclus, les modèles proposés par ces études semblent fortement soumis à un biais d'échantillonnage (par exemple, histotypes parfois complètement manquant ou sous-représentés). Néanmoins, deux études ont montré de bonnes performances d'approches radiomics pour cette question et indiquant un potentiel champ d'application en aide au diagnostic. La première par Juntu et *al.* (2010) a proposé d'extraire 300 indices de texture (de 1er ordre, 2nd ordre et ondelettes) depuis des ROI en 2D de 50 x 50 voxels des séquences pondérées T1 (avec plusieurs ROI possibles par patient, annotées bénin ou malin) et non sur des volumes tumoraux entiers. Après réduction du nombre de variables et entrainement en cross-validation, leur meilleur modèle était basé sur l'algorithme radial SVM apportant, sur la cohorte d'entrainement, une *accuracy* de 0.94 et une AUROC de 0.91, supérieures aux performances de radiologues experts (0.90 et 0.85, respectivement). Plus récemment, Wang et *al.* (2019) ont repris cette question sur une cohorte de 91 patients, cette fois ci sur des VOI en 3D segmentées sur séquences IRM T2 Fat Sat en scindant la cohorte en cohortes d'entrainement et de validation avec utilisation d'une régression logistique pénalisée en cross-validation pour construire un score radiomics. Les auteurs ont ensuite construit un nomogramme combinant des variables cliniques et radiologiques pertinentes et ce score radiomics a montré qu'il améliorait la prédiction de la malignité comparativement au score radiomics seul ou aux données cliniques seules.

(2) *La distinction entre un type de tumeur bénigne et un type de STM bien précis présentant une forte ressemblance radiologique voire anatomopathologique* - typiquement, la distinction entre LPS bien différencié et lipome ou entre M/RC-LPS parfaitement homogène et myxome (Figure 1-49).

Figure 1-49. Situations illustrant les limites de l'analyse radiologique classique sur IRM. Cas de la distinction entre un liposarcome bien différencié (**a.1**) et un lipome (**a.2**) devant une lésion adipeuse homogène – séquences axiales T1 spin echo du cuisse. Cas de la distinction entre des lésions myxoides homogènes: myxofibrosarcome de bas grade (**b.1**), myxome (**b.2**) , liposarcome myxoide et à cellules rondes (**b.3**), chondrosarcome myxoide extra-squelettique (**b.4**)– ici sur des séquences T2 STIR spin echo.



Concernant la distinction LPS bien différencié versus lipome, trois études rétrospectives ont été menées (Thornhill et al, 2014 ; Vos et al, 2019 ; Malinauskaite et al, 2020). Toutes sont unicentriques, pour un nombre total de 194 patients (entre 34 et 116) et montrent un apport significatif d'approches radiomics comparativement aux approches radiologiques classiques avec des AUROC pour les meilleurs modèles construits comprises entre 0.81 et 0.98 (cette dernière valeur n'ayant pas été obtenue sur une cohorte de test extérieure). A noter qu'une seule a pu présenter ces résultats sur une cohorte de validation.

L'étude de Vos et al. (2019) est la plus conséquente (116 patients analysés) avec la méthodologie la plus explicite et la plus complète, accessible sur *github* donnant un RQS de 13.5. Le *data-mining* propose une méthodologie originale consistant en un *work-flow* nommé WORC (*workflow for optimal radiomics classification*) incluant transformation – réduction – sélection – gestion du déséquilibre des classes et plusieurs algorithmes de *machine-learning* avec tuning des hyper-paramètres générant 100 000 modèles dont seuls les 50 meilleurs sont conservés et combinés (méthode Ensembliste) pour conclure au modèle final.

Deux autres études se sont intéressées à la distinction bénin/malin devant une tumeur myxoïde. La première s'est uniquement intéressée à la distinction myxome versus MFS sur une cohorte de 56 patients (Martin-Carreras et al, 2019). Cette question peut s'avérer pertinente dans le cas de tumeurs des parties molles présentant un signal myxoïde homogène. Cependant cette étude présentait 2 biais principaux à savoir : les effectifs majoritaires de MFS de moyen et haut grade pour lesquels la malignité en IRM analysée par un radiologue expert ne fait pas de doute, et l'analyse sur une séquence T1 seule qui parait peu adaptée à l'analyse de tumeurs myxoïdes. La seconde par Kim et *al*. (2017) s'est intéressée aux relations entre le statut bénin/malin de 40 tumeurs avec signal IRM myxoïde et 13 indices radiomics issus d'analyse de texture de 1er ordre et par ondelette. Si cette dernière étude n'incluait pas d'analyse multivariée, elle illustre cependant l'association univariée significative entre 6 de ces indices et ce statut.

1.6.1.2. Aide à la prédiction du grade FNCLCC

Six études, toutes rétrospectives, se sont intéressées à la corrélation d'indices radiomics avec le grade voire à la construction d'une signature radiomics pour prédire le grade histopathologique FNCLCC sur la base du scanner (Peeken et al, 2019a), de séquences IRM structurale - essentiellement T1, T2 Fat Sat et T1 Fat Sat après injection IV de chélates de Gadolinium (Peeken et al, 2019b ; Zhang et al, 2018 ; Wang et al, 2019 ; Xiang et al, 2019), et sur l'IRM de diffusion (Corino et al, 2017). Au total, 671 patients ont ainsi été inclus dont 459 en IRM.

Dans l'ensemble, toutes ont montré de bonnes performances de modèles radiomics avec des AUC comprises entre 0.62 et 0.96. Cependant, nous pouvons noter que (i) toutes étaient rétrospectives, (ii) seules 3 d'entre elles ont évalué les performances de leur modèle sur une cohorte de validation (hors centre dans 2 cas), (iv) aucune n'a

comparé ses performances à celles basées sur les références de la littérature radiologique (Zhao et al, 2014 ; Crombé et al, 2019) - seulement à des classifications AJCC ou TNM peu utilisées en routine - et (v) aucune ne partageait les détails de sa modélisation. Leur RQS moyen était de 9.6/35 (3 - 18). Enfin, la plus importante étude radiologique a évalué ce grade sur microbiopsie et non sur pièce opératoire finale comme recommandé, ce qui est à fort risque de biais d'échantillonnage. En effet, une sous-estimation du grade réel est estimé entre 12.3 et 55% des cas selon les séries.

1.6.1.3. Aide à la prédiction du pronostic

Au total, six études radiomics se sont concentrées sur la prédiction du pronostic de patients atteints de STM avec des résultats positifs, qu'il s'agisse de la prédiction de la rechute métastatique (classification sans analyse de survie) en combinant PET/CT et IRM du TCIA (*the cancer imagining atlas*) (Vallières et al, 2015 ; Vallières et al, 2017) , de la prédiction de la survie globale à partir de séquences pondérées T1 (Spraker et al, 2019), de l'évaluation de la valeur pronostique locale, métastatique ou globale de signature radiomics précédemment établie associée au grade FNCLCC sur la base du scanner (patients traités par radiothérapie néoadjuvante) ou de séquences IRM structurales (Peeken et al, 2019a ; Peeken et al, 2019b), de la prédiction de la survie globale de patients traités par anti-angiogéniques et radiothérapie néoadjuvante et évalués par scanner injecté (Hayano et al, 2015). Ces études ont toutes identifié un lien en analyse multivariée entre indices radiomics ou signature/nomogramme radiomics avec la variable à prédire pour un RQS moyen de 11.2 (5 - 18). Toutes étaient rétrospectives et 3/6 (50%) multicentriques, intéressant un total de 734 patients dont 232 (entre 20 et 212 par étude) au scanner et 502 (entre 51 [TCIA] et 226 par étude) en IRM. Il faut cependant souligner qu'aucune étude n'a intégré des variables radiologiques connues pour être associées au pronostic des patients, et qu'aucune n'a comparé les performances de son meilleur modèle à celles de modèles radiologiques sémantiques classiques.

Les études de Vallières et *al.* sont aussi des études méthodologiques utilisant le jeu de données sarcomes *open source* du TCIA (Vallières et al, 2015 ; Vallières et al, 2017). Elles s'attachent tout autant à démontrer l'importance des paramètres de post-traitement (en calculant les indices radiomics pour différentes tailles de voxel, algorithmes pour la discrétisation, filtres, pondération TEP et IRM lors de la fusion

des images, segmentation avec et sans oedème péritumoral...) et d'acquisition des examens (en comparant des modèles basés sur des séquences synthétiques différant par les temps d'écho et de répétition pour l'IRM et le nombre de *span* en PET). Si la prédiction à réaliser peut être critiquable (survenue d'une rechute métastatique sans prendre en compte le délai de survenue, très variable entre les 2 groupes « rechute » et « absence de rechute »), Vallières et *al.* apportent cette preuve de concept et obtiennent une AUROC maximale de 0.98 pour leur meilleur modèle. Dans le sous groupe de 30 patients de la 2^{ième} étude, ils démontrent que l'AUROC peut être significativement augmentée en optimisant synthétiquement les paramètres d'acquisition (0.89 versus 0.84, p = 0.04 selon le test de Delong).

Par ailleurs, les études du groupe de Eary *et al.* dédiées au TEP-CT ont proposé dès les années 2000 une autre méthode de quantification de l'hétérogénéité singulière propre à cette modalité d'imagerie, distincte des analyses de texture classique. Les auteurs quantifient la divergence voxel-à-voxel entre la distribution réelle et la distribution idéale de la fixation au ^{18}F-FDG supposée suivre un pattern ellipsoïde avec une décroissante progressive du SUV depuis le centre vers les bords de la tumeur (Eary et al, 2008 ; O'Sullivan et al, 2011 ; Wolszynski et al, 2018). Nous n'intégrons pas les études de ce groupe dans notre revue puisqu'elles mélangent STM et sarcomes osseux, mais celles-ci montrent que l'hétérogénéité de fixation ainsi quantifiée est significativement corrélée à l'OS et la survie sans progression en analyse univariée et multivariée ($p < 0.0001$). De plus, en extrayant l'importance des variables de plusieurs modèles de *machine-learning* basés, entre autres, sur cette quantification (nommée « *metabolic gradient* »), Wolszynski et *al.* illustrent son rôle clef dans les performances et ainsi son intérêt potentiel et complémentaire des autres indices classiques et radiomics (Wolszynski et al, 2018).

1.6.1.4. Aide à la prédiction de la réponse thérapeutique

Nous avons identifié trois études au total s'intéressant à l'évaluation de la réponse dans le cadre des STM dont une a été exclue car elle ne prenait en compte qu'un seul indice supposé refléter l'hétérogénéité tumorale (Tian et al, 2015). Toutes sont rétrospectives, unicentriques, sans validation ni analyse multivariée (Esser et al, 2018 ; Esser et al, 2019). Elles montrent tout au plus des corrélations entre indices

radiomics baseline et delta-radiomics (différence absolue) de type texture de 1er et 2nd ordre et le statut de la réponse selon des critères conventionnels type RECIST ou Choi. Cependant, il n'y avait pas de corrélation avec la réponse histologique et les traitements évalués correspondaient à des traitements de seconde ligne (pazopanib et trabectedine).

1.6.1.5. Autres analyses

Deux études peuvent être mentionnées pour leur objectif original à savoir la corrélation avec des marqueurs de prolifération cellulaire (Ki67) et l'identification de signes de rechute locale (Meyer et al, 2019 ; Tagliafico et al, 2019).

1.6.2. Synthèse des analyses radiomics appliquées aux STM

A l'issue de ce panorama, nous pouvons mettre en avant que:

- hormis d'éventuelle corrélations *radiogenomics*, la plupart des applications potentielles de nouveaux biomarqueurs radiologiques dédiés aux STM ont au moins été testées une fois avec une approche radiomics, cependant:
- aucune de ces études n'a été construite prospectivement;
- seules 4/27 (14.8%) études étaient multicentriques;
- seules 6/27 (22.2%) études possédaient une cohorte de validation (issue ou non du même centre)
- 15/27 (55.5%) avaient pour 1er auteur un radiologue et 20/27 (74.1%) au moins un radiologue parmi les auteurs, mais seulement 7/27 (25.9%) études ont proposé: (i) l'ajout de variables radiologiques d'intérêt rapportées dans la littérature scientifique pour améliorer les modèles radiomics et/ou (ii) la comparaison du meilleur modèle radiomics au meilleur modèle radiologique (voire clinique et/ou biologique)
- seules 2/27 (7.4%) ont proposé les documents nécessaires pour reproduire les résultats;
- aucune ne présentait dans sa méthodologie tous les éléments exposés dans les documentations d'IBSI - en sachant que les recommandations IBSI sont postérieures à la majorité de ces études.

Il en résulte un RQS moyen de 6.35/35 (-5 - 17), inférieur à celui des études globales, sur les tumeurs gliales et sur le cancer du rein.

1.6.3. Pistes pour l'amélioration des analyses radiomics appliquées aux STM

Bien que les sarcomes soient des cancers rares, nous pouvons constater le nombre relativement élevé d'études radiomics qui leur est dédié. En incluant celles issues de l'institut Bergonié, leur ordre de grandeur (n = 47) est similaire à celles des cancers du rein (n = 57) alors que leur prévalence est environ 3 fois moindre (source e-cancer.fr)
Pour expliquer ce faible RQS moyen des approches radiomics dédiées aux STM, les hypothèses suivantes peuvent être formulées:

- leur rareté, responsable de populations d'étude moindres, de la complexité à établir des cohortes de validation intra et extra-centre, de la tendance à mélanger des sous-types histologiques différents morphologiquement à risque de biaiser les résultats en l'absence (au minimum) d'ajustement ;
- leur méconnaissance par la communauté médicale hors centre spécialiste, responsable de prise en charge hors centre inadaptée et sans imagerie préalable ;
- l'absence de standardisation des protocoles d'imagerie dédiés aux STM, responsables de l'exclusion de patients par l'absence de la séquence d'imagerie la plus appropriée pour la question posée et de variabilité des paramètres d'acquisition d'imagerie fonctionnelle (variabilité des b-values, des paramètres temporels de DCE-MRI...) ;
- leur caractère ubiquitaire entrainant de grandes variabilités des paramètres d'acquisition des imageries et l'usage d'antennes différentes qui sont compliqués à corriger *a posteriori*.

Améliorer la qualité des études radiomics dédiées aux STM afin de leur trouver une place en pratique clinique devra passer par : (i) une meilleure connaissance de ces tumeurs, de leurs réseaux de prise en charge, du principe des études radiomics et des recommandations IBSI, (ii) l'édition de *guidelines* pour la réalisation d'une imagerie d'une masse des parties molles et leur diffusion à large échelle, (iii) l'intégration de l'imagerie dans les databases sarcomes cliniques et pathologiques préexistantes, (iv) le développement d'équipes pluridisciplinaires médicales, paramédicales, biostatistiques, informatiques et mathématiques impliquées - chacun de ces aspects nécessitant des

moyens financiers et humains conséquents ainsi qu'une réelle volonté de translation clinique.

1.7. Organisation de la thèse

La suite du manuscrit détaille les travaux de recherche issus des interactions entre l'équipe INRIA Bordeaux-Sud-Ouest Modélisation en ONCologie (MONC) et le département d'imagerie oncologique de l'institut Bergonié entre 2017 et 2020. Nous avons choisi une organisation synthétisant le cheminement du patient (de l'évaluation de la gravité initiale de la maladie à l'évaluation de la réponse aux traitements systémiques donnés pour les formes les plus graves) et le cheminement scientifique (des premiers résultats concluant à l'identification des limites de ses propres travaux pour en améliorer la qualité). Ainsi, la première partie sera dédiée à l'identification d'une combinaison d'indice radiomics permettant de mieux évaluer le pronostic initial de patients atteints de M/RC-LPS. La deuxième partie s'intéressera à l'amélioration de la prédiction de la réponse à la NAC par rapport aux critères de référence via les approches delta-radiomics. La troisième et dernière partie s'attachera à l'identification de biais en IRM structurale et avec les séquences DCE-MRI, avec pour cette dernière séquence, la recherche d'autres méthodes d'analyses pour optimiser l'extraction de l'information pronostique.
Chaque article scientifique contenu dans ces parties sera précédé d'une introduction en resituant le contexte et suivi d'une discussion en résumant l'apport et les limites.

2. PREDICTION PRONOSTIQUE

2.1. Introduction

Les M/RC-LPS sont des STM à translocation, atypiques parmi les STM pour plusieurs raisons. Ils prédominent chez les jeunes adultes et les adultes d'âge moyen, présentent une dissémination métastatique atypique et le grade FNCLCC s'applique mal, faisant qu'un autre critère pronostic lui est préféré en routine clinique: le pourcentage de cellules rondes sur l'analyse histologique de la pièce opératoire. Un pourcentage > 5% est associé à une moins bonne survie globale et un plus haut risque de rechute métastatique. Tout comme pour le grade FNCLCC, ce pourcentage a été établi sur des pièces opératoires entières. Avec la généralisation des microbiopsies, il existe un risque élevé de sous-estimation de ce pourcentage. Ces contingents à cellules rondes sont faits de nombreuses petites cellules dédifférenciées sans composante myxoïde. Bien qu'il n'y ait pas eu de corrélations voxel-à-voxel entre zones riches en cellules rondes et signal IRM, plusieurs observations sur l'institut nous amènent à penser qu'il existe une relation entre hétérogénéité sur les séquences IRM pondérées T2 et l'abondance de cellules rondes. Ainsi, nous avons formulé l'hypothèse qu'employer une analyse radiomics pour quantifier l'hétérogénéité du phénotype radiologique sur la séquence T2 de l'IRM initiale, pré-thérapeutique, de patients atteints de M/RC-LPS permettrait d'améliorer, voire de potentialiser la prédiction du pronostic de ces patients comparativement aux critères cliniques et radiologiques classiques.

2.2. Article 1

L'article issu de ce travail a été publié en janvier 2020 dans *European Radiology* (PMID 31953663, doi: 10.1007/s00330-019-06562-5 – rang B SIGAPS, IF = 4.027). Il s'agit ici du manuscrit final reformaté. Les *Supplementary Data* sont en Annexe 3.

European Radiology
https://doi.org/10.1007/s00330-019-06562-5

ONCOLOGY

Can radiomics improve the prediction of metastatic relapse of myxoid/round cell liposarcomas?

Amandine Crombé[1,2,3,4] · François Le Loarer[4,5] · Maxime Sitbon[1] · Antoine Italiano[6] · Eberhard Stoeckle[7] · Xavier Buy[1] · Michèle Kind[1]

Received: 27 July 2019 / Revised: 24 October 2019 / Accepted: 30 October 2019

Abstract

Objective The strongest adverse prognostic factor in myxoid/round cell liposarcomas (MRC-LPS) is the presence of a round cell component above 5% within the tumor bulk. Its identification is underestimated on biopsies and in the neoadjuvant setting. The aim was to improve the prediction of patients' prognosis through a radiomics approach.

Methods Thirty-five out of 89 patients with MRC-LPS managed at our sarcoma reference center from 2008 to 2017 were included in this IRB-approved retrospective study as they presented with a pre-treatment contrast-enhanced MRI (median age, 49 years old). Two radiologists reported usual conventional/semantic radiological variables. After signal intensity (SI) normalization, voxel size standardization of T2-WI, and whole tumor volume segmentation, 44 3D-radiomics features were extracted. Using least absolute shrinkage and selection operator penalized Cox regression on prefiltered features, a radiomics score based on 3 weighted radiomics features was generated. Four prognostic multivariate models for MRFS were compared using concordance index: (1) clinical model, (2) semantic radiological model, (3) radiomics model, and (4) radiomics + semantic radiological model.

Results Twelve patients showed a metastatic relapse. The radiomics score included FOS_Skewness, GLRLM_LRHGE, and SHAPE_Volume and correlated with MRFS (hazard ratio = 19.37, $p = 0.0009$) and visual heterogeneity on T2-WI ($p < 0.0001$). A high score indicated a poorer prognosis. After adjustment, the best predictive performances were obtained with model (4) (concordance index = 0.937) and the lowest with model (1) (concordance index = 0.637).

Conclusion Adding selected radiomics features that quantify tumor heterogeneity and shape at baseline to a conventional radiological analysis improves prediction of MRC-LPS patients' prognosis.

Key Points

- *Fourteen radiomics features quantifying shape and heterogeneity of myxoid/round cell liposarcomas on T2-WI were associated with metastatic relapse in univariate analysis.*
- *A radiomics score based on 3 selected and weighted radiomics features was a strong and independent prognostic factor for metastatic relapse-free survival.*
- *The best prediction of metastatic relapse-free survival for myxoid/round cell liposarcomas was achieved by combining the radiomics score to relevant radiological features.*

Keywords Sarcoma · Patient-specific modeling · Liposarcomas, myxoid · Prognosis · Magnetic resonance imaging

INTRODUCTION

Myxoid/round cells liposarcomas (M/RC-LPS) are the most frequent malignant mesenchymal tumors of adipocytic lineage in young and middle-aged adults (Fletcher et al, 2013). They are underlined by specific translocations (FUS/CHOP or EWS/CHOP) (Sreekantaiah et al, 1992; Turc-Carel et al, 1986). One third of patients will develop metastases, notably in soft-tissues and bones (Engström et al, 2008; Haniball et al, 2011; Asano et al, 2012; Fuglø et al, 2013; Fiore et al, 2007).

The most important prognostic factor for M/RC-LPS is the percentage of round cells on surgical specimen, in addition to size and depth (Antonescu et al, 2001). Patients with more than 5% of round cells have higher risks of developing metastases and lower survivals. The last guidelines of the European Society of Medical Oncology recommend that sarcoma patients should undergo biopsy before treatment (Casali et al, 2018). Depending on the biopsy analysis, non-metastatic patients can be treated with additional chemotherapy and/or radiotherapy. Hence, the therapeutic strategy for M/RC-LPS patients is mostly based on the histological assessment performed on biopsies, running the risk of underestimating the round cell component. Moreover, given the large volume of some sarcomas, even applying the last histopathological guidelines could lead to potential underestimation (Wardelmann et al, 2016).

M/RC-LPSs are known to exhibit fluid-like signal intensity (SI) on T2-weighted-imaging (T2-WI) because of their myxoid stroma containing water with mucopolysaccharide matrix. Fatty components usually account for less than 10% and also display high SI on T2-WI (Petscavage-Thomas et al, 2014). The MRI signal of round cells has not been clearly characterized so far. However, as they are organized in solid sheets, rich in numerous dedifferentiated small cells with high nuclear-to-cytoplasmic ratio, without myxoid stroma, they are supposed to show lower SI on T2-WI; and, consequently lead to increased heterogeneity on T2-WI (Crombé et al, 2018; Gimber et al, 2017). Two studies based on several sarcomas histotypes have demonstrated that heterogeneity on T2-WI is independently associated with histological grade, metastatic relapse-free survival (MRFS) and overall survival (Zhao et al, 2014; Crombé et al, 2019a). However, heterogeneity was qualitatively and subjectively assessed.

Radiomics consists in the extensive quantification of imaging phenotype beyond what the human eye can describe, through a mathematical processing of medical imaging that includes histogram, grey-level matrices, fractal, wavelet or shape analyses

(Gillies et al, 2016). Herein, we hypothesized that quantifying M/RC-LPS heterogeneity induced by round cells through a radiomics analysis of T2-WI on pre-treatment MRI could help predicting patients' MRFS. Thus, we built a radiomics score composed of selected radiomics features associated with MRFS and we compared the performances of usual prognostic models with models including this score.

MATERIALS AND METHODS

Study Population

This single-center study was IRB-approved. Requirement for informed consent was waived because of its retrospective nature.

We included all consecutive adult patients treated by curative surgery and followed-up in our Sarcoma Reference Center from January 2008 to December 2017, with histologically-proven M/RC-LPS, pre-treatment contrast-enhanced MRI including Turbo Spin Echo (TSE) T2-WI and without metastasis at diagnostic (assessed by whole-body MRI or chest CT-scan and spine MRI). Of the 89 patients with a diagnosis of M/RC-LPS in our pathological database, 44 were excluded because they did not have a baseline contrast-enhanced MRI in our PACS, 9 were excluded because the protocol included a fat suppressed T2-WI instead of simple T2-WI, and one was excluded because of metastases at diagnosis.

The following covariables were reported: gender, age, location, depth, World Health Organization performance-status (WHO-PS), initial biopsy, round cells on biopsy, surgical margins, adjuvant/neoadjuvant chemotherapy and/or radiotherapy and surgical margins.

Follow-ups consisted in clinical examinations and chest radiographs every 3 months for 2 years, then every 6 months for 5 years, and then annually, with supplementary local MRI, spine MRI and chest CT-scan in case of abnormal findings. All the relapses were histologically proven. Since 2015, patients now undergo whole-body MRIs every year.

Overall survival, MRFS and local relapse-free survival correspond to the time from surgery to disease-related death, metastatic and local relapses, respectively. Patients without event during the study period were censored.

MRI acquisition

MRI examinations were carried out on 1.5-T MR-systems with adjustment of the coils, field-of-view and matrix depending on tumors size and location. All examinations included: T1-WI (echo time/repetition time: 10-15/500-700msec), TSE T2-WI without fat-suppression (70-130/2400-6800msec) and contrast-enhanced T1-WI (CE-T1-WI), for which different sequences were used (different fat-suppression techniques, gradient echo or TSE, 2D or 3D, with different contrast agents). Section thickness ranged from 3 to 5mm.

Semantic radiological analysis

Two radiologists (one senior and one fellow with 4.5 years of experience in MRI including 6 months in our sarcoma reference center) double-blind reviewed the imaging dataset. Next, they performed a consensual reading on which the analysis was performed (Annexe 3 - Supplementary Data 1 provides inter-observer agreements). The following semantic radiological variables were reported according to published definitions (Zhao et al, 2014; Crombé et al, 2019a; Yoo et al, 2014; Lefkowitz et al, 2013): longest diameter; SI heterogeneity on T2-WI (semi-quantitatively categorized as 0%, 1-24%, 25-49%, 50-74% and ≥75% heterogeneous when each of these percentages of tumor volume showed areas with hypo-, iso- and hypersignal on T2-WI, respectively); necrotic signal (defined as high SI on T2-WI with irregular borders and no enhancement on CE-T1-WI, categorized as 0%, <50% and ≥50% of tumor volume); fatty signal (defined as high SI on T1-WI and T2-WI without enhancement, categorized as 0%, <50% and ≥50% of tumor volume); tail sign (defined as aponeurotic enhancement surrounding the tumor, categorized as absent, thin when <2mm and thick when ≥2mm); peritumoral edema (defined as high SI at T2-WI with infiltrative and feathery borders, distinguishable from the apparent tumor borders and without mass effects, categorized as absent or present); peritumoral enhancement (defined as infiltrative, non-nodular contrast enhancement with feathery borders at CE-T1-WI, beyond the apparent tumor borders without modification of the usual morphology of the local anatomy, categorized as absent or present).

Radiomics post-processing of T2-WI

All slices were resampled using bilinear interpolation to obtain a common isotropic in plane resolution of 1x1mm^2 and a thickness of 4mm. SIs on T2-WI were normalized

for non-uniform intensity using N4ITK bias correction (Tustison et al, 2010). The intensity ranges were standardized using histogram-matching with the acquisition of a healthy volunteer's thigh as reference (Crombé et al, 2019d; Nyúl et Udupa, 1999). Whole tumor volumes on T2-WI were manually segmented by a senior radiologist helped by the other sequences to adjust the boundaries and validated by another senior radiologist. The extraction of 44 3D-radiomics features was performed with LIFEx software (version 4.70) (Nioche et al, 2018). Standardized SIs were discretized in 128 fixed bins of a bin size of 0.0315. Three shape features, 9 first-order and 32 second-order texture features from grey-level co-occurrence matrix (GLCM, n=7 - with a 4-voxels distance to neighbors), grey-level run length matrix (GLRLM, n=11), neighborhood grey-level different matrix (NGLDM, n=3) and grey-level zone length matrix (GLZLM, n=11) were estimated. Values of the radiomics features were standard scaled. Definitions can be found on *https://www.lifexsoft.org/index.php/resources/19-texture/radiomic-features.*

Statistical analysis

Statistical analyses were performed with R, version 3.5.2. All tests were two-tailed. A P-value <0.05 was deemed significant.

We first performed univariate analyses. Regarding clinical and semantic radiological features, Kaplan-Meier curves for MRFS were drawn and differences in survivals were assessed using log-rank test ("survival" package). Regarding radiomics, we performed univariate Cox regressions between MRFS and each feature (: prefiltering step). Correlations between features were tested with Spearman test. The features with a p-value < 0.05 after prefiltering were entered as input for a least absolute shrinkage and selection operator (LASSO) penalized Cox regression using "glmnet" package (Simon et al, 2011). Only 3 radiomics features were selected in the models to avoid overfitting because of the limited population size. A ten-fold cross validation was performed to select the lambda that provides the minimum cross-validated errors and that helped determining the radiomics features. The coefficients of the 3 resulting features were used to calculate the radiomics score. The hazard ratio (HR) with 95% confidence interval (CI95%) of the radiomics score was estimated and Kaplan-Meier curves for MRFS were compared between patients with radiomics score <median and ≥median.

We built different prognostic models: (1) clinical model based on size (<5cm vs. ≥5cm) and depth (deep vs. superficial); (2) radiological model, in which semantic radiological variables with a p-value <0.05 at univariate analysis were included and selected following stepwise selection; (3) radiomics model based on the radiomics score; and (4) radiomics + radiological model, combining the variables from model (2) and (3). All models were adjusted for: surgical margins (R0 vs. R1–R2), adjuvant radiation therapy (present vs. absent) and adjuvant/neoadjuvant chemotherapy (present vs. absent).

The models were compared using integrated-AUC, which corresponds to a weighted mean of the AUC during the follow-up, and concordance index (c-index) using the "risksetROC" package (Heagerty et Zheng, 2005). A c-index between 0.7 and <0.8 corresponds to an acceptable discrimination, 0.8 to <0.9 to an excellent discrimination, 0.9 to <1 to an outstanding discrimination and 1 to perfect discrimination (Harrell et al, 1996). Comparisons of c-index were performed with "survcomp" package (Schröder et al, 2011).

RESULTS

Population characteristics

The cohort included 35 patients (17 women, median age: 49 years old). Most tumors were deep-seated (85.7%, 34/35) and located in the lower limb (85.7%, 30/35), with a median size of 136 mm (Table 2-1). Pre-treatment biopsy was performed in 85.7% (30/35) of patients, which showed round cells in 23.3% (7/30) of patients. There were 2 local relapses, 12 metastatic relapses, and 4 deaths. The 1-year and 5-year MRFS probabilities were 91.1% (CI95%=(82.6-100)) and 57% (CI95%=(39.9-81.5), respectively.

Univariate analysis

Table 2-2 shows the results of univariate analysis for clinical and radiological variables. The log-rank tests did not find significant association between the clinical variables and MRFS. Among the semantic radiological features, heterogeneous SI on T2-WI, peritumoral enhancement and fatty signal correlated with MRFS (log-rank test, p=0.01, 0.04 and 0.02 respectively).

Table 2-1. Patient characteristics

Characteristics	
Gender	
Male	51.4 (18/35)
Female	48.6 (17/35)
Age	49 (28 - 94)
Location	
Upper limb	2.9 (1/35)
Trunk	11.4 (4/35)
Lower limb	85.7 (30/35)
Depth	
Superficial	2.9 (1/35)
Deep	97.1 (34/35)
Size	136 (34 - 250)
< 5cm	5.7 (2/35)
≥5cm	94.3 (33/35)
WHO-PS	
WHO-PS 0	97.1 (34/35)
WHO-PS 1	2.9 (1/35)
Initial biopsy	
No	14.3 (5/35)
Imaging-guided microbiopsy	45.7 (16/35)
clinical microbiopsy	25.7 (9/35)
Surgical biopsy	14.3 (5/35)
Initial treatment	
Surgery	28.6 (10/35)
Neaodjuvant chemotherapy + surgery	71.4 (25/35)
Adjuvant radiotherapy	
No	14.3 (5/35)
Yes	85.7 (30/35)

Data are presented as number of patients with percentage in parentheses. Age and Size are given as median with range in parentheses.

Table 2-2. Prognostic value of clinical and semantic radiological variables for metastatic relapse-free survival according to Log-rank test.

Clinical Variables				Semantic Radiological Variables			
Variables	**Nb. of patients**	**Nb. of events**	**Log-rank p-value**	**Variables**	**Nb. of patients**	**Nb. of events**	**Log-rank p-value**
Gender				**Size**			
Female	17	7	0.6	**< 50mm**	2	0	0.4
Male	18	5		**≥ 50mm**	33	12	
WHO-PS				**Heterogeneous SI on T2-WI**			
WHO-PS 1	34	11	0.08	**0%, i.e. homogeneous**	10	0	**0.01***
WHO-PS 2	1	1		**1-24% of tumor volume**	2	0	
Age				**25-49% of tumor volume**	12	7	
< median (49 years old)	18	7	0.9	**50-74% of tumor volume**	6	2	
≥ median (49 years old)	17	5		**≥75% of tumor volume**	5	3	
Depth				**Necrotic signal**			
Deep	34	12	0.6	**No**	28	10	0.3
Superficial	1	0		**Yes (<50% tumor volume)**	7	2	
Adjuvant Radiotherapy				**Fatty signal**			
No	5	2	0.4	**No**	11	2	**0.02***
Yes	30	10		**Yes (<50% tumor volume)**	18	6	
Initial treatment				**Yes (≥50% tumor volume)**	6	4	
Surgery	11	2	0.3	**Tail sign**			
Neoadjuvant Chemotherapy	24	10		**Thick**	1	1	0.4
Location				**Thin**	5	2	
Lower limb	30	10	0.5	**Absent**	29	8	
Upper limb	1	0		**Peritumoral edema**			
Trunk	4	2		**No 0.**	10	5	0.1
Sample before				**Yes**	25	7	
No	5	1	0.6	**Peritumoral enhancement**			
Yes	30	11		**No**	26	7	**0.04***
Margins				**Yes**	9	5	
R0	16	4	0.2				
R1-R2	19	8					

NOTE. Median age of the series was 49 years-old. Abbreviations: SI: signal intensity, WHO-PS: world health organization performance status.

The univariate Cox regressions in the prefiltering step identified fourteen radiomics features that correlated with MRFS (range of p-values: 0.001-0.045, Table 3). The correlation plot is given in Annexe 3 - Supplementary Data 2.

Construction of the radiomics score

The 14 radiomics features selected after univariate analysis were reduced to 3 weighted radiomics features (for lambda=0.1168), namely FOS_Skewness, GLRLM_LRHGE and SHAPE_Volume (weights=0.4162, -0.1666 and 0.1071, respectively). The radiomics score corresponded to the sum of these weighted features (median=-0.0068, range=(-1.0930; +1.2269)). The radiomics score was significantly associated with MRFS (HR=19.37, CI95%=(3.39-110.7)) and was a valid prognostic variable after testing proportional hazards assumption (p=0.227). Figure 2-1 shows the prognostic modeling steps and Kaplan-Meier curves when the radiomics score was dichotomized per its median. Figures 2-2 and 2-3 show two opposite examples and their outcome: a heterogeneous tumor with high radiomics score and a homogeneous tumor with low radiomics score, respectively.

Significant correlations were found between heterogeneous SI on T2-WI, the 2 selected texture features (FOS_Skewness, GLRLM_LRHGE), and the radiomics score (ρ = 0.769, -0.561, and 0.784, $p<0.001$ – Figure 2-4).

Comparison of prognostic models

Figure 2-5 and Table 2-4 summarize the models performances. After stepwise selection, the radiological model (2) included: fatty signal ≥50% of tumor volume, peritumoral enhancement, and heterogeneous SI on T2-WI ≥25% of tumor volume (c-index=0.901, iAUC=0.898). The lowest performances were obtained with clinical model (1) (c-index=0.637, iAUC=0.686), and the highest performances with model (4) (c-index=0.937, iAUC=0.925). Models (2), (3) and (4) were all significantly better than model (1), but comparisons between them did not reach significance.

Table 2-3. Prognostic value of the radiomics features for metastatic relapse-free survival in univariate analysis.

Radiomics features	HR (CI95%)	p-value
FOS_minimum	0.964 (0.577-1.610)	0.8879
FOS_mean	0.366 (0.177-0.756)	**0.0066***
FOS_stdev	1.224 (0.704-2.130)	0.4736
FOS_maximum	0.963 (0.499-1.861)	0.9115
FOS_Skewness	3.005 (1.519-5.944)	**0.0016****
FOS_Kurtosis	0.408 (0.139-1.198)	0.1028
FOS_Entropy_log10	1.152 (0.553-2.400)	0.7049
FOS_Entropy_log2	1.152 (0.553-2.400)	0.7049
FOS_Energy	1.031 (0.460-2.311)	0.9403
SHAPE_Volume	1.494 (1.009-2.212)	**0.0451***
SHAPE_Sphericity	0.853 (0.515-1.413)	0.5376
SHAPE_Compacity	1.846 (0.968-3.520)	**0.0628**
GLCM_Homogeneity	0.894 (0.456-1.753)	0.7439
GLCM_Energy	1.083 (0.498-2.356)	0.8402
GLCM_Contrast	1.112 (0.658-1.877)	0.6925
GLCM_Correlation	1.449 (0.835-2.514)	0.1877
GLCM_Entropy_log10	1.416 (0.708-2.831)	0.3256
GLCM_Entropy_log2	1.416 (0.708-2.831)	0.3256
GLCM_Dissimilarity	1.130 (0.663-1.927)	0.6533
GLRLM_SRE	1.064 (0.560-2.021)	0.8504
GLRLM_LRE	0.946 (0.457-1.960)	0.8823
GLRLM_LGRE	2.458 (1.276-4.734)	**0.0072***
GLRLM_HGRE	0.341 (0.154-0.754)	**0.0079***
GLRLM_SRLGE	2.546 (1.293-5.016)	**0.0069***
GLRLM_SRHGE	0.373 (0.175-0.795)	**0.0107***
GLRLM_LRLGE	1.521 (0.834-2.774)	0.1712
GLRLM_LRHGE	0.390 (0.170-0.893)	**0.0258***
GLRLM_GLNU	1.851 (1.113-3.079)	**0.0177***
GLRLM_RLNU	1.347 (0.942-1.928)	0.1029
GLRLM_RP	1.045 (0.534-2.046)	0.8975
NGLDM_Coarseness	0.013 (0.000-0.699)	**0.0327***
NGLDM_Contrast	1.070 (0.600-1.909)	0.8178
NGLDM_Busyness	0.583 (0.337-1.009)	**0.0537**
GLZLM_SZE	0.630 (0.341-1.162)	0.1388
GLZLM_LZE	0.832 (0.424-1.632)	0.5927
GLZLM_LGZE	2.190 (1.054-4.550)	**0.0357***
GLZLM_HGZE	0.423 (0.194-0.921)	**0.0302***
GLZLM_SZLGE	1.715 (0.869-3.384)	0.1197
GLZLM_SZHGE	0.420 (0.200-0.881)	**0.0218***
GLZLM_LZLGE	1.088 (0.564-2.098)	0.8023
GLZLM_LZHGE	0.616 (0.203-1.874)	0.3938
GLZLM_GLNU	1.390 (0.996-1.939)	**0.0526**
GLZLM_ZLNU	1.935 (1.080-3.468)	**0.0265***
GLZLM_ZP	0.834 (0.425-1.637)	0.5982

NOTE. Abbreviations: HR : hazard ratio, CI95% : 95% confidence interval. *: $p<0.05$, **: $p<0.005$, ***: $p<0.001$.

Figure 2-1. Construction of the prognostic radiomics score for metastatic relapse free survival (MRFS). Least absolute shrinkage and selection operator (LASSO) regression analysis was performed on the subset of 14 radiomics features significantly associated with MRFS in order to identify exactly 3 weighted radiomics features. (**a**) shows the partial likelihood deviance as function of different shrinkage parameters. A selection of 3 features with lowest deviance corresponded to log(Lambda) ≈ -2.15 (black arrow), i.e. Lambda ≈ 0.1168. (**b**) Plot of the features coefficients against Lambda. Each line corresponds to a feature. The final radiomics score was made by the weighted sum of FOS_Skewness, GLRLM_LRHGE and SHAPE_Volume. (**c**) Kaplan-Meier analysis of two groups of patients - above and below median radiomics score – with MRFS, in addition to risk table and log-rank p-value.

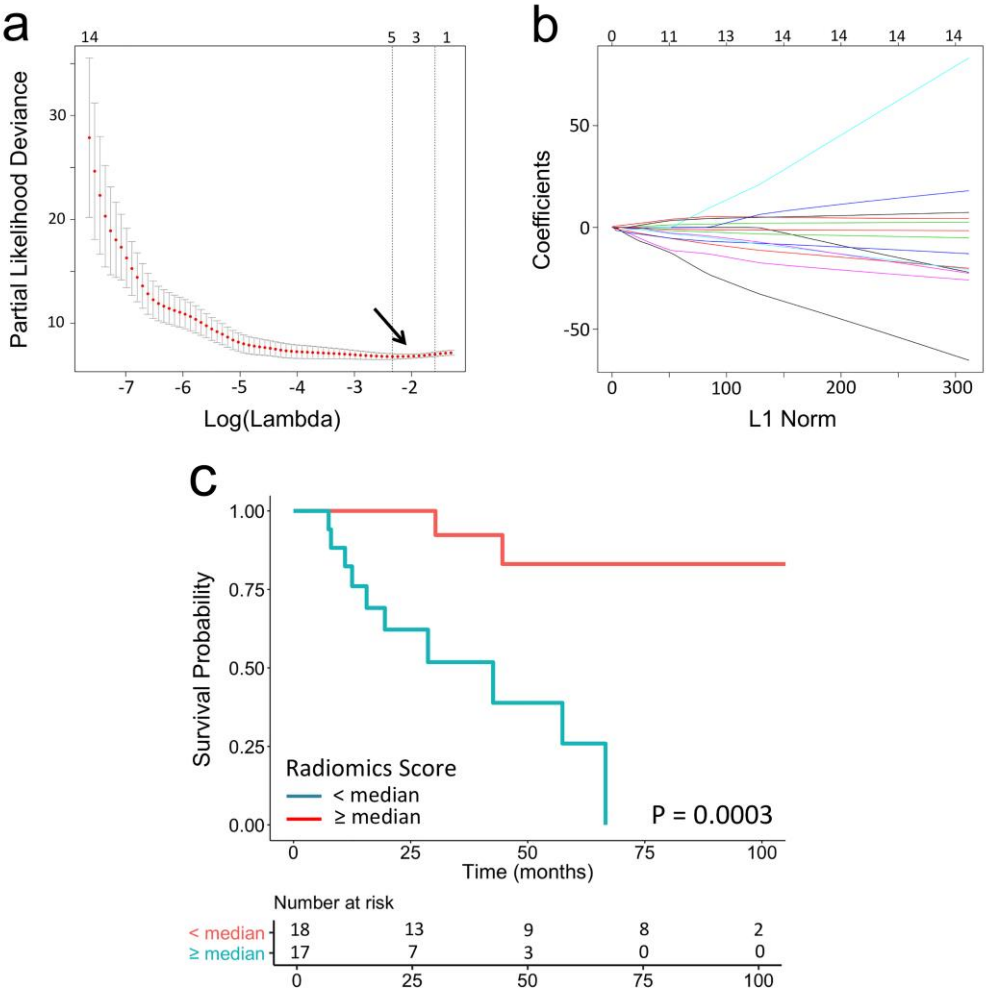


Figure 2-2. Example of high radiomics score. A 50-year-old patient presented with a deep-seated 167 mm-long myxoid/round cells liposarcoma of the thigh (white arrowhead) with heterogeneous signal intensity (SI) on sagittal short time inversion recovery (**a**) and T2 weighted imaging (-WI) (**b**), rather low SI on T1-WI (**c**) and heterogeneous enhancement on static contrast enhanced T1-WI (**d**), with a hemorrhagic area (white arrows). (**e**) The frequency histogram of normalized SI on T2-WI demonstrates a rather large distribution of SI. The radiomics score was 0.331 (median = -0.001). The patient undergoes 5 cycles of anthracycline based chemotherapy, surgery and adjuvant radiotherapy. Pathological analysis of the surgical specimen found only 5% of residual tumor cells, without round cells. Follow-up by whole-body MRI shows the occurrence of multiple soft-tissue and bone metastases 15.6 months after surgery.

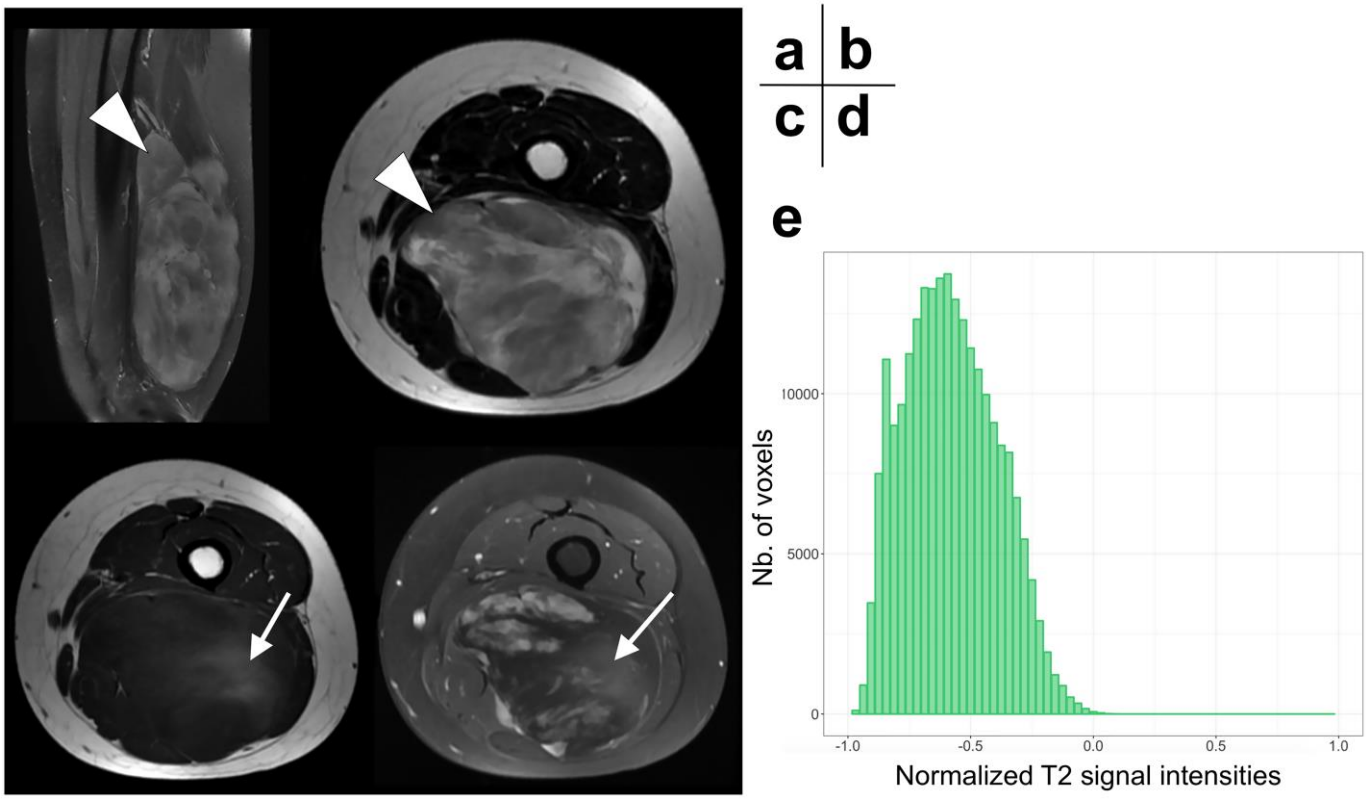


Figure 2-3. Example of low radiomics score. A 43-year-old patient presented with a deep-seated 107 mm-long myxoid/round cells liposarcoma of the arm (white arrow) with iso- to low signal intensity (SI) on T1 weighted-imaging (-WI), homogeneous SI on T2-WI (**b**) and rather homogeneous diffuse enhancement on static contrast enhanced T1-WI (**c**). (**d**) The frequency histogram of normalized SI on T2-WI demonstrates a narrow distribution of several voxels with high SI, and a large tail on the left side. The radiomics score was -0.783 (median = -0.001). The patient underwent 5 cycles of anthracycline based chemotherapy, surgery and adjuvant radiotherapy. Pathological analysis of the surgical specimen found 85% of residual tumor cells, without round cells. Follow-up did not show any recurrence 7 years later.

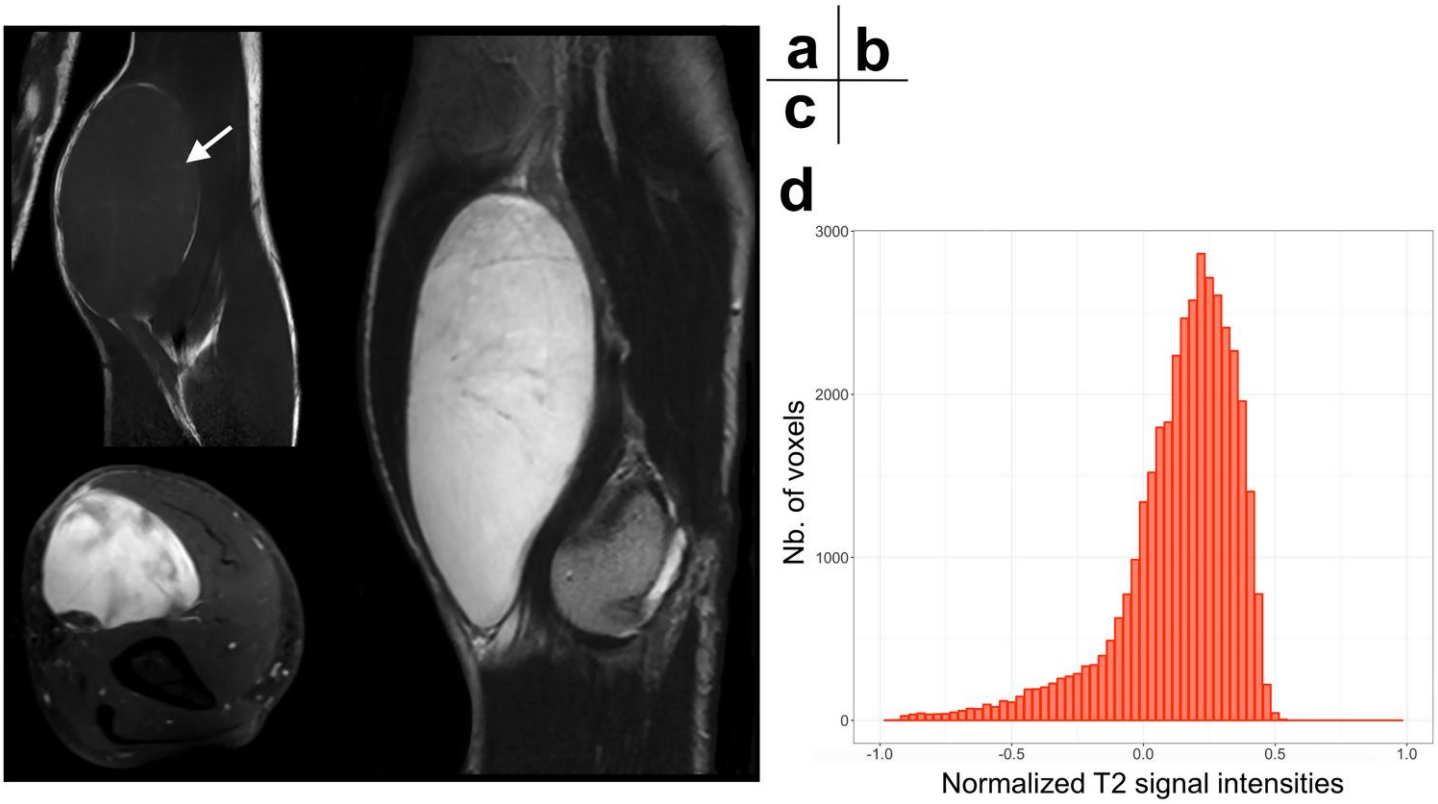


Figure 2-4. Correlations between the 2 relevant texture features and the radiomics score with the visual heterogeneity on T2-WI according to a 5-points scale. P-value corresponds to the p-value of the Spearman rank test with rho (ρ) in parentheses. *: p<0.05, **: p<0.001, ***: p<0.0001. Abbreviations: FOS: first-order histogram, GLRLM_LRHGE: Gray level run length matrix, long run high gray level emphasis, SI: signal intensity, -WI: weighted imaging

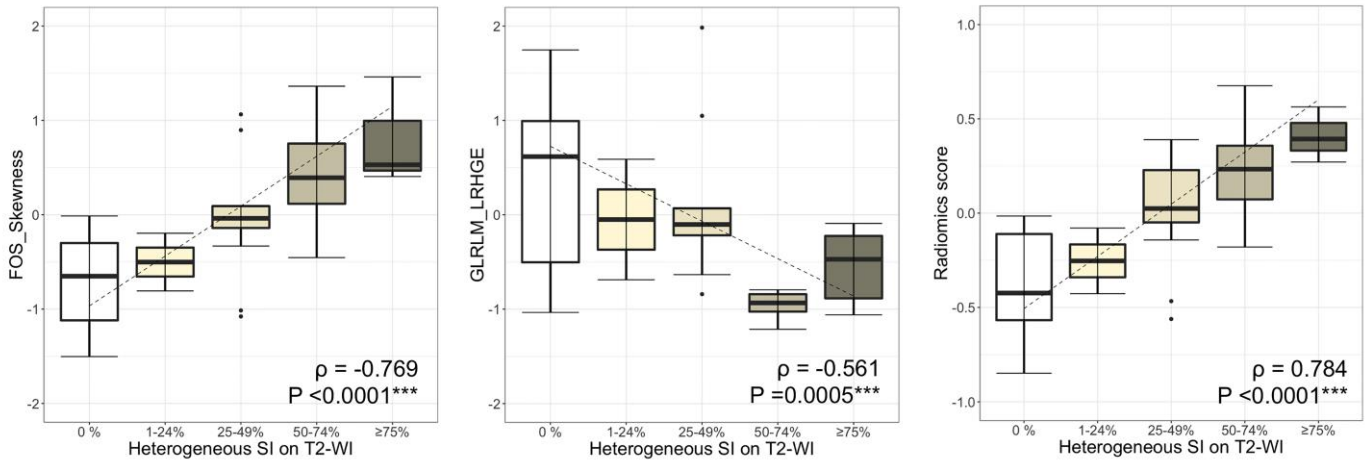


Figure 2-5. Survival analyses. (a) Areas under the time-dependent receiver operating characteristics curves (AUC) of the 4 models for metastatic relapse free survivals (MRFS). (b) Survival receiver operating characteristics curves for the four models at 2 years after the curative surgery. Model (1) corresponds to the clinical model, model (2) to the semantic radiological model, model (3) to the radiomics model, model (4) to the combination of radiomics and relevant semantic radiological variables.

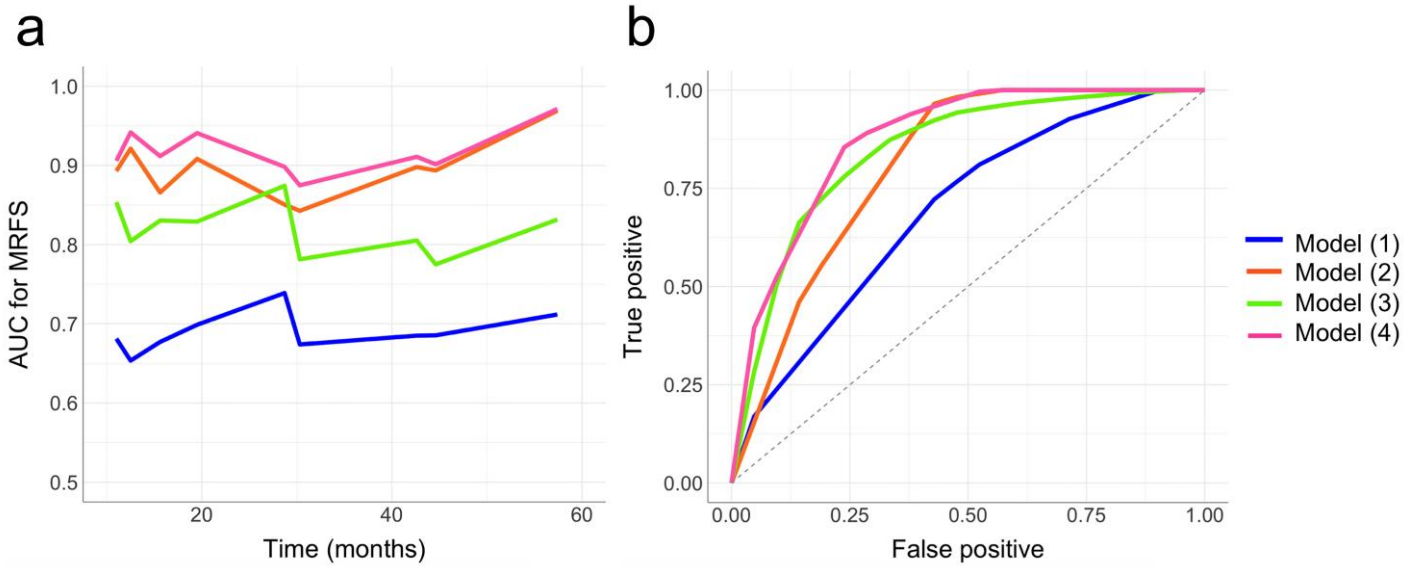


Table 2-4. Performances of the models measured by concordance index (C-index) and iAUC.

Model		C-index (CI95%)	iAUC	Comparison with	p-value
Model (1)	Clinical features	0.637 (0.476-0.798)	0.686	Model (2)	**0.007***
				Model (3)	**0.009***
				Model (4)	**0.015***
Model (2)	Semantic radiological features	0.901 (0.830-0.972)	0.898	Model (3)	0.428
				Model (4)	0.774
Model (3)	Radiomics features	0.817 (0.703-0.931)	0.828	Model (4)	0.487
Model (4)	Radiomics + radiological features	0.937 (0.874-1.000)	0.925	-	-

NOTE. Each model was adjusted for: margins of the surgical specimen, adjuvant/neoadjuvant chemotherapy and adjuvant radiotherapy. Model (1) corresponds to the clinical model, Model (2) to the model based on relevant semantic radiological variables after univariate analysis and stepwise selection; Model (3) to the model based on selected radiomics features, and Model (4) to the model combining radiomics score and radiological features from model (2). All the models were adjusted for the adjuvant/neoadjuvant chemotherapy, adjuvant radiotherapy and margins of the surgical specimen.
Abbreviations: CI95%: 95% confidence interval, iAUC: integrated area under the time-dependent receiver operating characteristic curve. *: $p<0.05$.

DISCUSSION

Our study illustrates that adding radiomics analysis on a relevant MRI sequence can improve the prediction of metastatic relapse in M/RC-LPS patients. We built a radiomics score based on carefully selected radiomics features that correlated with visual heterogeneity and deepened in a quantitative manner based on what human eyes find heterogeneous. We found that the best prognostic model combined radiomics score with relevant semantic radiological features.

Building a radiomics score with LASSO-penalized Cox regression method has already demonstrated excellent results in other cancers (Lu et al, 2019; Khorrami et al, 2019). Although a larger number of features was recommended for the optimal features selection according to LASSO, we limited the score to 3 radiomics features because of limited population study, to avoid overfitting. The radiomics score is easy to interpret, within a limited range with a median close to 0. Thus, a positive score indicates heterogeneous lesions and high risk of metastatic relapse, while a negative score indicates a rather homogeneous lesion of better prognosis. A close look on the 3 features of the score can help visualizing the radiological aspects of M/RC-LPS with a poorer prognosis. These M/RC-LPSs had higher volume. A high histogram skewness - becoming positive when the heterogeneity on T2-WI increased - indicates a more asymmetric distribution of the SIs on T2-WI in patients with worse MRFS - high SIs having larger and longer tails while low SIs provide the largest part of the histogram. In other words, it seems that myxoid content decreased in patients with lower MRFS to the benefit of non-myxoid that could correspond to transition towards round cells content. GLRLM_LRHGE, which estimates the proportions of higher grey-level values long runs of voxels, progressively decreased when the heterogeneity on T2-WI increased, suggesting that tumors with poorer MRFS show a lower proportion of long runs of voxels with myxoid signal.

The interpretation of the radiomics score concurs with the literature. A recent study has shown that heterogeneous SI on ≥ 50% of tumor volume was predictive of lower MRFS (Crombé et al, 2019a). This study included different histotypes and visual heterogeneity was generally due to infarction, necrosis, fibrosis, intratumoral compartments with varying SIs. However, reasons for intra-tumoral heterogeneity may depend on the histotype. Regarding MRC-LPS, heterogeneity on T2-WI is

strongly supposed to be due to round cells. Validating this hypothesis would require voxel-wise comparisons of histological and MRI slices. Our group is currently working on these voxel-wise approaches including multi-parametric MRI. The advantage of MRI over a biopsy sample is to enable a global, non-invasive view of the tumor and to provide the possibility to correct underestimation in grade, amount of round cells or other prognostic biomarkers. Moreover, imaging-related features could carry complementary and independent prognostic information.

Our series is comparable with other M/RC-LPS series regarding characteristics of the population and outcomes. The incidence of metastatic relapses usually ranges from 10 to 33% (Engström et al, 2008; Haniball et al, 2011; Asano et al, 2012; Fuglø et al, 2013; Fiore et al, 2007). Most relapses were located in soft-tissues and bone except. It should be noted that only one patient had a superficial tumor, which is in agreement with the usual repartition of M/RC-LPS (Fletcher et al, 2013). However, the prognostic value of the tumor depth is certainly far less important in our clinical model than in other series of sarcomas. Moreover, we identified 6 patients with over 50% of fatty signal, while the amount of fat inside M/RC-LPS rarely exceeds 20% [30].

The results of the semantic radiological analyses also concur with prior studies. Kuyumcu et *al*. found that M/RC-LPSs with high fat content had worse overall survival (Kuyumcu et al, 2018). The amount of necrosis on MRI has been previously identified as a negative prognostic factor (Tateishi et al, 2004; Löwenthal et al, 2014) and certainly contributed to radiological heterogeneity. However, necrosis can be difficult to estimate on MRI and confused with myxoid or cystic components, which could explain why the association between necrosis and MRFS did not reach significance. Myxoid stroma demonstrates progressive and delayed enhancement. Consequently, early acquisition delays could miss the contrast enhancement of some areas. Therefore, CE-T1-WI was not used to perform the radiomics analysis. Moreover, because of the retrospective nature of the study, several parameters of CE-T1-WI were not controlled and could have introduced biases in the radiomics features - for instance: 2D vs. 3D sequences, TSE vs. gradient-echo sequences, fat-suppression vs. not, different contrast agents or acquisition delays after injection. Interestingly, peritumoral enhancement was also significantly associated with MRFS in the present study. The underlying biological mechanism explaining the relationship between peritumoral enhancement and prognosis has not been elucidated. To our knowledge,

no study has investigated which genes could be abnormally expressed in sarcomas with pathological surrounding tissues. We hypothesize that peritumoral enhancement would correspond to more inflammatory tumors, satellite tumoral cells, higher propensity to spread in the tissues and to develop local and distant relapses, leading to poorer prognosis.

Some prognostic radiological features were not directly evaluated herein, such as lack of pseudocapsule, vascular encasement or bone infiltration. These features describe poorly defined, infiltrating tumors, such as MRI growth pattern or peritumoral enhancement (Crombé et al, 2019a; Nakamura et al, 2017). We purposely decided to limit the amount of features and to focus on the most relevant ones in the latest radiological studies (Crombé et al, 2019a; Zhao et al, 2014).

The prognostic performances of our best model are not perfect. Our results showed that the radiomics model alone (i.e. model (3)) did not perform better than the radiological model (2) based on semantic features aiming at qualitatively or semi-quantitatively describing the radiological phenotype of MRC-LPS. Both c-index and iAUC were slightly higher with the model (2) than with the model (3) - even if the comparisons did not reach significance. Though an advantage of radiomics approaches over radiological approaches is to provide an 'objective' assessment of tumor heterogeneity and shape with standardized tools, herein, radiomics features alone without a comprehensive visual judgment by expert radiologists did not improve the prediction of MFS for MRC-LPS. Moreover, computing the radiomics score is time-consuming and won't be feasible in clinical practice until segmentation and MRI post-processing will be automated. However, as it can be seen through the performances of the model (4), the performances of predictive models were increased by combining the relevant radiological and radiomics features from the models (2) and (3). Consequently, we believe that the next step would be to investigate techniques for objectively quantifying the relevant semantic features and to combine them with the assessment of intratumoral heterogeneity trough radiomics approaches. Hence, we hope these preliminary results will encourage researcher to pursue efforts in this field. In addition, the integration of molecular characteristics of MRC-LPS could enhance predictive models performances (Gillies et al, 2016). Other MRI sequences, such as diffusion-weighted-imaging, dynamic-contrast-enhanced-MRI or MR-spectroscopy could improve quantitative depiction of MRC-LPS. In particular, diffusion-weighted-imaging could help identifying MRC-LPS with a more

constraining water diffusion (i.e. with higher round cells content) as well as to better characterize and quantify the peritumoral abnormalities - such as a aponeurotic spreading and infiltrative growth pattern (Hong et al, 2019; Yoon et al, 2019). In addition, recent studies have found that dynamic-contrast enhanced MRI and diffusion-weighted imaging could help identifying sarcomas with high proliferation rates and be further used to improve the response prediction in case of neoadjuvant treatment (Lee et al, 2019; Crombé et al, 2019c).

Our pilot study has limits. First, the radiomics score was built on a small population as we only included in the study patients with TSE T2-WI. Nonetheless, it remains one of the largest radiological studies about MRC-LPS, since previous ones never exceeded 43 patients. Second, we did not test the radiomics score on external validation cohorts. However, details for the calculation of the score are provided, hence, other centers could test it on their own population. Third, even if the therapeutic guidelines were followed, there were some variations in the treatments, that is why we adjusted the models for surgical margins, chemotherapy and radiotherapy. Moreover, none of the finally included patients underwent neoadjuvant radiotherapy, though this treatment is routinely performed in other centers. Fourth, other radiomics features, prefiltering, selection or penalized methods, and survival models could have been used and compared. This was not the purpose of the study and we decided to restrict the methodology to statistical techniques that have proven to work well with radiomics.

To conclude, we found that integrating the quantification of shape and heterogeneity on MRI through a radiomics score can improve the prediction of metastatic relapse in MRC-LPS. In addition to clinical, pathological and radiological prognostic factors, our radiomics score helped quantifying these modifications, paving the way to an improvement of both patients' stratification and of the tailoring of their clinical management.

*

* *

2.3. Limites et ouvertures

Cette étude présente des limites méthodologiques importantes faisant qu'elle ne peut dépasser le stade de la preuve de concept: absence de filtrage des indices radiomics selon leur robustesse après segmentations multiples, absence d'étude de fantôme, absence de *test-retest*, absence de corrélation directe possible entre score radiomics et pourcentage réel de cellules rondes, absence de courbes de calibration, absence de cohorte de validation extérieure, caractère retrospectif, absence d'études cout-efficacité, absence de jeu de données *open source*. Nous devons aussi mentionner que les IRM provenaient de centres différents (bien que plus de la moitié aient été réalisées sur l'institut Bergonié). Son RQS est donc de +7.

Ce qui nous semble cependant important est la démonstration qu'il existe une information pronostique initiale dans l'imagerie de ces tumeurs et qu'il est possible de la quantifier par approche radiomics. De plus, les indices radiomics seuls n'apportent pas le meilleur indicateur de performance: le score radiomics s'avère même un peu moins performant que des critères radiologiques classiques. Il s'avère par contre que les indices radiomics et radiologiques se potentialisent pour apporter les meilleures performances. Ce résultat sous entend qu'il est nécessaire que les radiologues intègrent leur expérience dans les études radiomics, quitte à ce que des méthodes de quantification des variables radiologiques soient secondairement mises au point.

3. PREDICTION DE LA REPONSE

3.1. Introduction

Cette étude, réalisée conjointement avec Mme Cynthia Périer alors doctorante dans l'équipe MONC, a pour origine la constatation par les cliniciens et radiologues que, dès la 1ère évaluation par IRM à 2 cures post-NAC, les STM localement avancés pouvaient présenter des modifications de leur organisation intra-tumorale particulièrement bien visibles sur la séquence IRM pondérée T2. En effet, alors que certains STM peuvent rester stables morphologiquement, d'autres vont présenter une augmentation de leur hétérogénéité intra-tumorale liée à l'intrication de phénomènes de nécrose et de saignements (élevant le signal en T2) et de fibrose (diminuant le signal en T2), en plus de modifications de forme (vers un aspect d'aplatissement). Les hypothèses que nous avons cherchées à explorer étaient que ces modifications précoces pouvaient être capturées par une approche delta-radiomics et qu'elles corrélaient avec la réponse histologique sur la pièce opératoire finale post-NAC plus fortement que les critères de réponse les plus communément utilisés c'est-à-dire RECIST v1.1.

3.2. Article 2

L'article issu de ce travail a été publié en décembre 2018 dans *Journal of Magnetic Resonance Imaging* (PMID: 30569552 , doi: 10.1002/jmri.26589 – rang B SIGAPS, IF = 3.732). Il s'agit ici du manuscrit final reformaté.

ORIGINAL RESEARCH

T_2-Based MRI Delta-Radiomics Improve Response Prediction in Soft-Tissue Sarcomas Treated by Neoadjuvant Chemotherapy.

Amandine Crombé, MD, MS,[1,2*] Cynthia Périer, MS,[2] Michèle Kind, MD MS,[1] Baudouin Denis De Senneville, PhD,[2] François Le Loarer, MD PhD,[3] Antoine Italiano, MD PhD,[4] Xavier Buy, MD,[1] and Olivier Saut, PhD[2]

Background: Standard of care for patients with high-grade soft-tissue sarcoma (STS) are being redefined since neoadjuvant chemotherapy (NAC) has demonstrated a positive effect on patients' outcome. Yet response evaluation in clinical trials still relies on RECIST criteria.
Purpose: To investigate the added value of a Delta-radiomics approach for early response prediction in patients with STS undergoing NAC.
Study Type: Retrospective.
Population: Sixty-five adult patients with newly-diagnosed, locally-advanced, histologically proven high-grade STS of trunk and extremities. All were treated by anthracycline-based NAC followed by surgery and had available MRI at baseline and after two chemotherapy cycles.
Field Strength/Sequence: Pre- and postcontrast enhanced T_1-weighted imaging (T_1-WI), turbo spin echo T_2-WI at 1.5 T.
Assessment: A threshold of <10% viable cells on surgical specimens defined good response (Good-HR). Two senior radiologists performed a semantic analysis of the MRI. After 3D manual segmentation of tumors at baseline and early evaluation, and standardization of voxel-sizes and intensities, absolute changes in 33 texture and shape features were calculated.
Statistical Tests: Classification models based on logistic regression, support vector machine, k-nearest neighbors, and random forests were elaborated using crossvalidation (training and validation) on 50 patients ("training cohort") and was validated on 15 other patients ("test cohort").
Results: Sixteen patients were good-HR. Neither RECIST status ($P = 0.112$) nor semantic radiological variables were associated with response (range of P-values: 0.134–0.490) except an edema decrease ($P = 0.003$), although 14 shape and texture features were (range of P-values: 0.002–0.037). On the training cohort, the highest diagnostic performances were obtained with random forests built on three features: Δ_Histogram_Entropy, Δ_Elongation, Δ_Surrounding_Edema, which provided: area under the curve the receiver operating characteristic = 0.86, accuracy = 88.1%, sensitivity = 94.1%, and specificity = 66.3%. On the test cohort, this model provided an accuracy of 74.6% but 3/5 good-HR were systematically ill-classified.
Data Conclusion: A T_2-based Delta-radiomics approach might improve early response assessment in STS patients with a limited number of features.
Level of Evidence: 3
Technical Efficacy: 2

J. MAGN. RESON. IMAGING 2018.

INTRODUCTION

Standard of care for locally advanced high-grade soft-tissue sarcomas (STS) has been recently redefined, as phase 3 clinical trials have demonstrated improved overall and metastasis-free survivals in patients treated with anthracycline-based NAC (Saponara et al, 2017; Gronchi et al, 2017; Issels et al, 2010). Despite encouraging results of ^{18}Fluorodeoxyglucose position emission tomography (^{18}F-FDG-PET/CT), modified Choi criteria and dynamic-contrast enhanced MRI (DCE-MRI), evaluation of response to NAC still relies on RECIST 1.1 (Eisenhauer et al, 2009).

Non-invasive quantification of tumor heterogeneity and its changing phenotype during treatment is a recent, promising and challenging field of research referred to as radiomics. Radiomics techniques aim at leveraging big-data analytics and personalized medicine approaches in oncologic imaging (Lambin et al, 2017; Limkin et al, 2017; Gillies et al, 2016). To achieve this, several numeric features are extracted to quantify and to screen tumor phenotype and surrounding tissue on any available imaging modality. After a careful selection of features, machine learning algorithms can be designed and trained to answer crucial oncologic questions such as associations between imaging phenotypes and molecular subtypes with specific treatment and outcomes, prediction of response and patient outcome by including other –omics (genomic, transcriptomics) information within the model (Aerts et al, 2016).

Because of their complex morphology, architecture and changes during treatments, STS may be particularly appropriate to the radiomics approach. Indeed, radiomics on DWI may help to improve STS grading on microbiopsy (Corino et al, 2017). In addition, Hayano et al. have demonstrated that texture parameters on CT-scan were associated with neoangiogenesis and overall survival for STS treated with radiotherapy and bevacizumab (Hayano et al, 2015; Tian et al, 2015). STS heterogeneity assessed on ^{18}F-FDG-PET-CT may be more predictive of survival as compared to classical measure of maximal standardized uptake value (SUV_{max}) (Eary et al, 2008). Recently, composite texture features from MRI and from ^{18}F-FDG-PET/CT have enabled to identify aggressive tumors at risk of lung metastasis at baseline (Vallières et al, 2015). Together, these promising studies highlight the potential of radiomics applied to STS. However, to our knowledge, applications to response prediction to NAC have never been attempted.

Visual MRI evaluation of STS during NAC can highlight a wide range of morphologic alterations combining fibrotic and necrotic processes, infarction, bleeding, re-differentiation or selection of resistant component, leading to change in tumor heterogeneity that could be quantified with shape and texture features. As change in longest diameter (LD) is not a sufficient criterion to predict therapeutic response, we hypothesized that the changes in shape and texture features during treatment (i.e. delta-radiomics) could help predict NAC efficacy through the histologic response.

MATERIALS AND METHODS

Patients

The institutional review board approved this study and the requirement for informed consent was waived.

All consecutive adult patients between June 2007 and June 2017 were included, as they presented with histologically proven high-grade STS of extremities or trunk wall, without metastasis on chest CT-scan, eligible for an anthracycline-based NAC according to the regional sarcoma reference center board. High-grade was defined as grade III STS according to the French Federation of Cancer Centers Sarcoma Group grading system (Trojani et al, 1984).

Criteria for inclusion were: measurable tumor with MRI, available MRI performed <28 days before the first cycle of NAC (: baseline, MRI_0) and between cycle 2 and 3 of NAC (: early evaluation, MRI_1), 4 to 6 cycles of NAC, histological response assessment on surgical specimen by an expert pathologist following published guidelines (Wardelmann et al, 2016). A threshold of <10% of viable cells assessed on whole tumor defined good histological response (good-HR) (Cousin et al, 2017).

Of the 163 patients with a newly diagnosed STS of trunk wall and extremities who underwent NAC at our institution (according to the pharmacology department), 28 patients were excluded because of non-anthracycline-based NAC, 20 because of less than 4 cycles of NAC, 33 because T2-weighted-imaging (T2-WI) was not performed at baseline, 7 because T2-WI was not performed at early evaluation, 10 because of non-diagnostic MRI at baseline and/or early evaluation.

MR imaging

Images were acquired in daily practice using 1.5-Tesla MR-systems from different radiological centers. Ninety-three examinations (72%) were carried out on a Magnetom AERA, (Siemens Healthineers, Erlangen, Germany). Coils, field-of-view and matrices were adapted to tumor location and size. To be considered as 'diagnostic', MRI must include at least 2D T2-WI turbo-spin echo (TSE) sequence without fat-suppression, T1-WI before and after Gadolinium-chelates injection (contrast-enhanced T1-WI, CE-T1-WI) and 2 orthogonal acquisition plans. Section thickness ranged from 3 to 5 mm. Ranges of repetition time / echo time were: 500-700/10-15 msec for T1-WI and 2400-6860/100-130 msec for T2-WI.

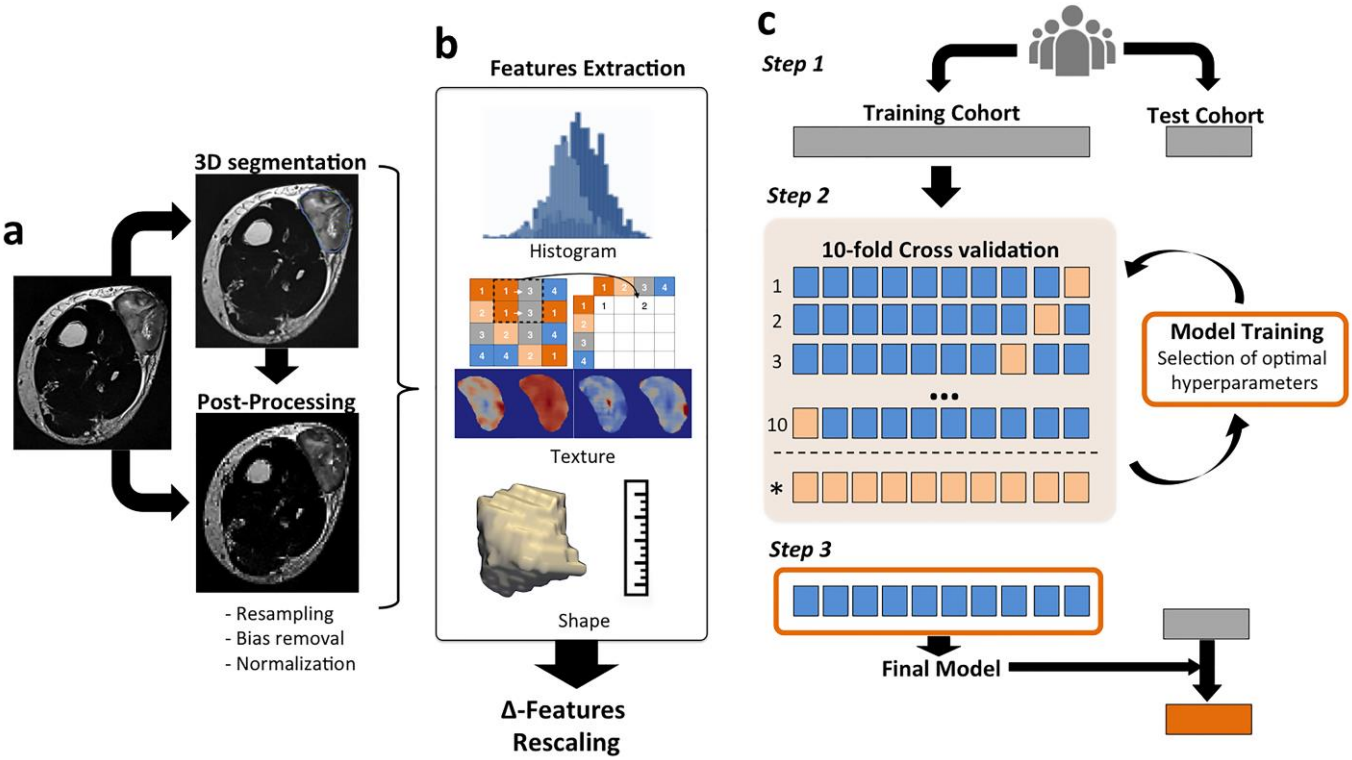
Semantic radiological features

Three senior radiologists (AC, XB and MK, with 3, 5 and 27 years of experience in STS imaging, respectively), independently reviewed the MRI blinded to patient data in a randomized fashion on a dedicated PACS workstation. They reported:

- LD in mm on MRI_0 and MRI_1, relative change in LD and RECIST response status.
- Percentage of tumor volume with changes compatible with fibrosis (low signal intensity (SI) on T2-WI, T1-WI, subtle enhancement) and/or necrosis (fluid-like SI on T2-WI, variable SI on T1-WI, no enhancement), as follows: 0%, <50% and ≥50%,
- Change in margin definition on CE-T1-WI (Δ_Margin_Definition), as follows: 'well-defined or better definition' versus 'stable ill-defined margins or worst',
- Change in surrounding edema on T2-WI without or with fat-suppression technique when available (Δ_Edema), as follows: 'none or decreased' versus 'stable or increased',
- Changes in peritumoral enhancement on CE-T1-WI (Δ_Peritumoral_enhancement), as follows: 'none or decreased' versus 'stable or increased'.

One radiologist (AC, with 3 years of experience in STS imaging) did a second reading 1.5 months later. Inter- and intra-observer agreements between the radiologists can be found in supplemental data. A consensual reading was performed 3 months after for the statistical analysis. The inter- and intra-observer agreements are given in Annexe 4 – Supplementary Data 1.

Figure 3-1: Radiomics pipeline. (**a**) First step consisted in MRI post-processing, including resampling (with a bi-linear interpolation), bias removal (N4) and normalization of signal intensities (with histogram-matching). The volume of interest was manually segmented, slice by slice, and then propagated on post-processed images, enabling the extraction of histogram-based, texture and shape features (**b**). This process was applied on baseline MRI and MRI after 2 cycles of neoadjuvant chemotherapy providing delta-radiomics features (Δ_features), which were rescaled (standard scaling). (**c**) Statistical method. In step 1, the whole data set was partitioned into a 'Training Cohort' and a 'Test Cohort'. In step 2, the 'Training cohort' was used to build the model. It was based on a 10-fold cross-validation that consisted in separating the 50 patients into 10 blocks of 5 patients. For each of the 10 combinations, the classifier was trained on the subset of 9 blocks (blue squares), then validated on the remaining block (in light orange). At the end of the cross-validation, each block has been used once for validation (*). This whole process was repeated with different tuning parameters proper to each type of classifier (: hyperparameters, *Supplemental Data*) and different methods for features selection and preprocessing, until obtaining a model with the highest accuracy and area under the ROC curve (AUROC). Those optimal metrics are shown in the cross-validation section of the results. In step 3, a model with the optimal combination of parameters was fitted on the whole training cohort. This final model was tested on the independent test cohort (dark orange) and its diagnostic performance (accuracy, AUROC, PPV, NPV, specificity, sensitivity, negative/predictive value) was calculated.



MRI post-processing (Figure 3-1)

Slice-by-slice 3D-delineation of whole tumor was manually made on T2-WI by one senior radiologist (AC with 3 years of experience in STS imaging) using the ROI manager of OSIRIX software.

All slices were resampled using bi-linear interpolation to obtain a common isotropic in plane 1x1 mm^2 pixel aspect. Signal intensities on T2-WI were normalized for non-uniform intensity (bias field correction) and the intensity ranges were standardized using histogram-matching with the acquisition of a healthy volunteer's thigh as reference (Tustison et al, 2010; Nyul & Udupa, 1999). Thirty-three first- and second-order texture and shape features were computed using in-house Python software based on the ITK library (Pedregosa et al, 2011). We calculated the absolute change of a given feature 'X' for each patient as follows: $\Delta\_X = X_{MRI\_1} - X_{MRI\_0}$.

Statistical analyses

First, a standardization scaling of the shape and texture features was performed. Comparisons between good-HR and poor-HR were assessed with Student or Mann-Whitney tests depending on results to the Shapiro-Wilk normality test. Association of categorical and ordinal variables with response was assessed with Chi-2 and Fischer tests. Correlations between features were assessed with Spearman's rank test. All tests were two-tailed. A p-value ≤ 0.05 was deemed significant.

To elaborate and validate the prediction model, the whole data set was partitioned in two: a training cohort (50 patients, included from June 2007 to June 2016) and a test cohort (15 patients, from July 2016 to June 2017 whose MRI were acquired after the initiation of the project). We initially selected only one feature per category (semantic, shape and texture categories) according to its lowest p-value at univariate analysis and lowest correlation with other significant features.

The selected combination of features was used to define models with 10-fold stratified cross validation on the training cohort. First, for each run and each set, the missing values were imputed with training features median and quantitative features were normalized by removing the training mean and scaling to unit variance. Several classification algorithms were evaluated using the scikit-learn library (Pedregosa et al, 2011): random forests (RF), k-nearest neighbors (KNN), support vector machine (SVM) and logistic regression (LR). The parameters of those estimators were optimized by cross-validated grid-search. The selected classifiers were then trained

with the whole 50-patients set and applied on the 15 patients from the test set using the same preprocessing method (Figure 3-1.c).

The cross-validation step was repeated 100 times with shuffled folds composition. The full process (including the final test) was also repeated with different random initialization seed for the RF algorithm. Average test metrics are reported for each step: accuracy, area under the ROC curve (AUROC), specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV) and train score.

Finally, we increased the number of features that were included in the model in a forward stepwise fashion according to their p-value at univariate analysis and we calculated the corresponding classifiers test metrics.

RESULTS

Patient characteristics (Table 3-1)

The cohort included 65 patients (27 females, mean age: 57.9 ± 12.8 years old), of which 16 (24.6%) were good-HR. The most frequent histotypes were undifferentiated sarcoma (50.8%), followed by myogenic sarcoma (leiomyosarcoma and rhabdomyosarcoma, 20%). Most of them were deep-seated (93.8%) in the lower limb (58.5%). Twenty-two patients (33.8%) received 4 cycles of NAC in total.

Standard radiological assessment (Table 3-2)

No association was found between baseline epidemiologic characteristics and histological response. LD at baseline was significantly higher in good-HR (146 ± 66 mm vs. 110 ± 51 mm, p=0.038). Relative change in LD at early evaluation was also significantly different between good-HR and poor-HR (-11.2 ± 20.8% versus 2.9 ± 19.5%, p=0.027), however, response status according to RECIST 1.1 was not associated with histological response (p=0.112) as most good-HR and poor-HR were classified as stable disease by these criteria (81.3% and 79.6%, respectively). Of all the semantic radiological features, only Δ_Edema was associated with response (p=0.003), with substantial inter- and intra-rater agreements (0.637 and 0.769, respectively).

Table 3-1. Epidemiologic characteristics

Characteristics		Patients (n=65)
Gender	Male	38 (58.5)
	Female	27 (41.5)
Age at diagnosis (y), mean ± sd		57.9 ± 12.8
Histotype	Undifferentiated sarcoma[1]	33 (50.8)
	Muscular sarcoma[2]	13 (20)
	M/RC liposarcoma[3]	5 (7.7)
	Other liposarcoma[4]	6 (9.2)
	Synovial sarcoma	7 (10.8)
	MPNST	1 (1.5)
Location	Trunk wall	12 (18.5)
	Pelvic Girdle	2 (3.1)
	Shoulder Girdle	6 (9.2)
	Upper limb	7 (10.8)
	Lower limb	38 (58.5)
Depth	Superficial	4 (6.2)
	Deep	61 (93.8)
LD at baseline (mm), mean ± sd		119 ± 56
Nb cycles	4 cycles	22 (33.8)
	5 or 6 cycles	43 (66.2)

NOTE. LD indicates longest diameter, sd indicates standard deviation, MPNST indicates malignant peripheral nerve sheath tumor.
Data are numbers of patients with percentages in parentheses, except for age and LD.
[1] : myxofibrosarcoma or undifferentiated sarcoma ;
[2] : leiomyosarcoma and rhabdomyosarcoma ;
[3] : myxoid/round cells liposarcoma ;
[4] : pleomorphic or dedifferentiated liposarcoma.

Table 3-2. Association between demographic and semantic radiological features and histological response.

	Good_HR	Poor_HR	p-value
Baseline clinico-radiological features			
Gender			
Male	11 (68.8)	27 (55.1)	0.393
Female	5 (21.2)	22 (44.9)	
Age at diagnosis (y)	58.8 ± 11.4	57.6 ± 13.3	0.873
Histotype			
Undifferentiated sarcoma[1]	10 (62.5)	23 (46.9)	0.257
Muscular sarcoma[2]	5 (21.2)	8 (16.3)	
M/RC liposarcoma[3]	1 (6.3)	4 (8.2)	
Other liposarcoma[4]	0 (0)	6 (12.2)	
Synovial sarcoma	0 (0)	7 (14.3)	
MPNST	0 (0)	1 (2.1)	
Location			
Trunk wall	2 (12.5)	10 (20.4)	0.146
Pelvic Girdle	2 (12.5)	0 (0)	
Shoulder Girdle	1 (6.3)	5 (10.2)	
Upper limb	2 (12.5)	5 (10.2)	
Lower limb	9 (56.2)	29 (59.2)	
Depth			
Superficial	1 (6.3)	3 (6.1)	1.000
Deep	15 (93.7)	46 (93.9)	
Nb cycles			
4 cycles	11 (68.8)	32 (65.3)	1.000
5 or 6 cycles	5 (21.2)	17 (34.7)	
LD on MRI_0 (mm)	146 (66)	110 (51)	**0.038 ***
MRI_0 to MRI_1			
Change in LD (%)	-11.2 ± 20.8	2.9 ± 19.5	**0.027 ***
RECIST 1.1			
Complete Response	0 (0)	0 (0)	0.112
Partial Response	3 (18.8)	3 (6.1%)	
Stable Disease	13 (81.2)	39 (79.6)	
Progressive Disease	0 (0)	7 (14.3)	
Objective Response			
Yes	3 (18.8)	3 (6.1)	0.154
No	13 (81.2)	46 (93.9)	
Δ_Margin_definition[§]			
Well- or better limited	5 (21.2)	9 (20)	0.490
stable or worst	11 (68.8)	36 (80)	
Δ_Edema			
None or decrease	12 (75)	15 (30.6)	**0.003 ****
Stable or increase	4 (25)	34 (69.4)	
Δ_Peritumoral enhancement[§]			
None or decrease	12 (80)	24 (57.1)	0.134
Stable or increase	3 (20)	18 (42.9)	
Fibro-Necrotic Changes			
No	2 (21.2)	14 (25.6)	0.430
< 50% tumor volume	9 (56.2)	23 (46.9)	
≥ 50% tumor volume	5 (31.3)	12 (24.5)	

NOTE. LD indicates longest diameter, sd insdicates standard deviation. MPNST indicates malignant peripheral nerve sheath tumor.
Data are numbers of patients with percentages in parentheses, except for age, LD and change in LD.
[1] : myxofibrosarcoma or undifferentiated sarcoma ;
[2] : leiomyosarcoma and rhabdomyosarcoma ;
[3] : myxoid/round cells liposarcoma ;
[4] : pleomorphic or dedifferentiated liposarcoma.
§: 8 patients had missing values for Δ_Peritumoral_enhancement and 4 for Δ_Margin_definition due to defective MR protocol (incomplete acquisition of edema on post contrast T1-WI, different acquisition plan on MRI_0 and MRI_1). * : $p \leq 0.05$; ** : $p \leq 0.005$.

Radiomics assessment

The population study was partitioned in a training cohort (50 patients, 11 Good-HR) and a test set (15 patients, 5 Good-HR). There was no statistical difference between the training and test cohorts regarding the baseline epidemiological characteristics (Annexe 4 - Supplemental Data 2).

Within the training cohort, changes in twelve first and second order textural indices were associated with response at univariate analysis: Δ_Histogram_Entropy (p=0.002), Δ_Stdev (p=0.008), Δ_ClusterProminence_5 (p=0.038), Δ_Energy_1 (p=0.015), Δ_Energy_2 (p=0.014), Δ_Energy_5 (p=0.010), Δ_Entropy_1 (p=0.005), Δ_Entropy_2 (p=0.004), Δ_Entropy_5 (p=0.003), Δ_Homogeneity_1 (p=0.037), Δ_Homogeneity_2 (p=0.022), Δ_Homogeneity_5 (p=0.014), as well as two shape features: Δ_Elongation (p=0.019) and Δ_Flatness (p=0.019) (Table 3-3). Correlation matrix demonstrated significant and strong correlations between all first and second order texture features (Table 3-4). Since the lowest p-value was obtained with Δ_Histogram_Entropy for texture features and Δ_Elongation for shape features, the initial selection for building the model included these two features and Δ_Edema.

Table 5 provides the performance of the classifiers for their optimal set of parameters and for this selection. On the training set, the highest mean accuracy after cross-Avalidation was obtained with RF (88.1%), followed by LR (85.8%), KNN (80.5%) and SVM (75.2%). In an objective response setting, RECIST 1.1 provided one of the lowest accuracy with 76.0% of correctly predicted patients. In descending order, AUROC were 0.87 with LR, 0.86 with RF, 0.81 with KNN, 0.67 with SVM and 0.66 for relative change in LD (Figure 3-2).

In the test set, the accuracy of the prediction of the 15 patients for classifiers trained with the whole training set on the 3 initial features were: 74.6% for RF, 66.7% for LR, 53% for SVM, 66.7% for KNN and 73.3% for an objective response according to RECIST 1.1. In details, while 9 (90%) of poor-HR were correctly predicted with RF, 3 (60%) good HR were systematically misclassified in the test set.

Table 3-3. Association between delta-radiomics features and response in training cohort

Variables	Good-HR	Poor-HR	p-value
1st order feature			
Δ_Histogram_Entropy	-0.185 ± 0.548	0.316 ± 0.406	**0.002 ****
Δ_Interval	-0.037 ± 0.179	0.017 ± 0.260	0.524
Δ_Kurtosis	2.967± 10.329	-8.16 ± 24.83	0.056
Δ_Mean	-0.071 ± 0.269	-0.078 ± 0.223	0.879
Δ_Skewness	-0.015 ± 1.725	0.488 ± 2.261	0.598
Δ_Stdev	-0.001 ± 0.094	0.090 ± 0.096	**0.008 ****
2nd order features			
Δ_ClusterProminence_1	496.6 ± 5280.8	6004 ± 10585	0.070
Δ_ClusterProminence_2	-366.7 ± 5075.7	5096 ± 8913	0.051
Δ_ClusterProminence_5	-816.9 ± 4070.6	3654 ± 6244	**0.038 ***
Δ_ClusterShade_1	-154.97 ± 376.63	-86.6 ± 480.8	0.666
Δ_ClusterShade_2	-134.0 ± 347.6	-74.3 ± 410.9	0.631
Δ_ClusterShade_5	-97.1 ± 278.1	-56.3 ± 294.4	0.648
Δ_Energy_1	0.066 ± 0.134	-0.06 ± 0.104	**0.015 ***
Δ_Energy_2	0.065 ± 0.126	-0.053 ± 0.096	**0.014 ***
Δ_Energy_5	0.059 ± 0.118	-0.047 ± 0.089	**0.010 ***
Δ_Entropy_1	-0.392 ± 1.474	0.945 ± 1.282	**0.005 ****
Δ_Entropy_2	-0.461 ± 1.559	0.984 ± 1.355	**0.004 ****
Δ_Entropy_5	-0.545 ± 1.596	0.988 ± 1.414	**0.003 ****
Δ_Homogeneity_1	0.001 ± 0.123	-0.083 ± 0.113	**0.037 ****
Δ_Homogeneity_2	0.014 ± 0.144	-0.093 ± 0.130	**0.022 ***
Δ_Homogeneity_5	0.030 ± 0.160	-0.100 ± 0.146	**0.014 ****
Δ_Inertia_1	0.916 ± 1.431	1.843 ± 2.316	0.256
Δ_Inertia_2	1.587 ± 3.089	3.799 ± 4.640	0.137
Δ_Inertia_5	1.579 ± 5.708	6.992 ± 8.608	0.056
Shape features			
Δ_Pixels_number	6695 ± 41169	2747 ± 80507	0.078
Δ_Elongation	-0.081 ± 0.181	0.064 ± 0.191	**0.019 ***
Δ_Equivalent_spherical_radius	-2.065 ± 7.797	0.328 ± 10.323	0.266
Δ_Roundness	-0.025 ± 0.051	-0.015 ± 0.083	0.714
Δ_Perimeter	-20.517 ± 12737	1303 ± 16842	0.810
Δ_Physical_size	-19716 ± 209957	31197 ± 391970	0.355
Δ_Flatness	0.200 ± 0.281	0.029 ± 0.249	**0.019 ***
Δ_Perimeter_on_border_ratio	-0.003 ± 0.006	0.002 ± 0.013	0.183
Δ_Feret_diameter	-3.033 ± 23.707	5.45 ± 33.16	0.202

NOTE. Data are given as mean and standard deviation.
*: $p<0.05$, **: $p<0.005$

Table 3-4 : Correlation matrix of the significant texture and shape features at univariate analysis

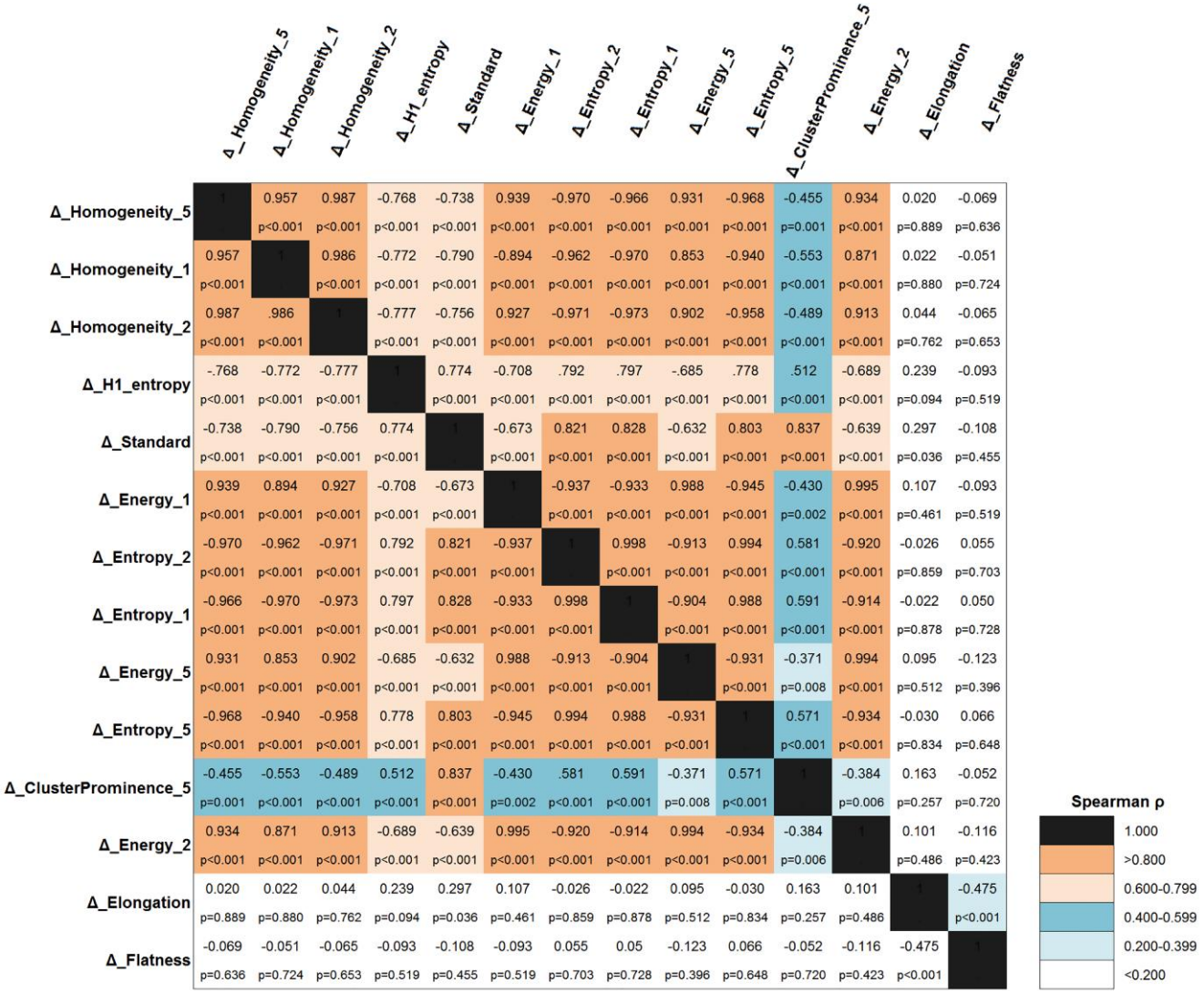
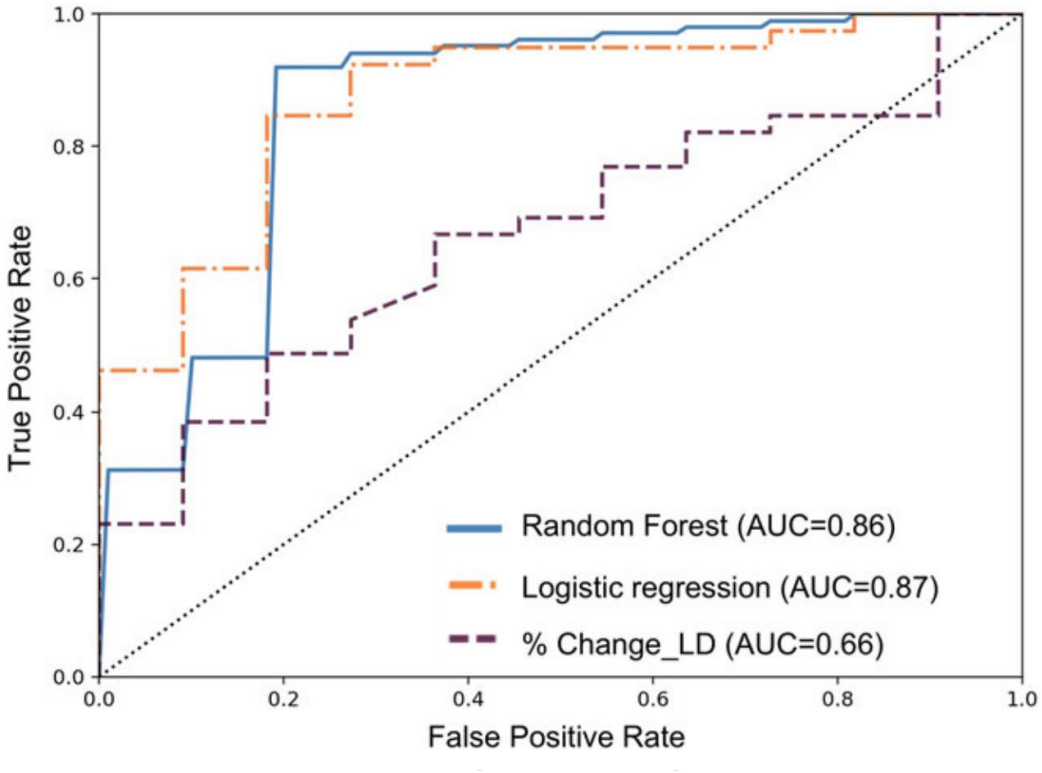


Table 3-5. Diagnostic performance of the classifiers on the 3 selected features for training and test cohorts (respectively cross-validation and final test steps).

Classifiers	Accuracy	AUROC	Sensitivity	Specificity	PPV	NPV	Train Score
Training Cohort							
Random Forest	88.1 ± 2.3 (87.6-88.5)	0.86 ± 0.01 (0.86-0.87)	94.1 ± 1.7 (93.8-94.4)	66.3 ± 7.9 (64.7-67.8)	90.9 ± 2.0 (90.5-91.2)	76.2 ± 6.0 (75.0-77.3)	0.98 ± 0.00
K-nearest neighbors	80.5 ± 1.2 (80.2-80.7)	0.81 ± 0.02 (0.81-0.82)	97.3 ± 0.7 (97.1-97.4)	20.5 ± 4.9 (19.5-21.4)	81.3 ± 0.9 (81.1-81.5)	66.9 ± 9.3 (65.1-68.8)	1.00 ± 0.00
Support Vector Machines	75.2 ± 3.4 (74.5-75.8)	0.67 ± 0.06 (0.66-0.68)	85.4 ± 3.6 (84.7-86.2)	37.9 ± 7.0 (36.5-39.3)	83.0 ± 1.8 (82.6-83.3)	42.8 ± 8.0 (41.3-44.4)	0.96 ± 0.0
Logistic Regression	85.8 ± 0.8 (85.6-86.0)	0.87 ± 0.01 (0.86-0.87)	94.8 ± 0.3 (94.8-94.9)	53.2 ± 3.5 (52.5-53.9)	87.8 ± 0.8 (87.6-87.9)	74.4 ± 1.6 (74.1-74.7)	0.87 ± 0.00
RECIST 1.1[§]	76.0	0.66	57.0	90.9	66.7	21.3	_
Test Cohort							
Random Forest	74.6 ± 4.5 (73.7-75.5)	0.63 ± 0.03 (0.62-0.63)	98.0 ± 4.0 (97.2-98.8)	27.8 ± 9.8 (25.9-29.7)	73.1 ± 3.0 (72.6-73.7)	90.8 ± 18.7 (87.2-94.5)	0.98 ± 0.01
K-nearest neighbours	66.7	0.53	100.0	0.0	66.7	0.0	1.00
Support Vector Machines	53.3	0.52	80.0	0.0	61.5	0.0	0.98
Logistic Regression	66.7	0.46	90.0	20.0	69.2	50.0	0.86
RECIST 1.1[§]	73.3	0.72	90.0	40.0	75.0	66.6	_

AUROC indicates area under the ROC curve, PPV indicates positive predictive value, NPV indicates negative predictive value.
Accuracy, sensitivity, specificity, PPV and NPV after cross- validation are given in percentage with standard deviation and 95% confidence interval in parentheses.
Regarding the test cohort, random forests were generated using 100 random seeds, which provided a set of 100 values for each statistics. Standard deviations are those of the 100 values.
[§] Statistics are given for RECIST 1.1 in an objective response setting, that is to say 'complete response or partial response' vs. 'stable disease or progressive disease'. AUROC corresponded to AUROC of relative change in longest diameter, on which RECIST 1.1 status is based.

Figure 3-2 ROC curves of random forest model, logistic regression model and relative change in longest diameter from baseline to post-2 cycles of chemotherapy (% Change_LD) at cross-validation. Random forest and logistic regression were based on the optimal selection of features (Change in surrounding edema, change in histogram-entropy, change in Elongation). For each classifier, the individual scores of each sample from all folds are sorted together into a single ROC curve and then averaged across the 100 repetitions.



Since the best compromise was obtained with the RF classifier, we investigated the impact of adding features in the RF model (Figure 3-3). Accuracy and AUROC were not improved in the training cohort and they decreased in the test cohort. In the training cohort, specificity was at its highest with 3 features while sensitivity remained constant. In the test cohort, higher sensitivity and specificity were obtained with 3 features. Figure 3-4 illustrates the added value of the final RF algorithm for two cases with a stable disease according to RECIST 1.1, one being a poor-HR, the other a good-HR. Results with other popular features selection methods and RF classifier are given in Annexe 4 - Supplemental Data 3.

Figure 3-3: Accuracy, AUROC, sensitivity and specificity of the random forest algorithm as functions of the numbers of features included in the model. These statistic metrics were calculated in the training cohort (**a**) and the test cohort (**b**). Features were added in the ascending order of their p-value (descending order of statistical significance) as listed in Table 1 and 2. The grey dashed vertical line emphasizes the initially selected 3-features model (changes in edema, histogram_entropy and elongation from baseline to post-2 cycles evaluation).

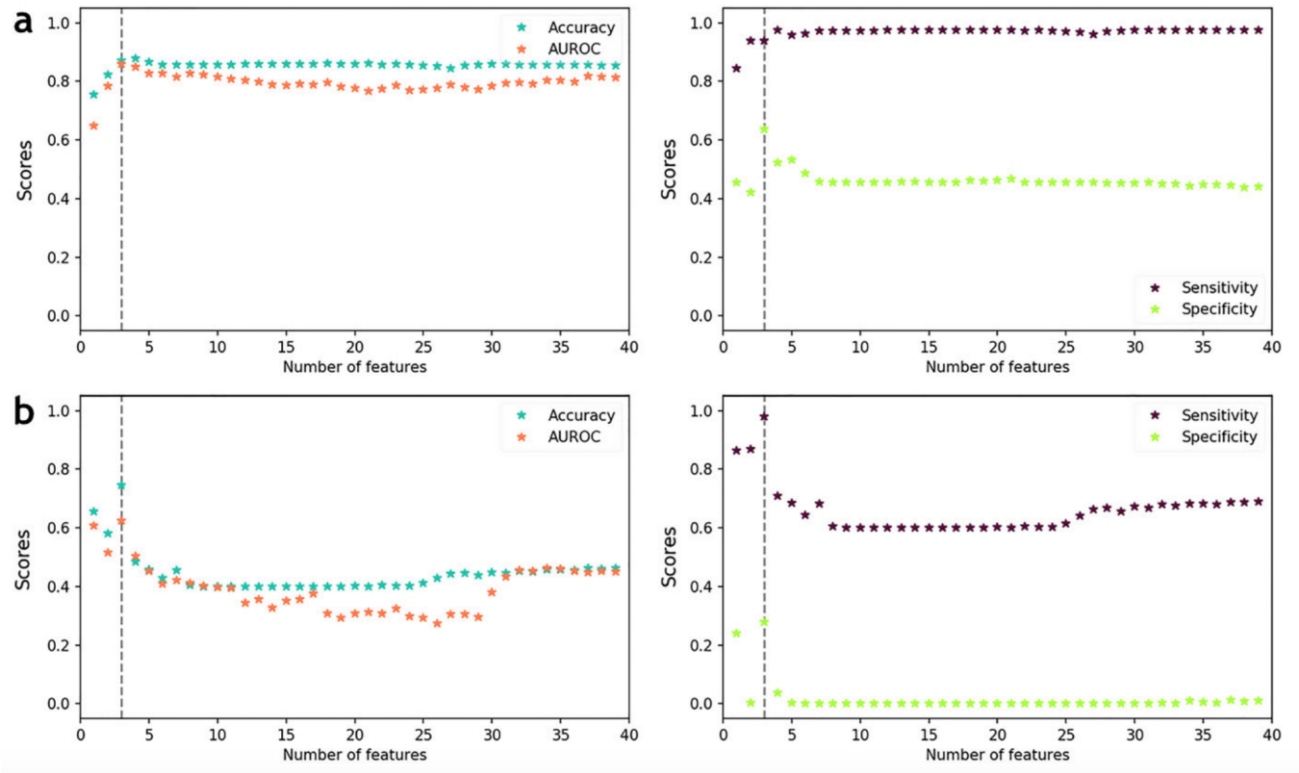
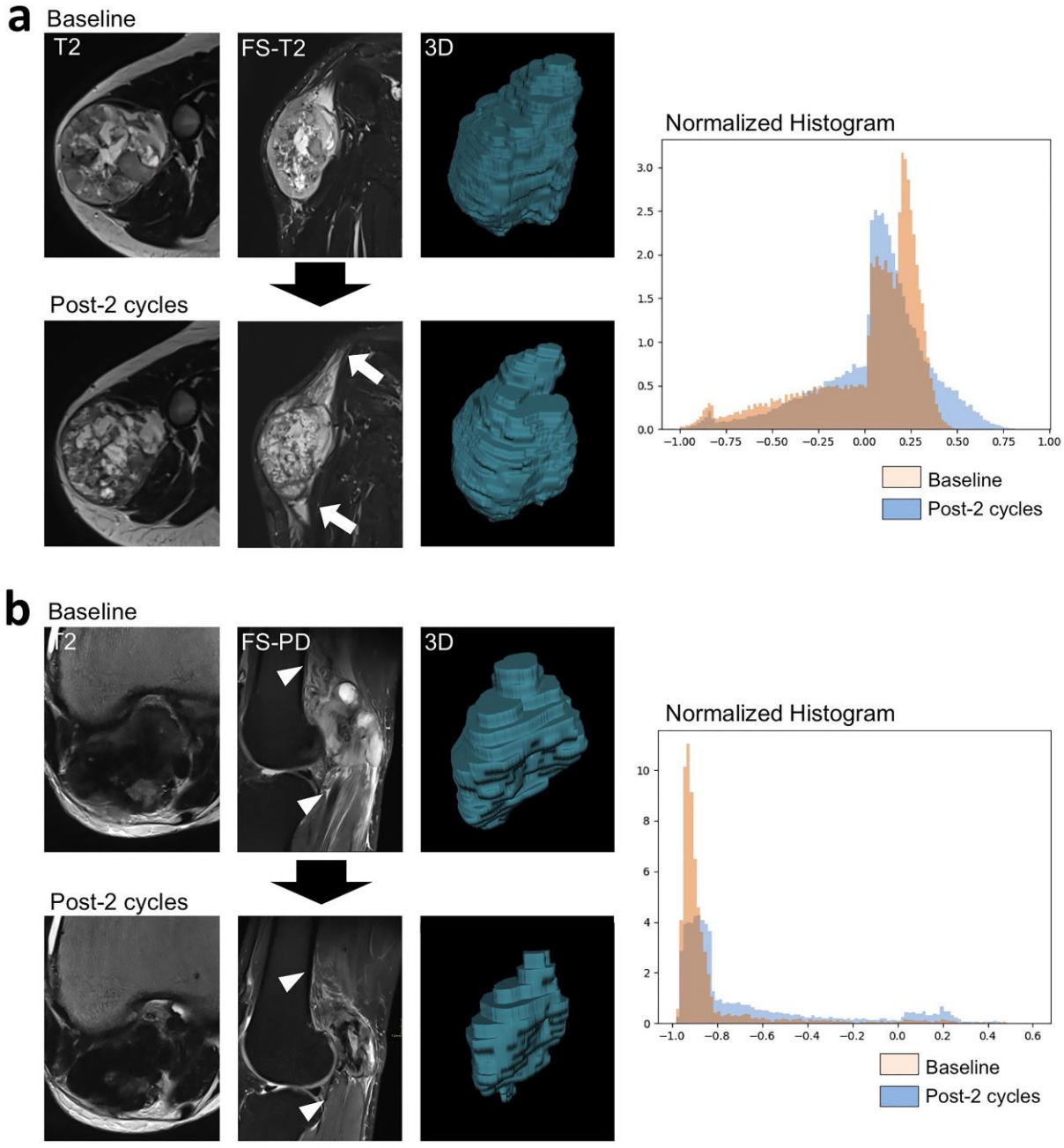


Figure 3-4: Added value of final random forest (RF) model for early response prediction. (**a**) 76 years-old male presented with a deep-seated grade III pleomorphic rhabdomyosarcoma of the shoulder. After 2 cycles of chemotherapy, the tumor was stable according to RECIST 1.1 criteria, but it demonstrated an increase of its surrounding edema (white arrows), stability of its shape and stable histogram entropy. Hence, the final RF model predicted a poor histological response that was confirmed on surgical specimen (70% residual viable cells). (**b**) 50 years-old male presented with a deep-seated grade III undifferentiated pleomorphic sarcoma of the popliteal region. After 2 cycles of chemotherapy, the tumor was stable according to RECIST 1.1 criteria. Surrounding edema markedly decreased (white arrow heads) with a retraction of its shape on 3D reconstruction and a decreased entropy on normalized histogram. The final RF model predicted a good response that was confirmed on surgical specimen (5% residual viable cells). T2: T2-weighted imaging, FS: fat-sat, PD: proton-density weighted-imaging



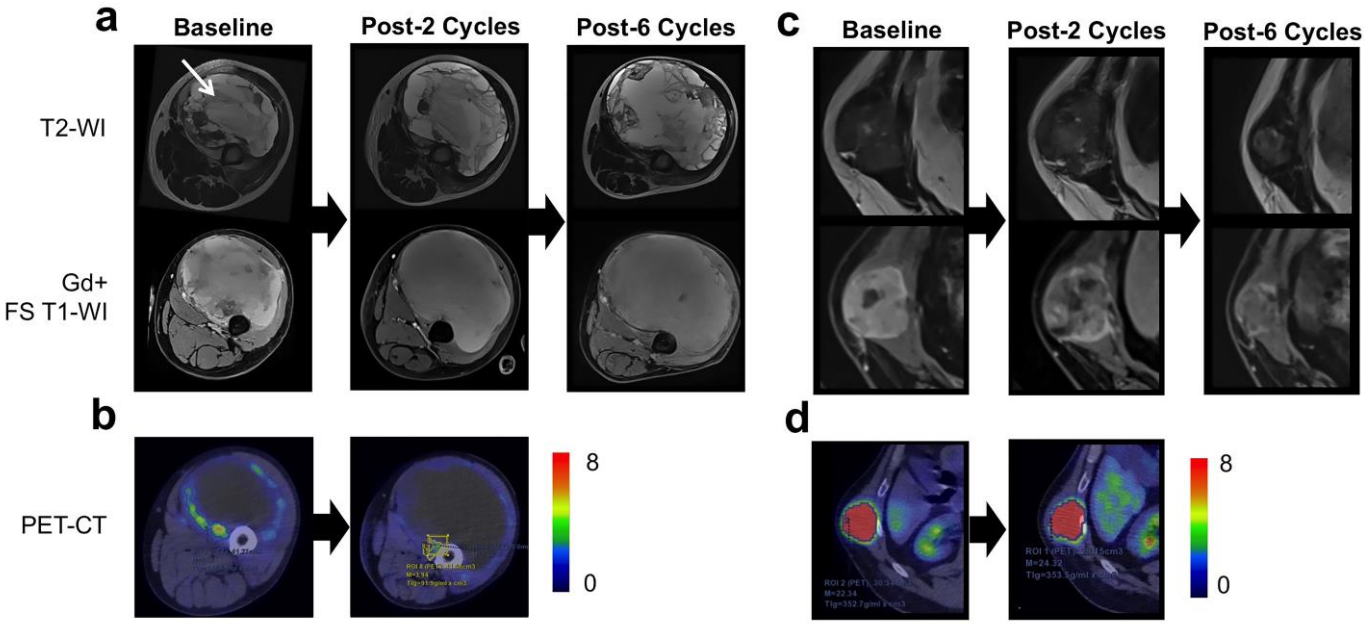
The retrospective analysis of the false positive predictions made by the RF model highlighted cases of massively necrotic-hemorrhagic tumors and late-responder profiles (Figure 3-5). Quantification of tumor heterogeneity was biased by heterogeneous large blood clots on baseline examination and their changes at early evaluation. Analysis based on other imaging modalities of 'late-responder' cases did not provide any clue to predict a good response after 2 cycles, whereas pre-surgical evaluation demonstrated extensive fibro-necrotic changes, strong decrease of DCE-MRI parameters and SUVmax.

DISCUSSION

In this study, we developed and evaluated radiomics models to predict the histological response of STS during NAC that were based on changes on T2-WI from baseline to early evaluation. Overall, our best model was obtained with RF classifiers on 3 relevant features from analysis of STS shape, heterogeneity and surrounding tissue. It performed better than RECIST 1.1 with an accuracy of 88.1% and AUROC of 0.86 at cross-validation and had the highest scores on the independent test cohort. However, those last results highlighted outliers requiring additional characterization.

Performances of our best predictive model were comparable or higher than those found for other imaging biomarkers in literature although we should be careful in making comparisons since different cut-offs may have been used to define a good histological response, different chemotherapy regimens may have been prescribed and imaging may have been performed at different stages in time. Stacchiotti et *al.* (2009) investigated CHOI criteria to predict a pathological very good response defined as <10% viable cells on surgical specimen in a series of 37 patients. They obtained an accuracy of 74.1% (14/22 CHOI partial responders being true good-HR and 6/6 CHOI non-responder being true poor-HR). In another studies, a decrease in contrast-enhancement of -30.5% between two MRI with optimized acquisition delay after contrast-agent injection provided an accuracy of 82.8% (Crombé et al, 2019c). On a retrospective series of 23 patients, multiparametric assessment combining qualitative evaluation of diffusion imaging and DCE-MRI provided a best AUROC of 0.833 (Soldatos et al, 2016). At early evaluation with ^{18}F-FDG-PET/CT, a decrease >35% of

Figure 3-5: Outliers patients who were misclassified as poor responders by the model. **(a)** Example of massively necrotic tumor at baseline: 52 years-old male presented with deep-seated grade III undifferentiated pleomorphic sarcoma of the left thigh. Blood clots and fibrinous septa were mixed with necrosis (white arrow), only small buds of tumor were seen against tumor wall. Therefore, changes in tumor heterogeneity were mostly due to change in structure and signal of the necrotic-hemorrhagic compartment. (**b**) This patient benefited from a ^{18}F-FDG-PET-CT at baseline and after two cycles showing a strong decrease of SUVmax (8.16 to 3.94, -51.7%) suggestive of chemotherapy efficacy. (**c**) Example of a 'late responder' profile: 66 years-old male presented with a deep-seated grade III pleomorphic rhabdomyosarcoma of the abdominal wall. No obvious change was seen by visual assessment at early evaluation. (**d**) ^{18}F-FDG-PET-CT demonstrated a slight paradoxical increase of SUVmax (22.34 to 24.32) although the patient was a good histological responder after 4 additional cycles of chemotherapy. T2-WI: T2 weighted imaging; Gd+ FS T1-WI: Fat-Sat T1 weighted imaging after Gadolinium-chelates injection.



SUVmax provided an AUROC of 0.83 in a prospective study of 50 patients (Benz et al, 2008; Benz et al, 2009).

Association between decrease in edema and good response did not surprise us. Surrounding edema is associated with high-grade STS and satellite tumor cells (Hanna et al, 1991; White et al, 2005; Zhao et al, 2014). NAC efficacy should logically go with reduced satellite tumor cells and thus a decrease of signal anomalies surrounding STS on MRI. A decrease in tumor cellularity turning into fibro-necrotic tissue could explain a tumor softening leading to changes in shape, towards retraction of its borders. Finally, these fibro-necrotic processes lead to a larger range of signal intensity values within tumors, that is to say a flattening of the SI histogram responsible for change in its entropy.

The performances of our prediction models were lower in the test set. This was explained by the lower number of patients compared with the training set and by the presence of 3 outliers in the test set, without equivalent in the training set. These 3 outliers were systematically ill classified by the models. A careful retrospective analysis of these tumors to identify ‘late-responder’ and ‘massively necrotic STS’ profiles. These last ones are difficult to image and their morphological changes during treatment can be complicated to interpret. Evaluation with RECIST 1.1 is biased as it mostly measures the necrosis and not the changes in viable tumor component. DCE-MRI and diffusion imaging are challenging because viable tissue generally consists in small buds attached to the tumor wall within a large hemorrhagic mass. In our case, ^{18}F-FDG-PET/CT correctly predicted a good response according to the previously published cut-off of a 35-38% decrease in SUVmax (Benz et al, 2009). These two observations from the test cohort provide insights into next features to add to the future predictive models. Partitioning dataset in independent training and test datasets enabled to have a larger view of the response patterns of STS, and to consider additional imaging features from other advanced imaging modalities for improvement of the future models.

Interestingly, our best models relied on a limited set of features from non-contrast enhanced sequences. Corino et al. also found that only 3 features from diffusion imaging provided the highest accuracy to predict histological grade of STS (Corino et al, 2017). The best model to predict lung metastatic relapse of STS according to Vallières et al. relied on 4 texture features (Vallières et al, 2015). In their studies, adding any other feature to the model did not improve prediction. In a context of

controversy about long-term effects of Gadolinium-chelates contrast agents, an imaging work-flow for response evaluation may be considered in which known outliers of the model or patients with an intermediate probability of response could be assessed in a second step with contrast agent injection and advanced imaging modalities such as DCE-MRI, DWI and ^{18}F-FDG-PET/CT.

Our study has limits. First, this is a retrospective study with a relatively small number of patients. Nevertheless, our series is one of the largest regarding STS and MRI, with uniformly treated patients with the chemotherapy of reference. No epidemiological data was added into the model because none was associated with the tumor response at univariate analysis. Indeed, the population study only included patients who shared all the epidemiological features associated with worse prognosis, namely high-grade, deep- or deep and superficial STS with LD above 5cm (Coindre et al, 1996). Beside, if modest, the cohort was significant enough to put a few data aside to form an independent validation set and control our results.

Second, imaging protocols were not designed for radiomics studies. 2D TSE T2-WI was used for features extraction because (i) it was the most commonly acquired sequence, (ii) it provided a large range of morphological changes during treatment, (iii) there was no change in the acquisition parameters during the study period. T2-WI can capture fibrotic and necrotic processes (decreased and increased T2 SI, respectively). T2-WI has already demonstrated good results in textural approaches applied to other tumor types (Dong et al, 2018; Henderson et al, 2017; Hocquelet et al, 2018; Nketiah et al, 2017; Gnep et al, 2017). Conversely, post-contrast T1-WI sequences showed heterogeneous acquisition protocols in our series: some were 3D gradient recalled echo imaging and others 2D TSE, different fat suppression techniques were used (Dixon method, fat-sat, short TI inversion recovery, subtraction with pre-contrast T1-WI), as well as different contrast agents. The acquisition delay after contrast agent injection was not standardized although it may have a significant effect on changes in tumor heterogeneity quantified on CE-T1-WI.

Third, we were not able to directly compare the performances of our model with those of modified Choi criteria for MRI. These Choi criteria were defined on subtracted CE-T1-WI (i.e. subtraction of CE-T1-WI and TE-WI before injection) in order to avoid confusion between high signal intensity on T1-WI due to hemorrhagic alteration during NAC and high signal intensity on T1-WI due to enhancement of viable tumor

component. Unfortunately, most patients of our study did not undergo the same T1-WI sequence before and after contrast-agent injection. Moreover, a previous study has demonstrated that the accuracy of modified Choi criteria depended on the acquisition delay after contrast-agent injection, ranging from 40% to 72.1% (Crombé at al, 2019c). Herein, this acquisition delay was not controlled.

Changes in surrounding edema helped predict response but its assessment could only be qualitative because of non-standardized sequence for its evaluation. Future studies should include automatic and quantitative assessment of edema and its changes, since it was one of the best predictor for the response in the current study. Adding another imaging modalities would have markedly decreased the population study and we made the decision to privilege one informative sequence and an acceptable population study. Nevertheless, this point stresses the urge for a standardized MRI protocol for STS.

Moreover, post-processing included voxel size standardization - with an acceptable voxel size to preserve the global shape of tumor - and signal intensity normalization. The aim was to limit the inherent bias due to MR acquisition at different stages on different MR-systems and to improve the reliability and the reproducibility of the extracted features. 3D Segmentation was manually performed, slice-by-slice. Automatic and semi-automatic methods were tried before the study with disappointing results as compared to those obtained by the expert radiologist from a sarcoma reference center.

We decided to limit the number of extracted textural features, despite the fact that several others could have been calculated from fractal analysis, wavelet, and other matrices. Therefore, we limited the risk of finding relevant features by chance and facilitated the understanding of our results. Those we calculated are widely used and can be easily found in open libraries. Due to the relatively small population study, we did not apply deep-learning purposely and focused on time-tested classifiers. Finally, one could question the outcome. The histological response is routinely used as an intermediate evaluator reflecting immediate efficacy of NAC and patient prognosis (Cousin et al, 2017). Nonetheless, it is a semi-quantitative assessment, with possible subjectivity and sample bias. Ultimately, our goal is to build predictive models for survival, but only 41 patients in this series have a follow-up of more than 2 years and 26 of more than 5 years.

To conclude, our preliminary results indicate that T2-based delta-radiomics approach applied to STS in the neoadjuvant setting is feasible, provides valuable information to predict response after only 2 cycles and improves evaluation compared to RECIST 1.1. Optimization of the model is still needed with the study of larger cohorts and inclusion of other categories of features, other imaging modalities and other '-omics' criteria.

*

* *

3.3. Limites et ouvertures

Cette étude présente aussi plusieurs limites méthodologiques l'empêchant de dépasser le stade de la preuve de concept. En reprenant les items du RQS, nous n'avions pas répété les segmentations pour sélectionner les indices radiomics les plus robustes, ni réalisé de *test-retest*, ni d'analyse de fantômes. Si les IRM de révaluation ont été réalisées sur l'institut Bergonié, les IRM basline ont aussi été réalisées sur d'autres centres. Bien que nous ayons corrélé nos résultats à une variable biologique, nous n'avons pas évalué la valeur pronostique des indices radiomics (avec et sans prendre en considération la réponse histologique) en raison du manque de recul suffisant pour près d'un tiers de la cohorte. Si nous avons comparé les performances de nos meilleurs modèles par rapport à RECIST v1.1, notre population d'étude ne nous a pas permis d'ajouter les critères d'évaluation de la réponse type mRECIST, EASL, Choi et ses variants ainsi que PERCIST. Enfin, la nature rétrospective et la petite taille de la cohorte de validation limitent la généralisation de nos résultats. Nous devons aussi souligner les résultats moyens des random forests dans la cohorte de validation devant faire suspecter de l'overfitting. Le RQS de cette étude est donc de 13.

Néanmoins, il ressort là encore qu'enrichir les modèles radiomics avec des variables radiologiques permet d'améliorer leurs performances et que les approches delta-radiomics méritent d'être approfondies dans le contexte des STM en situation néo-adjuvante. Les essais de recherche de biomarqueurs NEOSARCOMICS/ CIRSARC (NCT02789384) en cours d'inclusion à l'institut Bergonié et dans six autres centres de lutte contre le cancer français pourront permettre de mieux comparer les divers critères de réponse dont ceux basés sur le ^{18}F-FDG-TEP/CT, les patients ayant une IRM baseline (multi-paramètrique sur Bergonié), tous les 2 cycles et pré-chirurgicale, ainsi qu'un TEP/CT baseline et à 2 cures (injecté sur l'institut Bergonié).

4. AMELIORER LES PREDICTIONS EN OPTIMISANT LES PROCESSUS DE POST-TRAITEMENT

4.1. En IRM structurale

4.1.1. Introduction

Dans les études précédentes, basées sur des séquences morphologiques, nous avons utilisé une même technique de normalisation des intensités de signal passant par un *histogram-matchin*g avec une acquisition de référence. D'après la méta-analyse de Park et *al.* (2020a), l'IRM est la première modalité d'imagerie des approches radiomics. D'autres approches détaillées dans la partie 1 auraient pu être employées et auraient pu modifier les valeurs des indices radiomics et donc nos prédictions. Pour tester cette hypothèse, nous avons préparé les indices radiomics sur la séquence T2 de l'IRM initiale d'une cohorte de 70 patients atteints de STM localement avancés selon 5 méthodes répandues de normalisation des intensités de signal. Nous avons ensuite étudié 3 aspects de l'influence des normalisations: (i) à l'échelle des indices en recherchant quels indices radiomics étaient les plus influencés, (ii) sur les clusters identifiés par une approche non supervisées, et (iii) sur les performances de modèles prédictifs de la survie à 2 ans (après division de la cohorte totale en cohorte d'entrainement et de validation)

Le choix de la séquence T2 est justifié par les travaux antérieurs de notre groupe où nous avions montré que l'hétérogénéité (évaluée qualitativement) sur la séquence T2 de l'IRM initial était un marqueur pronostic du grade, de la MFS et de l'OS des patients atteints de STM (Crombé et al, 2019a).

4.1.2. Article 3 *(soumis en mars 2020)*

*

TITLE

Body-MRI Signal Intensity Harmonization Techniques Influence Radiomics Features and Can Enhance Radiomics-based Predictions in Sarcoma Patients

ABSTRACT

Objectives: Signal intensity (SI) harmonization techniques are needed to homogenize multicentric MRIs but may introduce bias in radiomics features (RFs). Heterogeneity of soft-tissue sarcoma (STS) on T2-weighted imaging (-WI) correlates with metastatic-relapse-free survival (MFS). Our aim was to investigate the influence of harmonization techniques on RFs and radiomics-based unsupervised and supervised predictive models.

Methods: Seventy patients with locally-advanced STS were included (median age: 58, study period: June 2006-November 2016). After voxel-size resampling and bias-field correction, pre-treatment spin echo T2-WI were independently post-processed per 5 methods: standardization per adipose tissue mean SI (*fat-standardization*), *basic-normalization*, histogram matching (HM) with the histogram of a randomly-chosen patient (*HM-1-patient*), HM with the average normalized histogram of the whole population (*HM-All-patients*), ComBat harmonization of RFs obtained with HM-All-patients (*HM-All-patients+ComBat*). Forty-five texture RFs were extracted from tumor volumes. Influence of harmonization techniques on RFs was assessed with repeated-measures ANOVAs. Prognostic values of hierarchical clustering results based on RFs obtained with the 5 techniques were evaluated with hazard ratios (HR) and concordance-indices. Various RFs-based supervised classifiers were trained on 50 patients to predict survival at 2years and evaluated on 20 different patients with AUC.

Results: All RFs were significantly influenced by harmonization techniques. Clusters obtained with *basic-normalization*, *HM-All-patients*, *HM-All-patients+ComBat* RFs correlated with MFS in multivariate Cox models (p=0.004, 0.02 and 0.007,

respectively). In the testing cohort, AUC of supervised models ranged from 0.80 (with *HM-1-patient* and *fat-standardization*) to 0.67 (with *basic-normalization*).
Conclusion: Radiomics postprocessing pipelines should include and evaluate various harmonization techniques because they influence models performances.

INTRODUCTION

Radiomics has now become a large field of research referring to the extraction and mining of several quantitative variables, named radiomics features (RFs), which extensively screen the shape and texture of objects of interests within medical images of any modality. In oncologic imaging, these RFs are integrated in predictive models based on machine-learning classifiers in order to answer key questions such as the discrimination between benign and malignant lesions, identification of molecular subgroups, prediction of patients' outcome, or building of radiogenomics signatures (Limkin et al, 2017; Lambin et al, 2017; Gillies et al, 2016). Regarding sarcomas, radiomics have improved predictions of grading, prognosis and response to chemotherapy/radiotherapy, based on CT-scans, structural MRI alone or combined with positron emission tomography, dynamic-contrast enhanced or diffusion MRI (Vallières et al, 2015; Peeken et al, 2019a; Peeken et al, 2019b; Crombé et al, 2019d; Spraker et al, 2019; Corino et al, 2018).

Though one aim of radiomics is to provide an objective assessment of tumor phenotype, several studies have shown the influence of pre- and postprocessing factors on the value of RFs (Berenguer et al, 2018; Crombé et al, 2019e; Bogowicz et al, 2016; Buch et al, 2018; Caramella et al, 2018; Ford et al, 2018). These findings question the validity of inter-site radiomics studies and show the need for a pipeline to harmonize medical imaging. This issue is even more prominent with MRI because of the absence of standard intensity scale. Thus, signal intensities (SIs) lack of fixed meaning, even on the same MR-scanner for a same sequence. If gray-levels discretization, voxel-size standardization and bias-field correction are easy to implement in postprocessing, there is a lack of consensus on the techniques for harmonizing intensities in MRI datasets. Several techniques for intensity harmonization have been proposed in the neuroimaging literature to enable robust analysis of structural and diffusion MRIs across different radiological centers and longitudinally. The available intensity harmonization methods regarding body-MRI

are scarcer. The most frequently encountered are global scaling (*e.g.* where SIs values are centered by removing the mean and scaled to unit variance, or transformed to range between 0 and 1), ratio with SIs of a same tissue that is not affected by the disease (for instance adipose tissue or muscle in musculoskeletal imaging), or histogram matching (HM, where the intensity histograms are transformed to match a reference intensity histogram) (Wang et al, 1998; Nyul et Udupa, 1999; Nyul et al, 2000). In addition, Orlhac et al. have recently shown that ComBat harmonization method, which was initially described in genomics to remove batch effect, could correct nonbiological differences related to the type of scanners in a phantom and patient study (Orlhac et al, 2019). Though the authors focused on CT-scanner, ComBat may help reduce unwanted variations in MRI-based radiomics datasets as well.

Thus, our aim was to investigate how the intensity harmonization methods could influence MRI-based radiomics analyses in a uniformly-treated cohort of soft-tissue sarcomas (STS) patients for whom heterogeneity on initial T2-weighted-imaging (-WI) has been correlated with metastatic-relapse free survival (MFS) (Crombé et al, 2019a). To do so, a 3-step analysis reflecting the different aspects of radiomics approaches was developed to investigate the influence of intensity harmonization techniques on: (i) the RFs values; (ii) the prognostic value of RFs-based unsupervised classifications; and (iii) the performances of supervised classifications for predicting early metastatic relapses.

MATERIALS AND METHODS

Study population

This single-center study was IRB-approved. The need for written informed consent was waived because of its retrospective nature.

Patients were consecutively recruited as they fulfilled the following inclusion criteria: newly-diagnosed, non-metastatic (according to chest CT-scan), histologically-proven high-grade STS of trunk wall or extremities (n=163), treated with 4-6 cycles of anthracycline-based neoadjuvant chemotherapy and curative surgery at our sarcoma reference center from June 2006 to November 2016 (n=133), available baseline MRI (n=95) with axial spin-echo T2-WI without artefacts (n=72), and available clinical and radiological follow-ups for at least 2 years after the surgery (n=70). There was a

subject overlap of 53 patients with a prior study (Crombé et al, 2019d). Follow-ups consisted in a clinical examination and chest radiograph every 3 months for 2 years, every 6 months for 5 years and annually until 10 years after surgery, which were complemented by chest CT-scans and MRIs in case of doubtful findings. All relapses were histopathologically confirmed.

MRI acquisition

The baseline MRI examinations were acquired on 1.5-Tesla MR-systems (Philips Signa [17/70, 24.3%], Siemens MAGNETOM Aera [41/70, 58.5%], General Electrics Healthcare Optima Jem MR450w [12/70, 17.1%]) with adjustment of coils, field-of-view and matrix depending on tumor size, location and depth. Regarding T2-WI, the range of repetition and echo times were 2400-4500msec and 70-130 msec, respectively. Slice thickness ranged from 3 to 5 mm.

MRI postprocessing (Figure 4.1-1)

After anonymizing MRIs, the postprocessing was performed with R (version 3.5.3) by using the "oro.nifti", "ANTsR" and "extranstr" packages (Muschelli et al, 2019).

First, T2-WIs were converted to nifti format. Voxel size resampling (with b-spline interpolator) and N4 bias field correction were applied to obtain a common spatial resolution of 1 x 1 x 4 mm^3 and to correct non-uniform intensities (Tustinson et al, 2010).

Second, a senior radiologist (A.C., with 4 years of experience in sarcoma imaging) manually segmented the whole tumor volume, slice-by-slice, using LIFEx freeware (version 5.10) (Nioche et al, 2018). The volumes-of-interests were all validated by a second senior radiologist (M.K., with 28 years of experience in sarcoma imaging).

Third, 4 distinct postprocessing methods were applied in parallel to the whole imaging dataset in order to harmonize the SIs of the T2-WI, providing 4 additional harmonized datasets, i.e.:

(1) *Fat-standardization,* which consisted in dividing all the SIs of a given T2-WI by the mean SI of fat on that T2-WI (*i.e.* $SI(x,y,z)_{Fat\text{-}standardization} = SI(x,y,z) / mean(SI(fat))$). To do so, the radiologist segmented a volume of at least 10 cm^3 of normally-appearing fat on each T2-WI in order to extract the mean fat SI per patient.

(2) *Basic-normalization*, which consisted in normalizing the SIs of a T2-WI according to the minimum and maximum of all voxels included in this T2-WI sequence (*i.e.* $SI(x,y,z)_{Basic\text{-}normalization} = [SI(x,y,z) – min(SIs)] / [max(SIs) – min(SIs)]$).

(3) *HM-1-patient*, which consisted in performing a non-linear matching of the intensity histogram of each T2-WI sequence with the intensity histogram of a same normalized T2-WI from the same randomly chosen patient in the MRI dataset (https://github.com/abdhigithub/hatch).

(4) *HM-all-patients*, which consisted in performing a non-linear matching of the intensity histogram of each T2-WI sequence with the average intensity histogram of the whole normalized MRI dataset.

HM-1-patient and *HM-All-patients* were trained on 100 histogram landmarks. Figure 4.1-2 shows the superimposed intensities distributions of the 70 patients depending on the harmonization techniques. Please note that these techniques were applied to the whole sequence and not on a volume of interest.

Radiomics features extraction

The tumor volumes were then propagated on the 4 new datasets enabling the extraction of 4 sets of 45 3D-RFs using LIFEx (Nioche et al, 2018). SIs were discretized into 128 fixed bins with a bin size of 0.0078125 (=[max(dataset)–min(dataset)]/128=[1-0]/128) for *Basic-normalization*, *HM-1-patient* and *HM-All-patients*; and with a bin size of 0.0509375 (=[5.95–(-0.57)]/128) for *Fat-standardization*. Thirteen histogram-based and 32 second-order texture features from grey-level co-occurrence matrix (GLCM, n=7 - with a 1-voxel distance to neighbors), grey-level run length matrix (GLRLM, n=11), neighborhood grey-level different matrix (NGLDM, n=3) and grey-level zone length matrix (GLZLM, n=11) were estimated (Annexe 5 - Supplemental Data 1).

Figure 4-1.1 : Study pipeline. Abbreviations: HM: histogram matching; RF: radiomics feature; WI: weighted imaging.

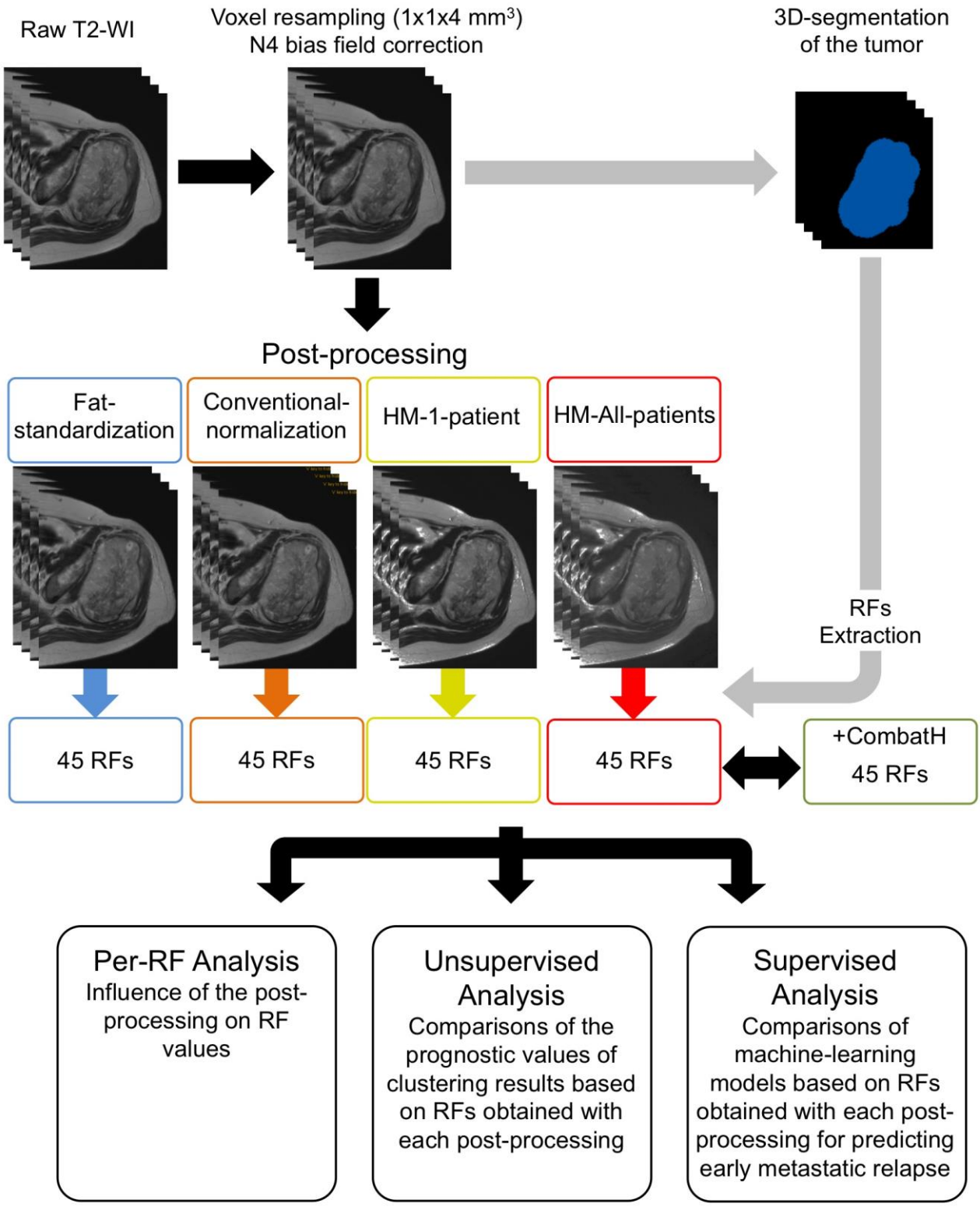
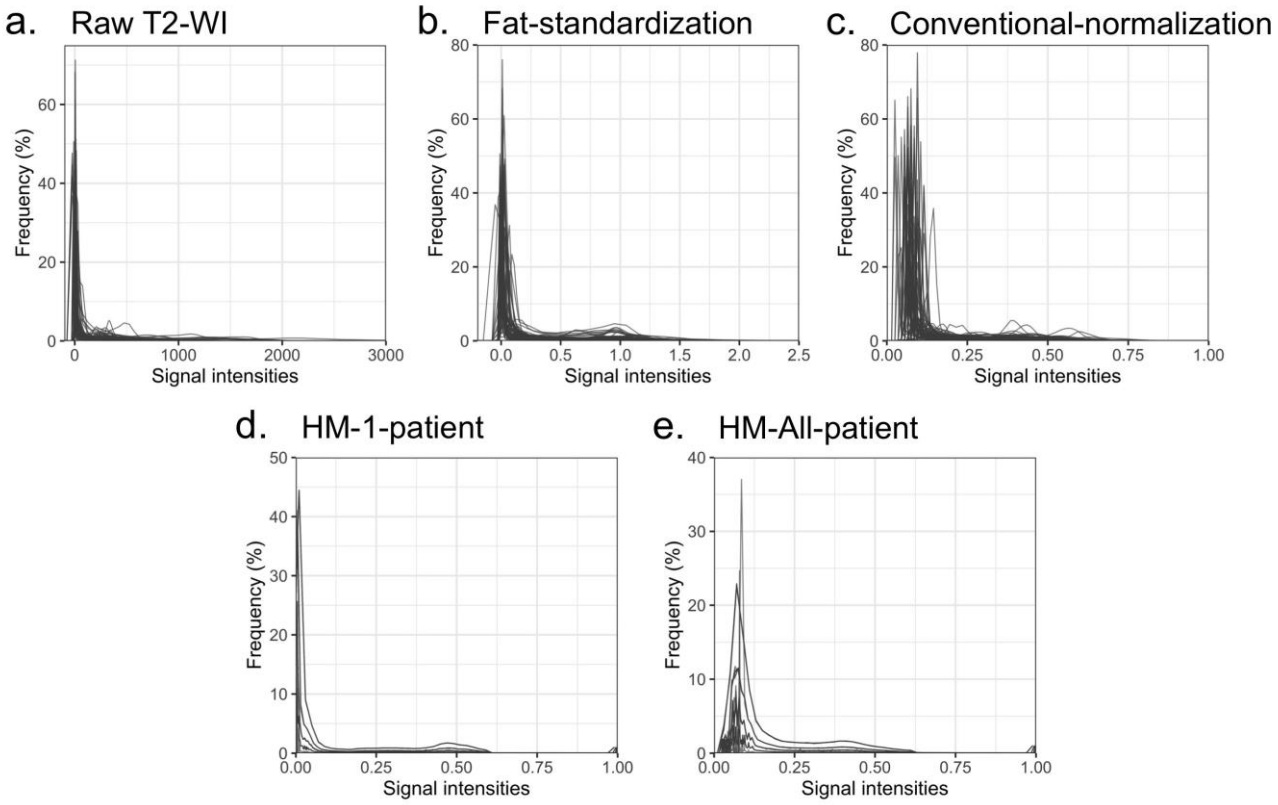


Figure 4.1-2. Distribution of the signal intensities in the whole population before applying intensity harmonization techniques (a); and after applying: *fat-standardization* method (b); *basic-normalization* method (c); histogram matching with a randomly chosen normalized histogram of a patient (*HM-1-patient*, d); and histogram matching with the average normalized histogram of the 70 included patients (*HM-All-patients*, e). Each line represents one patient.
*: Raw T2-WI were voxel-size resampled and N4 bias field corrected beforehand.



ComBat compensation

We applied the ComBat-Harmonization function in R (https://github.com/fortin1/ComBatHarmonization), with a non-parametric setting, to the 45 RFs extracted from the *HM-All-patients* dataset in order to remove unwanted noise due to technical between-MR-systems variations (Orlhac et al, 2019; Fortin et al, 2017). The tumor location and the MR-coil used were defined as covariables because they could influence the acquisition parameters of the T2-WI. Details regarding ComBat are given in Annexe 5 - Supplemental Data 2. The resulting RFs were labelled *HM-All-patients+ComBat.* Finally, 5 paired datasets of 45 RFs were obtained.

Statistical analysis

Statistical analysis was performed with R. All tests were two-tailed. A p-value<0.05 was deemed significant. A 3-steps approach was performed to evaluate the impact of harmonization techniques (Figure 4.1-1):

(1) *Per-RF analysis*: herein, RFs were all normalized in order to range from 0 to 1 and to facilitate comparisons. For each RF, the influence of the harmonization technique was evaluated with one-way repeated-measures ANOVA. Post-Hoc comparisons were assessed with Tukey test and Bonferroni corrections.

(2) *Unsupervised analysis:* A hierarchical clustering analysis with the Ward method was applied on each of the 5 subsets of RFs. RFs were centered and scaled by mean beforehand and the Euclidean distance between each pair of patients was computed. Visual inspection of the silhouette plot enabled to select 2 clusters of patients for each harmonization technique. The Baker's gamma coefficient between each pair of dendrograms was calculated with the "dendextend" package, as well as the Kappa index between each pair of clustering results, which enabled the quantification of their divergence depending on the harmonization technique (Galili et al, 2015). The associations between MFS and discovered clusters were assessed with Kaplan-Meier analysis and multivariable Cox models - after adjusting to the longest baseline diameter (< vs. ≥ 10 cm), performance status (0 vs. 1-2), histological type (undifferentiated sarcomas vs. other), number of chemotherapy cycles (4 vs. 5-6), chemotherapy type (anthracycline-ifosfamide vs doxorubicine), adjuvant radiotherapy, surgical margins (R0 vs. R1-R2) and histological response (goods vs. poor responder to chemotherapy with a cut-off of 10% viable cells on post-

chemotherapy surgical specimen). Prognostic performances of the multivariate models were evaluated and compared through concordance-index, which estimates the ability to correctly classify the subjects.

(3) *Supervised analysis*: A same supervised machine-learning approach was applied to the 5 datasets of RFs in order to identify patients with an occurrence of metastatic relapse 2years after treatment. This step was performed with the "caret" package (Kuhn et al, 2008). The population was randomly divided into one training cohort of 50 patients and one testing cohort of 20 patients with the same proportion of events. The training cohort was used to train different popular classifiers (classical logistic regression, penalized binomial logistic regression with combination of lasso and ridge regressions [glmnet], random forests and support vector machines) and to tune their hyperparameters (if needed) by using 10-fold cross validation, repeated 5 times. Details regarding the classifiers are given in Annexe 5 - Supplemental Data 3. The same partition of patients was used for the 5 datasets. The same clinical and pathological covariables as in the unsupervised analysis were included, in addition to the same 3 shape RFs (volume, compacity and sphericity – independent from the postprocessing technique). The performances of models that were built on each of the 5 RFs datasets were evaluated via accuracy and area under the ROC curves (AUC) with 95%CI. Finally, for each RFs dataset, the model with the highest AUC in cross-validation was applied on the whole training cohort and on the testing cohort.

RESULTS

Patients' population (Table 4.1-1)

Thirty-two of the 70 patients (45.7%) were women with a median age of 58 (range: 19-84). The most frequent histological types were high-grade undifferentiated sarcomas (31/70, 44.3%), with a median size of 116 mm (range 40-273) and mostly deep-seated in the lower limb (35/70, 50%).

Per-RF analysis

The influence of the harmonization technique was significant for all the RFs (p-values range: <0.001-0.007, Supplemental Data 4). The number of significant differences in the RFs comparisons for each pair of postprocessing techniques is given in Table 4.1-2.

Table 4.1-1: Clinical and pathological features of the study population.

Characteristics	No. Of patients
Age (years old)	
median (range)	58 (19-84)
Gender	
Men	38/70 (54.3)
Women	32/70 (45.7)
WHO Performance Status	
PS 0	55/70 (78.6)
PS 1	15/70 (21.4)
Histotype	
Undifferentiated sarcoma	31/70 (44.3)
Synovial sarcoma	8/70 (11.4)
Rhabdomyosarcoma	8/70 (11.4)
Leiomyosarcoma	6/70 (8.6)
Myxoid/round cells liposarcoma	6/70 (8.6)
Pleomorphic sarcoma	3/70 (4.3)
Other sarcomas	8/70 (11.4)
Longest diameter (mm)	
median (range)	106 (40-273)
Volume (cm3)	
median (range)	220 (10.2-3084)
Location	
Trunk	12/70 (17.1)
Shoulder girdle	9/70 (12.9)
Upper limb	9/70 (12.9)
Pelvic girdle	5/70 (7.1)
Lower limb	35/70 (50)
Depth	
Deep-seated	65/70 (92.9)
Superficial and aponeurotic	5/70 (7.1)
No. Of cycle	
4 cycles	18/70 (25.7)
5-6 cycles	52/70 (74.3)
Chemotherapy	
Anthracycline-ifosfamide	64/70 (91.4)
Doxorubicine	6/70 (8.6)
Adjuvant radiotherapy	
No	5/70 (7.1)
Yes	65/70 (92.9)
Margins	
R0	41/70 (58.5)
R1	29/70 (41.4)
Histological response	
Good	16/70 (22.9)
Poor	54/70 (77.1)

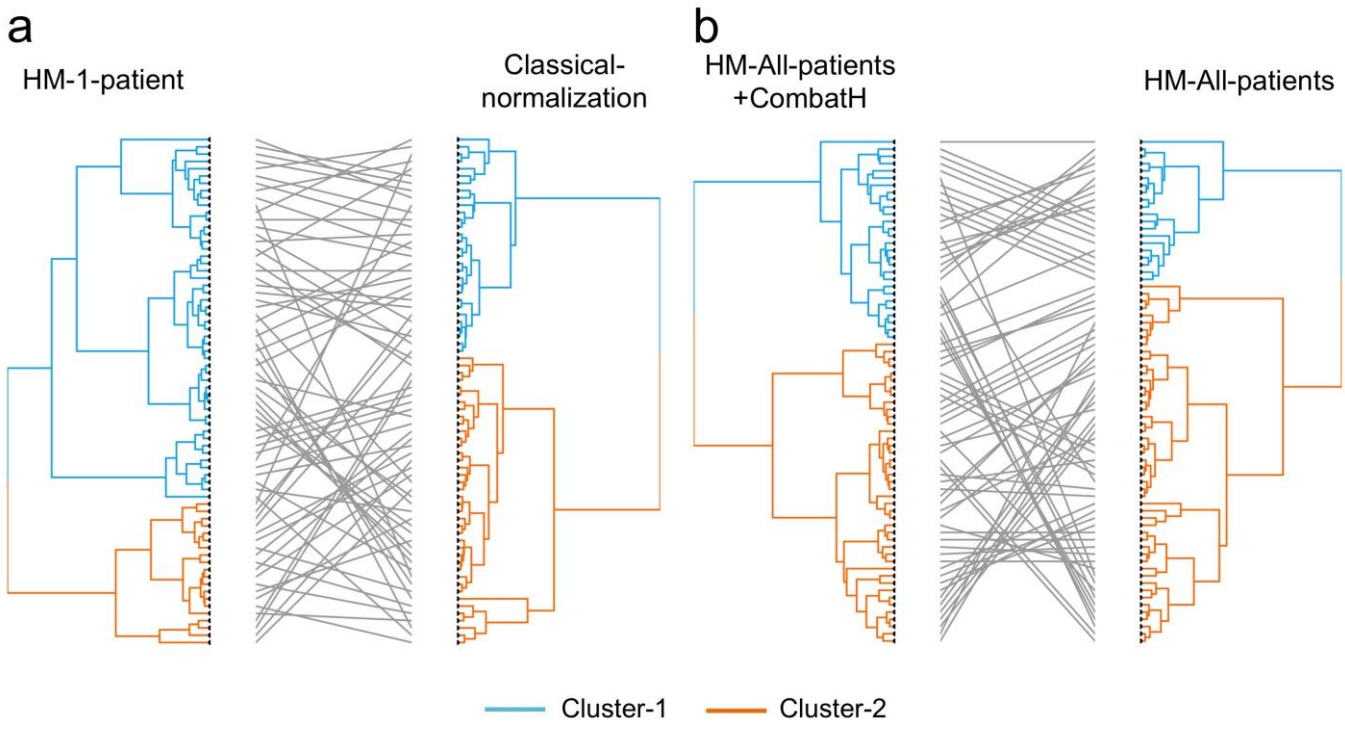
NOTE. Results are number of patients with percentage in parentheses, except for age, longest diameter and volume that are expressed as median with range in parentheses

Table 4.1-2. Summary of the per-radiomics features (RFs) analysis.

Post-Hoc Comparisons[1]			No. of significant differences[2]
HM-All-patients	vs.	Fat-standardization	35/45 (77.8%)
HM-1-patient	vs.	Fat-standardization	34/45 (75.6%)
HM-All-patients+ComBat	vs.	Fat-standardization	33/45 (73.3%)
Basic-normalization	vs.	HM-All-patients+ComBat	29/45 (64.4%)
Basic -normalization	vs.	HM-All-patients	29/45 (64.4%)
Basic -normalization	vs.	Fat-standardization	21/45 (46.7%)
Basic -normalization	vs.	HM-1-patient	20/45 (44.4%)
HM-1-patient	vs.	HM-All-patients+ComBat	17/45 (37.8%)
HM-All-patients	vs.	HM-All-patients+ComBat	13/45 (28.9%)
HM-All-patients	vs.	HM-1-patient	8/45 (17.8%)

NOTE. [1]: Post-Hoc comparisons correspond to the post-hoc Bonferroni-corrected Tukey tests for repeated-measures ANOVAs where the influence of the intensity harmonization techniques on the 45 RFs was investigated.
[2]: The number (no.) of significant differences corresponds to the number of RFs that were significantly different in a given post-hoc comparisons of 2 intensity harmonization techniques (with percentage over the total number of RFs in parentheses).
Abbreviation: HM: histogram matching, No.: number.

Figure 4.1-3: Comparisons of the hierarchical clustering results based on radiomics features with: the highest divergence (a), and the lowest divergence (b). The dendrograms were obtained according to the following intensity harmonization technique: histogram matching with a randomly-chosen normalized histogram of a patient (*HM-1-patient*) versus *basic-normalization*; and histogram matching with the average normalized histogram of the 70 included patients (*HM-All-patients*) versus *HM-All-patients* combined with ComBat harmonization method (*HM-All-patients+ComBat*). By convention, cluster-1 (in blue) corresponds to the group of patients with the best prognosis for metastatic-relapse free survival.



The highest and lowest amounts of differences were obtained for post-hoc comparisons between *HM-All-patients* vs. *Fat-standardization* techniques (35 statistically different RFs out of 45 [77.8%]) and HM-All-patients vs. HM-1-patient (8/45 [17.8%])), respectively. The most sensitive RFs to the postprocessing technique were GLZLM_LZE and Histogram_mean with 9 out of the 10 possible post-hoc comparisons (90%) that were significantly different. Conversely, the less sensitive was GLRLM_GLNU with only one significant difference among the 10 post-hoc comparisons (10%).

Unsupervised analysis

None of the 5 unsupervised classifications were the same. The most correlated pair of clustering was obtained with *HM-All-patients* vs. *HM-All-patients+ComBat* (Kappa=0.75, Baker coefficient=0.55) and the less correlated was obtained with *Basic-normalization* vs. *HM-1-patient* (Kappa=0.25, Baker coefficient=0.11), which are both represented in Figure 4.1-3 (Annexe 5 - Supplemental Data 5).

Regarding the survival analysis, the clusters obtained with *Basic-normalization*, *HM-All-patients* and *HM-All-patients+ComBat* were independently associated with MFS in the multivariate modeling (p=0.007, 0.004 and 0.02, respectively – Table 4.1-3) but none of those obtained with *Fat-standardization* and *HM-1-patient*. Kaplan-Meier curves for the 5 clustering analyses are given in Figure 4.1-4. Concordance-indices of the 5 prognostic models were between 0.71 (95%CI=(0.67-0.75)) - for *HM-1-patient* - and 0.75 (95%CI=(0.70-0.79)) - for *HM-All-patients*.

Table 4.1-3: Unsupervised analysis based on radiomics features (RFs) - Prognostic value of the clustering results depending on the intensity harmonization technique.

Intensity harmonization technique	Clustering result	No. Of patients	No. Of events	2-years survival probability	Multivariate Cox Modeling[1]		
					HR	p-value	Concordance-index
Fat-standardization	Cluster-1	53	23	62.3 (50.5-76.8)	-	-	0.72 (0.67-0.76)
	Cluster-2	17	9	58.8 (39.5-87.6)	1.65 (0.70-3.89)	0.3	
Basic-normalization	Cluster-1	30	11	70 (55.4-88.5)	-	-	0.75 (0.72-0.79)
	Cluster-2	40	21	55 (41.6-72.8)	3.26 (1.48-7.71)	0.007*	
HM-1-patient	Cluster-1	50	22	64 (52-78.8)	-	-	0.71 (0.67-0.75)
	Cluster-2	20	10	55 (37-81.8)	1.52 (0.66-3.49)	0.3	
HM-All-patients	Cluster-1	20	5	80 (64.3-99.6)	-	-	0.75 (0.70-0.79)
	Cluster-2	50	27	54 (41.8-69.7)	4.72 (1.64-13.56)	0.004**	
HM-All-patients+ComBat	Cluster-1	28	10	67.9 (52.6-87.6)	-	-	0.73 (0.68-0.77)
	Cluster-2	42	22	57.1 (44-74.3)	2.89 (1.19-7.05)	0.02*	

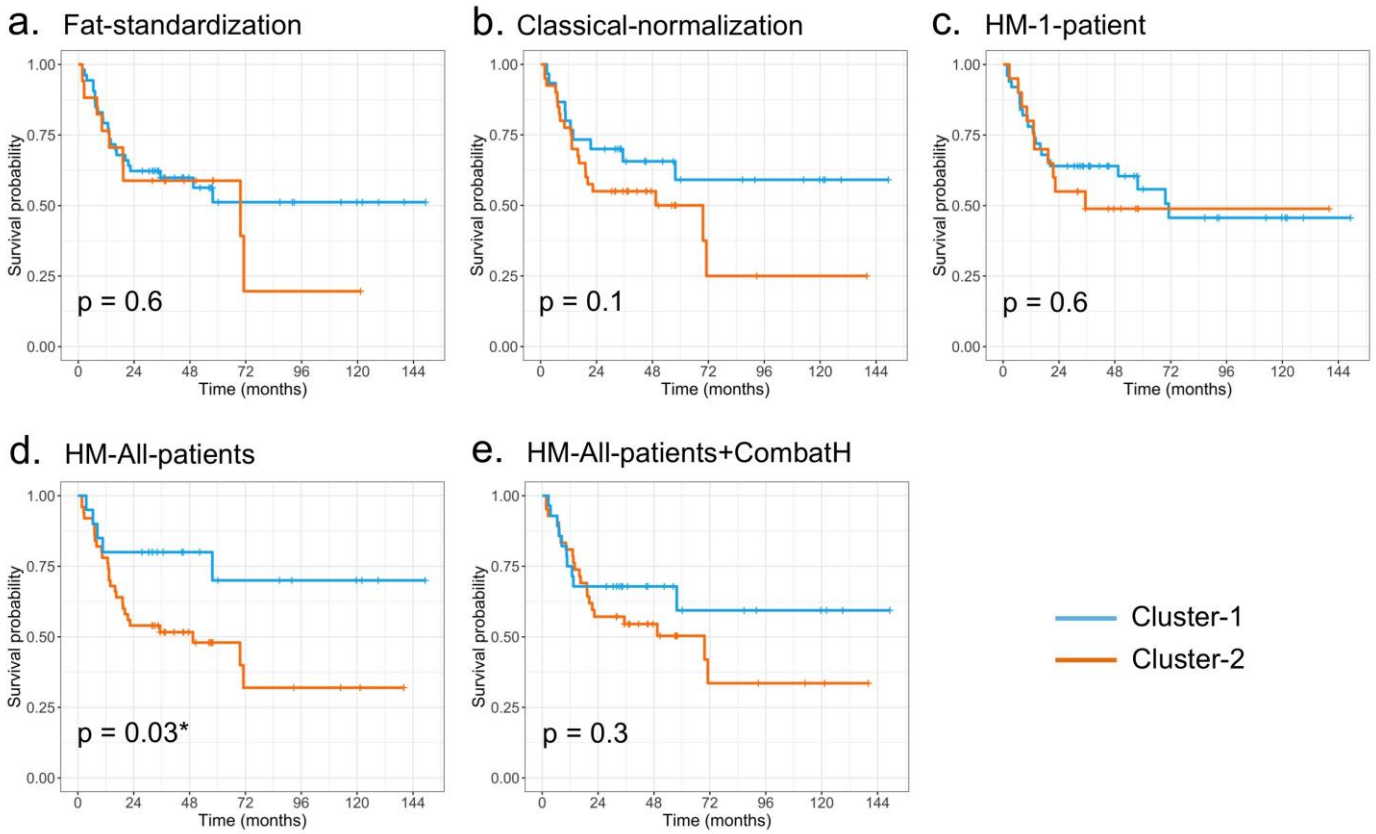
NOTE. Results for 2-years survival probability, hazard ratio and concordance-index are given with 95% confidence interval.

[1] Multivariate Cox modeling were adjusted for the following clinical and pathological covariables: performance status, histotype, initial longest diameter of the tumor, type of neoadjuvant chemotherapy, number of cycles of chemotherapy, surgical margins, histological response and adjuvant Radiotherapy,

Abbreviations: HM: histogram matching, HR: hazard ratio, No: number.

*: $p<0.05$, **: $p<0.005$, ***: $p<0.001$

Figure 4.1-4. Kaplan-Meier curves for metastatic-relapse free survival depending on unsupervised clustering results based on radiomics features obtained with the different intensity harmonization techniques. The following intensity harmonization techniques were used: *Fat-standardization* (a); *Basic-normalization* (b); histogram matching with a randomly-chosen normalized histogram of a patient (*HM-1-patient*, c); histogram matching with the average normalized histogram of the 70 included patients (*HM-All-patients*, d); *HM-All-patients* combined with ComBat harmonization method (*HM-All-patients+ComBat*). The p-value corresponds to the univariate log-rank test. *: p<0.05.



Supervised analysis

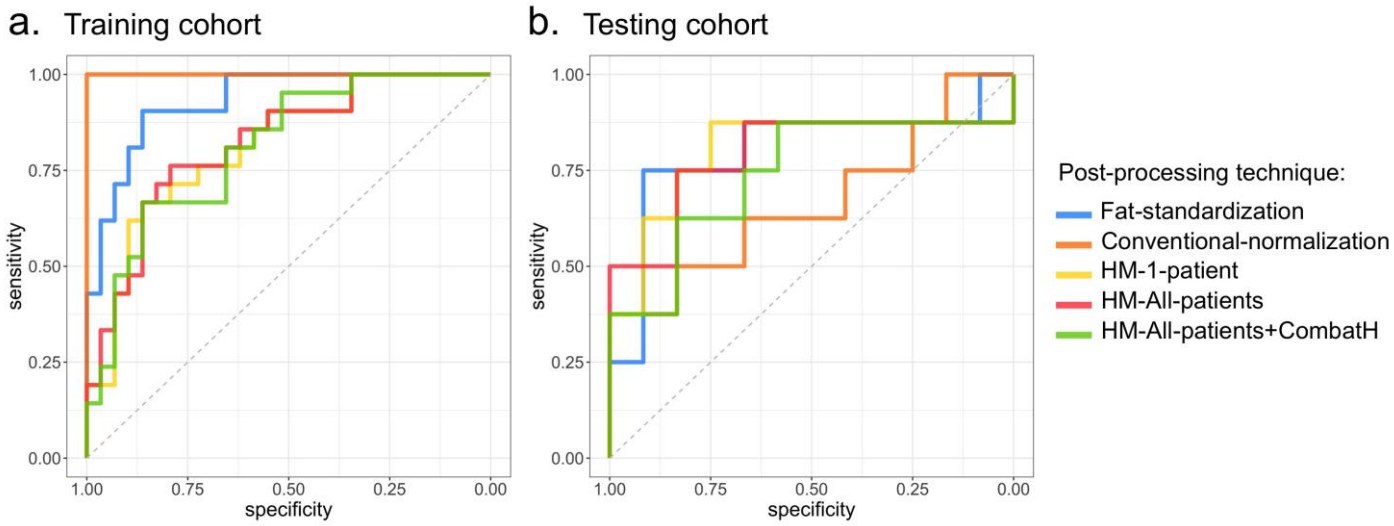
After training the machine-learning classifiers, the models with highest AUC were obtained with glmnet for *Fat-standardization*, *HM-1-patient, HM-All-patients* and *HM-All-patients+ComBat*, and random forests (after Wilcoxon prefiltering) for *Basic-normalization*. Performances in each cohort are given in Table 4.1-4 (details regarding each model in repeated cross-validation are given in Annexe 5 - Supplemental Data 6). In the training cohort, the best performances were found with *Basic-normalization* (AUC=1, 95%CI=(1-1)), followed by *Fat-standardization* postprocessings (AUC=0.93, 95%CI=(0.87-1)). In descending order, the final AUCs in the testing cohort were: 0.80 (95%CI=(0.55-1)) for *Fat-standardization* and *HM-1-patient*, 0.79 (95%CI=(0.54-1)) for *HM-All-patients*, 0.74 (95%CI=(0.48-1)) for *HM-All-patients+ComBat* and 0.67 (95%CI=(0.40-94)) for *Basic-normalization* (Figure 4.1-5).

Table 4.1-4: Supervised analysis based on radiomics features (RFs) - Performances of models based on radiomics features for predicting metastatic relapse 2 years after the end of treatment, depending on the intensity harmonization technique.

Intensity harmonization technique	Classifier	Training cohort		Testing cohort	
		AUC	Accuracy	AUC	Accuracy
Fat-standardization	Glmnet	0.93 (0.87-1.00)	0.84 (0.71-0.93)	0.80 (0.55-1.00)	0.75 (0.51-0.91)
Basic-normalization	Random forests	1.00 (1.00-1.00)	1.00 (0.93-1.00)	0.67 (0.40-0.94)	0.60 (0.36-0.81)
HM-1-patient	Glmnet	0.81 (0.69-0.93)	0.70 (0.55-0.82)	0.80 (0.55-1.00)	0.75 (0.51-0.91)
HM-All-patients	Glmnet	0.82 (0.70-0.94)	0.72 (0.57-0.84)	0.79 (0.54-1.00)	0.75 (0.51-0.91)
HM-All-patient+ComBat	Glmnet	0.81 (0.69-0.93)	0.64 (0.49-0.77)	0.74 (0.48-1.00)	0.60 (0.36-0.81)

NOTE. Results for AUC and accuracy are given with 95% confidence interval.
Models were adjusted for the following clinical and pathological covariables: performance status, histotype, initial longest diameter of the tumor, type of neoadjuvant chemotherapy, number of cycles of chemotherapy, surgical margins, histological response and adjuvant Radiotherapy.
Abbreviations: AUC: area under the receiver operating characteristic curve; glmnet: penalized binomial logistic regression with optimized combination of ridge and lasso regressions, HM: histogram matching.
Details regarding the assessment of the best classifier and its optimal hyperparameters are given in Supplemental Data 6.

Figure 4.1-5. ROC curves for the best models in 5-times repeated 10-fold cross-validation of the training cohort, in the whole training cohort (a) and in the testing cohort (b). The supervised machine-learning models aimed at predicting the occurrence of metastatic relapse 2 years after the end of treatment. They were developed on radiomics features obtained with the following intensity harmonization techniques: *Fat-standardization*, *Basic-normalization*, histogram matching with a randomly chosen normalized histogram of a patient (*HM-1-patient*), histogram matching with the average normalized histogram of the 70 included patients (*HM-All-patients*), and *HM-All-patients* combined with ComBat harmonization method (*HM-All-patients+ComBat*).



DISCUSSION

The postprocessing of medical images for the purpose of radiomics studies is mandatory to homogenize and make comparable multicentric datasets but it can result in additional bias at risk of altering the performances of predictive models. Because structural MRIs are acquired in arbitrary units, the intensity harmonization is crucial to enable the comparability of examinations acquired with different MR-systems, coils, and acquisition parameters. We found that all the 45 widely-used texture features were significantly influenced by the intensity harmonization techniques. Furthermore, prognostic predictions based on unsupervised clustering of radiomics features also changed from one intensity harmonization technique to another, some clustering being associated with metastatic-relapse free survival in the multivariate analysis (i.e. *Basic-normalization, HM-All-patients* and *HM-All-patients+ComBatH*) while others were not (i.e. *Fat-standardization* and *HM-1-patient*). Finally, a similar influence was noticed in supervised models aiming at predicting early metastatic relapse after 2 years.

Our results concur with previous studies that found significant influences of postprocessing steps on the absolute values of RFs (such as voxel size standardization, gray-levels discretization or manual segmentation) in addition to pre-processing steps (such as magnetic field strength, manufacturers, coils, acquisition parameters or filters). Recently, Scalco et *al.* found that the SI harmonization methods for T2-WI had a significant impact on the reproducibility of RFs and in the inter-observer reproducibility of RFs that were extracted from pelvic organs from 2 MRIs separated by months (2019). The originality of our study lies in the investigation of how much it could influence oncological predictions with clinical implications.

Previous studies have already emphasized the influence of intensity harmonization algorithms on segmentation and classification tasks but they mostly involved brain MRI for inflammatory or degenerative diseases, and not specifically their influence on radiomics analyses (Fortin et al, 2017; Fortin et al, 2016; Shinohara et al, 2014; Robitaille et al, 2012). The methods they proposed were not systematically transposable to body-imaging or not available in open source language (for instance, DeepHarmony) (Dewey et al, 2019). Herein, we decided to focus on techniques that were previously used in the body-imaging radiomics literature (i.e. global scaling, histogram-matching or ComBat) but further studies should consider translating other

popular intensity harmonization algorithms to body MRI. For instance, the RAVEL algorithm aims at estimating a voxel specific unwanted variation by using a control region (i.e. brain cerebro-spinal fluid) to estimate factors of unwanted variations (Fortin et al, 2016). Hence, by defining healthy adipose tissue as control, RAVEL algorithm may be applied to body-MR. An alternative to intensity harmonization at postprocessing could be to pre-process images in order to obtain raw comparable images since the acquisition, through the acquisition of standardized T1-mapping or T2-mapping. However, thousands of MRIs have already been stored and, logically, the radiological community expects to pool and include these images in retrospective radiomics studies.

Herein, none of the intensity harmonization techniques demonstrated an unequivocal superiority compared to the others. This observation suggests that the "best" technique may depend on the data and the study objectives and that there is no better universal technique. Though the unsupervised analysis highlighted the performances of *Basic-normalization*, *HM-All-patients* and *HM-All-patients+ComBat*, the supervised analysis rather emphasized *Fat-standardization*, *HM-1-patient* and *HM-All-patients+ComBat* in the testing cohort. It suggests that radiomics studies should investigate all the available intensity harmonization methods in an exploratory subset of the study population before selecting the one maximizing the predictions. Hence, the intensity harmonization techniques could be consider as an "hyperparameter" of the postprocessing pipeline. Interestingly, the constant moderately good performances of *HM-All-patients+ComBat* in unsupervised and supervised analysis (with similar results in training and testing cohort) lead us to believe that this method may provide the more realistic harmonization of intensities. However, the covariable arguments given to the ComBat function could be incomplete. A distinctive feature of sarcomas over other cancers is their ubiquity, hence, requiring adjusting several other acquisition parameters depending on the tumor location (for instance thoracic wall, thigh or wrist). Further studies should investigate the best covariables for ComBat for body-MRI.

Our results also concur with previous sarcoma studies. Together with peritumoral contrast-enhancement and presence of necrotic component, heterogeneous SIs on T2-WI (which was qualitatively assessed) is known to be predictive of MFS, irrespective of histological grading (Crombé et al, 2019a). Our results confirm this result in a quantitative way.

Our study has limits. First, this is a retrospective study based on a relatively small cohort although our study population was comparable with prior research regarding STS and radiomics. Second, we did not investigate the influence of harmonization techniques on other cancer types and other MRI sequences. It should be noted that T1-WI, contrast-enhanced T1-WI, DCE-MRI or diffusion imaging are also routinely performed at each step of the management of STS patients. We could have performed the same analysis on these other sequences but we purposely chose to focus on T2-WI to demonstrate the effect of harmonization methods, as a proof of concept. Although the performances of the T2-based radiomics models were correct in the present study, they do not appear sufficient to be used in practice. We believe that adding RFs and delta-RFs from these other sequences as well as other radiological features (such as peritumoral enhancement, aponeurotic enhancement or bone/vessel/nerves invasion) and molecular features would have improved the performances of the models (Crombé et al, 2019a; Crombé et al, 2019d; Crombé et al, 2020b; Spraker et al, 2019; Yoo et al, 2014; Holzapfel et al, 2015). Finally, other machine-learning classifiers could have been tuned, with less overfitting, but we purposely decided to focus on the most effective according to recent sarcoma studies.

To conclude, our study highlights that the intensity harmonization techniques can directly influence the values of MRI-based RFs leading to variations in the predictions of unsupervised and supervised models. These techniques need to be deepened regarding body-MRI and should be carefully explored and detailed when building radiomics models because they could significantly enhance or decrease models performances.

*

* *

4.1.3. Limites et ouvertures

Cet article à visée méthodologique illustre l'influence des techniques de normalisation pour chacun des trois aspects de l'étude. Les variables radiologiques habituellement pronostiques n'ont pour cette raison pas été ajoutées, alors qu'elles auraient pu améliorer les modèles, tout comme d'autres variables extra-radiologiques. Concernant la partie "prédiction de la rechute métastatique à 2 ans", nous avons estimé le RQS à 10 en raison de l'absence de *test-retest*, de fantômes, de données *sham*, de re-segmentation, d'analyse multivariée avec d'autres variables extra-radiologiques, de corrélation biologique, de recherche de seuils, de courbe de calibration, de comparaison à un gold standard, de validation sur plusieurs cohortes extérieures, de prospectivité, d'étude cout-efficacité, et de possibilité d'accès libre aux données.
Il en ressort aussi qu'aucune technique ne se détache clairement des autres en terme de performances. De notre point de vue, cela signifie qu'actuellement, il n'y a pas une meilleure technique valable en toutes circonstances parmi les 5 proposées, mais que les différentes techniques disponibles peuvent être testées comme le sont d'autres paramètres de post-traitement, afin d'identifier la méthode qui mettra le plus en valeur les données de l'examen. Cela signifie aussi que ce champ de recherche doit être davantage développé et davantage mis en avant dans les recommandations concernant les études radiomics, et que de nouvelles méthodes, éventuellement extrapolées des la neuro-imagerie, restent à découvrir et diffuser.

4.2. En DCE-MRI

4.2.1. Introduction

Les séquences DCE-MRI brutes et les cartes paramétriques qui en sont extraites permettent d'identifier des profils de rehaussement intra-tumoraux qualitativement très différents (Figure 1-17) pour une même valeur moyenne des paramètres semi-quantitatifs et quantitatifs, et pour un même type histologique. Nous avons aussi vu que les paramètres de ces séquences pourraient permettre d'identifier plus précocement les bons des mauvais répondeurs aux traitements systémiques mais,

selon les résultats publiés et notre expérience, il existe un important chevauchement des variations des valeurs moyennes des indices extraits. L'hypothèse que l'hétérogénéité de la néo-angiogenèse tumorale (alors essentiellement mesurée via des indices de texture de 1er ordre) puisse être une source de biomarqueurs d'imagerie a été résumée par Jackson et al (2007) et reprise dans plusieurs études facilitées par la généralisation de logiciels et packages dédiés aux analyses de texture. Des résultats encourageants ont été publiés concernant les cancers prostatiques, du rectum, du sein, ou encore ORL (Ginsburg et al, 2017; Nie et al; 2016; Braman et al, 2017; Bowen et al, 2018; Liu et al, 2019; Thomassin-Naggara et al, 2017; Fan et al, 2018). Plusieurs de ces études proposent d'extraire les indices de texture directement depuis les cartes paramétriques issues de la séquence DCE-MRI.

Alors que les paramètres d'acquisition de séquences IRM structurales simples peuvent déjà être une source de variations significatives des indices radiomics, nous avons supposé que la variation des paramètres temporels pourrait aussi être un facteur supplémentaire d'instabilité. Cette hypothèse avait déjà été émise par Bogowicz et *al.* pour le scanner de perfusion (2016) et nous avons souhaité la tester pour leur équivalent IRM à l'échelle des indices radiomics et aussi à l'échelle d'une prédiction simple c'est-à-dire la réponse histologique après NAC. Ce travail est présenté dans le paragraphe 4.2.2.

Nous nous sommes ensuite posés la question de la meilleure méthode pour extraire une information à visée pronostique des séquences DCE-MRI. En effet, les indices radiomics peuvent être extraits directement à chaque phase de la séquence DCE-MRI brute créant une profusion d'indices radiomics à chaque instant de l'acquisition. Il est possible de tracer ces courbes de variation des indices radiomics au cours du temps et de les intégrer - tout comme l'AUC_{90s} correspond à l'intégration de la valeur moyenne du signal dans le volume d'intérêt dans les 90 secondes suivant l'injection. Nous avons donc proposé 4 stratégies basées sur les séquences DCE-MRI pré-traitement d'une cohorte rétrospective de 50 patients atteints de STM de haut-grade uniformément traités sur l'institut afin de créer des modèles pronostics de la survie sans rechute métastatique en travaillant sur :

(1) les indices radiomics extraits des cartes paramétriques K^{trans} et AUC_{90s};

(2) l'intégralité des indices radiomics extraits à chaque phase brute;

(3) l'intégrale de la courbe "indice radiomics = f(t)"

(4) un sous-échantillon d'indices radiomics extraits à 30s, 60s et 90s post-injection. Deux modèles plus classiques ont servi de référence pour évaluer les performances de ces modèles: un modèle basé sur des critères purement issus de l'analyse radiologique classique, et un modèle radiomics basé sur la séquence pondérée T2. Les résultats sont présentés dans le paragraphe 4.2.3.

4.2.2. Article 4: influences des paramètres temporels sur les indices radiomics issus des séquences DCE-MRI et sur une prédiction clinique

L'article issu de ce travail a été publié en mai 2019 dans *Journal of Magnetic Resonance Imaging* (PMID: 30980697, doi: 10.1002/jmri.26753 - rang B SIGAPS, IF = 3.732). Il s'agit ici du manuscrit final reformaté.

ORIGINAL RESEARCH

Influence of Temporal Parameters of DCE-MRI on the Quantification of Heterogeneity in Tumor Vascularization

Amandine Crombé, MD, MS,[1,2*] Olivier Saut, PhD,[2] Jerome Guigui, RT,[1] Antoine Italiano, MD, PhD,[3] Xavier Buy, MD,[1] and Michèle Kind, MD, MS[1]

Background: Evaluating heterogeneity in tumor vascularization through texture analysis could improve predictions of patients' outcome and response evaluation.
Purpose: To investigate the influence of temporal parameters on texture features extracted from dynamic contrast-enhanced (DCE)-MRI parametric maps.
Study type: Prospective cross-sectional study.
Subjects: Twenty-five adults with soft-tissue sarcoma (STS), median age: 68 years.
Field Strength/Sequence: DCE-MRI acquisition using a CAIPIRINHA-Dixon-TWIST-VIBE sequence at 1.5T (temporal resolutions: 2 sec, duration: 5 min).
Assessment: The area under time–intensity curve (AUC) and K^{trans} maps were generated for several temporal resolution (dt = 2 sec, 4 sec, 6 sec, 8 sec, 10 sec, 12 sec, 20 sec) and scan durations (T = 3 min, 4 min, 5 min for a 6-sec sampling) by downsampling and truncating the initial DCE-MRI sequence. Tumor volume was manually segmented and propagated on all parametric maps. Thirty-two first- and second order-texture features were extracted per map to quantify the intratumoral heterogeneity.
Statistical Tests: The influence of temporal parameters on texture features was studied with repeated-measures analysis of variance (or nonparametric equivalent). The dispersion of each texture feature depending on temporal parameters was estimated with coefficients of variation (CVs). The performances of multivariate models to predict the response to chemotherapy (ie, binary logistic regression based on the baseline texture features) were compared.
Results: The temporal resolution had a significant influence on 12/32 (37.5%) and 14/32 (43.8%) texture features evaluated on AUC and K^{trans} maps, respectively (range of $P < 0.0001$–0.0395). Scan duration had a significant influence on 23/32 (71.9%) texture features from K^{trans} map (range of $P < 0.0001$–0.0321). Dispersion was high (mean CV >0.5) with sampling for 2/32 (6.3%) and 10/32 (31.3%) features from AUC and K^{trans} maps, respectively; and with truncating for 6/32 (18.8%) features from K^{trans} map. The area under the receiver operating characteristics curve of predictive models ranged from 0.77 (95% confidence interval [CI] = [0.54–1.00], with dt = 6 sec T = 4 min) to 0.90 (95% CI = [0.74–1.00], with dt = 6 sec T = 5 min).
Data Conclusion: The values of texture features extracted from DCE-MRI parametric maps can be influenced by temporal parameters, which can lead to variations in performance of predictive models.
Level of Evidence: 2
Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2019.

INTRODUCTION

The extensive quantification of tumor heterogeneity on medical imaging is a growing field of research in oncology referred to as radiomics. The underlying hypothesis of radiomics is that the imaging phenotype of a tumor could reflect its intrinsic molecular identities and aggressiveness (O'Connor, 2017; Limkin et al, 2017). Texture analyses consist in the mathematical processing of images in order to extract numeric indices that objectively measure the heterogeneity, named texture features. The most commonly encountered ones in the medical literature are 1st order features, which are based on frequency histograms without spatial information, and 2nd order features, which quantify the 2D and/or 3D rearrangements of voxels of different gray-levels. Predictive radiomics approaches based on texture features, machine learning algorithms and the potential combination with clinical characteristics and other –omics data (i.e. genomics, transcriptomics, proteomics and metabolomics) could help better stratify the therapeutic strategy for cancer patients and evaluate treatment responses (Limkin et al, 2017).

Radiomics approaches can be applied to every imaging modality including dynamic contrast-enhanced MRI (DCE-MRI). DCE-MRI aims at providing a non-invasive macroscopic assessment of tumor perfusion and neo-angiogenesis – a key pro-oncogenetic process – through the rapid acquisition of a time series of T1-weighted imaging (Jackson et al, 2007). DCE-MRI parameters have been used to discriminate benign from malignant tumors or to monitor treatment efficacy especially anti-angiogenic regimens – the area under the time intensity curve (AUC) and the influx volume transfer constant (K^{trans}) being the most studied) (O'Connor et al, 2017) . However, these studies were based on average values of the DCE-MRI parameters that do not reflect the complexity of tumors. Indeed, homogeneous poorly vascularized tumors could have the same mean AUC and K^{trans} values as heterogeneous tumors with both hypervascularized and large avascular necrotic areas. Consequently, quantifying the spatial heterogeneity in vascularization of the whole tumor volume from DCE-MRI data may be more informative and realistic. In that sense, radiomics approaches on DCE-MRI have recently shown encouraging results, alone or with other MRI sequences, in order to improve the detection of prostate cancer, to distinguish benign and malignant adnexal masses, to identify relevant molecular subtypes of breast cancers, to detect lymph node metastases in breast cancers, or to predict response to neoadjuvant treatment for rectum, breast and

nasopharyngeal cancers (Ginsburg et al, 2017; Nie et al, 2016; Braman et al, 2017; Bowen et al, 2018; Liu et al, 2019; Thomasson-Nagara et al, 2017; Fan et al, 2018; Rose et al, 2009) . Soft-tissue sarcomas (STS) are malignant mesenchymal tumors with important inter and intra-tumoral heterogeneity known to be associated with high grade (Zhao et al, 2014; Crombé et al, 2019a). Previous studies have shown that DCE-MRI could be useful to predict response to chemotherapy in high-grade STS as well as radiomics approaches (Soldatos et al, 2016; Meyer et al, 2013; Crombé et al, 2019c; Huang et al, 2016; Tian et al, 2015; Hayano et al, 2015). Thus, it can be hypothesized that combining radiomics and DCE-MRI could enhance the early prediction of tumor response to treatment.

Robust radiomics models require the inclusion of a large number of patients, divided into training, validation and external test cohorts, before being implemented into clinical practice. Pooling data from different centers is necessary but it runs the risk of introducing bias to the values of texture features. Indeed, each step of the radiomics process can introduce variability independently from the intrinsic heterogeneity of the tumor, for instance: MRI field strength, manufacturers, coils, acquisition parameters, segmentation, voxel-size resampling, normalization techniques or grey-level discretization (Buch et al, 2018; Mayerhoefer et al, 2009; Collewet et al, 2004; Ford et al, 2018). Previous studies have demonstrated that temporal parameters (i.e. scan duration and temporal resolution) could significantly modify the ability to discriminate benign from malignant prostate or breast lesions (Othman et al, 2016; Othman et al, 2016; Hao et al, 2015), but they were based on average values of DCE-MRI indices or morphology of the time-intensity curves. Only one study has focused on the stability of texture features extracted from computed tomography perfusion maps identifying an influence of temporal resolution (Bogowicz et al, 2016). Hence, data regarding the influence of temporal parameters on texture features extracted from DCE-MRI parametric maps are lacking.

Thus, our aims were to investigate the influence of temporal parameters: (i) on the values of widely used texture features extracted from DCE-MRI parametric maps of STS, (ii) on the dispersion of these parameters, and (iii) on the performance of predictive models for chemotherapy responses.

MATERIALS AND METHODS

Patient population

In this prospective single-center study, the institutional review board waived the requirements of informed patient consent. From November 2017 to June 2018, 30 consecutive adult patients were included as they presented at our sarcoma reference center for the management of a histologically-proven high-grade (according to the French *Federation Nationale des Centres de Lutte contre le Cancer* [FNCLCC] grading system) STS of the trunk wall or extremities of more than 4cm, and required a contrast-enhanced MRI for diagnostic and/or therapeutic management. Age, gender, histological type, tumor depth relative to superficial fascia, tumor location, longest diameter, specific treatments and short-term patients' outcome were retrieved from medical records. We defined a good responder as: (i) <10% of stainable viable tumor cells on surgical specimen after neo-adjuvant chemotherapy for locally-advanced non-metastatic STS (28), and (ii) a partial or complete response 6 months after performing DCE-MRI according to RECIST 1.1 criteria for metastatic or inoperable patients, without changing treatment – this last empirical definition being proposed regarding the usual median progression-free survivals of metastatic STS patients (Ray-Coquard et al, 2017).

Data acquisition

Patients underwent MRI scans in the same 1.5T MR-system (MAGNETOM Aera; Siemens Healthineers, Erlangen, Germany) with adapted coils depending on tumor locations and sizes, i.e. 15 channels transmit/receiver coil for knee and extremities and 18 channels transmit/receiver body coil for trunk wall and thighs. Patients were examined in supine position. The protocol followed the Quantitative Imaging Biomarkers Alliance recommendations (QIBA®) and consisted in a T1-mapping followed by the DCE-MRI acquisition with the same field-of-view (350 x 320 mm) and spatial resolution (1.1 x 1.1 x 4 mm^3), in order to optimize the conversion of the signal in Gadolinium chelates concentration. We used a CAIPIRINHA (Controlled Aliasing in Parallel Imaging Results in Higher Acceleration) Dixon TWIST (Time-resolved angiography With Stochastic Trajectories) VIBE (Volume Interpolated Breath hold Examination) sequence (Michaely et al, 2013). The T1-mapping used a similar spoiled gradient-echo sequence with variable flip angles (2° and 15°) and with

echo and repetition times of 1.41 ms and 3.79 ms, respectively. In brief, the principle of TWIST is to divide the k-space in a central region, which encodes information about contrast, and a peripheral region, which encodes information about shapes, edges and details and is undersampled (Michaely et al, 2013). The suppression of the fat signal was performed using a Dixon-based water-fat separation (echo times were 2.39 ms for in-phase and 4.77 ms for opposite-phase conditions). A repetition time of 6.89 ms and a flip angle of 25° were used. The sampling was accelerated with a parallel acquisition technique with an undersampling factor of 2. The value of the TWIST view-sharing parameters A (i.e. the percentage sizes of the central portion of the k-space) and B (i.e. the percentage sizes of the peripheral portion of the k-space) were 15% and 20%, respectively. The temporal resolution resulted in dt = 2s, except for the first TWIST phase (6.9 s for the full k-space sampling). Five phases were acquired before the intra-venous power-injection of 0.1 mmol/kg of gadoteric acid (Dotarem, Guerbet, Villepinte, France) at a rate of 2 mL/s followed by a 20mL flush of 0.9% of NaCl solution thanks to a MRI-compatible automatic injector (Sonic Shot 7, Nemoto Kyorindo, Tokyo, Japan). The total scan duration was 5 min and included 144 phases (full-dataset).

Data reconstruction

Data reconstruction and post-processing were achieved with Olea Sphere®, v3.0 SP14 Software (Olea Medical, La Ciotat, France) by a senior radiologist with 7 years of experience in MRI blinded to clinical data. First, motion artifacts were systematically corrected with a rigid body co-registration method. For each patient, the full-dataset was downsampled and truncated to obtain the datasets to assess the effect of scan duration and temporal resolution. In any case, the 5 first phases were kept in the datasets to ensure a similar baseline for all maps. In total, 7 DCE-MRI datasets per patient were generated to evaluate temporal resolution, namely: dt = 2s (raw data), 4s, 6s, 8s, 10s, 12s, 20s, with a scan duration of 5min. Three DCE-MRI datasets were generated to evaluate the influence of scan duration, namely: T=3min, 4min, 5min, all with a temporal resolution of dt = 6s. Figure 1 shows how the datasets were built.

The next step consisted in the calculation of the parametric maps using the permeability plug-in of Olea Sphere®. First, the senior radiologist manually chose the largest feeding artery within the field-of-view of the raw DCE-MRI dataset, with the

exclusion of the first and the last slices to avoid artifacts. A fixed voxel was manually placed in this artery and the arterial input function (AIF) was measured as the average of 4 directly adjacent voxels within this artery that demonstrated an arterial enhancement and the less noise. The same artery and the same voxels were used for the other reconstructed datasets. Hence, even if the baseline of the AIF was preserved, the AIF was also downsampled and truncated (Figure 4.2-1). Second, the AIF signal intensity time-course was converted into blood R1 time-course, with the hematocrit = 0.45 and the gadoteric acid relaxivity at 1.5T = 3.6 L/mmol.s^{-1}. We chose to focus on the most widely used parameters: the area under the time-intensity curve (AUC), at 90s (units: mmol/L/s) and the pharmacokinetic parameter K^{trans} (influx volume transfer constant from plasma to extra-vascular, extracellular space, which represents the capillary permeability, units: min^{-1}). K^{trans} maps were calculated using the Extended Tofts model (Tofts et al, 1999). Of note, AUC was not evaluated for scan duration because we focused on the first 90 seconds of the time-intensity curve.

Five out of the 30 MRI examinations were excluded due to Dixon fat-water swap artifacts (n=3) and motion artifacts (that were too large to be well corrected with the motion correction, n=2) biasing the quantitative assessment.

Table 4.2-1. MRI texture features extracted from K^{trans} and AUC maps.

First-order Feature	Grey level co-occurrence matrix	Grey level run length matrix	Grey level size zone matrix
Average	Cluster prominence	GL non-uniformity	GL non-uniformity
Energy	Cluster shade	GL variance	GL variance
Entropy	Cluster tendency	High GL run emphasis	Large area emphasis
Inter-quartile range	Contrast	Low GL run emphasis	Small area emphasis
Kurtosis	Correlation	Long run emphasis	Size zone non-uniformity
Skewness	Inverse difference moment	Run entropy	Zone entropy
Standard deviation	Joint average	Run length non-uniformity	Zone variance
	Joint energy	Run variance	
	Joint entropy	Short run emphasis	

NOTE. Abbreviations: GL: grey level

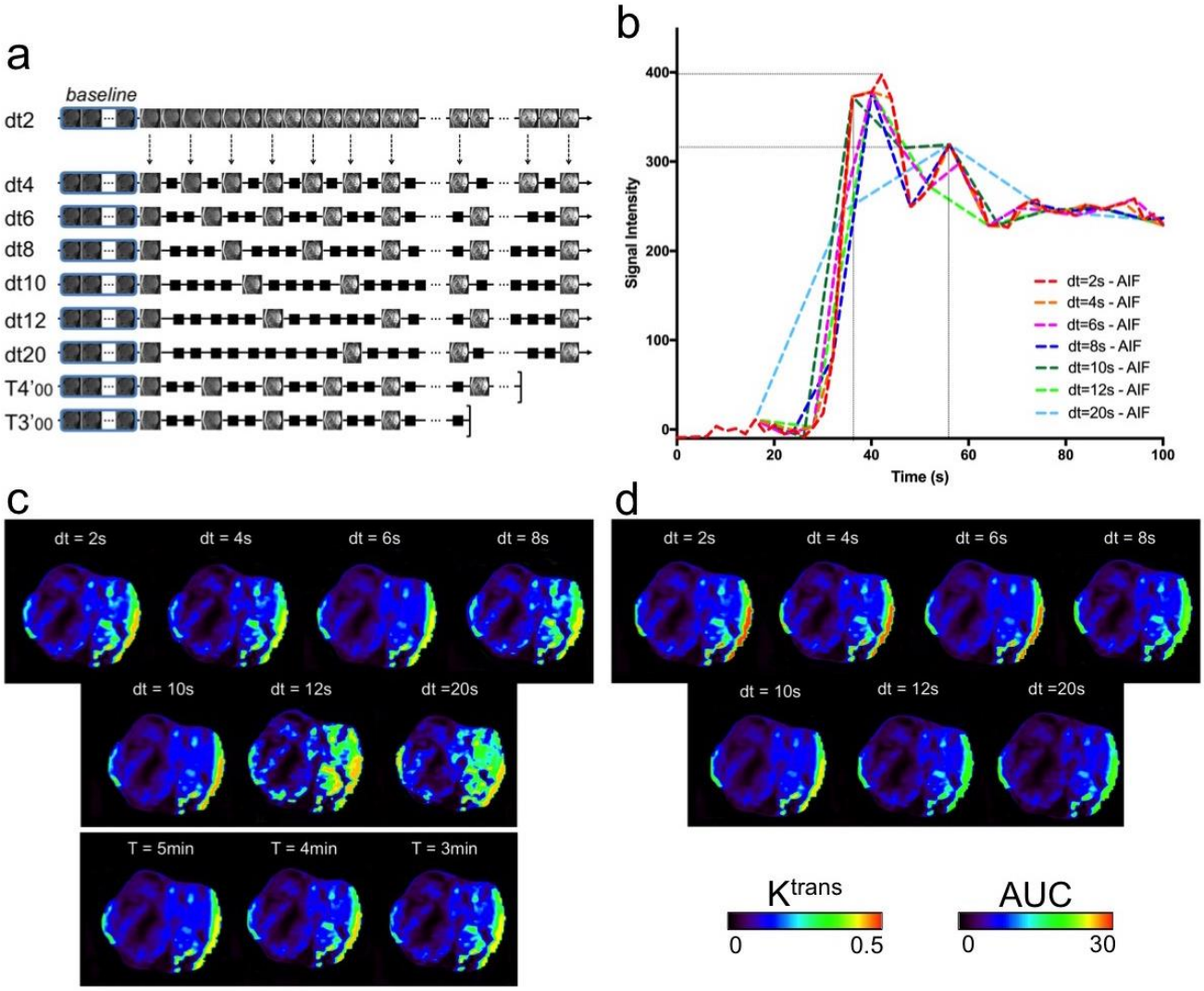
Data analysis

Texture analysis. For each patient, the entire tumor volume was manually segmented by the same senior radiologist using OleaSphere®, slice-by-slice, on the last phase of the DCE-MRI acquisition (because the tumor demonstrated the best contrast with surrounding tissues compared with the other phases) and with the help of the conventional sequences acquired during the same MRI examination, i.e.: axial turbo spin echo (TSE) T2-weighted imaging (refocusing angle = 150°, repetition time = 6860 ms, echo time = 120 ms, slice thickness = 4 mm, field-of-view = 250 x 250 mm), coronal or sagittal TSE short time inversion recovery T2-weighted imaging (refocusing angle = 150°, repetition time = 4150 ms, echo time = 69 ms, inversion time = 150 ms, slice thickness = 4 mm, field-of-view = 350 x 350 mm), axial TSE T1-weighted imaging (refocusing angle = 150°, repetition time = 420 ms, echo time = 10 ms, slice thickness = 4 mm, field-of-view = 250 x 250 mm), pre- and post-contrast agent injection axial Fat-Sat TSE T1-weighted imaging (refocusing angle = 180°, repetition time = 581 ms, echo time = 10 ms, slice thickness = 4 mm, field-of-view = 250 x 250 mm).

The voxels located at the extreme edge of the tumor were excluded to avoid partial volume effect. The volume of interest (VOI) was then propagated on all the parametric maps (K^{trans} and AUC for the 7 samplings, K^{trans} for the 3 truncations). Texture analysis consisted in the 3D extraction of 32 quantitative features, provided in Table 1, from histogram analysis (: first order statistics, FOS), grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRM) and grey-level size zone matrix (GLSZM) (Haralick et al, 1973; Galloway et al, 1975; Thibault et al, 2013). Before, K^{trans} and AUC values were discretized in 256 fixed bins between the minimum and maximum observed in the series for each parameter. The spatial offset was fixed to a displacement of 4 pixels and an angle of 45°. No technique for resampling or standardization was applied because the voxel size was initially the same for all patients and texture analysis was performed on scaled parametric maps and not on raw MRI data.

For each patient 'p' ($i \in \{1;2;\ldots;N_p\}$ where N_p is the total number of included patients), for each feature 'F_i' ($i \in \{1;2;\ldots;32\}$) and for each DCE-MRI map 'X ($X \in \{K^{trans};AUC\}$), we extracted 7 paired values of $F_{p,i,X}$ regarding sampling and 3 paired values of $F_{p,i,X}$ regarding scan duration ($F_{p,i,X}(dt)$ and $F_{p,iX}(T)$, respectively).

Figure 4.2-1 : Reconstruction of data acquisition. (**a**) Retrospective downsampling and truncating were performed to obtain post-processed data with a temporal resolution of 2s (: dt2), 4s (: dt4), 6s (: dt6s)… and 20s (: dt20), and with a scan duration of 4min (: T4'00 – with a temporal resolution of 6s) and 3min (: T3'00 - with a temporal resolution of 6s). Of note, the baseline was not re-sampled. (**b**) The same voxels in the same artery were used to determine the artery input function (AIF) that was used in the extended Tofts model. However, AIFs were also truncated and downsampled leading to variations in time to reach the peak (from 36s to 54s) and in maximum signal intensity of the peak (from 319 to 396.8) in this example of high-grade synovial sarcoma of the popliteal region. (**c**) Next, parametric maps of K^{trans} (: influx volume transfer constant) and (**d**) AUC (: area under the time intensity curves at 90s after arrival of the contrast agent bolus in the acquisition volume) were generated as functions of 'dt' and 'T'.



Coefficient of variation (CV). We calculated the mean CV and its standard deviation (SD) on each feature F_i in order to assess the influence of sampling (: CV-dt, with AUC or K^{trans}), and scan duration (: CV-T, with K^{trans}) on the dispersion of the texture features, as follows:

CV-dt= $\sum_{p=1}^{Np}(SD(F_{i,p,dt2}; F_{i,p,dt4}; \ldots; F_{i,p,dt20})/mean(F_{i,p,dt2}; F_{i,p,dt4}; \ldots; F_{i,p,dt20}))/Np$

CV-T= $\sum_{p=1}^{Np}(SD(F_{i,p,T3}; F_{i,p,T4},; F_{i,p,T5})/mean(F_{i,p,T3}; F_{i,p,T4}; F_{i,p,T5}))/Np$

We defined a texture feature as very poorly variable if CV<0.1, poorly variable if CV $\in$ [0.1-0.2[, mildly variable if CV $\in$ [0.2-0.5[, highly variable if CV $\in$ [0.5-1[and extremely variable if CV≥1.0.

Statistical analysis

Statistical analyses were performed using SPSS (IBM corp, version 21.0, Armonk, NY) and GraphPad Prism (GraphPad Software, version 7, San Diego, CA). Variables were expressed as average, standard deviation, median, and range, as appropriate. A p-value < 0.05 was deemed significant. All tests were two-tailed.

Normality was assessed for each continuous value by using the Shapiro-Wilk test.

Repeated-measures (rm-) ANOVA with Geisser-Greenhouse correction for non-sphericity and post-hoc Tukey tests with corrections for multiple comparisons using statistical hypothesis testing (or non-parametric equivalent rm-Friedman tests with post-hoc Dunn tests) were used to study the influence of sampling and truncating on texture features extracted from K^{trans} and AUC maps.

For texture features with a statistical influence of temporal resolution, a linear model was applied for correlation analysis. A Bonferroni correction was applied for multiple comparisons and a p-value of 0.01 was considered significant.

Finally, we investigated the influence of temporal parameters on the performance of a model aiming at predicting the treatment response based on texture features from K^{trans} and AUC maps. Twenty patients were analyzable for this part of the study; the 5 others were either operated directly after the MRI without chemotherapy or were not sufficiently followed-up. Texture features that were associated with the response at univariate level according to Student t-tests (or Mann-Whitney tests) with a p-value <0.05 were entered in a multivariate binary logistic regression using a backward stepwise selection method based on the probability of the Wald statistics. The area under the receiver operating characteristics curve (AUROC) of the models were

calculated according to scan durations and temporal resolutions and compared according to the Delong methods (Delong et al, 1988).

RESULTS

Patients (Table 4.2-2)

Twenty-five patients were finally included (9/25 (36%) women, median age 68 years old, range: 31-94). The most frequent histotype was undifferentiated pleomorphic sarcoma. The median size was 81 mm (range: 42 - 180) and the median tumor volume was 0.357 L (range: 0.023 – 1.048).

Table 4.2-2. Epidemiological features of the population study.

Characteristics	Patients
Gender	
Men	16/25 (64%)
Women	9/25 (36%)
Age (years)	68 (31-94)
Histological types	
Undifferentiated pleomorphic sarcoma	13/25 (52%)
Myxoid/round cells liposarcoma	4/25 (16%)
Rhabdomyosarcoma	2/25 (8%)
Myxofibrosarcoma	1/25 (4%)
Synovial sarcoma	2/25 (8%)
Pleomorphic liposarcoma	1/25 (4%)
Undifferentiated sarcoma - others	2/25 (8%)
Location	
Upper limb	4/25 (16%)
Shoulder girdle	1/25 (4%)
Trunk wall	2/25 (8%)
Lower limb	18/25 (72%)
Size (mm)	81 (42-180)
Number of voxels	356 959 (22 816-1 048 136)
Volume (L)	0.357 (0.023-1.048)
Tumor depth	
Deep	16/25 (64%)
Deep and superficial	8/25 (32%)
Superficial	1/25 (4%)

NOTE. Data are number of patients with percentage in parentheses, except for age, size, number of voxels and volume of the tumor, given as median and range.

Influence of temporal parameters on DCE-MRI texture features

Table 4.2-3 provides the results of the rm-ANOVA (or rm-Friedman test). Regarding the effect of temporal resolution, 12/32 (37.5%) AUC-based texture features were significantly influenced: 5/7 FOS features, 1/9 GLCM features, 2/9 GLRLM features and 4/7 GLSZM features (range of p-value=0.0395 to <0.0001). Fourteen out of 32 (43.8%) K^{trans}-based texture features were significantly influenced by temporal resolution: 1/7 FOS features, 5/9 GLCM features, 4/9 GLRLM features and 4/7 GLSZM features (range of p-value = 0.0331 to 0.0007).

Regarding the effect of scan duration, 23/32 (71.9%) K^{trans}-based texture features were significantly influenced: 7/7 FOS features, 8/9 GLCM features, 6/9 GLRLM features and 2/7 GLSZM features (range of p-value = 0.0321 to < 0.0001).

Three texture features were influenced by temporal parameters in the 3 settings: GLRLM_GL_Variance, GLSZM_Large_area_emphasis, and GLSZM_Zone_variance. Six texture features were not, no matter the setting: GLCM_Correlation, GLRLM_Gray_level_non-uniformity, GLRLM_Run_entropy, GLRLM_Run_variance, GLSZM_Small_area_emphasis and GLSZM_Zone_entropy.

A summary of post-hoc tests is given in Table 4.2-4. Regarding the influence of temporal resolution on AUC maps, the highest number of texture features that statistically differed was observed in post-hoc comparisons between dt = 20s and dt = 2s (10/32, 31.3%) and dt = 20s and dt = 4s (9/32, 28.1%). On K^{trans} maps, the highest differences were seen between dt = 20s and dt = 2s (10/32, 59.4%), followed by dt = 20s versus dt = 6s (6/32, 18.8%). Regarding the influence of truncating on K^{trans} maps, 23/32 (71.9%) and 17/32 (53.1%) texture features were significantly different between T = 5min and T = 4min, and between T = 5min and T = 3min, respectively.

Significant linear correlations were found between sampling on AUC and GLSZM_Size_zone_non-uniformity ($p = 0.0018$, $r^2 = 0.879$) and GLSZM_Zone_variance ($p = 0.0043$, $r^2 = 0.830$). Similarly, there were significant linear correlations between sampling on K^{trans} and 4 texture features: GLSZM_non-uniformity ($p < 0.0001$, $r^2 = 0.978$), GLSZM_Large area emphasis ($p = 0.0009$, $r^2 = 0.909$), GLSZM_Size_zone_non-uniformity ($p = 0.0002$, $r^2 = 0.954$), GLSZM_Zone_variance ($p = 0.0002$, $r^2 = 0.950$) (Figure 4.2-2).

Table 4.2-3. Assessment of the influence of temporal resolution (sampling) and scan duration (truncating) on texture parameters extracted from DCE-MRI parametric maps (K^{trans} and AUC).

Texture features	AUC - Sampling		K^{trans} - Sampling		K^{trans} - Truncating	
	F-value	p-value	F-value	p-value	F-value	p-value
First-order Feature						
Average	18.54	***0.0050***	5.12	0.5290	14.60	***0.0022***
Energy	25.55	***0.0003****	2.13	0.9072	15.00	***0.0018***
Entropy	7.15	0.3069	16.46	***0.0115****	19.19	***0.0002****
Inter-quartile range	8.607	0.1969	5.61	0.4688	15.05	***0.0018***
Kurtosis	50.89	***<0.0001****	6.35	0.0958	15.49	***0.0168****
Skewness	41.57	***<0.0001****	10.88	0.0921	8.80	***0.0321****
Standard deviation	26.12	***0.0002****	1.77	0.9395	14.50	***0.0023***
GLCM						
Cluster prominence	17.07	***0.0090***	11.46	0.0751	16.20	***0.0010***
Cluster shade	5.07	0.5354	12.15	0.0587	12.55	***0.0057***
Cluster tendency	10.30	0.1127	9.96	0.1262	23.85	***<0.0001****
Contrast	9.27	0.1589	13.7	***0.0331****	17.85	***0.0005****
Correlation	2.78	0.8358	5.59	0.4709	4.25	0.2357
Inverse difference moment	6.58	0.3617	13.70	***0.0331****	17.85	***0.0005****
Joint average	1.94	0.9253	13.96	***0.0301****	18.75	***0.0003****
Joint energy	6.45	0.3750	13.83	***0.0316****	18.40	***0.0004****
Joint entropy	6.87	0.3327	16.64	***0.0233****	19.55	***0.0002****
GLRLM						
GL non-uniformity	0.96	0.9869	5.83	0.4427	1.05	0.7892
GL variance	20.74	***0.0020***	17.7	***0.0070****	29.30	***<0.0001****
High GL run emphasis	5.74	*0.4525*	19.16	***0.0039***	29.30	***<0.0001****
Low GL run emphasis	3.27	0.7746	19.16	***0.0039***	29.30	***<0.0001****
Long run emphasis	11.00	0.0884	17.05	***0.0091***	31.50	***<0.0001****
Run entropy	0.32	0.8632§	1.40	0.2556§	0.9251	0.4056§
Run length non-uniformity	7.32	0.2918	11.46	0.0753	14.85	***0.0019***
Run variance	10.47	0.1062	11.36	0.0778	5.95	0.1141
Short run emphasis	13.23	***0.0395****	1.56	0.2133	22.95	***<0.0001****
GLSZM						
GL non-uniformity	9.69	0.1384	15.31	***0.0180****	10.94	0.012
GL variance	25.72	***0.0003****	11.89	0.0645	10.18	0.0171
Large area emphasis	19.29	***0.0037***	21.41	***0.0015***	15.30	***0.0016***
Small area emphasis	9.16	0.1649	4.91	0.5555	4.40	0.2214
Size zone non-uniformity	16.11	***0.0132****	16.87	***0.0098****	5.87	0.118
Zone entropy	3.08	0.2190§	4.746	0.5768	1.01	0.3797§
Zone variance	20.26	***0.0025***	23.46	***0.0007****	15.30	***0.0016***

NOTE. Data are F-values and p-values for the repeated measures (rm-) ANOVA (: §) or non parametric equivalent rm-Friedman test.
Abbreviations: GL : grey-level; GLCM : grey-level occurrence matrix; GLRLM: grey-level run length matrix; GLSZM: grey level size zone matrix
*: $p \leq 0.05$; **: $p < 0.005$, ***: $p < 0.001$

Table 4.2-4. Summary of the post-hoc tests: number of texture features that were significantly different when comparing 2 distinct temporal resolution (: sampling) for AUC (**a**) and K^{trans} (**b**), or 2 distinct scan durations (: truncating) for K^{trans} (**c**).

a. AUC - Sampling

	dt2	dt4	dt6	dt8	dt10	dt12	dt20
dt2	-	0	0	2	4	1	10
dt4	0	-	0	2	2	0	9
dt6	0	0	-	0	0	0	2
dt8	2	2	0	-	0	0	0
dt10	4	2	0	0	-	0	0
dt12	1	0	0	0	0	-	0
dt20	10	9	2	0	0	0	-

b. K^{trans} - Sampling

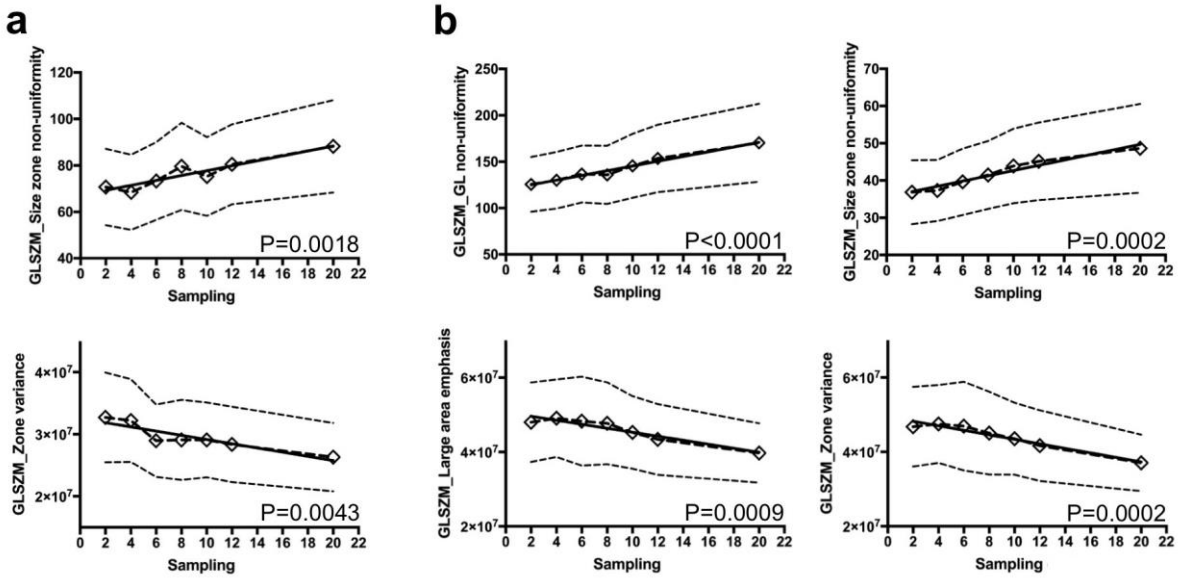
	dt2	dt4	dt6	dt8	dt10	dt12	dt20
dt2	-	0	0	0	0	0	10
dt4	0	-	0	0	0	0	2
dt6	0	0	-	0	0	0	6
dt8	0	0	0	-	0	0	0
dt10	0	0	0	0	-	0	0
dt12	0	0	0	0	0	-	0
dt20	10	2	6	0	0	0	-

c. K^{trans} - Truncating

	T3'00	T4'00	T5'00
T3'00	-	0	17
T4'00	0	-	23
T5'00	17	23	-

NOTE. Abbreviations: dtx corresponds to a sampling of 'x' seconds; Ty corresponds to a scan duration of 'y' minutes.

Figure 4.2-2: Linear correlations between texture features from DCE-MRI parametric maps and temporal resolution. Only significant correlations for AUC (**a**) and K^{trans} (**b**) after corrections to allow multiple comparisons are shown with their p-value. Sampling is given in seconds. Texture features have arbitrary units. Abbreviations: GL: grey level; GLSLZM: grey level size zone matrix.



Variations in scan duration did not reveal any linear correlations. Details for linear regressions are given in Annexe 6 - Supplementary Data 1. Figure 4.2-3 shows an example of variation of parametric maps and corresponding histograms with changes in scan duration and temporal resolution.

Effect of temporal parameters on the dispersion of DCE-MRI texture features

Figure 4.2-4 and Table 4.2-5 summarize the analysis of dispersion of texture features according to the 3 configurations (i.e. AUC-sampling, K^{trans}-sampling, K^{trans}-truncating). Most texture features extracted from AUC maps remained poorly to mildly variable with changes in sampling. Only GLCM_Cluster_prominence was categorized as highly variable (CV-dt = 0.50), while 11 out of 32 (34.4%) texture features extracted from K^{trans} were highly to extremely variable (range of CV-dt = 0.50 - 1.11).

Five texture features extracted from K^{trans} maps were highly variable with changes in scan duration: FOS_Average, FOS_Energy, FOS_Inter-quartile_range, FOS_Standard_deviation and GLSZM_Zone_entropy (range of CV-T = 0.54 - 0.96). Annexe 6 - Supplementary Data 2 provides all the values of CV with SD.

Figure 4.2-3: Illustrated case of the influence of temporal parameters on DCE-MRI parametric maps and texture features of sarcoma. A 63 years old male with a high-grade, deep and superficial, myxofibrosarcoma of the left thigh underwent his baseline MRI examination including conventional sequences (**a**) and DCE-MRI acquisition. (**b**) The whole tumor volume was manually segmented, slice-by-slice, on the last phase of the DCE-MRI acquisition, i.e. 300s after Gadolinium chelates intravenous injection. (**c**) The AUC (: area under the time intensity curve at 90s after arrival of the contrast agent bolus in the acquisition volume, units: mmol/L/s, on the right) and K^{trans} (influx volume transfer constant, units: /s, on the left) parametric maps were reconstructed with the different scan durations and temporal resolutions. After the whole tumor volume segmentation, frequency histograms were reconstructed for (**d**) AUC depending on the different temporal resolution, (**e**) Ktrans depending on the different temporal resolution, and (**f**) Ktrans according to the different scan duration. Abbreviation: STIR T2-WI: short time inversion recovery T2-weighted imaging, T2-WI: T2 weighted imaging; FS CE-T1-WI: fat saturation contrast enhanced T1 weighted imaging; DCE-MRI t=300s: last phase of the dynamic contrast enhanced MRI, on which was segmented the tumor volume, 'dtx' corresponds to a sampling of 'x' seconds; 'Ty' corresponds to a scan duration of 'y' minutes.

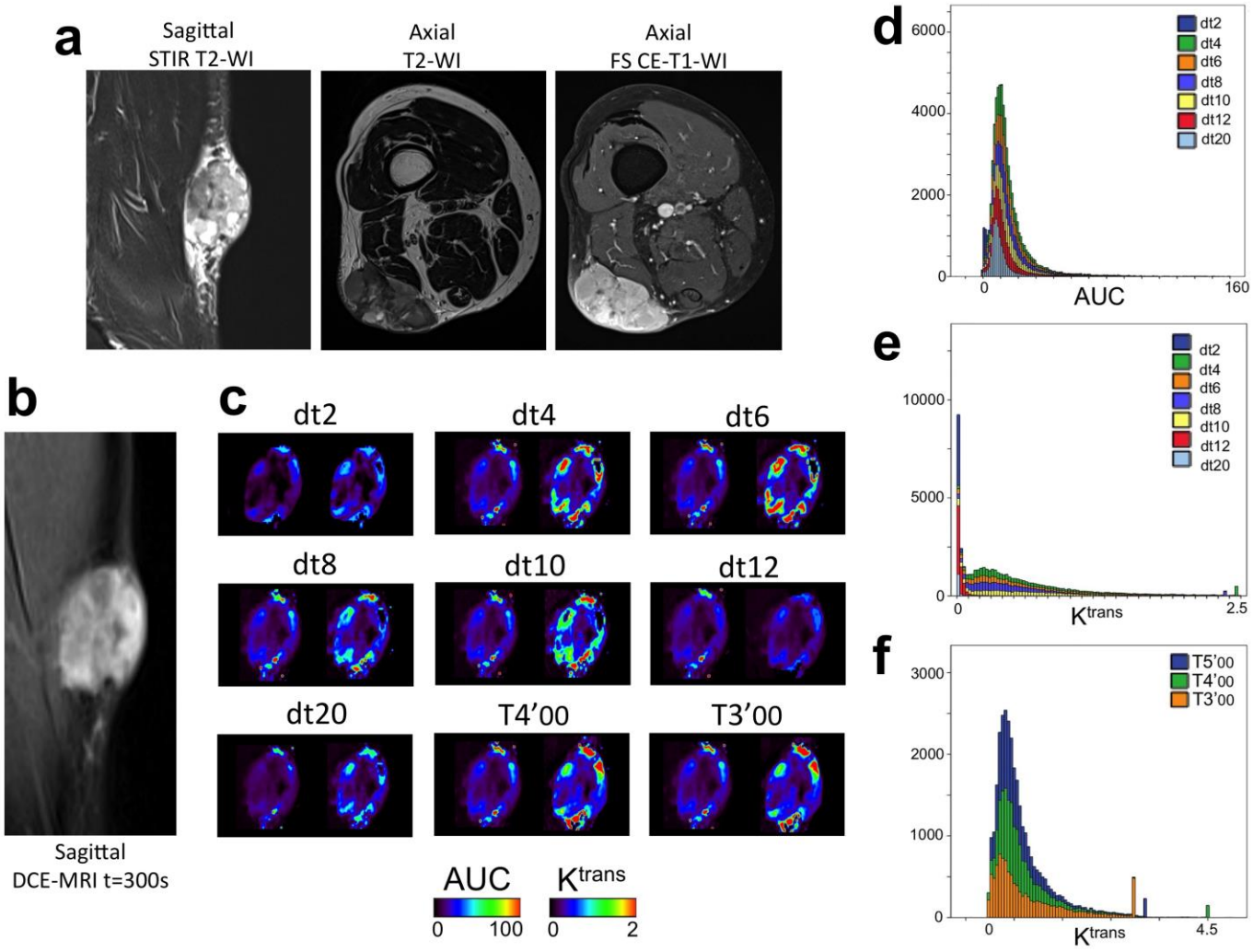


Figure 4.2-4 Coefficient of variation of the texture features extracted from DCE-MRI depending on temporal parameters. Influence of the temporal resolution of DCE-MRI acquisition (: sampling) on the dispersion of each texture feature from (**a**) AUC (: area under the time intensity curve) and (**b**) K^{trans} (: influx volume transfer constant) maps. (**c**) Influence of the scan duration (: truncating) on the dispersion of each texture feature extracted from K^{trans} map. Results are given with standard deviation. GL: grey level.

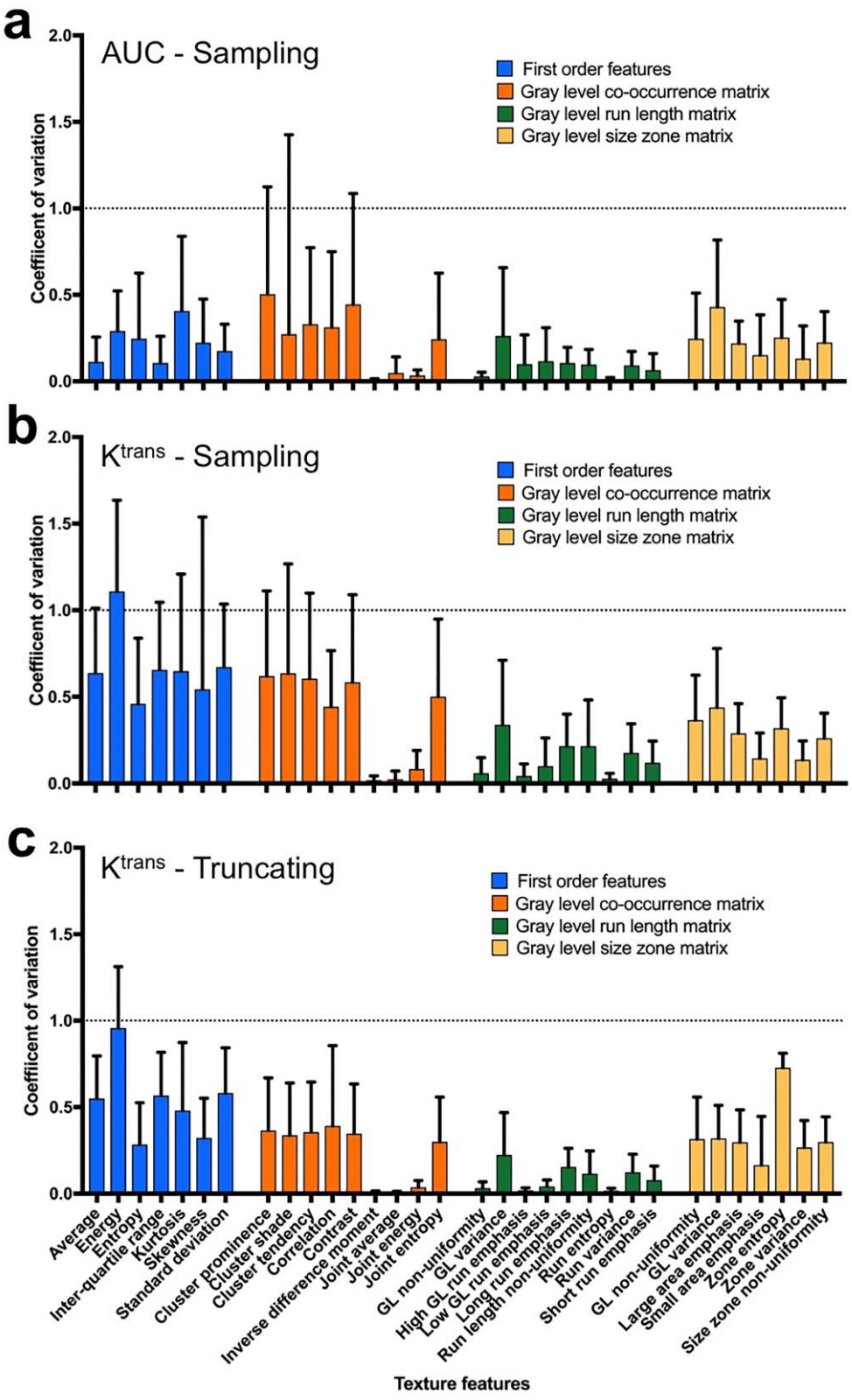


Table 4.2-5. Degree of dispersion of the texture features from K^{trans} and AUC maps according to temporal resolution (: sampling) and scan duration (: truncating).

	Very poorly variable CV < 0.10	**Poorly variable 0.10 ≤ CV < 0.20**	**Mildly variable 0.20 ≤ CV < 0.50**	**Highly variable 0.50 ≤ CV ≤ 1**	**Extremely variable CV ≥ 1.00**
AUC - Sampling	GLCM_Inverse Difference Moment GLCM_Joint average GLCM_Joint energy GLRLM_GL non-uniformity GLRLM_High GL run emphasis GLRLM_Run length non-uniformity GLRLM_Run entropy GLRLM_Run variance GLRLM_Short run emphasis	FOS_Average FOS_Inter-quartile range FOS_Standard deviation GLRLM_Low GL run emphasis GLRLM_Long run emphasis GLSZM_Small area emphasis GLSZM_Zone entropy	FOS_Energy FOS_Entropy FOS_Kurtosis FOS_Skewness GLCM_Cluster shade GLCM_Cluster tendency GLCM_Correlation GLCM_Contrast GLCM_Joint entropy GLRLM_GL variance GLSZM_GL non-uniformity GLSZM_GL variance GLSZM_Large area emphasis GLSZM_Zone variance GLSZM_Size zone non-uniformity	GLCM_Cluster prominence	
K^{trans} - Sampling	GLCM_Inverse Difference Moment GLCM_Joint average GLCM_Joint energy GLRLM_GL non-uniformity GLRLM_High GL run emphasis GLRLM_Low GL run emphasis GLRLM_Run entropy	GLRLM_Run variance GLRLM_Short run emphasis GLSZM_Small area emphasis GLSZM_Zone entropy	FOS_Entropy GLCM_Correlation GLRLM_GL variance GLRLM_Long run emphasis GLRLM_Run length non-uniformity GLSZM_GL non-uniformity GLSZM_GL variance GLSZM_Large area emphasis GLSZM_Zone variance GLSZM_Size zone non-uniformity	FOS_Inter-quartile range FOS_Kurtosis FOS_Average FOS_Skewness FOS_Standard deviation GLCM_Cluster prominence GLCM_Cluster shade GLCM_Cluster tendency GLCM_Contrast GLCM_Joint entropy	FOS_Energy
K^{trans} - Truncating	GLCM_Inverse Difference Moment GLCM_Joint average GLCM_Joint energy GLRLM_GL non-uniformitY GLRLM_High GL run emphasis GLRLM_Low GL run emphasis GLRLM_Run entropy GLRLM_Short run emphasis	GLRLM_Long run emphasis GLRLM_Run length non-uniformity GLRLM_Run variance GLSZM_Small area emphasis	FOS_Entropy FOS_Kurtosis FOS_Skewness GLCM_Cluster prominence GLCM_Cluster shade GLCM_Cluster tendency GLCM_Correlation GLCM_Contrast GLCM_Joint entropy GLRLM_GL variance GLSZM_GL non-uniformity GLSZM_GL variance GLSZM_Large area emphasis GLSZM_Zone variance GLSZM_Size zone non-uniformity	FOS_Average FOS_Energy FOS_Inter-quartile range FOS_Standard deviation GLSZM_Zone entropy	

NOTE. Abbreviations: CV: coefficient of variation; FOS: First order statistics; GL: grey level; GLCM: grey level co-occurrence matrix; GLRLM: grey level run length matrix; GLSZM: grey level size zone matrix

Effect of temporal parameters on a prediction based on DCE-MRI texture features

Twenty patients were treated with chemotherapy and were analyzable for this sub-part of the study. There were 5/20 (25%) good responses (4 good histological responses and 1 partial response after 6 months without changing treatment). Table 4.2-6 shows the results of univariate analyses and performances of the models quantified by AUROC. After univariate analyses, the selected variables entered in the binary logistic regression were not systematically the same when temporal resolution and scan duration changed, even if some were frequently encountered: AUC_GLRLM_Run_entropy (in 8 models), K^{trans}_GLSZM_Small_area_emphasis (in 3 models). AUROC ranged from 0.77 (CI95% = (0.54 - 1.00)) with a scan duration of 5min and a temporal resolution of 6s, to 0.90 (CI95% = (0.74 - 1.00)) with a scan duration of 4min and a temporal resolution of 6min. Details of the univariate analyses for each model and comparisons between AUROC can be found in Annexe 6 - Supplementary Data 3 and Annexe 6 - Supplementary Data 4, respectively.

DISCUSSION

In this study, we examined the influence of temporal parameters on a large set of widely used and easily available statistical texture features extracted from whole tumor volumes segmented on K^{trans} and AUC parametric maps. We found that a large number of them were significantly dependent on the scan duration and temporal resolution. Moreover, merely half of them remained poorly variable with changes in temporal parameters. This led to non-negligible variations in the AUROCs of the models for a response prediction to chemotherapy, even though the comparisons did not provide statistical differences.

A major challenge for radiomics approaches is to ensure a good quality of the quantitative data on which they rely. Besides reproducibility, repeatability, non-redundancy and validity, quality here means controlling the bias due to slight variations in the imaging acquisition parameters. Standardizing the imaging protocols between radiological centers is unavoidable. However, thousands of MRI examinations (and DCE-MRI sequences) have already been acquired and the temptation to pool data from different MR-systems in order to begin to build and test

Table 4.2-6. Area under the ROC curves of models for response prediction to chemotherapy based on texture features extracted from AUC and K^{trans} maps with varying temporal resolution (sampling) and scan duration (truncating)

Temporal parameters	Univariate analysis		Model	
	Significant features	p-value	AUROC (CI95%)	p-value
◆ **Sampling**				
dt2	K^{trans}_GLRLM_Run entropy	0.021	0.85 (0.68-1.00)	0.021
	AUC_GLRLM_Run entropy	0.050		
dt4	AUC_GLRLM_Run entropy	0.050	0.84 (0.62-1.00)	0.026
	K^{trans}_FOS_Inter-quartile range	0.026		
	K^{trans}_FOS_Standard deviation	0.013		
	K^{trans}_GLRLM_Long run emphasis	0.021		
	K^{trans}_GLRLM_Run entropy	0.032		
dt6	AUC_GLRLM_Run entropy	0.050	0.77 (0.54-1.00)	0.074
dt8	AUC_GLCM_Correlation	0.032	0.83 (0.61-1.00)	0.032
dt10	K^{trans}_GLRLM_Long run emphasis	0.032	0.83 (0.63-1.00)	0.032
dt12	AUC_GLCM_Correlation	0.040	0.83 (0.58-1.00)	0.032
	AUC_GLRLM_Run entropy	0.032		
	K^{trans}_GLSZM_Small area emphasis	0.050		
dt20	AUC_GLRLM_Run entropy	0.049	0.81 (0.58-1.00)	0.121
◆ **Truncating**				
T3'00	AUC_GLRLM_Run entropy	0.050	0.88 (0.71-1.00)	0.021
	K^{trans}_GLSZM_Small area emphasis	0.021		
T4'00	AUC_GLRLM_Run entropy	0.050	0.90 (0.74-1.00)	0.018
	K^{trans}_GLSZM_Small area emphasis	0.016		
T5'00	AUC_GLRLM_Run entropy	0.050	0.77 (0.54-1.00)	0.074

NOTE. Abbreviations: AUROC: area under the receiver operating characteristics curve; CI95%: 95% confidence interval; GL: grey level; GLCM: grey level co-occurrence matrix; GLRLM: grey level run length matrix; GLSZM: grey level size zone matrix; 'dtx' corresponds to a sampling of 'x' seconds; 'Ty' corresponds to a scan duration of 'y' minutes.

predictive models for key oncological questions is understandable. Our study focused on STS as a tumor model to examine the influence of temporal parameters. High-grade STS are characterized by complex architectures and changes during treatments, making them particularly appropriate for radiomics. In particular, Vallières *et al.* stressed the influence of the post-processing parameters of imaging on a prediction model of occurrence of lung metastases in STS patients (Vallières et el, 2015).

Our results are in agreement with previous studies that investigated the influence of temporal parameters on dynamic acquisitions. Othman *et al.* showed that shorter scan duration was responsible for overestimation of pharmacokinetic parameters and lead to incorrect classifications of benign prostate lesions as malignant (Othman et al, 2016). A similar influence on prediction models was found with breast lesions (Hao et al, 2015). This could be explained by the Tofts model itself because it assumes an immediate equilibrium between the compartments though it requires up to 2min in case of breast imaging (Kuhl et al, 1999). Poor temporal resolutions can also lead to incorrect assessments of pharmacokinetic parameters according to pre-clinical and clinical studies regarding prostate lesions – even if it did not significantly modify the ability of these parameters to discriminate benign and malignant tumors (Othman et al, 2016). Heisen *et al.* (2010) showed that the K^{trans} variations could go up to 25% of its initial value with varying temporal resolutions. In the single study that focused on the acquisition parameters at risk of influencing perfusion maps, Bogowicz *et al.* found that the percentage of unstable texture features ranged from 56 to 98% with different artery contouring, and from 58 to 75% with different temporal resolution, which is higher than in our study (Bogowicz et al, 2016). This highest variability in their study may be explained by the difference in histological types, in imaging modality (CT vs. MRI), in the texture features that were calculated and in the perfusion maps that were studied (blood flow, mean time transit and blood volume vs. AUC and K^{trans} in our studies).

Regarding STS, our results also highlight potential cut-offs for temporal resolution and scan duration beyond which statistical differences in texture features from DCE-MRI maps may occur. Indeed, even if temporal resolution had an influence on several texture features, post-hoc tests with correction for multiple comparisons showed that data with temporal resolution below 8s could be pooled, as well as data with temporal resolution above 8s. On the contrary, we found that data obtained with a scan duration of 5 min should not be pooled with those with a scan duration of 3 or 4 min.

After correction of multiple tests, some radiomics features extracted from AUC and K^{trans} maps demonstrated linear relationships with sampling. All belonged to the GLSZM category. GLSZM_Size_zone_non_uniformity and GLSZM_Grey_level_non_uniformity increased with downsampling, suggesting more heterogeneity in intensity values and in size zone volumes of the same grey level voxels with low temporal resolutions (high dt). Conversely, GLSZM_Large_area_emphasis and GLSZM_Zone_variance tended to decrease, suggesting that poor temporal resolution (high dt) led to smaller size zones. No linear correlation was found with truncating, probably because we only generated 3 time points. Poor temporal resolution (high dt) can be responsible for less accurate estimations, especially when the tumor – or areas in the tumor –demonstrates a rapid and strong enhancement (wash-in and peak on time-intensity curves). Consequently, hypervascularized intra-tumoral areas with high K^{trans} and AUC values could be missed, resulting in areas with lower values of K^{trans} and AUC.

All the categories of texture features seemed equally influenced by temporal resolution, but the dispersion was more marked with FOS, then GLCM, GLRLM and finally GLSZM. The high sensitivity of FOS could be due to their lack of spatial information. Indeed, all the voxels are pooled without considering the fact that adjacent voxels may react similarly and proportionally to changes in temporal parameters, which is the case of GLCM, GLSZM and GLRLM.

Regarding the influence of temporal parameters on the performance of a model for response prediction based on AUC and K^{trans} of STS, we identified 6 relevant features that were selected across the different models, namely: FOS_Inter_quartile_range, FOS_Standard_deviation, GLCM_Correlation, GLRLM_Long_run_emphasis, GLRLM_Run_entropy and GLSZM_Small_area_emphasis. Interestingly, GLRLM_Run_entropy and GLRLM_Small_area_emphasis were not significantly influenced by temporal resolution and showed low CVs. Thus, these texture features could be good candidates for multi-center studies based on DCE-MRI sequences of STS. However, these results do not mean that ultra-fast DCE-MRI acquisitions are unnecessary. The aim of DCE-MRI is to represent and estimate at best the vascular characteristics of tumors. Hypervascularized tumors require excellent temporal resolutions. If some researchers want to increase the statistical power of their radiomics study by putting together DCE-MRIs with different temporal parameters,

then it should be carefully done because it could introduce a significant bias in their results.

We did not investigate the test-retest reproducibility of DCE-MRI-based texture features because it was hardly justifiable to inject a contrast agent twice in cancer patients. However, further radiomics studies performed on perfusion phantoms could help analyzing this crucial aspect. It should be noted that we designed our study to limit bias that could have been introduced in the quantification of heterogeneity: we utilized the same 1.5T MR-system, the same DCE-MRI sequences with the same acquisition parameters following QIBA® recommendations, the same contrast agent, the same imaging filters, the same software for post-processing and the same feature extraction parameters. Nevertheless, we had to adjust the coils because of variations in tumor sizes and locations, which may have introduced some bias. Moreover, truncating and downsampling the AIFs probably contributed to the variations in texture features from DCE-MRI parametric maps, even if we kept the same voxels in the same artery of interest. Indeed, the AIF is crucial in pharmacokinetic modeling (Tofts et al, 1999). For instance, a recent study showed that K^{trans} could range from 0.25/min to more 2/min in prostate cancers in a same series of DCE-MRI acquisitions depending on the method to determine the AIF from different cancer centers, which led to variations in times to peak and peak amplitudes (Huang et al, 2016b). Herein, poor temporal resolutions could lead to missing the real AIF peak though the K^{trans} estimation strongly relies on it, as well as to superposition of the AIF and the tumor enhancement curve during the first part of the acquisition.

Our study has several limits. First, the population was small and made of heterogeneous histological types, though all were high-grade. Some tumors enhanced rather homogeneously and progressively (for instance myxoid/round cell liposarcomas) while others showed more heterogeneous enhancements, with the time-intensity curves of some components displaying a strong wash-in, followed by a peak and a wash-out (for instance, undifferentiated pleomorphic sarcomas). Temporal parameters were certainly more influential in the second case. Nevertheless, until now, the inclusion of STS patients in clinical trials relies on tumor grade and not on histotypes. Second, we did not investigate other semi-quantitative and pharmacokinetic DCE-MRI parameters (such as Time-to-peak, Wash-in, Wash-out, K^{ep} [efflux rate constant from the extra-cellular, extra-vascular space to the plasma

compartment], K^{el} [contrast agent elimination rate constant], V_e [extra-cellular, extra-vascular space volume], V_p [plasmatic volume]), or other perfusion models than extended Tofts. We decided to focus on the most studied parameters in the literature. However, given that some areas in STS can show a wash-out, it is also possible that the pharmacokinetics that quantify the decreasing part of the time-intensity curve are also influenced by temporal parameters. Third, alternative methods for downsampling could have been applied by recombining k-space data instead of removing some phases (Heisen et al, 2010). In a clinical setting, there is a compromise between temporal resolution, signal-to-noise ratio and spatial resolution. A decrease in temporal resolution will benefit signal-to-noise-ratio (by averaging twice if there is a down-sampling by a factor 2, for instance). Fourth, our multivariate models for the outcome sub-study can be questioned since only 20 patients were included with 2 definitions of good treatment responses depending on the patient's operability. We used a classical statistical approach without validation cohort to build the predictive models. More advanced selection methods could have been used (for instance: least absolute shrinkage and selection operator, ElasticNet, supervised principal component analysis), as well as supervised machine-learning classifiers (for instance: random forest, k-nearest neighbors, support vector machines) or deep learning. Consequently, we did not test if the influence of temporal parameters was still present with these other statistical methods. The aim of this part of the study was to illustrate the influence of temporal parameters on a prediction and not to validate a radiomics model for response prediction. However, we hope that further prospective studies will take into account this research aiming at improving the quality of radiomics methods in order to build clean and robust models to answer key oncological questions.

To conclude, our study screened several aspects of the influence of temporal parameters on texture features extracted from DCE-MRI parametric maps of STS. We showed that both scan duration and temporal resolution introduced a non-negligible variability in the quantification of heterogeneity that could lead to a decreased performance of prediction models for response to chemotherapy. In addition to all the other acquisition and post-processing parameters, standardizing the scan duration and temporal resolution of DCE-MRI must be considered in prospective multi-centric trials to build reliable radiomics approaches.

*

* *

4.2.3. Article 5: influences des paramètres temporels sur les indices radiomics issus des séquences DCE-MRI et influence sur une prédiction clinique

L'article issu de ce travail a été publié en janvier 2020 dans *Journal of Magnetic Resonance Imaging* (PMID: 31922323, doi: 10.1002/jmri.27040 - rang B SIGAPS, IF = 3.732). Il s'agit ici du manuscrit final reformaté.

ORIGINAL RESEARCH

High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models?

Amandine Crombé, MD MSc,[1,2,3*] David Fadli, MD,[1] Xavier Buy, MD,[1] Antoine Italiano, MD PhD,[3,4] Olivier Saut, PhD,[2,3] and Michèle Kind, MD MSc[1]

Background: Heterogeneity on pretreatment dynamic contrast-enhanced (DCE)-MRI of sarcomas may be prognostic, but the best technique to capture this characteristic remains unknown.
Purpose: To investigate the best method to extract prognostic data from baseline DCE-MRI.
Study Type: Retrospective, single-center.
Population: Fifty consecutive uniformly-treated adults with nonmetastatic high-grade sarcomas.
Field Strength/Sequence: 1.5T; T_2-weighted-imaging, fat-suppressed fast spoiled gradient echo DCE-MRI.
Assessment: Ninety-two radiomics features (RFs) were extracted at each DCE-MRI phase (11, from t = 0–88 sec). Relative changes in RFs (rRFs) since the acquisition baseline were calculated (11 × 92 rRFs). Curves of rRF as function of time post-injection were integrated (92 integrated-rRFs [irRFs]). K^{trans} and area under the time–intensity curve at 88-sec parametric maps were computed and 2 × 92 parametric-RFs (pRFs) were extracted. Five DCE-MRI-based radiomics models were built on: an RFs subset (32 sec, 64 sec, 88 sec); all rRFs; all irRFs; and all pRFs. Two models were elaborated as reference, on: conventional radiological features; and T_2-WI RFs.
Statistical Tests: A common machine-learning approach was applied to radiomics models. Features with $P < 0.05$ at univariate analysis were entered in a LASSO-penalized Cox regression including bootstrapped 10-fold cross-validation. The resulting radiomics scores (RScores) were dichotomized per their median and entered in multivariate Cox models for predicting metastatic relapse-free survival. Models were compared with integrative area under the curve (AUC) and concordance index.
Results: Only dichotomized RScores from models based on rRFs subset, all rRFS and irRFS correlated with prognostic ($P = 0.0107–0.0377$). The models including all rRFs and irRFs had the highest c-index (0.83), followed by the radiological model. The radiological model had the highest integrative AUC (0.87), followed by models including all rRFs and irRFs. The radiological and full rRFs models were significantly better than the T_2-based radiomics model ($P = 0.02$).
Data Conclusion: The initial DCE-MRI of STS contains prognostic information. It seems more relevant to make predictions on rRFs instead of pRFs.
Evidence Level: 3
Technical Efficacy: 3

J. MAGN. RESON. IMAGING 2020.

INTRODUCTION

Radiomics refer to the extraction and analysis of several quantitative radiomics features (RFs) that extensively screen texture and shape of objects of interest of medical images. The use of radiomics in oncology is predicated on the hypothesis that the heterogeneity of a tumor's phenotype (quantified with RFs) correlates with its molecular heterogeneity and patients' outcomes. These RFs are integrated in predictive models relying on machine-learning algorithms in order to answer key oncological questions such as the identification of molecular subgroups of interest, the prediction of histological responses or the identification of prognostic radiomics-based signatures.

Radiomics can be applied to any type of medical image, including conventional MRI and dynamic-contrast-enhanced MRI (DCE-MRI). DCE-MRI aims at a non-invasive quantifying of tumor angiogenesis through the rapid acquisition of a time series of T1-weighted imaging (-WI) following the intravenous injection of a contrast agent. Among the DCE-MRI indices, two have emerged in pre-clinical and clinical oncological studies as potential biomarkers: the area under the time-intensity curve (AUC) and the influx volume transfer constant (K^{trans}) based on perfusion model. Though most of these studies were based on average values of AUC and K^{trans} inside tumors, others have stressed their spatial heterogeneity and thus investigated radiomics analyses based on DCE-MRI parametric maps (Rose et al, 2009). Such approaches have recently demonstrated significant improvement in distinguishing benign from malignant tumors and predicting histological response after neoadjuvant treatment, patients' prognosis, or lymph node extensions in various tumor types (Nie et al, 2016; Bowen et al, 2018; Liu et al, 2019; Thomassin-Nagara et al, 2017; Crombé et al, 2019e).

However, using RFs directly extracted from AUC, K^{trans} or other pharmacokinetic parametric maps could lead to simplifications and data losses. The perfusion models synthetize the information that is contained in the raw DCE-MRI acquisitions and none are specific to a particular histological type (Sourbron et al, 2013). Moreover, other factors could bring bias to the estimation of DCE-MRI parameters, for instance the estimation method of arterial input function, software and convolution functions (Beuzit et al, 2016; Huang et al, 2019; Cheng et al, 2008).

Two complementary studies involving lung cancers and pulmonary nodules have shown that RFs that were directly calculated on raw DCE-MRI acquisitions (i.e. at

each phase) were significantly influenced by the delay between injection and acquisition (Kim et al, 2016; Yoon et al, 2016). In addition, the authors found that the survival predictions based on RFs were not equivalent depending on the acquisition delay. The plots of the relative change in RFs (rRFs) since DCE-MRI baseline as a function of time following injection particularly drew our attention because they showed different profiles depending on the radiomics features and may contain predictive information. Consequently, we hypothesized that an alternative to the extraction of RFs from parametric maps for building predictive models could be the direct extraction of RFs and rRFs from raw DCE-MRI acquisitions, as well as the integration of the function rRF(time).

Hence, the aim of this study was to investigate the way to optimize the post-processing of DCE-MRI acquisition in the perspective of a prognostic radiomics study with the metastatic relapse-free survival (MFS) prediction as an end point. To do so, we built and compared different models, namely a classical radiological model, a radiomics model based on conventional T2-WI, and various models depending on the post-processing of DCE-MRI acquisitions of high-grade soft-tissue sarcomas (STS), for which radiomics analyses have already improved the predictions of patients' outcome and the response to treatment (Spraker et al, 2019; Vallières et al, 2015; Peeken et al, 2019a; Peeken et al, 2019b; Crombé et al, 2019d).

MATERIALS AND METHODS

Study design

This single-center study was approved by our institutional review board. The need for informed consent was waived by its retrospective nature.

From May 2012 to January 2018, we included all consecutive adult patients who were referred to our Sarcoma Reference Center for the therapeutic management of a biopsy-proven high-grade STS of the trunk wall or extremities, without metastasis on whole-body contrast-enhanced, computed tomography. Patients were uniformly treated at our institution with a neoadjuvant anthracycline based-chemotherapy (4 to 6 cycles), curative surgery at our Sarcoma Reference Center and adjuvant radiotherapy (50 Gy in 2 Gy fractions) (n = 133).

We excluded patients: (i) without a baseline MRI including at least T2-weighted imaging (-WI) and DCE-MRI with T1-mapping and a temporal resolution of dt = 2s, 4s or 8s (n = 79); and (ii) without a follow-up of at least 1 year after treatment (n = 4). The cohort included 50 adult patients (median age: 64.5 years, 22 women). Half of the tumors corresponded to undifferentiated pleomorphic sarcomas.

The following variables were retrieved from medical reports: age, sex, histological type, tumor location, tumor depth relative to the superficial facia, margins on the surgical specimen, histological response (defined as <10% viable cells on surgical specimen), occurrence of metastatic relapse and time from surgery to metastatic relapse, defining MFS (Cousin et al, 2017).

Routine follow-ups consisted of a clinical examination and chest radiograph every 3 months for 2 years, every 6 months for 5 years and then annually until 10 years after the surgery. These examinations were complemented by chest CT scans (with 1mm thick reconstructions) and MRIs for local evaluation in case of abnormal or doubtful findings. All relapses were histopathologically confirmed.

MRI acquisition

All patients underwent MRI examination on the same 1.5T MR-system (Magnetom AERA, Siemens Healthineers, Erlangen, Germany) with adapted position and coils depending on tumor size and location (either 15-channel or 18-channels body transmitter / receiver coils). All MRI examination included an axial turbo spin echo T2-WI (refocusing angle = 150°, repetition time (TR) = 6860 msec, echo time (TE) = 120 msec, slice thickness = 3-5 mm, field-of-view (FOV) = 250-350 x 250-320 mm), a T1-mapping and a DCE-MRI sequence following the Quantitative Imaging Biomarkers Alliance recommendations. Three types of 3D fast spoiled gradient echo sequence DCE-MRI sequences were used as their temporal resolution could divide 8 msec, i.e. VIBE_8s (n = 44), TWIST-VIBE_4s (n = 2) and TWIST-VIBE_2s (n = 4). VIBE_8s consisted of a Fat-Sat VIBE (volume interpolated breath-hold examination) sequence with TR/TE = 4.3 / 1.7 msec, flip angle = 25° and a temporal resolution of 8s. TWIST-VIBE_2s and TWIST-VIBE_4s consisted of CAIPIRINHA (controlled aliasing in parallel imaging results in higher acceleration), Dixon TWIST (time-resolved angiography with stochastic trajectories) VIBE sequences with TR/ in-phase TE / opposite-phase TE = 6.89 / 2.39 / 4.77 msec, flip angle = 25°, undersampling factor = 2, view-sharing parameters A and B = 15% and 20%, resulting in temporal

resolutions of 2s and 4s, respectively. The FOV and matrices of these 3 sequences were adjusted for an in-plane resolution of 1.1 x 1.1 mm², with a thickness = 4mm.
The T1-mappings used similar gradient echo sequences, with variable flip angles (2° and 15°).
A minimum of 3 phases (: baseline) were acquired before the intravenous injection of 0.1 mmol/kg of gadoteric acid (Dotarem, Guerbet, Villepinte, France, n = 28) or 0.1 mmol/kg of gadobenate dimeglunine (Multihance, Bracco imaging, Milan, France, n = 22), at a rate of 2 mL/s followed by a flush of 0.9% NaCl solution by an MR-compatible automatic injector (Sonic Shot 7, Nemoto Kyorindo, Tokyo, Japan).
Other sequences were available but their acquisition parameters were not standardized, namely: pre-contrast T1-WI, fat-suppressed T2-WI, and fat-suppressed contrast-enhanced T1-WI (CE-T1-WI).

Radiological analysis

Three radiologists performed a double-blinded review of the whole imaging dataset (two senior radiologists [A.C., with 8 years of experience in MRI including 3 years in a sarcoma reference center, and M.K., with 27 years of experience in a sarcoma reference center], and one fellow [D.F., with 3 years of experience including a 6-month internship in a sarcoma reference center]) for inter- an intra-observer agreements (Annexe 7 - Supplementary Data 1) followed by a consensual reading according to which the statistical analysis was performed. They reported:
(1) the longest diameter (LD);
(2) heterogeneous signal intensities (SIs) on T1-WI (categorized as homogeneous, <25%, 26-50% heterogeneous and >50% heterogeneous depending on the percentage of tumor volume displaying low, iso- and high SIs on T1-WI);
(3) heterogeneous SIs on T2-WI (with the same categorization as T1-WI);
(4) heterogeneous SIs on CE-T1-WI (with the same categorization as T1-WI);
(5) necrotic signal (defined as area with bright SI on T2-WI and no contrast enhancement, categorized as: absent, <25%, 26-50%, 51-89%, ≥90% of tumor volume);
(6) haemorrhagic signal (defined as non-fatty high SI on T1-WI, without contrast enhancement, categorized as absent or present);

(7) peritumoral edema at T2-WI (defined as high, infiltrative area with high SI on T2-WI, without mass effect and distinguishable from apparent tumor border, categorized as absent, limited or extensive) (Zhao et al, 2014; Crombé et al, 2019a);
(8) peritumoral enhancement at CE-T1-WI (defined as contrast enhancement beyond apparent tumor borders, without mass effect) (Zhao et al, 2014; Crombé et al, 2019a);
(9) MRI-growth pattern at CE-T1-WI (defined as pushing-type [when the tumor was entirely well-circumscribed], focal-infiltrating or diffuse-infiltrating [when the tumor was ill-defined on <25% and ≥25% of its circumference, respectively]) (Nakamura et al, 2017);
(10) tail sign at CE-T1-WI (defined as contrast-enhanced aponeurotic enhancement, categorized as absent, thin <2mm and thick ≥2mm) (Yoo et al, 2014);
(11) vessel and/or nerve invasion (categorized as absent, encasement when <180° and invasion when ≥180° or luminal extension) (Holzapfel et al, 2015);
(12) bone invasion (categorized as absent, periosteal contact and invasion when bone +/- medulla were invaded).

Radiomics post-processing (Figure 4.3-1)

The post-processing was entirely performed with Olea Sphere, v.3.0 SP16 Software (Olea Medical, La Ciotat, France). No matter the sequence and/or post-processing, the same 92 3D texture RFs were extracted (19 based on histogram, 23 on gray-level co-occurrence matrix [GLCM], 16 on gray-level run-length matrix [GLRLM], 15 on gray-level size zone matrix [GLSZM], 5 on neighboring gray tone difference matrix [NGTDM] and 14 on gray-level dependence matrix [GLDM]).

T2-based RFs. A senior radiologist (A.C., with 8 years of experience in MRI including 3 years in a sarcoma reference center) manually segmented the whole tumor volume with possible margins' adjustments based on the other conventional sequences. All the T2-WIs were resampled using a B-spline interpolation to obtain common voxels of 1 x 1 x 4 mm^3. The SIs of each sequence were standard-scaled and then discretized into 128 fixed bins between -1 and +1. The 92 texture RFs were extracted as well as 16 additional shape RFs.

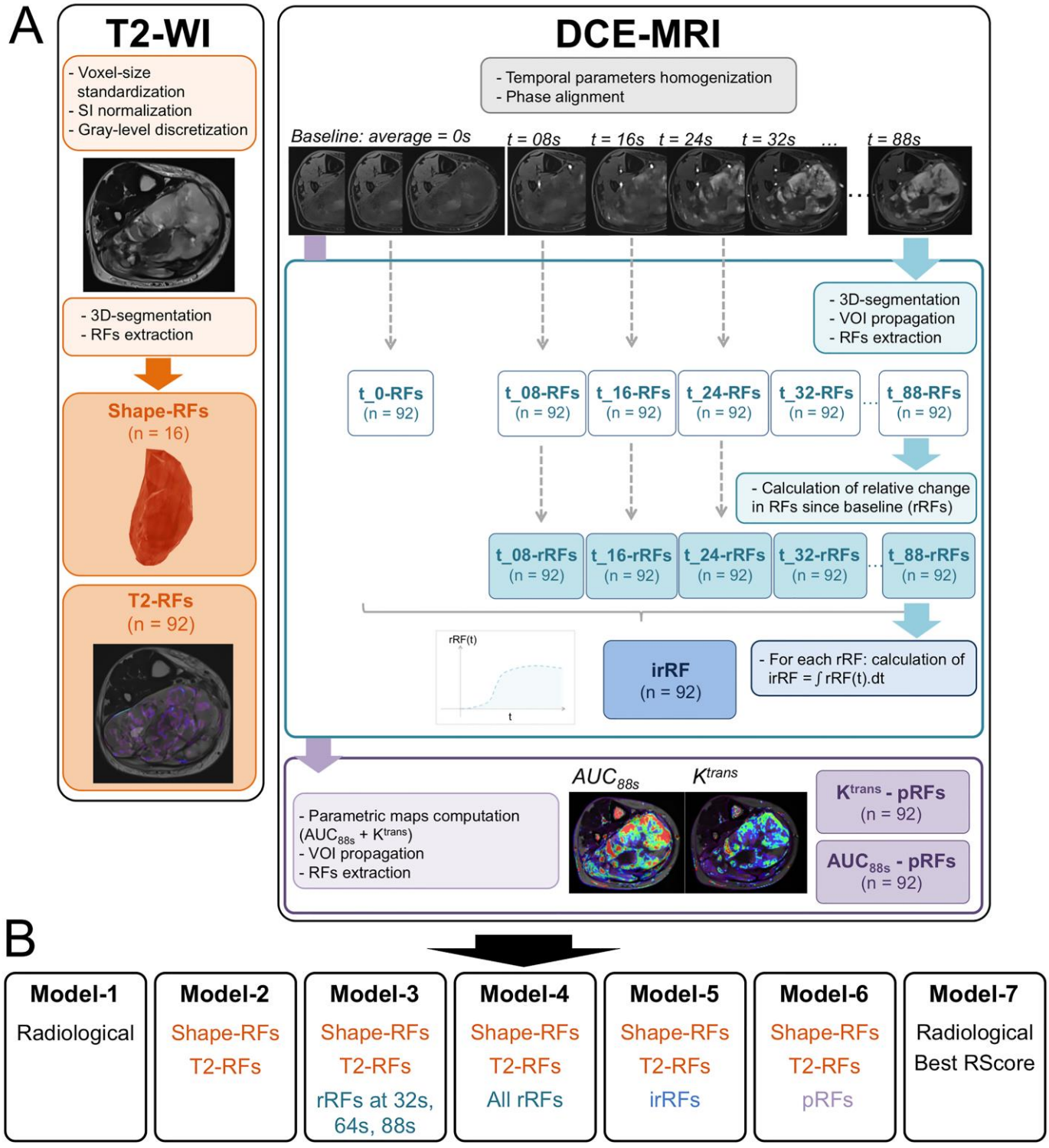
DCE-MRI-based RFs. Because of the different temporal resolutions, baseline and scan durations, the first step consisted in a homogenization of the temporal parameters. TWIST_2s and TWIST_4s were downsampled in order to obtain a common temporal resolution of 8 seconds by keeping 1 DCE-MRI phase out of 4 and

out of 2, respectively. Since the duration of baselines could vary, we performed a phase realignment, i.e. we defined t = 0s as the 1st phase where the contrast agent arrived in the feeding arteries of the tumor. Next, we truncated the downsampled and realigned DCE-MRI acquisitions in order to obtain a common scan duration of 88 sec for all patients. A radiologist (A.C.) manually segmented the whole tumor volume on the last phase (t = 88s) with the help of the other conventional sequences. This volume was propagated in all the other phases. The SIs were discretized into 128 fixed bins. Hence, the 92 RFs were extracted at each phase (i.e. 92 RFs at t = 8s, 92 RFs at t = 16s… and 92 RFs at t = 88s). For each RF, RF(t = 0s) was defined as the average value of this RF for all the phases before and including t = 0s. The relative change in RF (rRF) at each of these 11 phases for each of these 92 features since t = 0s was calculated as follows: rRF(t) = (RF(t) – RF(t = 0s)) / RF(t = 0s) x 100. Next, we integrated the 92 time-rRF functions with the "auc" function in the "flux" R package (using the 'jack-validate' method), as follows: integrated-rRF (irRF) = $\int_{t=0s}^{t=88s} rRF(t).dt$. In total, we calculated 92 irRFs and 1012 (11 x 92) rRFs.

K^{trans} and AUC_{88s}-based parametric-RFs (pRFs). We used the permeability plug-in of Olea Sphere to estimate the parametric maps. A radiologist (A.C.) manually chose the largest feeding artery within the FOV of the DCE-MRI acquisition. A fixed voxel was manually placed in this artery and the arterial input function was calculated with the average of 4 directly adjacent voxels. This arterial input function time-course was converted into blood R1 time-course with haematocrit = 0.45, gadoteric acid and gadobenate dimeglunine relaxivities at 1.5T = 3.6 and 8.3 L/mmol.s^{-1}, respectively. We generated the area under the time-intensity curve at 88s (AUC_{88s}, in mmol/L/s), the pharmacokinetic parameter K^{trans} (influx volume transfer constant from plasma to extravascular, extracellular space, which represents the capillary permeability, in min^{-1}) and the parametric maps according to the extended Tofts model.[9] The volume of interest segmented on the last DCE-MRI phase was propagated on these maps and, after discretizing the SIs into 128 fixed bins, we extracted 92 K^{trans} RFs and 92 AUC_{88s} pRFs.

Of note, voxel sizes for all DCE-MRI-related RFs were not standardized because the DCE-MRI sequences had the same spatial resolution. All the segmentations were validated by a second senior radiologist (M.K.).

Figure 4.3-1. Study design (**A**) Imaging post-processing. (**B**) Modelling. Seven models were built depending on the origins of the radiomics variables. A final Model-7 was proposed as the mix between best radiological Model-1 and radiomics models.
Abbreviations: AUC_{88s}: area under the time-intensity curve at 88s post-injection, DCE-MRI: dynamic contrast enhanced MRI, irRF: integrated-relative RF, K^{trans}: influx volume transfer constant, pRF: parametric RF, rRF: relative RF, RF: radiomics features, SI: signal intensity, t_n: DCE-MRI phase at t = n sec.
Color encoding: The post-processing steps are color-encoded depending on the RFs that they enable to obtain, i.e. orange for T2-based RFs, light blue for rRFs, dark blue for irRFs and violet for pRFs.



Statistical analysis

The statistical analyses were conducted in R, version 3.5.2 (R Foundation for Statistical Computing, Vienna, Austria). All tests were two-tailed. A p-value <0.05 was deemed significant. All multivariate predictive models were adjusted for: age (< vs. ≥ median), sex, histological type (undifferentiated sarcoma vs. myxoid/round cells liposarcoma vs. others), surgical margins (R0 vs. R1-R2) and histological response (good vs. poor response). Figures were performed with "ggplot2" and "survminer" R packages.

Radiological model (Model 1). Kaplan-Meier curves for MFS were drawn and differences in survivals were assessed using log-rank test ("survival" R package). All variables with a p-value <0.200 at univariate analysis were entered into a stepwise Cox regression model. The combination of variables that minimized the Akaike information criterion was included in Model-1.

Radiomics models (Models 2-7). All the radiomics-related features were standard-scaled before the statistical analysis. Average correlations between main categories of features - rRFs, irRFs and pRFs - were calculated and evaluated with the Spearman test and the correlation plot was drawn ("corrplot" R package).

We performed univariate Cox regression between MFS and each feature. The features with a p-value < 0.05 were entered in a least absolute shrinkage and selection operator (LASSO) Cox regression using "glmnet" R package (Simon et al, 2011).

Our aim was to build 5 different models depending on the main categories of features (Fig. 1). The Model-2 included T2-WI RFs and shape RFs. The Model-3 included T2-WI RFs, shape RFs and a subset of rRFs at t = 32s, 64s and 88s. The Model-4 included T2-WI RFs, shape RFs and all rRFs (from t = 8s to t = 88s). The Model-5 included T2-WI RFs, shape RFs and irRFs. The Model-6 included T2-WI RFs, shape RFs and pRFs.

For each radiomics model, a ten-fold cross validation was performed to select the lambda (the regularization parameter from the LASSO, i.e. a numeric value defining the amount of shrinkage) that provided the minimum cross-validated errors and helped determine the relevant features and their coefficient. This step was iterated 1000 times on different subsamples ("boot" R package). The final coefficients of the resulting features were used to calculate the radiomics scores (RScore, from 2 to 6). Each RScore was dichotomized per its median. Their hazard ratio (HR) with 95% confidence interval (95%CI) was estimated before and after covariables adjustment.

The performances of the models were quantified by using concordance index (c-index) and integrated area under the time-dependent ROC curves (iAUC). The iAUC corresponds to a weighted mean of the area under the ROC curves calculated at different time-points during the follow-up ("survAUC" R package) (Uno et al, 2007). The survival time-dependent ROC-curves as functions of the time following the end of treatments were plotted for all models to facilitate the visualization of the models performances over time. The value of c-indices and iAUCs with 95%CI were estimated by using bootstrap on 10000 resamplings of the study population. We defined a final Model-7 that combined the radiological variables from Model-1 and the RScore from the radiomics model with the highest performances. To compare iAUCs and c-indices, the differences in these metrics between each pair of models were calculated by using a bootstrapping method (with 10000 replicates). The difference was considered significant if its 95%CI did not include 0.

RESULTS

Study population (Table 4.3-1)

The median LD and volume were 106 mm and 276 cm^3, respectively. Twelve patients showed a good histological response on post-chemotherapy surgical specimen (12/50, 24%). The 2-year and 5-year MFS probabilities were of 62.9% (95%CI = (49.5–80)) and 55.9% (95%CI = (40.1–78.1)), respectively. There were 17 metastatic relapses, 16 occurring within the first two years after treatment.

Radiological model (Table 4.3-2)

After stepwise selection method, 3 radiological variables (among the 10 variables with a p-value <0.200 at univariate level) were finally included in the Model-1, namely: 'necrotic signal >25% of tumor volume', 'diffuse infiltrating MRI growth pattern', and 'encasement or invasion of vessel/nerve'. The c-index and iAUC of Model-1 were 0.82 (95%CI = (0.67–0.88)) and 0.87 (95%CI = (0.78–1)).

Table 4.3-1. Clinical and pathological features of the study population.

Variables	Patients
Sex	
Women	22/50 (44%)
Men	28/50 (56%)
Age (years)	64.5 (21 - 84)
Histological types	
Undifferentiated pleomorphic sarcoma	25/50 (50%)
Myxoid/round cells liposarcomas	7/50 (14%)
Rhabdomyosarcoma	6/50 (12%)
Synovial sarcoma	3/50 (6%)
Other undifferentiated sarcoma	3/50 (6%)
Pleomorphics liposarcoma	2/50 (4%)
Dedifferentiated liposarcoma	1/50 (2%)
Leiomyosarcoma	1/50 (2%)
Myxofibrosarcoma	1/50 (2%)
Malignant peripheral nerve sheath tumor	1/50 (2%)
Location	
Upper limb	5/50 (10%)
Shoulder girdle	6/50 (12%)
Trunk wall	6/50 (12%)
Pelvic girdle	2/50 (4%)
Lower limb	3/50 (6%)
Tumor depth	
Superficial	0/50 (0%)
Aponeurotic + Superficial	3/50 (6%)
Aponeurotic + Deep	10/50 (20%)
Aponeurotic + Deep + Superficial	12/50 (24%)
Strictly deep	25/50 (50%)
Longest diameter (mm)	106 (50 - 265)
Tumor volume (cm3)	276 (30 - 3029)
Histological response	
Good response (<10% viable cells)	12/50 (24%)
Poor response	38/50 (76%)
Surgical margins	
R0	30/50 (60%)
R1	20/50 (40%)
R2	0/50 (0%)

NOTE. Data are numbers of patients with percentage in parentheses, except for age, longest diameter and volume, which are given as median and range.

Table 4.3-2. Univariate analysis of the conventional radiological features for the prediction of metastatic relapse-free survival (MFS).

Variables	No. at risk	No. of events	2-years survival probability	p-value
Longest diameter				
< Median	25	8	65.3 (46.8 - 91.1)	0.6
≥ Median	25	9	60.8 (43.1 - 85.7)	
Volume				
< Median	25	7	65.5 (47 - 91.2)	0.3
≥ Median	25	10	60.2 (42.4 - 85.4)	
Heterogeneous SI on T1-WI				
0% - Homogeneous	3	0	100 (100 - 100)	*0.1*
< 25%	22	6	66.9 (47 - 95.3)	0.2
26 - 50%	12	3	71.4 (48.2 - 100)	0.3
> 50%	13	8	44 (23.3 - 83)	**0.03***
Heterogeneous SI on T2-WI				
0% - Homogeneous	1	0	100 (100 - 100)	*0.1*
< 25%	5	0	100 (100 - 100)	0.6
26 - 50%	11	3	72.7 (50.6 - 100)	0.1
> 50%	33	14	54.4 (38.3 - 77.3)	0.1
Heterogeneous SI on CE-T1-WI				
0% - Homogeneous	9	2	77.8 (54.9 - 100)	*0.8*
< 25%	16	5	58.7 (34.4 - 100)	0.4
26 - 50%	25	10	60.3 (42.9 - 84.6)	0.2
> 50%	1	0	100 (100 - 100)	0.2
Necrotic signal				
Absent	5	1	80 (51.6 - 100)	*0.2*
< 25%	17	2	85.6 (68.6 - 100)	0.7
26 - 50%	12	6	55.6 (32.5 - 95)	**0.01***
51 - 89%	9	4	38.9 (13.7 - 100)	0.09
≥ 90%	7	4	42.9 (18.2 - 100)	0.2
Haemorragic signal				
Absent	8	3	50 (20.4 - 100)	0.5
Present	42	14	63.7 (44.7 - 91.4)	
Peritumoral edema				
Absent	5	0	100 (100 - 100)	*0.09*
Limited	38	13	62 (46.4 - 82.8)	1.
Extensive	7	4	42.9 (18.2 - 100)	0.08
Peritumoral enhancement				
No	10	3	70 (46.7 - 82.8)	0.9
Yes	40	14	42.9 (18.2 - 100)	
MRI-growth pattern				
Pushing-type	5	2	60 (29.3 - 100)	*0.04*
Focal-infiltrating	19	2	88.4 (74.5 - 100)	0.5
Diffuse-infiltrating	26	13	45.9 (28.4 - 74)	**0.03***
Tail sign				
Absent	20	7	64.3 (46.2 - 89.5)	1.
Thin (<2 mm)	18	6	66.7 (45.9 - 96.8)	0.8
Thick (≥2 mm)	12	4	55.6 (29.9 - 100)	0.9
Vessel and/or nerve invasion				
Absent	28	7	75.5 (59.7 - 95.6)	*0.2*
Encasement	13	5	54.2 (30.2 - 97.2)	0.1
Invasion	9	5	40 (17.1 - 93.8)	0.1
Bone invasion				
No	26	6	71.3 (52.1 - 97.5)	*0.07*
Periosteal contact	22	11	47.5 (29.9 - 75.3)	0.4
Invasion	2	0	100 (100 - 100)	0.07

NOTE. The p-value corresponds to the log-rank test p-value. The 2-years MFS probability is given in percentage with 95% confidence interval.
For variables with more than 2 modalities, the first p-value (in italic) corresponds to the p-value when all modalities were considered together. The p-values below correspond to the p-values of the dichotomized variable according to its different levels.
Abbreviations: CE- contrast enhanced, No.: number, SI: signal intensity, -WI: weighted imaging.
*: $p<0.05$.

Radiomics models

The correlation plot of all RFs highlighted two clusters of features: rRFs and irRFs on one side, and AUC_{88s} and K^{trans} pRFs on the other side (Annexe 7 - Supplementary Data 2). Overall, rRFs extracted from t = 8s to t = 88s phases were more correlated with themselves and with iRFs than with pRFs (range of average Spearman rho: 0.030-0.086, 0.044-0.078 and -0.009-0.025, respectively). Figure 4.3-2 provides a visual assessment of the assessment of significant associations between RFs and MFS depending on their origin, at univariate analysis. In total, 252 features correlated with MFS and were subsequently entered in the LASSO Cox regressions, namely: 7 T2-WI RFs, 2 shape RFs, 158 rRFs, 16 irRFs, 38 AUC_{88s} pRFs and 31 K^{trans} pRFs.

Figure 4.3-2. Summary of the univariate Cox regression analysis between each of the radiomics features (RFs), relative RFs (rRFs), integrated rRF (irRFs) and parametric RFs (pRFs) and metastatic relapse-free survival.
The x-axis corresponds to the hazard ratio and the y-axis to the p-value. Each point represents a (ir/r/p)RF. The horizontal line indicates the threshold for significance (i.e. 0.05). The data in parentheses in the legend indicate the number and percentage of significant RFs for each features origin. Abbreviations: -WI: weighted imaging.

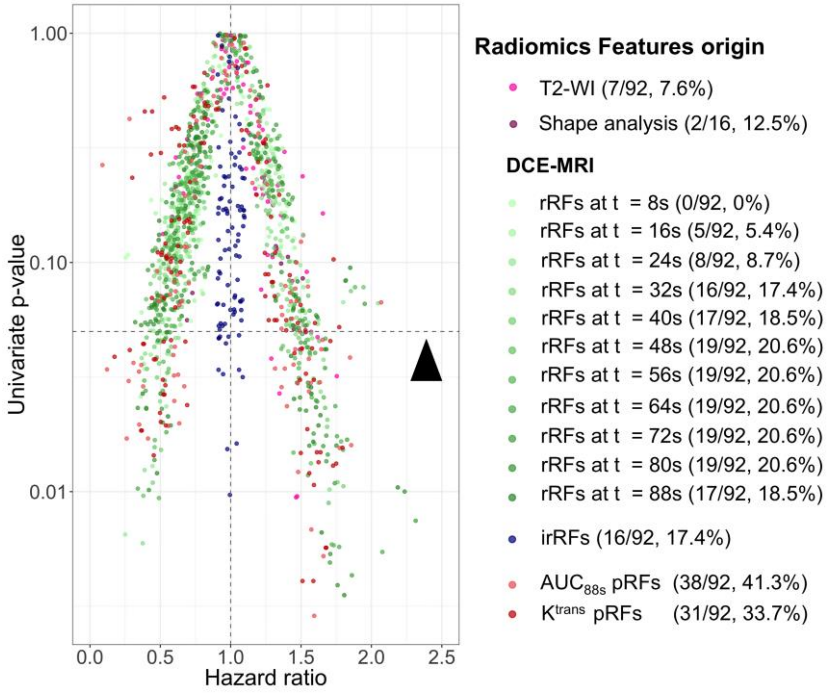


Table 4.3-3 summarizes the features (and their bootstrapped weight) that were finally selected to calculate each RScore. Each RScore corresponded to the weighted sum of these features. Some features were systematically selected across models, namely: T2-WI_GLDM_Dependence_Non_Uniformity (in all models), Histogram_10th_Percentile (in all models) and Histogram_Median (in Model-2 to Model-5).

After dichotomization, RScore_Model-4 (based on all rRFs) and RScore_Model-5 (based on iRFs) led to the same predictions, and consequently, to the same performances. A high RScore indicated a worse prognosis. Figure 4.3-3 shows the Kaplan-Meier curves for MFS for each RScore after dichotomization by its median. The largest differences between the curves were found with RScore_Model-4 and RScore_Model-5 (log-rank p-values = 0.0002), followed by RScore_Model-3 (p-value = 0.0023). However, no association was found at univariate level between RScore_Model-2 and MFS (p-value = 0.11).

Table 4.3-4 summarizes the performance for all models in uni- and multivariate analyses.

After adjustment with confounding covariables, RScore_Model-3, RScore_Model-4 and RScore_Model-5 remained significantly associated with MFS (p = 0.0377, 0.0107 and 0.0107, respectively), while RScore_Model-2 and RScore_Model-6 were not (p = 0.7950 and 0.0853, respectively).

Figure 4.3-4 illustrates the applications of the alternative post-processing methods for DCE-MRI-based radiomics models in 2 patients, one with an early metastatic relapse and one in complete remission 3.5 years after the end of treatment.

Table 4.3-3. Summary of the prognostic multivariate models for prediction of metastatic relapse-free survival (MFS).

Model Name	Finally selected variables[§§]	Median (range)
Model-1[§] *(Radiological)*	Necrotic SI (>25% of tumor volume) MRI-growth pattern (diffuse infiltrating) Vessel and/or nerve invasion (invasion)	-
Model-2 *(Shape + T2-WI RFs)*	T2_GLRLM_Run_Entropy (0.0924) T2_GLDM_Dependence_Non_Uniformity (0.1896) T2_GLDM_LDLGLE (0.1550)	-0.015 (-0.367 - 0.841)
Model-3 *(Shape + T2-WI RFs + subset rRFs)*	T2_GLRLM_Run_Entropy (0.0605) T2_GLDM_Dependence_Non_Uniformity (0.1298) T2_GLDM_LDLGLE (0.1152) t_32s_FOS_10th_percentile (-0.1593) t_32s_FOS_Median (-0.0676) t_64s_NGTDM_Busyness (0.0727) t_88s_GLDM_LDLGLE (0.1826)	-0.046 (-0.885 - 1.085)
Model-4 *(Shape + T2-WI RFs+ full rRFs)*	T2_GLRLM_Run_Entropy (0.1253) T2_GLDM_Dependence_Non_Uniformity (0.2009) T2_GLDM_LDLGLE (0.1834) t_32s_FOS_10th_percentile (-0.1126) t_32s_FOS_Median (-0.1257) t_48s_GLSZM_Zone_Entropy (-0.0627) t_72s_GLSZM_LALGLE (0.1002) t_72s_NGTDM_Busyness (0.2147) t_72s_GLDM_Dependence_Variance (0.3149) t_72s_GLDM_LDLGLE (0.0478) t_80s_GLSZM_LALGLE (0.1252)	-0.069 (-1.263 - 2.354)
Model-5 *(Shape + T2-WI RFs + irRFs*	T2_GLRLM_Run_Entropy (0.1370) T2_GLDM_Dependence_Non_Uniformity (0.1756) T2_GLDM_LDLGLE (0.1590) auc_FOS_10th_Percentile (-0.1520) auc_FOS_Median (-0.0333) auc_GLSZM_Zone_Entropy (-0.0578) auc_NGTDM_Busyness (0.1475) auc_GLDM_Dependence_Variance (0.0836)	-0.052 (-0.981 - 1.373)
Model-6 *(Shape + T2-WI RFs+ pRFs)*	T2_GLDM_Dependence_Non_Uniformity (0.1195) AUC_FOS_10th_Percentile (-0.0304) AUC_GLCM_Joint_Average (-0.0754) AUC_GLCM_Sum_Average (-0.0045) AUC_GLRLM_LRLGLE (0.1267) KTRANS_GLCM_Joint_Average (-0.0269)	-0.029 (-0.527 - 0.713)
Model-7	Variables from Model-1 and Model-4	-

NOTE. Only variables with a p-value less than 0.05 (named 'relevant') were initially entered in the multivariate radiomics modelling (Model-2 to Model- 6).
Abbreviations: DCE-MRI: dynamic contrast enhanced MRI, -WI: weighted imaging.
§: the variables included in Model 1 were selected after Stepwise regression among all radiological features with a p-value <0.200 at univariate analysis.
§§: the value in parentheses corresponds to the weight for each selected variable following, as given by the LASSO features selection.

Figure 4.3-3. Kaplan-Meier curves for the prediction of metastatic relpase-free survival depending on the different radiomics scores (RScores) from Model-2 (A), Model-3 (B), Model-4 (C), Model-5 (D) and Model-6 (E) Beforehand, each score was dichotomized per its median. *: p<0.05, **: p<0.005, ***: p<0.001

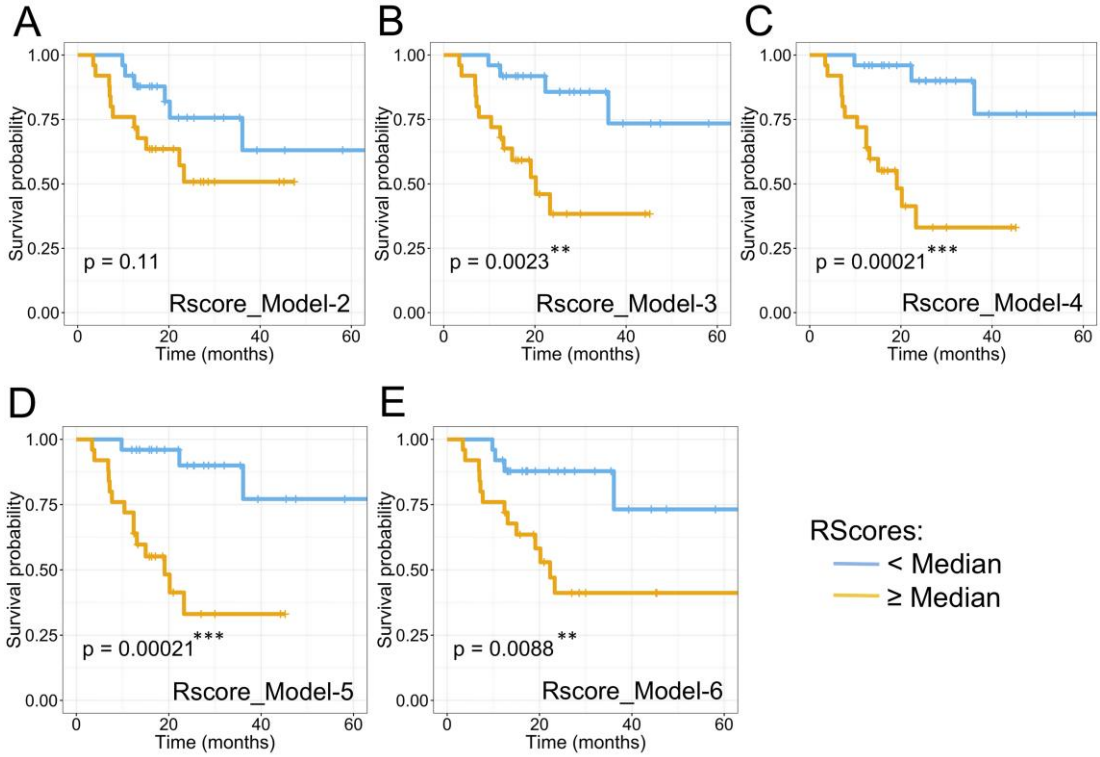
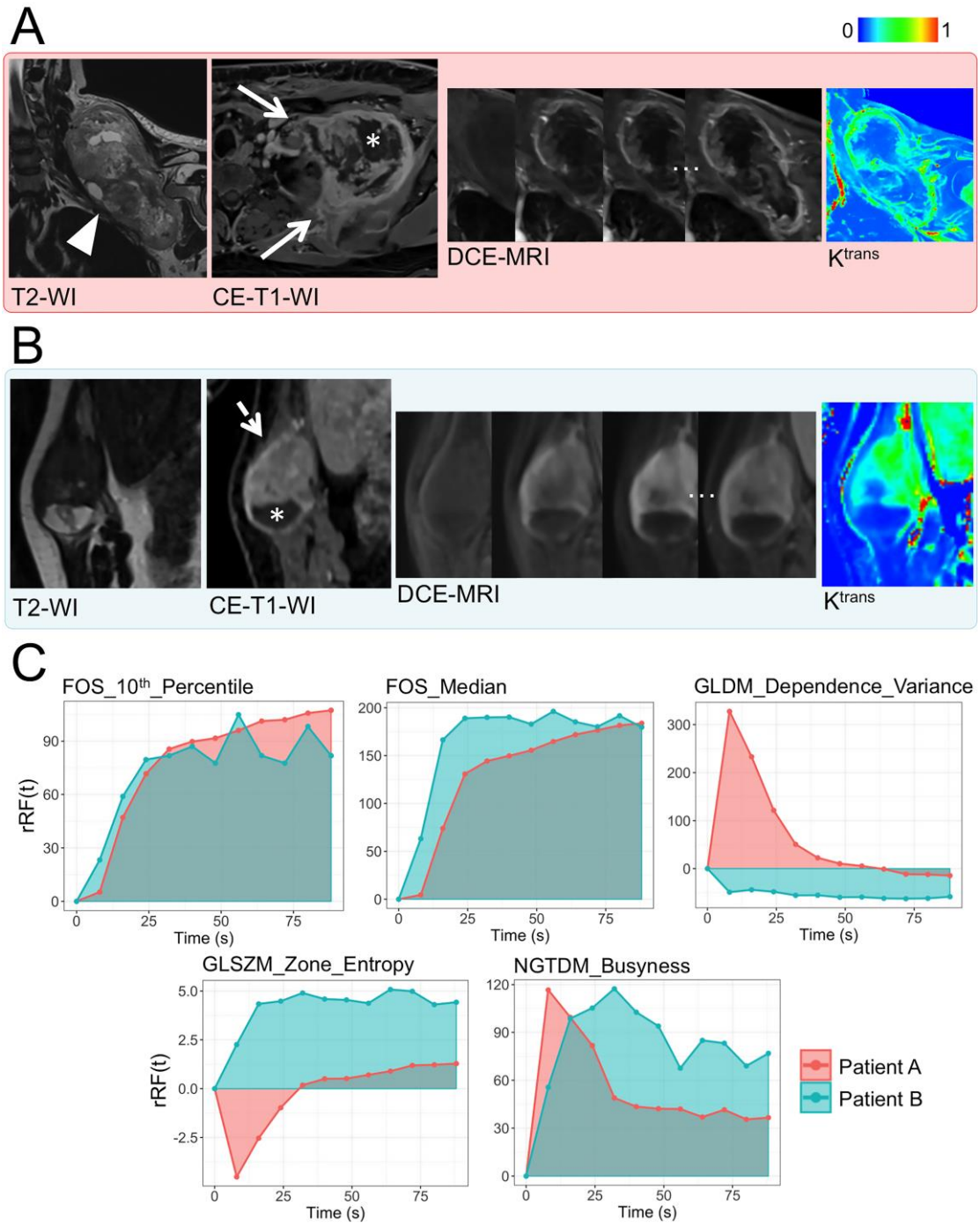


Table 4.3-4. Performances of the radiomics score (RScores) before and after adjustment, and corresponding predictive performances estimated by concordance-index (c-index) and integrated area under the curve (iAUC) at 5 years.

	Model	Univariate HR - dichotomized RScore	p-value	Multivariate HR - dichotomized RScore	p-value	c-index	iAUC
1	**Radiological**	-	-	-	-	0.82 (0.67 - 0.88)	0.87 (0.78 - 1)
2	**Shape + T2 RFs**	2.22 (0.82 - 6.02)	0.1130	1.16 (0.37 - 3.68)	0.7950	0.74 (0.57 - 0.81)	0.76 (0.62 - 0.99)
3	**+ subset rRFs**	4.94 (1.59 - 15.38)	**0.0023****	3.70 (1.08 12.71)	**0.0377***	0.79 (0.67 - 0.87)	0.84 (0.75 - 1)
4	**+ full rRFs**	7.78 (2.20 - 27.55)	**0.0002*****	6.42 (1.54 - 26.77)	**0.0107***	0.83 (0.71 - 0.91)	0.86 (0.78 - 1)
5	**+ irRFs**	7.78 (2.20 - 27.55)	**0.0002*****	6.42 (1.54 - 26.75)	**0.0107***	0.83 (0.71 - 0.91)	0.86 (0.78 - 1)
6	**+ pRFs**	4.00 (1.30 - 12.32)	**0.0088***	2.98 (0.86 - 10.33)	0.0853	0.79 (0.65 - 0.88)	0.82 (0.70 - 1)
7	**Radiological + Shape + T2 RFs + full rRFs**	-	**-**	-	**-**	0.84 (0.70 - 0.87)	0.89 (0.82 - 1)

NOTE. The p-values correspond to the uni- and multivariate Cox models (before and after adjustment for: age, sex, histological type, surgical margins (and histological response. Each RScore was dichotomized per its own median.
Hazard ratio, c-index and iAUCs are given with 95% confidence intervals.
Abbreviations: AUC: area under the curve; c-index: concordance-index; DCE-MRI: dynamic contrast enhanced MRI; HR: hazard ratio; iAUC: integrated AUC; irRFs: integrated relative radiomics features; pRFs: parametric radiomics features; RFs: radiomics features; rRFs: relative radiomics features. *: p<0.05, **: p<0.005, ***: p<0.001

Figure 4.3-4. Added value of alternative post-processing methods for DCE-MRI-based radiomics models. (**A**) A 39 years old male presented with a deep-seated malignant peripheral nerve sheath tumor invading the left subclavian vessels and the brachial plexus on coronal T2-weighted imaging (-WI, white arrow head). On contrast-enhanced T1-weighted imaging (CE-T1-WI), this tumor demonstrated a large area of central necrosis (white asteroid) and a diffuse infiltrative MRI-growth pattern (white arrow). DCE-MRI raw data and Ktrans map demonstrated thick and heterogeneous peripheral enhancements. (**B**) A 66 years old male presented with a high-grade, deep-seated, pleomorphic rhabdomyosarcoma of the right abdominal wall. The tumor was better delimited on CE-T1-WI (dashed white arrow) with small area of necrosis at its lower pole (white asteroid). DCE-MRI data showed an early, intense and rapidly homogeneous contrast enhancement in non-necrotic area. (**C**) For each patient, the curves of the relative radiomics features (rRFs, in %) as functions of time were plotted, integrated and superimposed for 5 illustrative RFs that were highlighted across the radiomics models. The RScores were all above their median for patient A, and all under for patient B. The patient A had a metastatic relapse 7 months after the surgery. The patient B is still in complete remission 3.5 years after the end of treatments.



Comparisons of prognostic models

Figure 4.3-5 shows the survival ROC curves for each model as functions of the time following surgery. The curve from Model-2 was constantly below the curves from the other models. Though the curves of these last 6 models were close during the 24 first months of follow-up – meaning comparable performances, the Model-1 and Model-7 showed higher AUC after this delay.

The highest c-index and iAUC among the radiomics models were reached with Model-4 and Model-5 (c-index = 0.83 [95%CI = (0.71–0.91)] and iAUC = 0.86 [95%CI = (0.78–1)], respectively) (Table 4.3-4). The lowest performances were obtained with Model-2 and with Model-6. Model-7 was built by joining RScore_Model-4 and the three radiological variables from Model-1 and provided the highest c-index and iAUC of the study (c-index = 0.84 [95%CI = (0.70–0.97)] and iAUC = 0.89 [95%CI = (0.82–1)], respectively).

Comparisons of c-index and iAUC are shown in Figure 4.3-6. Regarding c-index, the radiological Model-1, the radiomics Model-4 and Model-5, as well as the final Model-7 were significantly better than Model-2. Regarding iAUC, statistical difference was only obtained when comparing Model-2 with Model-7. None of the other comparisons were otherwise significant. Details for the c-indices and iAUC comparisons are given in Annexe 7 - Supplementary Data 3.

DISCUSSION

In this study, we proposed different methods to post-process and analyze raw DCE-MRI acquisitions for the purpose of a radiomics analysis, and we compared the performances of their corresponding prognostic model. Overall, the highest prognostic performances for DCE-MRI-related models were achieved with RScores built on full rRFs or irRFs. Both were completely equivalent. The lowest performances were reached with pRFs, albeit the differences with other models were not significantly different per iAUC or c-index metrics. We also found that performances of the conventional radiological model relying on a semi-quantitative or qualitative assessment by radiologists were not outdone and rivaled that of the best existing radiomics models, thus indicating potential ways of improving these.

Figure 4.3-5. Area under the time-dependent receiver operating characteristic curves (AUC) for metastatic relapse-free survival depending on the different predictive models. It can be noted that Model-4 and Model-5 entirely superimposed.

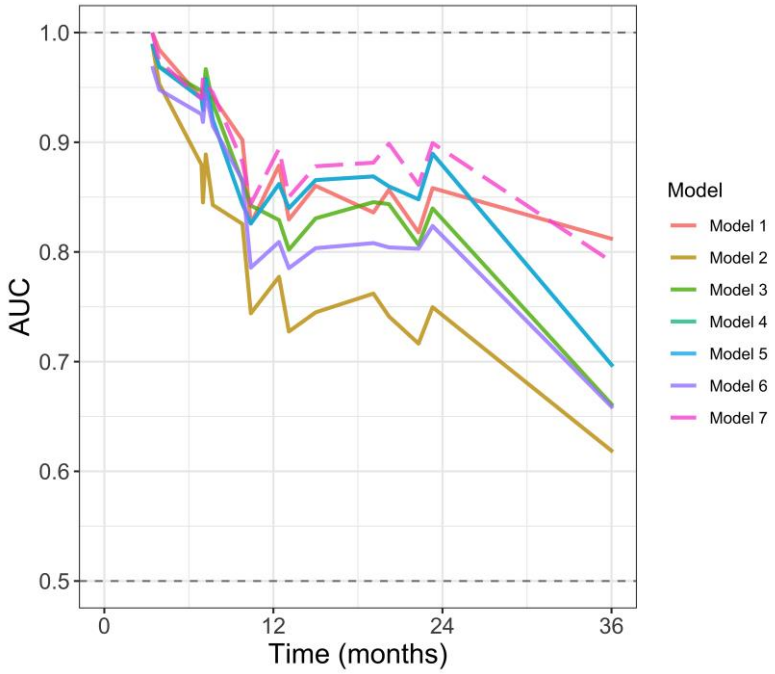
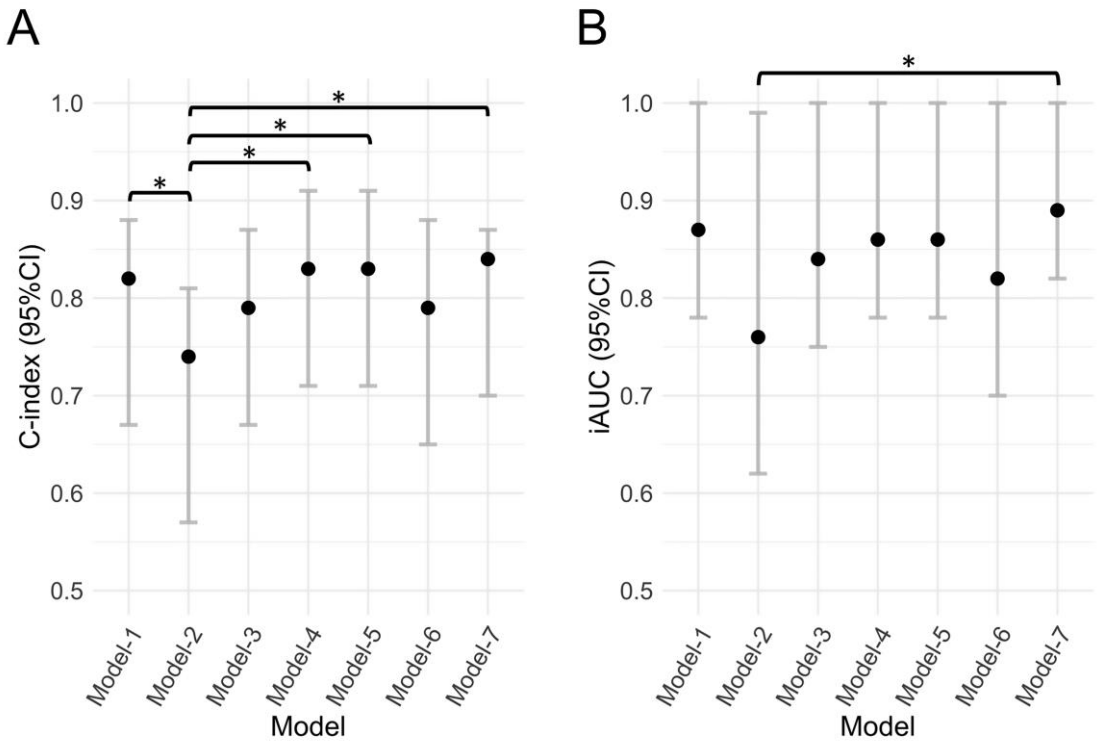


Figure 4.3-6. Comparisons of the performances of the predictive models, measured with concordance-index (c-index) and integrated area under the curve (iAUC) with 95% confidence interval (95%CI). *: $p < 0.05$.



The present cohort is comparable to previous prognostic studies involving STS regarding clinical data, histological types of high-grade tumors, histological response and outcome. The most frequent histological types were undifferentiated pleomorphic sarcomas and liposarcomas, which are typically over-represented in the high-grade subgroups. The rate of histological response was low (24%) and about one third of patients showed a metastatic relapse, mostly within the first two years after treatment, which concurs with prior studies focusing on patients treated in reference centers (Cousin et al, 2017; Crombé et al, 2019a; Blay et al, 2019). It should be noted that the performance status was not included among the covariables because all patients were graded 0 at diagnosis, according to the World Health Organization.

DCE-MRI has already proven its value in diagnosing malignant soft-tissue tumors and predicting grade and histological response of STS (Xia et al, 2017; Huang et al, 2016a; Meyer et al, 2013; Gondim Teixeira et al, 2018; Ahlawat et al, 2019). So far, the heterogeneity in tumor vascularization has been assessed on model- and non-model-based DCE-MRI parametric maps, mostly K^{trans} and AUC at 60s or 90s. However, previous studies have stressed several risks of bias in the estimation of RFs specific to DCE-MRI acquisition, besides usual bias in radiomics studies (for instance: MR-manufacturers, coils, acquisition parameters segmentation methods, post-acquisition filters, voxel size standardization, gray-level discretization, signal intensities normalization…) (Buch et al, 2018; Mayerhoefer et al, 2009; Collewet et al, 2004; Ford et al, 2018). Indeed, for both MRI- and CT-DCE acquisitions, the temporal resolution and scan duration may have a significant influence on several RFs that were directly extracted on parametric maps (Crombé et al, 2019e; Bogowicz et al, 2016). Therefore, we aimed at proposing alternative techniques to exploit the information within DCE-MRI acquisitions, without relying on parametric maps. A first idea was to directly study the RFs extracted from each phase of raw DCE-MRI acquisitions and to calculate their relative change since the baseline of the acquisition. Yoon et *al.* (2016) first described this methodology on a small set of histogram-based RFs They found that each time-rRFs curves had their own profile with significant temporal changes. In addition, they highlighted the fact that the prognostic information of the rRFs also depended on the acquisition delay after contrast-agent injection. Hence, selecting the optimal acquisition delay of relevant rRFs may optimize predictions without being affected by the additional biases of RFs quantified on parametric maps. This approach provides several strongly correlated rRFs and

requires a careful selection of variables, which are now available within open source statistical packages.

To deepen and condensate this approach within a unique metrics per type of RFs, we also calculated the area under the time-rRF curves for each of the 92 rRFs, i.e. irRFs. We found that, after dichotomizing the corresponding RScores per their median, these 2 approaches provided the exact prognostic information. Since (i) it introduced an additional computation with potential bias in the estimation of this integral and (ii) showed slightly lower performance before dichotomization, we believe it may be more prudent to retain the Model-4 with all rRFs.

We also proposed a 'simpler' model (Model-3) based on a subset of rRFs at about 30s, 60s and 90s post-injection. However, the corresponding dichotomized RScore_Model-3 was not an independent prognostic factor and performances were lower than with Model-4 and Model-5. Finally, the RScore based on pRFs also demonstrated a lack of independent prognostic value once included in multivariate Cox modeling.

Altogether, these results suggest that going through a post-processing step of parametric maps computation could be responsible for loss in information because of the necessary simplification when applying a general permeability model. It should be noted that the comparisons of multivariate prognostic Model-3, Model-4, Model-5 and Model-6 did not reach significance, for both iAUC and c-index estimators. Hence, although we cannot formally conclude that one of the 4 corresponding methods is better than the others, the trends that we identified in our study indicate that researcher should continue to investigate alternative methods for post-processing raw DCE-MRI acquisition for the purpose of radiomics approaches.

Our study included a conventional radiological analysis and a T2-WI based radiomics model as references for the radiomics models. The Model-1 demonstrated good performances, comparable to the best radiomics models. After stepwise selection method and adjustment, it included 3 radiological variables: (i) necrotic signal >25%, (ii) diffuse infiltrating MRI-growth-pattern and (iii) vessel/nerve encasement or invasion. These features have not been evaluated in this exact setting, but they are known to be associated with rather poorer prognosis (Crombé et al, 2019a; Nakamura et al, 2017; Yoo et al, 2014; Holzapfel et al, 2015). Other radiological features have been correlated with patients' prognosis but were not kept in the final model, such as:

heterogeneous SI on T2-WI, peritumoral enhancement or tail sign. We hypothesized that it was due to our inclusion criteria. Indeed, we only included high-grade STS though these criteria were significantly associated with MFS or overall survival in patients with low-, mid- and high grade tumors and not only high grades. The disadvantage of these conventional radiological features is their lack of objectivity and quantification. They may not be reproducible from one cancer center to another. Nevertheless, theses 3 radiological features could help improve our current results. Instead of simply adding these variables to our best RScore (as in Model-7), the next step should be to quantify these features objectively and automatically. Hence, automatic segmentation of STS on CE-T1-WI followed by a dedicated shape analysis could quantify the MRI-growth pattern. An automated habitat imaging on multiparametric-MRIs could improve the detection and estimation of the amount of necrosis.

The worse model was the T2-WI-based radiomics Model-2. Yet, it should be noted that its performance remained better than a random model and that some T2-WI-based RFs were systematically kept in DCE-MRI-based RScores. This illustrates that heterogeneity in T2-WI did have prognostic information but that it was lost during the radiologists' semi-quantitative and visual assessment.

It should be noted that we systematically adjusted our multivariate models with the histological subtype as a confounding covariable. We distinguished 3 categories, namely: undifferentiated sarcomas (because it was the largest group of patients), myxoid/round cells liposarcomas, and other sarcomas. Indeed, myxoid/round cells liposarcomas are characterized by specific vascular patterns on histological slices, atypical metastatic locations, and slightly distinct time-rRF curves compared with the 2 other categories - this whatever the family of textural features (not shown) (Gorelik et al, 2018). These ancillary results suggest that radiomics approaches based on DCE-MRI data could help diagnosing some subtypes of sarcomas, but it was beyond the scope of our study.

Although our prognostic models worked well, they are not perfect and need to be validated on external test cohorts. Other variables could be added to the models, from other MRI sequences (such as diffusion-weighted imaging, MR-spectroscopy) or other imaging modalities (such a positron emission tomography) or molecular analysis of STS. It is known that increased molecular complexity of STS is a poor

prognosis factor (Chibon et al, 2010). Consequently, it could be worth combining prognostic molecular signature with prognostic RFs.

Our study has limits. First, this is a retrospective single center study. However, this is the largest radiological study involving DCE-MRI and STS. Moreover, this pilot study formalizes the definitions of new types of RFs (rRFs and irRFs). Second, there was no validation cohort and the potential predictors were much more numerous than the included patients, which was at risk of overfitting the radiomics models. However, we combined two resampling methods (cross-validation and bootstrapping) to better estimate the performances of our models and we used penalized Cox regression, which is more adapted to multidimensional data. Nevertheless, these limits made our results exploratory and preliminary. Our aim was not to propose a definitive prognostic model for STS based on DCE-MRI data, but rather to investigate if there may be alternative methods to post-process DCE-MRI data in order to enhance prognostic radiomics models. Third, the DCE-MRI acquisitions were not standardized and we had to homogenize their temporal parameters by downsampling and truncating. The nature of the DCE-MRI sequence itself (i.e. TWIST or VIBE) and the initial temporal resolution may have introduced some noise, because it can be expected that acquisitions with an excellent temporal resolution may have lower signal-to-noise ratio leading to increased heterogeneity. Moreover, we used 2 different contrast agents, which may have introduced bias in the estimation of the textural features. To our knowledge, no study has already investigated the influence of Gadolinium-chelates contrast agents on the values of radiomics models. However, since gadobenate dimeglunine (MultiHance) can interact with serum proteins, it could have had a subtle effect (Liang et al, 2010). Fourth, we dichotomized the RScore per their median while optimized cut-offs could have been used. We purposely avoided this technique to limit overfitting. Finally, other radiomics analyses of other MR-sequences could have been performed and may have improved the performances.

In conclusion, our study indicates that the initial DCE-MRI acquisition of patients with locally-advanced STS contains significant prognostic information that can be revealed through a radiomics approach. The highest performances of DCE-MRI-based models were obtained by including rRFs (or irRFs) and not pRFs. Hence, optimizing

the post-processing of DCE-MRI acquisitions can increase the performance of radiomics models.

*

* *

4.3. Limites et ouvertures

Il s'agit ici d'articles méthodologiques comportant une partie clinique servant à la démonstration d'une (1) influence des paramètres temporels de la séquence DCE-MRI et (2) de la technique de post-traitement des données brutes par analyse radiomics sur les performances de modèles prédictifs.
Chacun des deux articles présentent des biais méthodologiques importants: absence de fantôme, de *test-retest* (en particulier, pour l'AIF), absence de segmentation multiples, absence de cohorte de validation, rétrospectivité pour la 2^{ième} étude, absence de courbes de calibration, absence d'études cout-efficacité, absence de corrélations biologiques avec des marqueurs histologiques et moléculaires de la néoangiogenèse... ce qui aboutit à des RQS faibles de 12 et 6.5 respectivement. Nous devons aussi souligner que les indices radiomics calculés par le logiciel Olea pourraient ne pas satisfaire les recommandations IBSI et que ce logiciel n'a pas participé à l'étude de Zwanenburg et al. (2020), ce qui pourrait constituer un biais supplémentaire. Il aurait pu aussi être pertinent de mettre en perspective les valeurs brutes moyennes des AUC et Ktrans (i.e. sans pre-processing) de ces cohortes avec les indices radiomics extraits des imageries post-traitées.

A l'issue de ces résultats, nous avons uniformisé toutes les séquences DCE-MRI en oncologie musculo-squelettique pour une durée de 5mn et une résolution temporelle de 4s et prévu de poursuivre prospectivement l'évaluation de la valeur ajoutée de l'hétérogénéité perfusionnelle et des ses variations dans le cadre des études CIRSARC / NEOSARCOMICS. Il est intéressant de noter que le meilleur des modèles pronostics s'est avéré, encore, celui comportant un mélange de variables radiologiques et de variables radiomics sélectionnées.

5. CONCLUSION

Les approches radiomics suscitent un engouement exponentiel dans la communauté scientifique radiologique, en particulier en cancérologie, pour de multiples raisons : elles sont au carrefour de domaines innovants et attractifs (traitement et analyse d'image, sciences des données / big data, apprentissage statistique, intelligence artificielle, « -omics ») pour un cout relativement modéré, sans nécésssiter d'investissement dans de nouveaux systèmes d'acquisition.

Dans le cas des sarcomes, la plupart des grands axes d'application a été exploré même si les corrélations radiomics-génomique / transcriptomique / protéomique / métabolomique, la transformation et l'intégration des connaissances radiologiques sous forme de variables numériques complémentaires (par exemple : quantification des variations de l'oedème péritumoral), et l'exploration de l'hétérogénéité inter-site en situation métastatique restent à étudier.

Ces approches radiomics arrivent ainsi à un moment décisif de leur évolution : les groupes y travaillant doivent dès-à-présent considérer le franchissement de la preuve de concept pour que ces approches soient validées prospectivement et viennent enrichir l'arsenal diagnostic et thérapeutique pour les patients atteints de sarcomes - si leur valeur ajoutée se confirme. La crainte est que les potentiels marqueurs issus des approches radiomics ne parviennent à compléter tous les éléments définissant un biomarqueur d'imagerie dédié à la cancérologie et se retrouvent, tout comme les précédents innovations radiologiques (DWI, MRS, DCE-MRI), cantonnées dans un entre-deux défini par la preuve qu'il existe une corrélation avec un phénomène physiopathologique mais l'impossibilité de généraliser cette preuve en pratique.

Cette étape de validation risque d'être ici encore plus difficile que pour les autres biomarqueurs tant la méthodologie particulière des approches radiomics s'avère complexe et sujette à de multiples facteurs d'influence. Néanmoins, des initiatives indépendantes internationales telles IBSI regroupant des membres des principaux groupes de recherche en radiomics ont pris la mesure de ces difficultés et s'efforcent de standardiser les définitions des indices radiomics, de recenser et d'expliquer ces méthodes. Cette prise de conscience transparait dans notre thèse puisqu'à l'enthousiasme des premiers travaux - suite à l'identification de signatures prédictives du pronostic (**article 1**) et de la réponse histologique après NAC (**article 2**) - ont

succèdé des questionnements méthodologiques relatifs aux techniques à employer pour uniformiser des données IRM multicentriques tout en limitant la perte d'information. Concernant l'IRM structurale, nous avons illustré l'influence majeure et sous-estimée de la technique de normalisation des intensités de signal sur chaque aspect des études radiomics (indices brutes, classification non-supervisée, analyse supervisée) (**article 3**). A ce titre, nous poursuivons actuellement nos travaux d'investigation des techniques de normalisation mais cette fois-ci prospectivement, comprenant des acquisitions multiples et des séquences de référence par mapping T1 et T2. Concernant les séquences DCE-MRI, nous avons d'abord montré l'influence supplémentaire de certains des paramètres propres de ces séquences sur des prédictions basées sur les cartes de K^{trans} et d'AUC_{90s} (**article 4**), puis montré que passer par le calcul de ces cartes, physiopathologiquement intéressantes, pourraient en pratique diminuer les performances de modèles prédictifs par rapport à des modèles construits sur des données DCE-MRI brutes (**article 5**).

Les résultats des 3 derniers articles et les publications, entretemps, d'IBSI incitent à considérer avec prudence les résultats des 2 premiers. Si les résultats sont cohérents avec les intuitions radiologiques qui ont motivé ces études, les signatures prédictives proposées ne seront très probablement pas définitives. L'apport de ces 2 premiers articles aura aussi été de susciter l'intérêt de la communauté radiologique et surtout oncologique et bioinformatique. Ces preuves de concept incitent ces autres disciplines, rompues aux essais prospectifs et à l'analyse de données « -omics », à investir dans le domaine de la radiomics. Cela se concrétise, à l'institut Bergonié, par l'obtention d'aides humaines et matérielles pour nos projets. Nous travaillons actuellement sur le croisement de données du micro-environnement tumoral, transcriptomiques et radiomics IRM de STM afin, *in fine*, de produire une signature radiogenomics. Nos résultats préliminaires tendent à montrer que les données radiomics (à l'échelle de la tumeur entière) et moléculaires CINSARC (à l'échelle d'une biopsie) sont finalement peu corrélées entre elles mais, par contre, qu'elles pourraient se compléter pour accroitre les performances pronostiques. Enfin, d'autres projets prospectifs dédiés aux STM sont en cours d'élaboration concernant l'étude de l'hétérogénéité inter-tumorale au cours de l'évolution métastatique et sa signification pronostique, ainsi que la recherche d'une hétérogénéité intra-tumorale de la signature CINSARC et, si présente, des moyens de l'identifier par TEP-IRM multimodale (2 financements régionaux obtenus).

La facilitation et l'accélération des analyses radiomics sont un autre axe de développement et plusieurs équipes travaillent à des pipelines d'automatisation des étapes chronophages et complexes de ces approches incluant des algorithmes (i) de segmentation multi-séquences, multi-modalités en 3D par deep-learning de la tumeur elle-même et de la périphérie tumorale, (ii) d'harmonisation des images, (iii) de l'extraction des indices radiomics intra- et inter-sites, et (iv) de l'intégration dans des modèles pronostics pour fournir une prédiction individuelle. Participer à la mise en place de tels pipelines est l'objectif de ma future mobilité à l'université de Cambridge dans l'équipe du Pr Sala.

En définitive, les résultats de cette thèse soutiennent l'hypothèse que la quantification et l'intégration des phénotypes radiologiques dans des modèles prédictifs a le potentiel d'améliorer les prises en charge diagnostique et thérapeutique des patients atteints de sarcome. En dépit de la relative rareté de ces tumeurs, la structuration en réseau du groupe sarcome français et l'organisation précoce en base de données nationales clinicobiologiques pourraient être une opportunité d'accélérer l'évaluation multicentrique pour les chercheurs en radiomics des sarcomes. Cependant, transformer cette hypothèse en un outil pratique et concret pour le patient nécessitera un effort collectif considérable, humain et financier, passant par un investissement concret et conséquent dans l'élaboration de pipelines d'automatisation des taches reposant sur l'intelligence artificielle et le respect des recommandations d'instances internationales indépendantes.

6. BIBLIOGRAPHIE

Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.

Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol. 2016;2:1636–1642.

Ahlawat S, Fritz J, Morris CD, Fayad LM. Magnetic resonance imaging biomarkers in musculoskeletal soft tissue tumors: Review of conventional features and focus on nonmorphologic imaging. J Magn Reson Imaging. 2019 ;50(1):11-27.

Alic L, van Vliet M, Wielopolski PA, et al. Regional heterogeneity changes in DCE-MRI as response to isolated limb perfusion in experimental soft-tissue sarcomas. Contrast Media Mol Imaging. 2013 Jul-Aug;8(4):340-9.

Altazi BA, Zhang GG, Fernandez DC, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J Appl Clin Med Phys. 2017;18(6):32-48.

Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans. Syst. Man Cybern. 1989;19:1264–74.

Andersen KF, Fuglo HM, Rasmussen SH, et al. Semi-Quantitative Calculations of Primary Tumor Metabolic Activity Using F-18 FDG PET/CT as a Predictor of Survival in 92 Patients With High-Grade Bone or Soft Tissue Sarcoma. Medicine (Baltimore). 2015a;94(28):e1142.

Andersen KF, Fuglo HM, Rasmussen SH, et al. Volume-Based F-18 FDG PET/CT Imaging Markers Provide Supplemental Prognostic Information to Histologic Grading in Patients With High-Grade Bone or Soft Tissue Sarcoma. Medicine (Baltimore). 2015b;94(51):e2319.

Antonescu CR, Tschernyavsky SJ, Decuseara R, et al. Prognostic impact of P53 status, TLS-CHOP fusion transcript structure, and histological grade in myxoid liposarcoma: a molecular and clinicopathologic study of 82 cases. Clin Cancer Res 2001;7:3977–3987

Asano N, Susa M, Hosaka S, et al. Metastatic patterns of myxoid/round cell liposarcoma: a review of a 25-year experience. Sarcoma 2012:345161.

Bagher-Ebadian H, Siddiqui F, Liu C, Movsas B, Chetty IJ. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. Med Phys. 2017;44(5):1755-1770.

Becker AS, Wagner MW, Wurnig MC, Boss A. Diffusion-weighted imaging of the abdomen: Impact of b-values on texture analysis features. NMR Biomed. 2017;30(1).

Benz MR, Allen-Auerbach MS, Eilber FC, et al. Combined assessment of metabolic and volumetric changes for assessment of tumor response in patients with soft-tissue sarcomas. J Nucl Med. 2008;49(10):1579-84.

Benz MR, Czernin J, Allen-Auerbach MS, et al. FDG-PET/CT imaging predicts histopathologic treatment responses after the initial cycle of neoadjuvant chemotherapy in high-grade soft-tissue sarcomas. Clin Cancer Res. 2009;15(8):2856-63.

Benz MR, Dry SM, Eilber FC, et al. Correlation between glycolytic phenotype and tumor grade in soft-tissue sarcomas by 18F-FDG PET. J Nucl Med. 2010;51(8):1174-81.

Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407-415.

Bianchi J, Gonçalves JR, Ruellas ACO, et al. Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles. Dentomaxillofac Radiol. 2019;48(6):20190049.

Blay JY, Soibinet P, Penel N, et al. Improved survival using specialized multidisciplinary board in sarcoma patients. Ann Oncol. 2017;28(11):2852-2859.

Blay JY, Honoré C, Stoeckle E, et al. Surgery in reference centers improves survival of sarcoma patients: a nationwide study. Ann Oncol. 2019;30(8):1407.

Benjamin RS. Pharmacokinetics of adriamycin (NSC-123127) in patients with sarcomas. Cancer Chemother Rep. 1974;58(2):271-3.

Benjamin RS, Choi H, Macapinlac HA, et al. We should desist using RECIST, at least in GIST. J Clin Oncol. 2007;25(13):1760-4.

Beuzit L, Eliat P-A, Brun V, et al. Dynamic contrast-enhanced MRI: Study of inter-software accuracy and reproducibility using simulated and clinical data. J Magn Reson Imaging 2016;43(6):1288–1300.

Bogowicz M, Riesterer O, Bundschuh RA, et al. Stability of radiomic features in CT perfusion maps. Phys Med Biol 2016; 61:8736–8749.

Bologna M, Corino V, Mainardi L. Technical Note: Virtual phantom analyses for preprocessing evaluation and detection of a robust feature set for MRI-radiomics of the brain. Med Phys. 2019;46(11):5116-5123.

Bowen SR, Yuh WTC, Hippe DS, et al. Tumor radiomic heterogeneity: Multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. J Magn Reson Imaging 2018; 47:1388–1396.

Braman NM, Etesami M, Prasanna P, et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res 2017; 19:57.

Branchini M, Zorz A, Zucchetta P, et al. Impact of acquisition count statistics reduction and SUV discretization on PET radiomic features in pediatric 18F-FDG-PET/MRI examinations. Phys Med. 2019;59:117-126.

Breiman L. Random Forests. Machine Learning 2001;45:5–32

Brier. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review. 1950;78: 1–3.

Bruix J, Sherman M, Llovet JM, et al; EASL Panel of Experts on HCC. Clinical management of hepatocellular carcinoma. Conclusions of the Barcelona-2000 EASL conference. European Association for the Study of the Liver. J Hepatol. 200;35(3):421-30.

Buch K, Li B, Qureshi MM, et al. Quantitative Assessment of Variation in CT Parameters on Texture Features: Pilot Study Using a Nonanatomic Phantom. AJNR Am J Neuroradiol. 2017;38(5):981-985.

Buch K, Kuno H, Qureshi MM, Li B, Sakai O. Quantitative variations in texture analysis features dependent on MRI scanning parameters: A phantom model. J Appl Clin Med Phys. 2018;19(6):253-264.

Buvat I, Orlhac F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. J Nucl Med. 2019;60(11):1543-1544.

Caramella C, Allorant A, Orlhac F, et al. Can we trust the calculation of texture indices of CT images? A phantom study. Med Phys 2018;45(4):1529–1536.

Casali PG, Abecassis N, Aro HT, et al; ESMO Guidelines Committee and EURACAN. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2018;29(Suppl 4):iv268-iv269.

Chapiro J, Duran R, Lin M, et al. Transarterial chemoembolization in soft-tissue sarcoma metastases to the liver - the use of imaging biomarkers as predictors of patient survival. Eur J Radiol. 2015;84(3):424-430.

Chawla NVBKW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2003;2002:341–78.

Cheng H-LM. Investigation and optimization of parameter accuracy in dynamic contrast-enhanced MRI. J Magn Reson Imaging 2008;28(3):736–743.

Cherezov D, Hawkins SH, Goldgof DB, et al. Delta radiomic features improve prediction for lung cancer incidence: A nested case-control analysis of the National Lung Screening Trial. Cancer Med. 2018;7(12):6340-6356.

Chibon F, Lagarde P, Salas S, et al. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. Nat Med. 2010;16(7):781-7.

Choi H, Charnsangavej C, Faria SC, et al. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. J Clin Oncol. 2007;25(13):1753-9.

Conners RW, Trivedi MM, Harlow CA. Segmentation of a high-resolution urban scene using texture operators. Computer Vision, Graphics, and Image Processing. 1984;25:273–310.

Coindre JM, Terrier P, Bui NB, et al. Prognostic factors in adult patients with locally controlled soft tissue sarcoma. A study of 546 patients from the French Federation of Cancer Centers Sarcoma Group. J Clin Oncol. 1996 Mar;14(3):869-77.

Coindre JM, Terrier P, Guillou L, et al. Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French Federation of Cancer Centers Sarcoma Group. Cancer. 2001;91(10):1914-26.

Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging. 2004;22(1):81-91.

Collins DL, Zijdenbos AP, Kollokian V, et al. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging. 1998;17(3):463-8.

Corino VDA, Montin E, Messina A, et al. Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. J Magn Reson Imaging. 2018;47(3):829-840.

Cousin S, Crombe A, Stoeckle E, et al. Clinical, radiological and genetic features, associated with the histopathologic response to neoadjuvant chemotherapy (NAC) and outcomes in locally advanced soft tissue sarcoma (STS) patients (pts). Journal of Clinical Oncology 2017 35:15_suppl, 11014-11014

Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) 1972;34(2):187-220

Crombe A, Alberti N, Stoeckle E, et al. Soft tissue masses with myxoid stroma: Can conventional magnetic resonance imaging differentiate benign from malignant tumors? Eur J Radiol. 2016;85(10):1875-1882.

Crombé A, Loarer FL, Alberti N, et al (2018) Homogeneous myxoid liposarcomas mimicking cysts on MRI: A challenging diagnosis. Eur J Radiol 102:41–48.

Crombé A, Marcellin PJ, Buy X, et al. Soft-Tissue Sarcomas: Assessment of MRI Features Correlating with Histologic Grade and Patient Outcome. Radiology. 2019a Jun;291(3):710-721.

Crombé A, Brisse HJ, Ledoux P, et al. Alveolar soft-part sarcoma: can MRI help discriminating from other soft-tissue tumors? A study of the French sarcoma group. Eur Radiol. 2019b Jun;29(6):3170-3182.

Crombé A, Le Loarer F, Cornelis F, et al. High-grade soft-tissue sarcoma: optimizing injection improves MRI evaluation of tumor response. Eur Radiol. 2019c Feb;29(2):545-555.

Crombé A, Périer C, Kind M, et al. T2 -based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. J Magn Reson Imaging. 2019;50(2):497-510.

Crombé A, Saut O, Guigui J, et al. Influence of temporal parameters of DCE-MRI on the quantification of heterogeneity in tumor vascularization. J Magn Reson Imaging. 2019;50(6):1773-1788.

Crombé A, Fadli D, Buy X, et al. High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models? J Magn Reson Imaging. 2020. doi: 10.1002/jmri.27040.

Crombé A, Le Loarer F, Sitbon M, et al. Can radiomics improve the prediction of metastatic relapse of myxoid/round cell liposarcomas? Eur Radiol. 2020. doi: 10.1007/s00330-019-06562-5.

Cuenod CA, Balvay D. Perfusion and vascular permeability: basic concepts and measurement in DCE-CT and DCE-MRI. Diagn Interv Imaging. 2013;94(12):1187-204.

De La Hoz Polo M, Dick E, Bhumbra R, et al. Surgical considerations when reporting MRI studies of soft tissue sarcoma of the limbs. Skeletal Radiol. 2017;46(12):1667-1678.

Depeursinge A and Julien Fageot. Biomedical Texture Operators and Aggregation Functions. In Adrien Depeursinge, Julien Fageot, and Omar Al-Kadi, editors, Biomedical texture analysis, chapter 3, pages 63–101. Academic Press, London, UK, 1st edition, 2017.

Del Grande F, Subhawong T, Weber K, et al. Detection of soft-tissue sarcoma recurrence: added value of functional MR imaging techniques at 3.0 T. Radiology. 2014;271(2):499-511.

DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.

Dewey BE, Zhao C, Reinhold JC, et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging 2019:64:160–170.

Dong Y, Feng Q, Yang W, et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. Eur Radiol. 2018;28:582–591.

Du J, Li K, Zhang W, Wang S, et al. Intravoxel Incoherent Motion MR Imaging: Comparison of Diffusion and Perfusion Characteristics for Differential Diagnosis of Soft Tissue Tumors. Medicine (Baltimore). 2015;94(25):e1028.

Dudeck O, Zeile M, Pink D, et al. Diffusion-weighted magnetic resonance imaging allows monitoring of anticancer treatment effects in patients with soft-tissue sarcomas. J Magn Reson Imaging. 2008;27(5):1109-13.

Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One. 2019;14(3):e0213459.

Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. J Nucl Med. 2008;49(12):1973-9.

Eary JF, Conrad EU, O'Sullivan J, Hawkins DS, Schuetze SM, O'Sullivan F. Sarcoma mid-therapy [F-18]fluorodeoxyglucose positron emission tomography (FDG PET) and patient outcome. J Bone Joint Surg Am. 2014;96(2):152-8.

Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.

Elias DA, White LM, Simpson DJ, et al. Osseous invasion by soft-tissue sarcoma: assessment with MR imaging. Radiology. 2003;229(1):145-52.

Engström K, Bergh P, Gustafson P, et al. Liposarcoma: outcome based on the Scandinavian Sarcoma Group register. Cancer 2008;113:1649–1656.

Erfanian Y, Grueneisen J, Kirchner J, et al. Integrated 18F-FDG PET/MRI compared to MRI alone for identification of local recurrences of soft tissue sarcomas: a comparison trial. Eur J Nucl Med Mol Imaging. 2017;44(11):1823-1831.

Esser M, Kloth C, Thaiss WM, et al. CT-response patterns and the role of CT-textural features in inoperable abdominal/retroperitoneal soft tissue sarcomas treated with trabectedin. Eur J Radiol. 2018t;107:175-182.

Esser M, Kloth C, Thaiss WM, et al. CT-morphologic and CT-textural patterns of response in inoperable soft tissue sarcomas treated with pazopanib-a preliminary retrospective cohort study. Br J Radiol. 2019;92(1103):20190158.

Evilevitch V, Weber WA, Tap WD, et al. Reduction of glucose metabolic activity is more accurate than change in size at predicting histopathologic response to neoadjuvant therapy in high-grade soft-tissue sarcomas. Clin Cancer Res. 2008;14(3):715-20.

Fan M, Cheng H, Zhang P, et al. DCE-MRI texture analysis with tumor subregion partitioning for predicting Ki-67 status of estrogen receptor-positive breast cancers. J Magn Reson Imaging 2018; 48:237–247.

Finette S, Bleier AR, Swindell W, Haber K. Breast tissue classification using diagnostic ultrasound and pattern recognition techniques: II. Experimental results. Ultrason Imaging. 1983 Jan;5(1):71-86.

Finette S, Bleier A, Swindell W. Breast tissue classification using diagnostic ultrasound and pattern recognition techniques: I. Methods of pattern recognition. Ultrason Imaging. 1983 Jan;5(1):55-70.

Fiore M, Grosso F, Lo Vullo S, et al. Myxoid/round cell and pleomorphic liposarcomas: prognostic factors and survival in a series of patients treated at a single institution. Cancer 2007;109:2522–2531.

Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. WHO Classification of Tumours of Soft Tissue and Bone, 2013, 4th edition, Vol 5. IARC Press, Lyon, France

Ford J, Dogan N, Young L, Yang F. Quantitative Radiomics: Impact of Pulse Sequence Parameter Selection on MRI-Based Textural Features of the Brain. Contrast Media Mol Imaging. 2018;2018:1729071.

Fornacon-Wood I, Mistry H, Ackermann CJ, Blackhall F, McPartlin A, Faivre-Finn C, Price GJ, O'Connor JPB. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. 2020 Jun 1. doi: 10.1007/s00330-020-06957-9.

Fortin JP, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data.

Neuroimage. 2017;161:149-170.

Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage. 2018;167:104-120.

Foy JJ, Armato SG 3rd, Al-Hallaq HA. Effects of variability in radiomics software packages on classifying patients with radiation pneumonitis. J Med Imaging (Bellingham). 2020;7(1):014504.

Friedman J1, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22.

Fuglø HM, Maretty-Nielsen K, Hovgaard D, et al. Metastatic pattern, local relapse, and survival of patients with myxoid liposarcoma: a retrospective study of 45 patients. Sarcoma 2013:548628.

Galloway MM. Texture analysis using gray level run lengths Pattern recognition Letters 1975 ;11:172-179.

Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. Clin Radiol. 2010;65(7):517-21.

Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278:563–577.

Gimber LH, Montgomery EA, Morris CD, Krupinski EA, Fayad LM. MRI characteristics associated with high-grade myxoid liposarcoma. Clin Radiol. 2017;72(7):613.e1-613.e6.

Ginsburg SB, Algohary A, Pahwa S, et al. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: Preliminary findings from a multi-institutional study. J Magn Reson Imaging 2017; 46:184–193.

Gnep K, Fargeas A, Gutiérrez-Carvajal RE, et al. Haralick textural features on T2 -weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. J Magn Reson Imaging. 2017;45:103–117.

Gondim Teixeira PA, Renaud A, Aubert S, et al. Perfusion MR imaging at 3-Tesla: Can it predict tumor grade and histologic necrosis rate of musculoskeletal sarcoma? Diagn Interv Imaging 2018;99(7–8):473–481.

Gorelik N, Reddy SMV, Turcotte RE, et al. Early detection of metastases using whole-body MRI for initial staging and routine follow-up of myxoid liposarcoma. Skeletal Radiol. 2018;47(3):369-379.

Gouin F, Renault A, Bertrand-Vasseur A, et al. Early detection of multiple bone and extra-skeletal metastases by body magnetic resonance imaging (BMRI) after treatment of Myxoid/Round-Cell Liposarcoma (MRCLS). Eur J Surg Oncol. 2019;45(12):2431-2436.

Gortzak E, Azzarelli A, Buesa J, et al; E.O.R.T.C. Soft Tissue Bone Sarcoma Group and the National Cancer Institute of Canada Clinical Trials Group/Canadian Sarcoma Group. A randomised phase II study on neo-adjuvant chemotherapy for 'high-risk' adult soft-tissue sarcoma. Eur J Cancer. 2001;37(9):1096-103.

Goya-Outi J, Orlhac F, Calmon R, et al. Computation of reliable textural indices from multimodal brain MRI: suggestions based on a study of patients with diffuse intrinsic pontine glioma. Phys Med Biol. 2018 May 10;63(10):105003.

Groisberg R, Hong DS, Holla V, et al. Clinical genomic profiling to identify actionable alterations for investigational therapies in patients with diverse sarcomas. Oncotarget. 2017;8(24):39254-39267.

Gronchi A, Ferrari S, Quagliuolo V, et al. Histotype-tailored neoadjuvant chemotherapy versus standard chemotherapy in patients with high-risk soft-tissue sarcomas (ISG-STS 1001): an

international, open-label, randomised, controlled, phase 3, multicentre trial. Lancet Oncol. 2017;18(6):812-822.

Grootjans W, Tixier F, van der Vos CS, et al. The Impact of Optimal Respiratory Gating and Image Noise on Evaluation of Intratumor Heterogeneity on 18F-FDG PET Imaging of Lung Cancer. J Nucl Med. 2016;57(11):1692-1698.

Grünwald V, Litière S, Young R, et al; EORTC STBSG. Absence of progression, not extent of tumour shrinkage, defines prognosis in soft-tissue sarcoma - An analysis of the EORTC 62012 study of the EORTC STBSG. Eur J Cancer. 2016;64:44-51.

Guo Z et al. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging IEEE TRANSACTIONS ON RADIATION AND PLASMA MEDICAL SCIENCES, 2019 ;3(2)

Haniball J, Sumathi VP, Kindblom L-G, et al. Prognostic factors and metastatic patterns in primary myxoid/round-cell liposarcoma. Sarcoma 2011:538085.

Hanna SL, Fletcher BD, Parham DM, Bugg MF. Muscle edema in musculoskeletal tumors: MR imaging characteristics and clinical significance. J Magn Reson Imaging. 1991;1:441–449.

Hao W, Zhao B, Wang G, Wang C, Liu H. Influence of scan duration on the estimation of pharmacokinetic parameters for breast lesions: a study based on CAIPIRINHA-Dixon-TWIST-VIBE technique. Eur Radiol 2015; 25:1162–1171.

Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, 1973;SMC-3(6) :610-621.

Harrell FE, Kerry LL, Mark DB. Multivariable prognostic models : issues in developing models, evaluating assumtions and adequacy, and measuring ad reducing errors. 1996;15(4):361-387.

Hayano K, Tian F, Kambadakone AR, et al. Texture Analysis of Non-Contrast-Enhanced Computed Tomography for Assessing Angiogenesis and Survival of Soft Tissue Sarcoma. J Comput Assist Tomogr. 2015;39(4):607-12.

Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92-105.

Heisen M, Fan X, Buurman J, et al. The influence of temporal resolution in determining pharmacokinetic parameters from DCE-MRI data. Magn Reson Med 2010; 63:811–816.

Henderson S, Purdie C, Michie C, et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. Eur Radiol. 2017;27:4602–4611.

Hermessi H, O Mourali, E Zagrouba. Deep feature learning for soft tissue sarcoma classification in MR images via transfer learning. Expert Systems with Applications, 2019 ;120 :116-127.

Herrmann K, Benz MR, Czernin J, Allen-Auerbach MS, Tap WD, Dry SM, Schuster T, Eckardt JJ, Phelps ME, Weber WA, Eilber FC.
18F-FDG-PET/CT Imaging as an early survival predictor in patients with primary high-grade soft tissue sarcomas undergoing neoadjuvant therapy.
Clin Cancer Res. 2012 Apr 1;18(7):2024-31. doi: 10.1158/1078-0432.CCR-11-2139. Epub 2012 Feb 14.

Hocquelet A, Auriac T, Perier C, et al. Pre-treatment magnetic resonance-based texture features as potential imaging biomarkers for predicting event free survival in anal cancer treated by chemoradiotherapy. Eur Radiol. 2018;28:2801–2811.

Holbrook M, Blocker SJ, Mowery MD, Badea CT. Multi-modal MRI segmentation of sarcoma tumors using convolutional neural networks proceedings Volume 10948, Medical Imaging 2019: Physics of Medical Imaging; 109484D (2019)

Holzapfel K, Regler J, Baum T, et al. Local Staging of Soft-Tissue Sarcoma: Emphasis on Assessment of Neurovascular Encasement-Value of MR Imaging in 174 Confirmed Cases. Radiology. 2015;275(2):501-9.

Hong JH, Jee WH, Jung CK, et al. Soft tissue sarcoma: adding diffusion-weighted imaging improves MR imaging evaluation of tumor margin infiltration. Eur Radiol 2019 ;29(5):2589-2597.

Honoré C, Méeus P, Stoeckle E, Bonvalot S. Soft tissue sarcoma in France in 2015: Epidemiology, classification and organization of clinical care. J Visc Surg. 2015;152(4):223-30.

Huang W, Beckett BR, Tudorica A, et al. Evaluation of Soft Tissue Sarcoma Response to Preoperative Chemoradiotherapy Using Dynamic Contrast-Enhanced Magnetic Resonance Imaging. Tomography. 2016a;2(4):308-316.

Huang W, Chen Y, Fedorov A, et al. The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge. Tomography 2016b; 2:56–66.

Isaksson LJ, Raimondi S, Botta F, et al. Effects of MRI image normalization techniques in prostate cancer radiomics. Phys Med. 2020;71:7-13

Issels RD, Lindner LH, Verweij J, et al; European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group (EORTC-STBSG); European Society for Hyperthermic Oncology (ESHO). Neo-adjuvant chemotherapy alone or with regional hyperthermia for localised high-risk soft-tissue sarcoma: a randomised phase 3 multicentre study. Lancet Oncol. 2010;11(6):561-70.

Italiano A, **Stoeckle E**. Role of perioperative chemotherapy in soft-tissue sarcomas: It's time to end a never-ending story. Eur J Cancer. 2018;97:53-54.

Jackson A, O'Connor JPB, Parker GJM, Jayson GC. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. Clin Cancer Res 2007; 13:3449–3459.

James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning – with applications in R. 2017, Springer edition.

Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics. 2007; 8(1):118–127.

Karavasilis V, Seddon BM, Ashley S, et al. Significant clinical benefit of first-line palliative chemotherapy in advanced soft-tissue sarcoma: retrospective analysis and identification of prognostic factors in 488 patients. Cancer. 2008;112(7):1585-91.

Khorrami M, Khunger M, Zagouras A, et al. Combination of Peri- and Intratumoral Radiomic Features on Baseline CT Scans Predicts Response to Chemotherapy in Lung Adenocarcinoma. Radiology: Artificial Intelligence 2019;1:180012.

Kikuta K, Kubota D, Yoshida A, et al. An analysis of factors related to the tail-like pattern of myxofibrosarcoma seen on MRI. Skeletal Radiol. 2015;44(1):55-62.

Kim H, Park CM, Park SJ, et al. Temporal Changes of Texture Features Extracted From Pulmonary Nodules on Dynamic Contrast-Enhanced Chest Computed Tomography: How Influential Is the Scan Delay? Invest Radiol 2016;51(9):569–574.

Kim HS, Kim JH, Yoon YC, Choe BK. Tumor spatial heterogeneity in myxoid-containing soft tissue using texture analysis of diffusion-weighted MRI. PLoS One. 2017;12(7):e0181339.

Kuhl CK, Mielcareck P, Klaschik S, et al. Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? Radiology. 1999;211(1):101-10.

Kuhn, M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 2008;28(5) doi :10.18637/jss.v028.i05

Kuyumcu G, Rubin BP, Bullen J, Ilaslan H (2018) Quantification of fat content in lipid-rich myxoid liposarcomas with MRI: a single-center experience with survival analysis. Skeletal Radiol 2018;47:1411–1417.

Labarre D, Aziza R, Filleron T, et al. Detection of local recurrences of limb soft tissue sarcomas: is magnetic resonance imaging (MRI) relevant? Eur J Radiol. 2009;72(1):50-3.

Lacroix M, Frouin F, Dirand AS, Nioche C, Orlhac F, Bernaudin JF, Brillet PY, Buvat I. Correction for Magnetic Field Inhomogeneities and Normalization of Voxel Values Are Needed to Better Reveal the Potential of MR Radiomic Features in Lung Cancer. Front Oncol. 2020 Jan 31;10:43. doi: 10.3389/fonc.2020.00043. eCollection 2020.

Lakkaraju A, Sinha R, Garikipati R, Edward S, Robinson P. Ultrasound for initial evaluation and triage of clinically suspicious soft-tissue masses. Clin Radiol. 2009;64(6):615-21.

Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749-762.

Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. Radiology. 1986;161(2):401-7.

Le Bihan D. What can we see with IVIM MRI? Neuroimage. 2019;187:56-67.

Le Doussal V, Coindre JM, Leroux A, et al. Prognostic factors for patients with localized primary malignant fibrous histiocytoma: a multicenter study of 216 patients with multivariate analysis. Cancer. 1996;77(9):1823-30.

Ledoux P, Kind M, Le Loarer F, et al. Abnormal vascularization of soft-tissue sarcomas on conventional MRI: Diagnostic and prognostic values. Eur J Radiol. 2019;117:112-119.

Lee JH, Yoon YC, Seo SW, Choi YL, Kim HS. Soft tissue sarcoma: DWI and DCE-MRI parameters correlate with Ki-67 labeling index. Eur Radiol. 2020;30(2):914-924.

Lefkowitz RA, Landa J, Hwang S, et al. Myxofibrosarcoma: prevalence and diagnostic value of the "tail sign" on magnetic resonance imaging. Skeletal Radiol. 2013;42(6):809-18.

Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Semin Liver Dis. 2010;30(1):52-60.

Lerski RA, Straughan K, Schad LR, et al. MR image texture analysis--an approach to tissue characterization. Magn Reson Imaging. 1993;11(6):873-87.

Liang J, Sammet S, Yang X, et al. Intraindividual in vivo comparison of gadolinium contrast agents for pharmacokinetic analysis using dynamic contrast enhanced magnetic resonance imaging. Invest Radiol 2010;45:233–244

Limkin EJ, Reuzé S, Carré A, et al. The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. Sci Rep. 2019;9(1):4329.

Liu C, Ding J, Spuhler K, et al.: Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI. J Magn Reson Imaging 2019; 49:131–140.

Lo Gullo R, Daimiel I, Morris EA, Pinker K. Combining molecular and imaging metrics in cancer: radiogenomics. Insights Imaging. 2020;11(1):1.

Löwenthal D, Zeile M, Niederhagen M, et al. Differentiation of myxoid liposarcoma by magnetic resonance imaging: a histopathologic correlation. Acta Radiol 2014;55:952–960.

Lu H, Arshad M, Thornton A, et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. Nat Commun 2019;10:764.

Lucchesi C, Khalifa E, Laizet Y, et al. Targetable Alterations in Adult Patients With Soft-Tissue Sarcomas: Insights for Personalized Therapy. JAMA Oncol. 2018;4(10):1398-1404.

Lynch JA, Hawkes DJ, Buckland-Wright JC. Analysis of texture in macroradiographs of osteoarthritic knees using the fractal signature. Phys Med Biol. 1991;36(6):709-22.

Mackin D, Fave X, Zhang L, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol. 2015;50(11):757-65.

Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, Court L. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One. 2017;12(9):e0178524.

Macpherson RE, Pratap S, Tyrrell H, et al. Retrospective audit of 957 consecutive 18F-FDG PET-CT scans compared to CT and MRI in 493 patients with different histological subtypes of bone and soft tissue sarcoma. Clin Sarcoma Res. 2018;8:9.

Malinauskaite I, Hofmeister J, Burgermeister S, et al. Radiomics and Machine Learning Differentiate Soft-Tissue Lipoma and Liposarcoma Better than Musculoskeletal Radiologists. Sarcoma. 2020;2020:7163453.

Manjon JV, Coupé P, Buades A, et al. Adaptative non-local means denoising of MR images with spatially varying noise levels. J Magn Reson Imaging 2010 ;31(1) :192-203.

Martin-Carreras T, Li H, Cooper K, Fan Y, Sebro R. Radiomic features from MRI distinguish myxomas from myxofibrosarcomas. BMC Med Imaging. 2019;19(1):67.

Meier A, Veeraraghavan H, Nougaret S et al. Association between CT-texture-derived tumor heterogeneity, outcomes, and BRCA mutation status in patients with high-grade serous ovarian cancer. Abdom Radiol (NY). 2019;44(6):2040-2047.

Messiou C, Bonvalot S, Gronchi A, et al. Evaluation of response after pre-operative radiotherapy in soft tissue sarcomas; the European Organisation for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group (EORTC-STBSG) and Imaging Group recommendations for radiological examination and reporting with an emphasis on magnetic resonance imaging. Eur J Cancer. 2016;56:37-44.

Meyer HJ, Renatus K, Höhn AK, et al. Texture analysis parameters derived from T1-and T2-weighted magnetic resonance images can reflect Ki67 index in soft tissue sarcoma. Surg Oncol. 2019;30:92-97.

Meyer JM, Perlewitz KS, Hayden JB, et al. Phase I trial of preoperative chemoradiation plus sorafenib for high-risk extremity soft tissue sarcomas with dynamic contrast-enhanced MRI correlates. Clin Cancer Res. 2013;19(24):6902-11.

Michaely HJ, Morelli JN, Budjan J, et al. CAIPIRINHA-Dixon-TWIST (CDT)-volume-interpolated breath-hold examination (VIBE): a new technique for fast time-resolved dynamic 3-dimensional imaging of the abdomen with high spatial resolution. Invest Radiol 2013; 48:590–597.

Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. Cancer. 1981;47(1):207-14.

Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys. 2020;21(1):179-190.

Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

Muschelli J, Gherman A, Fortin J-P, et al. Neuroconductor: an R platform for medical imaging analysis. Biostatistics 2019;20(2):218–239.

Nakamura T, Matsumine A, Matsubara T, et al. Infiltrative tumor growth patterns on magnetic resonance imaging associated with systemic inflammation and oncological outcome in patients with high-grade soft-tissue sarcoma. PLoS One. 2017;12(7):e0181787.

Nie K, Shi L, Chen Q, et al. Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. Clin Cancer Res 2016; 22:5256–5264.

Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res. 2018;78(16):4786-4789.

Nketiah G, Elschot M, Kim E, et al. T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. Eur Radiol. 2017;27:3050–3059.

Noebauer-Huhmann IM, Weber MA, Lalam RK, et al. Soft Tissue Tumors in Adults: ESSR-Approved Guidelines for Diagnostic Imaging. Semin Musculoskelet Radiol. 2015;19(5):475-82.

Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med. 1999;42(6):1072-81.

Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19:143–150.

O'Connor JPB. Cancer heterogeneity and imaging. Semin Cell Dev Biol 2017; 64:48–57.

O'Connor JP, Jackson A, Parker GJ, et al. Dynamic contrast-enhanced MRI in clinical trials of antivascular therapies. Nat Rev Clin Oncol. 2012;9(3):167-77.

O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169-186.

O'Sullivan F, Wolsztynski E, O'Sullivan J, et al. A statistical modeling approach to the analysis of spatial patterns of FDG-PET uptake in human sarcoma. IEEE Trans Med Imaging. 2011;30(12):2059-71

Orlhac F, Boughdad S, Philippe C, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med. 2018;59(8):1321-1328.

Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology. 2019;291(1):53-59.

Othman AE, Falkner F, Martirosian P, et al. Optimized Fast Dynamic Contrast-Enhanced Magnetic Resonance Imaging of the Prostate: Effect of Sampling Duration on Pharmacokinetic Parameters. Invest Radiol 2016; 51:106–112.

Othman AE, Falkner F, Weiss J, et al. Effect of Temporal Resolution on Diagnostic Performance of Dynamic Contrast-Enhanced Magnetic Resonance Imaging of the Prostate. Invest Radiol 2016; 51:290–296.

Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020a;30(1):523-536.

Park JE, Kim HS, Kim D, et al. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer. 2020b;20(1):29.

Park SY, Chung HW, Chae SY, Lee JS. Comparison of MRI and PET-CT in detecting the loco-regional recurrence of soft tissue sarcomas during surveillance. Skeletal Radiol. 2016;45(10):1375-84.

Panicek DM, Go SD, Healey JH, et al. Soft-tissue sarcoma involving bone or neurovascular structures: MR imaging prognostic factors. Radiology. 1997;205(3):871-5.

Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

Peeken JC, Spraker MB, Knebel C, et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. EBioMedicine. 2019a;48:332-340.

Peeken JC, Bernhofer M, Spraker MB, et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. Radiother Oncol. 2019b;135:187-196.

Penel N, Coindre JM, Giraud A, et al. Presentation and outcome of frequent and rare sarcoma histologic subtypes : a study of 10262 patients with localized visceral / soft tissue sarcoma managed in reference centers. Cancer 2018 ; 124(6) :1179-1187.

Peng Y, Bi L, Guo Y, Feng D, Fulham M, Kim J. Deep multi-modality collaborative learning for distant metastases predication in PET-CT soft-tissue sarcoma studies. Conf Proc IEEE Eng Med Biol Soc. 2019;2019:3658-3688.

Petitprez F, de Reyniès A, Keung EZ, et al. B cells are associated with survival and immunotherapy response in sarcoma.Nature. 2020;577(7791):556-560.

Petscavage-Thomas JM, Walker EA, Logie CI, et al. Soft-tissue myxomatous lesions: review of salient imaging features with pathologic comparison. Radiographics 2014;34:964–980.

Portera CA Jr, Ho V, Patel SR, et al. Alveolar soft part sarcoma: clinical course and patterns of metastasis in 70 patients treated at a single institution. Cancer. 2001;91(3):585-91.

Puri A, Gulia A, Hawaldar R, Ranganathan P, Badwe RA. Does intensity of surveillance affect survival after surgery for sarcomas? Results of a randomized noninferiority trial. Clin Orthop Relat Res. 2014;472(5):1568-75.

Ray-Coquard I, Collard O, Ducimetiere F, et al. Treatment patterns and survival in an exhaustive French cohort of pazopanib-eligible patients with metastatic soft tissue sarcoma (STS). BMC Cancer 2017; 17:111.

Ray-Coquard I, Serre D, Reichardt P, Martín-Broto J, Bauer S. Options for treating different soft tissue sarcoma subtypes. Future Oncol. 2018;14(10s):25-49.

Robinson K, Li H, Lan L, Schacht D, Giger M. Radiomics robustness assessment and classification evaluation: A two-stage method demonstrated on multivendor FFDM. Med Phys 2019;46(5):2145–2156.

Robitaille N, Mouiha A, Crépeault B, Valdivia F, Duchesne S, The Alzheimer's Disease Neuroimaging Initiative. Tissue-based MRI intensity standardization: application to multicentric datasets. Int J Biomed Imaging. 2012:347120.

Rose CJ, Mills SJ, O'Connor JPB, et al. Quantifying spatial heterogeneity in dynamic contrast-enhanced MRI parameter maps. Magn Reson Med 2009; 62:488–499.

Rothermundt C, Whelan JS, Dileo P, et al. What is the role of routine follow-up for localised limb soft tissue sarcomas? A retrospective analysis. Br J Cancer. 2014;110(10):2420-6.

Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

Saponara M, Stacchiotti S, Casali PG, Gronchi A. (Neo)adjuvant treatment in localised soft tissue sarcoma: The unsolved affair. Eur J Cancer. 2017;70:1-11.

Scalco E, Belfatto A, Mastropietro A, et al. T2w-MRI signal normalization affects radiomics features reproducibility. Med Phys. 2020 Jan 23. doi: 10.1002/mp.14038.

Schad LR, Blüml S, Zuna I. MR tissue characterization of intracranial tumors by means of texture analysis. Magn Reson Imaging. 1993;11(6):889-96.

Schnapauff D, Zeile M, Niederhagen MB, 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. 2009;29(6):1355-9.

Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44(3):1050-1062.

Shafiq-Ul-Hassan M, Latifi K, Zhang G, et al. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep. 2018;8(1):10545.

Shinohara RT, Sweeney EM, Goldsmith J, et al; Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing; Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin. 2014;6:9-19.

Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. Survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27:3206–3208.

Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software. 2010 ;33(1)

Soldatos T, Ahlawat S, Montgomery E, et al. Multiparametric Assessment of Treatment Response in High-Grade Soft-Tissue Sarcomas with Anatomic and Functional MR Imaging Sequences. Radiology. 2016;278(3):831-40.

Sourbron SP, Buckley DL. Classic models for dynamic contrast-enhanced MRI. NMR Biomed. 2013;26(8):1004-27.

Spraker MB, Wootton LS, Hippe DS, et al. MRI Radiomic Features Are Independently Associated With Overall Survival in Soft Tissue Sarcoma. Adv Radiat Oncol. 2019;4(2):413-421.

Sreekantaiah C, Karakousis CP, Leong SP, Sandberg AA. Cytogenetic findings in liposarcoma correlate with histopathologic subtypes. Cancer 1992;69:2484–2495

Stacchiotti S, Collini P, Messina A, et al. High-grade soft-tissue sarcomas: tumor response assessment--pilot study to assess the correlation between radiologic and pathologic response by using RECIST and Choi criteria. Radiology. 2009;251(2):447-56.

Stacchiotti S, Verderio P, Messina A, et al. Tumor response assessment by modified Choi criteria in localized high-risk soft tissue sarcoma treated with chemotherapy. Cancer. 2012;118:5857–5866.

Stevenson JD, Watson JJ, Cool P, et al. Whole-body magnetic resonance imaging in myxoid liposarcoma: A useful adjunct for the detection of extra-pulmonary metastatic disease. Eur J Surg Oncol. 2016;42(4):574-80.

Stevenson MG, Been LB, Hoekstra HJ, et al. Volume of interest delineation techniques for 18F-FDG PET-CT scans during neoadjuvant extremity soft tissue sarcoma treatment in adults: a feasibility study. EJNMMI Res. 2018;8(1):42.

Subhawong TK, Wang X, Durand DJ, et al. Proton MR spectroscopy in metabolic assessment of musculoskeletal lesions. AJR Am J Roentgenol. 2012;198(1):162-72.

Tagliafico AS, Bignotti B, Rossi F, Valdora F, Martinoli C. Local recurrence of soft tissue sarcoma: a radiomic analysis. Radiol Oncol. 2019;53(3):300-306.

Taïeb S, Penel N, Vanseymortier L, Ceugnart L. Soft tissue sarcomas or intramuscular haematomas? Eur J Radiol. 2009;72(1):44-9.

Taieb S, Saada-Bouzid E, Tresch E, et al; French Sarcoma Group. Comparison of response evaluation criteria in solid tumours and Choi criteria for response evaluation in patients with advanced soft tissue sarcoma treated with trabectedin: a retrospective analysis. Eur J Cancer. 2015;51(2):202-9

Tateishi U, Yamaguchi U, Seki K, Terauchi T, Arai Y, Hasegawa T. Glut-1 expression and enhanced glucose metabolism are associated with tumour grade in bone and soft tissue sarcomas: a prospective evaluation by [18F]fluorodeoxyglucose positron emission tomography. Eur J Nucl Med Mol Imaging. 2006;33(6):683-91.

Tateishi U, Hasegawa T, Beppu Y, et al. Prognostic significance of MRI findings in patients with myxoid-round cell liposarcoma. AJR Am J Roentgenol 2004;182:725–731.

Tibshirani, R. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 1996;58(1): 267-288.

Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000;92(3):205-16.

Thibault G, J. Angulo and F. Meyer. Advanced Statistical Matrices for Texture Characterization: Application to Cell Classification. IEEE Transactions on Biomedical Engineering, 2014 ; 61(3) :630-637.

Thomassin-Naggara I, Soualhi N, Balvay D, Darai E, Cuenod C-A. Quantifying tumor vascular heterogeneity with DCE-MRI in complex adnexal masses: A preliminary study. J Magn Reson Imaging 2017; 46:1776–1785.

Thornhill RE, Golfam M, Sheikh A, et al. Differentiation of lipoma from liposarcoma on MRI using texture and shape analysis. Acad Radiol. 2014;21(9):1185-94.

Tian F, Hayano K, Kambadakone AR, Sahani DV. Response assessment to neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion. Abdom Imaging. 2015;40(6):1705-12.

Tirkes T, Hollar MA, Tann M, Kohli MD, Akisik F, Sandrasegaran K. Response criteria in oncologic imaging: review of traditional and new criteria. Radiographics. 2013;33(5):1323-41.

Tofts PS, Brix G, Buckley DL, 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 1999; 10:223–232

Toulmonde M, Penel N, Adam J, et al. Use of PD-1 Targeting, Macrophage Infiltration, and IDO Pathway Activation in Sarcomas: A Phase 2 Clinical Trial. JAMA Oncol. 2018;4(1):93-97.

Trojani M, Contesso G, Coindre JM, et al. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. Int J Cancer. 1984;33(1):37-42.

Turc-Carel C, Limon J, Dal Cin P, et al. Cytogenetic studies of adipose tissue tumors. II. Recurrent reciprocal translocation t(12;16)(q13;p11) in myxoid liposarcomas. Cancer Genet Cytogenet 1986;23:291–299

Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010;29(6):1310-20.

Uno H, Cai TX, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc. 2007;102(478):527–37.

Ursprung S, Beer L, Bruining A, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. Eur Radiol. 2020. doi: 10.1007/s00330-020-06666-3.

Vallières M, Freeman CR, Skamene SR, 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. 2015;60(14):5471-96.

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. 2017;62(22):8536-8565.

Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med. 2018;59(2):189-193.

Van der Graaf WT, Blay JY, Chawla SP, et al; EORTC Soft Tissue and Bone Sarcoma Group; PALETTE study group. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. Lancet. 2012;379(9829):1879-86.

Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017;77(21):e104-e107.

Van Rijswijk CS, Geirnaerdt MJ, Hogendoorn PC, et al. Dynamic contrast-enhanced MR imaging in monitoring response to isolated limb perfusion in high-grade soft tissue sarcoma: initial results. Eur Radiol. 2003;13(8):1849-58.

Van Vliet M, Kliffen M, Krestin GP, van Dijke CF. Soft tissue sarcomas at a glance: clinical, histological, and MR imaging features of malignant extremity soft tissue tumors. Eur Radiol. 2009;19(6):1499-511.

Vargas HA, Veeraraghavan H, Micco M, et al. A novel representation of inter-site tumour heterogeneity from pre-treatment computed tomography textures classifies ovarian cancers by clinical outcome. Eur Radiol. 2017;27(9):3991-4001.

Viglianti BL, Lora-Michiels M, Poulson JM, et al. Dynamic contrast-enhanced magnetic resonance imaging as a predictor of clinical outcome in canine spontaneous soft tissue sarcomas treated with thermoradiotherapy. Clin Cancer Res. 2009;15(15):4993-5001.

Vos M, Starmans MPA, Timbergen MJM, et al. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. Br J Surg. 2019;106(13):1800-1809.

Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50 Suppl 1:122S-50S.

Wang L, Lai HM, Barker GJ, Miller DH, Tofts PS. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. Magn Reson Med 1998;39(2):322–327.

Wang H, Chen H, Duan S, Hao D, Liu J. Radiomics and Machine Learning With Multiparametric Preoperative MRI May Accurately Predict the Histopathological Grades of Soft Tissue Sarcomas. J Magn Reson Imaging. 2020;51(3):791-797.

Warburg O. On the origin of cancer cells. Science. 1956;123(3191):309-14.

Wardelmann E, Haas RL, Bovée JV, et al. Evaluation of response after neoadjuvant treatment in soft tissue sarcomas; the European Organization for Research and Treatment of Cancer-Soft Tissue and Bone Sarcoma Group (EORTC-STBSG) recommendations for pathological examination and reporting. Eur J Cancer. 2016;53:84-95.

Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, Huang SH, Purdie TG, O'Sullivan B, Aerts HJWL, Jaffray DA. Vulnerabilities of radiomic signature development: The need for safeguards. Radiother Oncol. 2019 Jan;130:2-9.

White LM, Wunder JS, Bell RS, et al. Histologic assessment of peritumoral edema in soft tissue sarcoma. Int J Radiat Oncol Biol Phys. 2005;61:1439–1445.

Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

Wolsztynski E, O'Sullivan F, Keyes E, et al. Positron emission tomography-based assessment of metabolic gradient and other prognostic features in sarcoma. J Med Imaging (Bellingham). 2018;5(2):024502

Wu G, Liu X, Xiong Y, Ran J, Li X. Intravoxel incoherent motion and diffusion kurtosis imaging for discriminating soft tissue sarcoma from vascular anomalies. Medicine (Baltimore). 2018;97(50):e13641.

Xia W, Yan Z, Gao X. Volume fractions of DCE-MRI parameter as early predictor of histologic response in soft tissue sarcoma: A feasibility study. Eur J Radiol. 2017;95:228-235.

Xiang P, Zhang X, Liu D, et al. Distinguishing soft tissue sarcomas of different histologic grades based on quantitative MR assessment of intratumoral heterogeneity. Eur J Radiol. 2019;118:194-199.

Yang F, Simpson G, Young L, Ford J, Dogan N, Wang L. Impact of contouring variability on oncological PET radiomics features in the lung. Sci Rep. 2020;10(1):369.

Yoo HJ, Hong SH, Kang Y, et al. MR imaging of myxofibrosarcoma and undifferentiated sarcoma with emphasis on tail sign; diagnostic and prognostic value. Eur Radiol. 2014;24(8):1749-57.

Yoon SH, Park CM, Park SJ, et al. Tumor Heterogeneity in Lung Cancer: Assessment with Dynamic Contrast-enhanced MR Imaging. Radiology 2016;280(3):940–948.

Yoon MA, Chee CG, Shin MJ, et al. Added value of diffusion-weighted imaging to conventional MRI for predicting fascial involvement of soft tissue sarcomas. Eur Radiol 29:1863-1873.

Zhang L, Fried DV, Fave XJ, et al. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42(3):1341-53.

Zhang Y, Zhu Y, Shi X, et al. Soft Tissue Sarcomas: Preoperative Predictive Histopathological Grading Based on Radiomics of MRI. Acad Radiol. 2019;26(9):1262-1268.

Zhang X, Zhong L, Zhang B, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: evaluation with two disease groups. Cancer Imaging. 2019 ;19(1):89.

Zhao F, Ahlawat S, Farahani SJ, et al. Can MR imaging be used to predict tumor grade in soft-tissue sarcoma? Radiology. 2014;272(1):192-201.

Zwanenburg A, Stefan Leger, Martin Vallières, Steffen Löck. Image biomarker standardisation initiative. 2019 arXiv:1612.07003

7. ANNEXES

Annexe 1 : Récapitulatif des tables et figures

TABLES

PART 1. INTRODUCTION

PART 2. PREDICTION PRONOSTIQUE

PART 3. PREDICTION DE LA REPONSE

PART 4. AMELIORER LES PREDICTIONS EN OPTIMISANT LES PROCESSUS DE POST-TRAITEMENT

FIGURES

PART 1. INTRODUCTION

PART 2. PREDICTION PRONOSTIQUE

PART 3. PREDICTION DE LA REPONSE

PART 4. AMELIORER LES PREDICTIONS EN OPTIMISANT LES PROCESSUS DE POST-TRAITEMENT

Annexe 2 : Indices radiomics utilisés

Etudes utilisant LIFEx

First-order texture features	Shape features	Gray-level co-occurrence matrix features (GLCM)
HISTO_min	SHAPE_Volume	GLCM_Homogeneity
HISTO_mean	SHAPE_Sphericity	GLCM_Energy
HISTO_std	SHAPE_Compacity	GLCM_Contrast
HISTO_max		GLCM_Correlation
HISTO_Skewness		GLCM_Entropy_log10
HISTO_Kurtosis		GLCM_Entropy_log2
HISTO_ExcessKurtosis		GLCM_Dissimilarity
HISTO_Entropy_log10		
HISTO_Entropy_log2		
HISTO_Energy		

Gray-level run length matrix features (GLRLM)	Neighborhood gray-level different matrix features (NGLDM)	Gray-level zone length matrix features (GLZLM)
GLRLM_SRE	NGLDM_Coarseness	GLZLM_SZE
GLRLM_LRE	NGLDM_Contrast	GLZLM_LZE
GLRLM_LGRE	NGLDM_Busyness	GLZLM_LGZE
GLRLM_HGRE		GLZLM_HGZE
GLRLM_SRLGE		GLZLM_SZLGE
GLRLM_SRHGE		GLZLM_SZHGE
GLRLM_LRLGE		GLZLM_LZLGE
GLRLM_LRHGE		GLZLM_LZHGE
GLRLM_GLNU		GLZLM_GLNU
GLRLM_RLNU		GLZLM_ZLNU
GLRLM_RP		GLZLM_ZP

Etudes utilisant OLEA

First-order feature	Gray level co-occurrence matrix	Gray level run length matrix	Gray level size zone matrix
Energy	Autocorrelation	GL non-uniformity	GL non-uniformity
Entropy	Cluster prominence	GL non-uniformity normalized	GL non-uniformity normalized
Inter-quartile range	Cluster shade	GL variance	GL variance
Kurtosis	Cluster tendency	High GL run emphasis	High GL zone emphasis
Maximum	Contrast	Long run emphasis	Large area emphasis
Mean	Correlation	Long run high GL emphasis	Large area low GL emphasis
Mean absolute deviation	Difference average	Long run low GL emphasis	Large area high GL emphasis
Median	Difference entropy	Low GL run emphasis	Low GL zone emphasis
Minimum	Difference variance	Run entropy	Small area emphasis
Range	Informal Measure of Correlation 1	Run length non-uniformity	Small area low GL emphasis
Robust mean absolute deviation	Informal Measure of Correlation 2	Run length non-uniformity normalized	Small area high GL emphasis
Root mean squared	Inverse difference	Run percentage	Size zone non-uniformity
Skewness	Inverse difference normalized	Run variance	Size zone non-uniformity normalized
Standard deviation	Inverse difference moment	Short run emphasis	Zone entropy
Total Energy	Inverse difference moment normalized	Short run high GL emphasis	Zone percentage
Variance	Inverse variance	Short run low GL emphasis	Zone variance
Uniformity	Joint average		
10^{th} percentile	Joint energy		
90^{th} percentile	Joint entropy		
	Maximum probability		
	Sum average		
	Sum entropy		
	Sum of squares		

Neighboring gray tone difference matrix	Gray level dependence matrix	Shape features
Coarseness	Dependence entropy	Compactness 1
Contrast	Dependence non-uniformity	Compactness 2
Busyness	Dependence non uniformity normalized	Flatness
Complexity	Dependence variance	Least axis
Strength	GL non-uniformity	Major axis
	GL variance	Maximum 2D column
	High GL emphasis	Maximum 2D diameter row
	Large dependence emphasis	Maximum 2D diameter slice
	Large dependence high GL emphasis	Maximum 3D diameter
	Large dependence low GL emphasis	Minor axis
	Low GL emphasis	Spherical disproportion
	Small dependence emphasis	Sphericity
	Small dependence high GL emphasis	Surface area
	Small dependence low GL emphasis	Surface area to volume ratio
		Volume

Etudes utilisant un algorithme in-house

First order texture features	Second Order texture features (d = 2 & 5, θ = {0°, 45°, 90°, 135°})	Shape features
Interval	Inertia	Elongation
Kurtosis	Energy	Flatness
Skewness	entropy	Sphericity
Entropy	Homogeneity	Equivalent spherical radius
	Cluster shade	Feret diameter
	Cluster prominence	

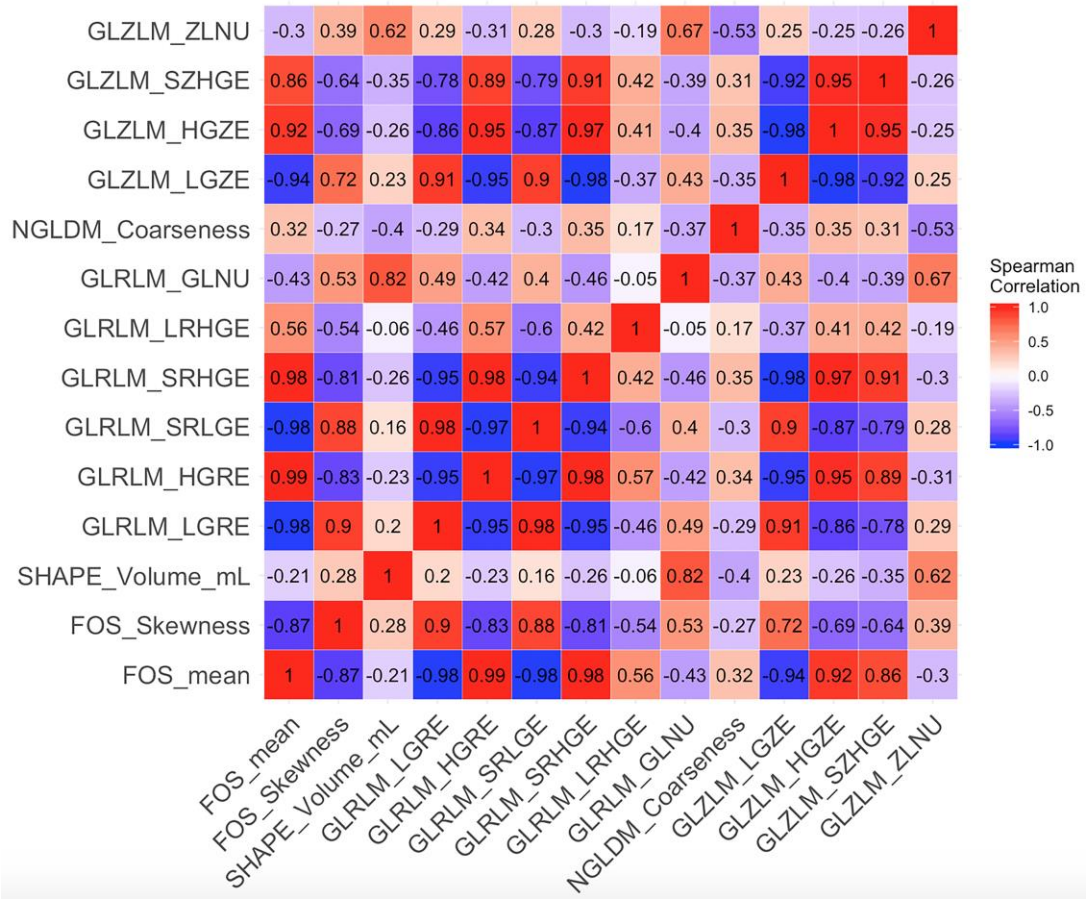
Annexe 3 : Supplementary Data - Article 1

Supplementary Data 1. Inter-observer agreements for the semantic radiological variables.

Variables	P-value	Inter-observer agreement
Longest diameter[1]	<0.0001***	0.967
Heterogeneous SI on T2-WI[2]	<0.0001***	0.756
Necrotic SI on T2-WI[2]	<0.0001***	0.925
Fatty SI on T2-WI[2]	<0.0001***	0.909
Tail sign[2]	<0.0001***	0.866
Peritumoral edema[2]	<0.0001***	0.913
Peritumoral enhancement[3]	<0.0001***	0.904

NOTE. Abbreviations: SI: signal intensity, WI: weighted imaging, CE: contrast-enhanced. *: p< 0.05, **: p<0.001, ***: p<0.0001.The inter-observer agreement was assessed using intraclass correlation coefficient for numeric variables (1), weighted kappa for ordinal variables (2) and Cohen kappa for categorical variables (3)

Supplementary Data 2. Correlation plot of the radiomics features associated with metastatic relapse-free survival in univariate analysis.



Annexe 4 : Supplementary Data - Article 2

Supplementary Data 1. Inter- and intra-observer agreements for semantic radiological features

The inter- and intra-observers reliabilities were assessed using interclass correlation coefficient (ICC) for continuous values (with two-way mixed for inter-rater assessment). The Cohen's Kappa (κ) for pairwise agreement was used for dichotomous variables. Weighted kappa statistics (κ_W) was used for ordinal variables. The agreement for ICC was defined as good (>0.75), moderate (0.5-0.75) or poor (<0.5). The agreement for classical κ and κ_W was defined as slight (0-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect (0.81-0.99).

	Intra-rater agreement (R1)	**Inter-rater agreements**		
		R1 - R2	**R1 - R3**	**R2 - R3**
Relative Change_LD[1]	-	0.75 (0.60-0.85)	0.80 (0.67-0.88)	0.76 (0.61-0.86)
RECIST 1.1 response status[2]	-	0.44 (0.12-0.76)	0.48 (0.17-0.79)	0.46 (0.21-0.71)
Δ_Edema[3]	0.77 (0.61-0.93)	0.64 (0.44-0.83)	0.71 (0.61-0.81)	0.67 (0.56-0.78)
Δ_Peritumoral enhancement[3]	0.63 (0.43-0.82)	0.43 (0.20-0.71)	0.75 (0.66-0.84)	0.63 (0.52-0.74
Δ_Margin_definition	0.43 (0.18-0.68)	0.46 (0.21-0.71)	0.56 (0.43-0.69)	0.53 (0.39-0.67)
Fibro-necrotic changes[3]	0.75 (0.57-0.94)	0.66 (0.45-0.88)	0.68 (0.56-0.80)	0.86 (0.78-0.94)

NOTE – LD: longest diameter.
Results are given with 95% confidence interval.
Statistical tests: [1]: interclass correlation coefficient with two-way mixed model; [2]: weighted kappa for ordinal values; [3]: Cohen's kappa.
R1: AC, R2: MK, R3: XB, with 3, 27 and 5 years of experience in MRI respectively

Supplementary Data 2. Comparison of epidemiological data of the training and test cohorts :

		Training set (n=50)	Test set (n=15)	p-value
Gender	Male	30 (60)	8 (53.3)	0.767
	Female	20 (40)	7 (46.7)	
Age at diagnosis (y)		56.8 ± 12.9	61.4 ±12.1	0.224
Histotype	Undifferentiated sarcoma [1]	25 (50)	8 (53.3)	0.904
	Muscular sarcoma [2]	10 (20)	3 (20)	
	M/RC Liposarcoma [3]	3 (6)	2 (13.3)	
	Other liposarcoma [4]	5 (10)	1 (6.7)	
	Synovial sarcoma	6 (12%)	1 (6.7)	
	MPNST	1 (1.6)	0 (0)	
Location	trunk wall	7 (14)	5 (33.3)	0.282
	Pelvic Girdle	2 (4)	0 (0)	
	Shoulder Girdle	6 (12)	0 (0)	
	Upper limb	6 (12)	1 (6.7)	
	Lower limb	29 (58)	9 (60)	
Depth	Superficial	3 (6)	1 (6.7)	1.000
	Deep	47 (94)	14 (93.3)	
LD on MRI_0 (mm)		116 ± 53	128 ± 53	0.466
Nb cycles	4 cycles	16 (32)	6 (40)	0.757
	5 or 6 cycles	34 (68)	9 (40)	

NOTE.- LD: longest diameter, sd: standard deviation.
Data are numbers of patients with percentages in parentheses, except for age, LD and change in LD. [1]: myxofibrosarcoma or undifferentiated sarcoma; [2]: leiomyosarcoma and rhabdomyosarcoma ; [3]: myxoid/round cells liposarcoma ; [4]: pleomorphic or dedifferentiated liposarcoma. MPNST: malignant peripheral nerve sheath tumor.

Supplementary Data 3. Comparisons of the diagnostic performances of random forest classifiers based on features that were selected with other popular features selection methods (LASSO and ElasticNet)

Radiomics methods	Training Set (n=50) (cross-validation)		Test Set (n=15)	
	Accuracy	AUROC	Accuracy	AUROC
RF + WLCX (3 features)	**0.870**	**0.860**	**0.750**	**0.620**
RF + WLCX (16 features)	0.860	0.790	0.400	0.360
RF + LASSO (16 features)	0.858	0.789	0.510	0.435
RF + ElasticNet (17 features)	0.860	0.800	0.496	0.405

NOTE.- AUROC: area under the receiver operating characteristic curve, LASSO: Least Absolute Shrinkage and Selection Operator, RF: random forest, WLCX: Wilcoxon-test based selection method. The 'RF + WLCX (3 features)' method corresponds to the method that is detailed in the manuscript. A total of 16 and 17 features were selected by LASSO and ElasticNet, respectively (including these same 3 features i.e. Δ_Histogram_Entropy, Δ_Elongation, Δ_Surrounding_Edema). The 'RF +WLCX (16 features)' method was built with the 16 features that were commonly selected by LASSO and ElasticNET, for comparison.

It can be noted that the highest accuracy and AUROC were obtained with RF + WLCX (3 features) for both training and test sets

Annexe 5 : Supplementary Data Article 3

Supplementary Data 1. Textural radiomics features (RFs) used in the study. RFs were extracted on 3D tumor volume with LIFEx freeware (Inserm, Orsay, France, www.lifexsoft.org). Details regarding the formula can be found at: https://www.lifexsoft.org/index.php/resources/19-texture/radiomic-features?filter_tag[0]=

First-order texture features		Grey-level co-occurrence matrix features (GLCM)
HISTO_min	HISTO_Quartile_1	GLCM_Homogeneity
HISTO_mean	HISTO_Quartile_2	GLCM_Energy
HISTO_std	HISTO_Quartile_3	GLCM_Contrast
HISTO_max		GLCM_Correlation
HISTO_Skewness		GLCM_Entropy_log10
HISTO_Kurtosis		GLCM_Entropy_log2
HISTO_ExcessKurtosis		GLCM_Dissimilarity
HISTO_Entropy_log10		
HISTO_Entropy_log2		
HISTO_Energy		

Grey-level run length matrix features(GLRLM)	Neighborhood grey-level different matrix features (NGLDM)	Grey-level zone length matrix features (GLZLM)
GLRLM_SRE	NGLDM_Coarseness	GLZLM_SZE
GLRLM_LRE	NGLDM_Contrast	GLZLM_LZE
GLRLM_LGRE	NGLDM_Busyness	GLZLM_LGZE
GLRLM_HGRE		GLZLM_HGZE
GLRLM_SRLGE		GLZLM_SZLGE
GLRLM_SRHGE		GLZLM_SZHGE
GLRLM_LRLGE		GLZLM_LZLGE
GLRLM_LRHGE		GLZLM_LZHGE
GLRLM_GLNU		GLZLM_GLNU
GLRLM_RLNU		GLZLM_ZLNU
GLRLM_RP		GLZLM_ZP

Supplementary Data 2. Definitions of ComBat Harmonization method.

Details can be found at: https://github.com/Jfortin1/ComBatHarmonization.
The aim of ComBat in a radiological/radiomics setting is to compensate variations in imaging datasets of a same imaging modality due to variations in imaging protocol while preserving biological variability, and notably when there are only a few patients per site. ComBat is classically applied at the end of the postprocessing pipeline, herein, after the extraction of radiomics features (RFs) obtained with the intensity harmonization technique that was hypothesized to be the more relevant among the 4 proposed techniques (i.e. Histogram matching with average normalized histogram of the population. We provided to the ComBat function: the MR-system as main variable and, as covariables, the coils and the tumor location.
This data-driven method identifies the protocol effect assuming that the value of each feature, y, measured in volume of interest, j, with imaging protocol, i, can be written as: $y_{ij} = \alpha + \gamma_i + \delta_i \times \varepsilon_{ij}$ (in which α is the average value for features y_{ij}; γ_i is an additive protocol effect and δ_i is a multiplicative protocol effect affected by an error term ε_{ij}). The compensations consists in estimating the model parameters α, γ_i and δ_i, and by using a maximum likelihood approach on the basis of the set of available observations: $y_{ij}^{ComBat} = \hat{\alpha} + \frac{y_{ij} - \hat{\alpha} - \hat{\gamma}_i}{\hat{\delta}_i}$, in which $\hat{\alpha}$, $\hat{\gamma}_i$ and $\hat{\delta}_i$ are estimators of α, γ_i and δ_i. Parametric and non-parametric forms of ComBat have been developed. The non-parametric form does not assume law followed by the parameters and has been used in the present study.

Supplementary Data 3. Basic concepts regarding the classifiers used in the study

We used 5 popular supervised machine-learning approaches in the "caret" R package:

1. *Binomial logistic regression:*
   - This classifier predicts the probability p of occurrence of a binary event (herein metastatic relapse 2 years after the end of treatment) by using a logit function, as follows: $p = \frac{1}{1 + \exp(-(\beta_O + \sum_{i=1}^{n} \beta_i \times X_i))}$
   Where X_i are explanatory variables, with a total of n variables.
   - No grid search was needed.

2. *Penalized binomial logistic regression with optimal combination of ridge and lasso regressions (or elastic net regression)*
   - The penalized version of a binomial logistic regression consists of reducing the number and the importance of variables in order to optimize the performances of the classification model. The coefficients of the less contributive variables are shrunken towards 0 (: ridge regression) or even set to 0 (: lasso, for least absolute shrinkage and selection operator). Herein, we used combination of both lasso and ridge, also named elastic net, with some

coefficients of predictors are shrunken towards 0 and others set to 0. The amount of ridge and lasso regression was investigated with a grid search.
- A grid search was performed to estimate the best value for α (: mixing percentage) and λ (: regularization parameter).

3. *Random forests:*
   - Random forests belong to decision trees. Each node is a test of a learning feature. Those tests result in clustering training set into subgroups (: branches). Thresholds can be fixed to end the partitioning (minimum subgroup size, maximum depth of the tree…). In random forests algorithm, multiple decision trees are developed with a subset of patients and a subset of randomly-chosen features for each one. The average prediction of the trees is then used to improve the prediction accuracy.
   - A grid search was performed to estimate the best value for mtry (i.e. the numbers of randomly selected predictors).

4. *Radial support vector machine:*
   - Support vector machines attempt to partition a space of features into 2 groups by finding an optimal way of separating these groups either by lines or (hyper)planes. When the problem cannot be solved with linear means, non-linear kernel functions can be used in order to transform the input space to a higher dimensional space where observations can be more easily linearly separated. Herein, we used a "radial" kernel.
   - A grid search was performed to estimate the best value for cost (: or penalty parameter, which informs on how much error is acceptable) and σ (: which determines how far the influence of a single training example reaches - and the risk of overfitting).

The following variables selection or filtering methods were used:
- "none" (i.e. all the variables were used in the models) – for all classifiers;
- principal component analysis (in which we kept the components that explained 95% of the variance) – for logistic regression, random forests and radial support vector machine;
- selection by filters using Wilcoxon test (in which we kept the variables that were associated with the outcome at univariable level with a p-value of less than 0.1) – for logistic regression, random forests and radial support vector machine;
- recursive feature elimination – for random forests and logistic regression

Supplementary Data 4. Detailed results of the repeated-measures ANOVAs

Radiomics Features	F-value	P-value	P-values of the post-Hoc comparisons									
			HM-All-patients vs. Fat-standardization	HM-All-patients+ComBat vs. Fat-standardization	Hm-1-patient vs. Fat-standardization	Basic-normalization vs. Fat-standardization	HM-All-patients+ ComBat vs. HM-All-patients	HM-1-patient vs. HM-All-patients	Basic-normalization vs. HM-All-patient	HM-1-patient vs. HM-All-patients+ComBat	Basic-normalization vs. HM-All-patients+ComBat	Basic-normalization vs. HM-1-patient
HISTO_min	1438.7	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0076	1.	<0.0001	0.0134	<0.0001	<0.0001
HISTO_mean	33.4	<0.0001	<0.0001	<0.0001	<0.0001	1.	1.	1.	<0.0001	0.1783	<0.0001	<0.0001
HISTO_std	5.7	0.0002	0.0054	0.0006	0.0015	0.5654	1.	1.	1.	1.	0.3417	0.5998
HISTO_max	200.4	<0.0001	<0.0001	0.9581	<0.0001	<0.0001	0.0004	1.	<0.0001	0.0001	<0.0001	<0.0001
HISTO_Q1	32.7	<0.0001	<0.0001	<0.0001	<0.0001	1.	1.	1.	<0.0001	1.	<0.0001	<0.0001
HISTO_Q2	23	<0.0001	<0.0001	<0.0001	<0.0001	1.	1.	0.771	<0.0001	1.	<0.0001	<0.0001
HISTO_Q3	5.2	0.0005	1.	1.	0.0002	0.8045	1.	0.0434	1.	0.0052	1.	0.1114
HISTO_Skewness	137.7	<0.0001	<0.0001	<0.0001	<0.0001	1.	0.4191	1.	<0.0001	1.	<0.0001	<0.0001
HISTO_Kurtosis	75.9	<0.0001	<0.0001	<0.0001	<0.0001	1.	0.1824	1.	<0.0001	0.1817	<0.0001	<0.0001
HISTO_ExcessKurtosis	75.9	<0.0001	<0.0001	<0.0001	<0.0001	1.	0.1824	1.	<0.0001	0.1817	<0.0001	<0.0001
HISTO_Entropy_log10	5.4	0.0004	0.0003	0.0148	0.0023	0.1232	1.	1.	0.9379	1.	1.	1.
HISTO_Entropy_log2	5.4	0.0004	0.0003	0.0148	0.0023	0.1232	1.	1.	0.9379	1.	1.	1.
HISTO_Energy	4.2	0.0024	0.0261	0.7609	0.0048	0.0103	1.	1.	1.	0.858	1.	1.
GLCM_Homogeneity	28	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.1933	1.	1.	0.3847	0.2743	1.
GLCM_Energy	13	<0.0001	0.0939	1.	0.0056	<0.0001	0.0009	1.	0.2115	<0.0001	<0.0001	1.
GLCM_Contrast	8.9	<0.0001	<0.0001	1.	0.0385	1.	0.0009	0.2062	<0.0001	1.	1.	0.1805
GLCM_Correlation	15.1	<0.0001	<0.0001	0.1642	1.	1.	0.0009	<0.0001	<0.0001	1.	0.0194	0.2866
GLCM_Entropy_log10	5.3	0.0004	0.0005	0.0528	0.0016	0.0097	1.	1.	1.	1.	1.	1.
GLCM_Entropy_log2	5.3	0.0004	0.0005	0.0528	0.0016	0.0097	1.	1.	1.	1.	1.	1.
GLCM_Dissimilarity	19.2	<0.0001	<0.0001	<0.0001	<0.0001	1.	0.2468	0.0548	0	1.	0.0001	0.0012
GLRLM_SRE	33.4	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.1068	1.	0.0534	0.0032	<0.0001	0.8251
GLRLM_LRE	24.5	<0.0001	0.0432	<0.0001	1.	0.1685	0.0009	0.0648	<0.0001	<0.0001	<0.0001	0.1168

GLRLM_LGRE	52.4	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.2845	0.9464	0.003	1.	1.	0.5144
GLRLM_HGRE	37	<0.0001	<0.0001	<0.0001	<0.0001	1.	1	0.1927	<0.0001	0.2338	0	<0.0001
GLRLM_SRLGE	43.4	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.2182	0.0158	0.0534	1.	1.	1.
GLRLM_SRHGE	27.5	<0.0001	<0.0001	<0.0001	<0.0001	1.	1	0.1734	<0.0001	0.226	<0.0001	<0.0001
GLRLM_LRLGE	25	<0.0001	<0.0001	0.0001	<0.0001	<0.0001	0.3704	0.3	0.7324	0.0002	0.0011	1.
GLRLM_LRHGE	14.8	<0.0001	1.	<0.0001	1.	0.0001	<0.0001	1.	0.0005	<0.0001	0.9691	0.0123
GLRLM_GLNU	3.6	0.0068	1.	0.8925	0.5531	1.	0.3798	1.	1.	0.003	0.0635	1.
GLRLM_RLNU	13	<0.0001	0.6813	0.0017	0.2572	0.4403	<0.0001	1.	0.0012	<0.0001	0.7968	0.0002
GLRLM_RP	18.2	<0.0001	0.0035	0.5182	0.0001	<0.0001	1.	1.	0.0001	0.1413	<0.0001	0.0024
NGLDM_Coarseness	10.4	<0.0001	<0.0001	0.0012	<0.0001	1.	1.	1.	0.0011	1.	0.0174	0.0003
NGLDM_Contrast	6.6	<0.0001	0.1143	0.1844	1.	1.	<0.0001	0.5545	0.0063	0.0297	1.	1.
NGLDM_Busyness	9.4	<0.0001	1.	1.	0.0333	<0.0001	1.	0.3063	0.0004	0.018	<0.0001	0.5405
GLZLM_SZE	11.3	<0.0001	0.3686	0.0001	0.4083	0.6214	0.2052	1.	0.0008	0.1833	<0.0001	0.0009
GLZLM_LZE	79.1	<0.0001	<0.0001	<0.0001	1.	0.0185	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0016
GLZLM_LGZE	121	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.8874	0.0165	<0.0001	1.	0.0004	0.0755
GLZLM_HGZE	34.4	<0.0001	<0.0001	<0.0001	<0.0001	1.	1.	1.	<0.0001	1.	<0.0001	<0.0001
GLZLM_SZLGE	101.5	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.9753	0.0016	<0.0001	0.3413	<0.0001	0.1129
GLZLM_SZHGE	11.6	<0.0001	<0.0001	0.0009	<0.0001	1.	1.	1.	0.0001	1.	0.0018	<0.0001
GLZLM_LZLGE	182.1	<0.0001	<0.0001	<0.0001	0.9747	0.0059	1.	<0.0001	<0.0001	<0.0001	<0.0001	0.7531
GLZLM_LZHGE	40.7	<0.0001	1.	<0.0001	1.	0.0649	<0.0001	1.	0.0048	<0.0001	<0.0001	0.1017
GLZLM_GLNU	10.6	<0.0001	1.	<0.0001	1.	0.0268	<0.0001	1.	0.6516	0.0001	0.0313	1.
GLZLM_ZLNU	112.4	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	1.	0.0438	1.	0.2178	0.4308	0.0002
GLZLM_ZP	87.3	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	1.	1.	<0.0001	<0.0001	1.

NOTE. Post-hoc tests corresponded to Tukey tests with Bonferroni corrections for multiple comparisons.
The intensity harmonization techniques were: *Fat-standardization*; *Basic-normalization*; histogram matching with a randomly-chosen normalized histogram of a patient (*HM-1-patient*); histogram matching with the average normalized histogram of the 70 included patients (*HM-All-patients*) HM-All-patients combined with ComBat harmonization method *(HM-All-patients+ComBat*).

Supplementary Data 5. Comparisons of hierarchical clustering results based on radiomics features obtained with different intensity harmonization techniques.

	Fat-standardization	Basic-normalization	HM-1-patient	HM-All-patients	HM-All-patients +ComBat
Fat-standardization	-	0.14	0.15	0.17	0.18
Basic-normalization	0.33 (0.5-0.17), p<0.001***	-	0.11	0.30	0.42
HM-1-patient	0.23 (0.48-0.02), p=0.052	0.25 (0.43-0.06), p=0.01*	-	0.26	0.29
HM-All-patients	0.36 (0.25-0.47), p=0.003**	0.51 (0.31-0.71), p<0.001***	0.4 (0.29-0.51), p<0.001***	-	0.55
HM-All-patients+ComBat	0.43 (0.29-0.57), p<0.001***	0.67 (0.48-0.86), p<0.001***	0.44 (0.28-0.6), p<0.001***	0.75 (0.91-0.59), p<0.001***	-

NOTE. Results in the upper diagonal correspond to the Baker's gamma coefficient, which quantifies the correlation between 2 dendrograms, ranging from 0 (completely different) to 1 (exactly the same dendrograms). Results in the lower diagonal correspond to the Cohen's kappa index with 95% confidence interval and p-value, in order to quantify the inter-clustering results agreements.

Supplementary Data 6. Detailed results of the different supervised machine learning classifiers developed on 5 independent subsets of radiomics features depending on the intensity harmonization technique.

Intensity harmonization technique	Classifier	Features selection / reduction method	Best hyperparameter tuning	Training cohort (results in repeated cross-validation)					Testing cohort				
				Accuracy	AUC	Se	Sp	Kappa	Accuracy	AUC	Se	Sp	Kappa
Fat-standardization	Radial SVM	SBF::Wilcoxon	sigma=10, C=1	0.619	0.594	0.963	0.142	0.110	0.55	0.538	0.917	0.	-0.098
		PCA	sigma=0.1, C=100	0.606	0.661	0.840	0.285	0.127	0.65	0.948	1.	0.125	0.146
		None	sigma=0.1, C=10	0.591	0.658	0.895	0.172	0.069	0.700	0.719	1.	0.25	0.286
	Glmnet Logistic regression		alpha=0.21, lambda=0.05	0.601	0.676	0.695	0.465	0.161	0.750	0.802	0.917	0.500	0.444
		RFE	size=2 features	0.582	0.592	0.817	0.260	0.078	0.600	0.406	0.917	0.125	0.048
		SBF::Wilcoxon	-	0.602	0.580	0.627	0.563	0.190	0.700	0.740	0.667	0.750	0.400
		PCA	-	0.560	0.607	0.642	0.447	0.093	0.700	0.760	0.750	0.625	0.375
		None	-	0.509	0.516	0.535	0.468	0.005	0.300	0.708	0.333	0.250	-0.400
	Random forests	RFE	mtry=2, size=53 features	0.492	0.530	0.600	0.345	-0.053	0.750	0.677	0.917	0.500	0.444
		SBF::Wilcoxon	mtry=3	0.579	0.587	0.648	0.483	0.131	0.700	0.677	0.750	0.625	0.375
		PCA	mtry=5	0.581	0.579	0.732	0.368	0.101	0.700	0.719	0.917	0.375	0.318
		None	mtry=6	0.509	0.533	0.632	0.340	-0.031	0.800	0.724	0.917	0.625	0.565
Basic-normalization	Radial SVM	SBF::Wilcoxon	sigma=10, C=0.1	0.636	0.575	0.983	0.152	0.143	0.550	0.432	0.917	0.	-0.098
		PCA	sigma=0.1, C=1	0.574	0.636	0.767	0.303	0.073	0.700	0.719	1.	0.250	0.041
		None	sigma=100, C=0.1	0.580	0.500	1.	0.	0	0.600	0.5	1.	0.	0.
	Glmnet Logistic regression		alpha=0.48, lambda=0.1	0.584	0.646	0.705	0.417	0.128	0.650	0.677	0.833	0.375	0.222
		Recursive feature elimination	size=5 features	0.528	0.559	0.683	0.312	-0.002	0.550	0.469	0.750	0.250	0.
		Wilcoxon selection	-	0.608	0.634	0.648	0.552	0.201	0.700	0.760	0.667	0.750	0.400
		PCA	-	0.585	0.585	0.610	0.427	0.037	0.750	0.688	1.	0.375	0.419
	Random forests	None	-	0.479	0.479	0.437	0.537	-0.026	0.350	0.604	0.167	0.625	-0.182
		RFE	mtry = 2 , size = 51	0.556	0.544	0.675	0.392	0.068	0.650	0.625	0.833	0.625	0.222
		SBF::Wilcoxon	mtry=3	0.591	0.666	0.645	0.510	0.156	0.600	0.667	0.667	0.500	0.167
		PCA	mtry=4	0.545	0.548	0.665	0.382	0.047	0.650	0.641	0.833	0.375	0.222
		None	mtry=3	0.598	0.554	0.700	0.462	0.162	0.750	0.656	0.917	0.500	0.444
HM-1-patient	Radial SVM	SBF::Wilcoxon	sigma=10, C=10	0.580	0.541	1.	0.	0.	0.600	0.552	1.	0.	0.
		PCA	sigma=10, C=0.1	0.58	0.516	1.	0.	0.	0.600	0.563	1.	0.	0.
		None	sigma=1, C=0.1	0.580	0.591	1.	0.	0.	0.600	0.782	1.	0.	0.
	Glmnet Logistic regression		alpha=0.89, lambda=0.13	0.625	0.715	0.827	0.347	0.182	0.750	0.802	1.	0.375	0.419
		RFE	size=3	0.570	0.590	0.752	0.322	0.079	0.800	0.729	0.917	0.625	0.565
		SBF::Wilcoxon	-	0.547	0.555	0.612	0.455	0.069	0.550	0.583	0.750	0.250	0.000
		PCA	-	0.545	0.548	0.643	0.398	0.039	0.650	0.740	0.750	0.500	0.255

	Random forests	None	-	0.473	0.467	0.487	0.457	-0.056	0.300	0.729	0.417	0.125	-0.458
		RFE	mtry=11, size=11	0.502	0.509	0.578	0.390	-0.031	0.600	0.662	0.500	0.667	0.167
		SBF::Wilcoxon	mtry=10	0.582	0.608	0.698	0.415	0.115	0.550	0.698	0.583	0.500	0.082
		PCA	mtry=4	0.523	0.467	0.680	0.300	-0.018	0.750	0.792	0.750	0.750	0.490
		None	mtry=4	0.519	0.502	0.640	0.348	-0.011	0.550	0.562	0.667	0.375	0.043
HM-All-patients	Radial SVM	SBF::Wilcoxon	sigma=10, C=100	0.580	0.561	1.	0.	0.	0.600	0.521	1.	0.	0.
		PCA	sigma=1, C=0.1	0.567	0.561	0.963	0.015	-0.023	0.650	0.646	1.	0.125	0.146
		None	sigma=0.1, C=1	0.559	0.557	0.877	0.118	-0.004	0.750	0.875	0.917	0.500	0.444
	Glmnet		alpha=0.68, lambda=0.16	0.614	0.708	0.834	0.302	0.148	0.750	0.792	1.	0.375	0.419
	Logistic regression	RFE	size=3	0.567	0.598	0.730	0.338	0.072	0.600	0.563	0.917	0.125	0.048
		SBF::Wilcoxon	-	0.587	0.585	0.647	0.502	0.146	0.450	0.526	0.583	0.250	-0.170
		PCA	-	0.515	0.523	0.623	0.362	-0.014	0.700	0.750	0.833	0.500	0.348
		None	-	0.510	0.510	0.517	0.485	0.003	0.550	0.438	0.417	0.750	0.151
	Random forests	RFE	mtry=2, size=3	0.516	0.498	0.572	0.435	0.004	0.500	0.458	0.417	0.625	0.038
		SBF::Wilcoxon	mtry=9	0.578	0.621	0.722	0.373	0.100	0.650	0.740	0.667	0.625	0.286
		PCA	mtry=6	0.496	0.447	0.632	0.302	-0.062	0.650	0.698	0.750	0.500	0.255
		None	mtry=8	0.479	0.472	0.613	0.290	-0.096	0.550	0.646	0.583	0.500	0.082
HM-All-patients+ComBat	Radial SVM	SBF::Wilcoxon	sigma=0.1, C=1	0.580	0.560	1.	0.	0.	0.6	0.578	1.	0.	0.
		PCA	sigma=1, C=10	0.580	0.584	0.997	0.005	0.002	0.6	0.578	1.	0.	0.
		None	sigma=1, C=0.1	0.580	0.615	1.	0.	0.	0.600	0.661	1.	0.	0.
	Glmnet		alpha=0.62, lambda=0.21	0.572	0.716	0.882	0.143	0.027	0.600	0.740	1.	0.	0.
	Logistic regression	RFE	size=8	0.536	0.581	0.640	0.395	0.034	0.500	0.510	0.500	0.500	0.
		SBF::Wilcoxon	-	0.521	0.495	0.570	0.443	0.014	0.650	0.656	0.583	0.750	0.314
		PCA	-	0.522	0.542	0.580	0.437	0.015	0.550	0.573	0.583	0.500	0.082
		None	-	0.517	0.508	0.483	0.553	0.035	0.550	0.505	0.667	0.375	0.043
	Random forests	RFE	mtry=2, size=29	0.543	0.571	0.640	0.410	0.052	0.7	0.635	0.833	0.500	0.348
		SBF::Wilcoxon	mtry=3	0.641	0.651	0.742	0.502	0.248	0.650	0.615	0.750	0.500	0.255
		PCA	mtry=5	0.546	0.551	0.678	0.358	0.035	0.700	0.625	0.917	0.375	0.318
		None	mtry=4	0.547	0.564	0.635	0.423	0.061	0.650	0.542	0.700	0.500	0.255

NOTE. Abbreviations: AUC: area under the ROC curves, glmnet: penalized binomial logistic regression with elastic net regularization, HM: histogram matching, RFE: recursive feature elimination, SBF: select by filter, Se: sensitivity, Sp: specificity, SVM: support vector machines.

Annexe 6 : Supplementary Data Article 4

Supplementary Data 1: Assessment of linear correlations for all the texture features with a significant influence of temporal parameters according to repeated measures ANOVA (or non-parametric equivalent)

Texture features	Linear regression	
	R2	p-value §
Temporal resolution		
AUC_FOS_Average	0.1530	0.3856
AUC_FOS_Energy	0.2804	0.2215
AUC_FOS_Kurtosis	0.5652	0.0513
AUC_FOS_Skewness	0.6680	**0.0248***
AUC_FOS_Standard deviation	0.5671	0.0507
AUC_GLCM_Cluster prominence	0.1445	0.4002
AUC_GLRLM_GL variance	0.7712	**0.0130***
AUC_GLRLM_Short run emphasis	0.1150	0.4567
AUC_GLSZM_GL variance	0.6195	**0.0357***
AUC_GLSZM_Large area emphasis	0.7054	**0.0180***
AUC_GLSZM_Size zone non-uniformity	0.8787	**0.0018****
AUC_GLSZM_Zone variance	0.8297	**0.0043****
Ktrans_FOS_Entropy	0.6224	**0.0350***
Ktrans_GLCM_Contrast	0.5078	0.0723
Ktrans_GLCM_Inverse difference moment	0.5078	0.0723
Ktrans_GLCM_Joint average	0.297	0.2059
Ktrans_GLCM_Joint entropy	0.5992	**0.0411***
Ktrans_GLRLM_GL variance	0.3359	0.1727
Ktrans_GLRLM_High GL run emphasis	0.5227	0.0664
Ktrans_GLRLM_Low GL run emphasis	0.5227	0.0664
Ktrans_GLRLM_Long run emphasis	0.7080	**0.0176***
Ktrans_GLSZM_GL non-uniformity	0.9777	**<0.0001****
Ktrans_GLSZM_Large area emphasis	0.9094	**0.0009****
Ktrans_GLSZM_Size zone non-uniformity	0.9542	**0.0002*****
Ktrans_GLSZM_Zone variance	0.9502	**0.0002*****
Scan Duration		
Ktrans_FOS_Average	0.1418	0.7542
Ktrans_FOS_Energy	0.03273	0.8842
Ktrans_FOS_Entropy	0.6786	0.3837
Ktrans_FOS_Inter-quartile range	0.08587	0.8107
Ktrans_FOS_Kurtosis	0.9286	0.1722
Ktrans_FOS_Skewness	0.7519	0.3319
Ktrans_FOS_Standard deviation	0.2067	0.6995
Ktrans_GLCM_Cluster prominence	0.8606	0.2436
Ktrans_GLCM_Cluster shade	0.8433	0.2591
Ktrans_GLCM_Cluster tendency	0.7591	0.3266
Ktrans_GLCM_Contrast	0.6212	0.4221
Ktrans_GLCM_Inverse difference moment	0.6212	0.4221
Ktrans_GLCM_Joint average	0.7169	0.3571
Ktrans_GLCM_Joint entropy	0.6801	0.3827
Ktrans_GLRLM_GL variance	0.7169	0.3571
Ktrans_GLRLM_High GL run emphasis	0.5996	0.4361
Ktrans_GLRLM_Low GL run emphasis	0.5996	0.4361
Ktrans_GLRLM_Long run emphasis	0.5181	0.4885
Ktrans_GLRLM_Run length non-uniformity	0.5322	0.4795
Ktrans_GLRLM_Short run emphasis	0.4187	0.552
Ktrans_GLSZM_Large area emphasis	0.1157	0.7791
Ktrans_GLSZM_Zone variance	0.1879	0.7146

NOTE. Abbreviations : AUC: area under the 'time-concentration curve' at 90s after arrival of the contrast agent bolus in the acquisition volume ; Ktrans: influx volume transfer constant ; FOS: first order statistics ; GL: grey level ; GLCM: grey level co-occurrence matrix, GLRLM: grey level run

length matrix ; GLSZM: grey level size zone matrix. §: uncorrected p-values. The significant results after correction for multiple comparisons are underlined. * : $p \leq 0.050$, **: $p \leq 0.005$, ***: $p \leq 0.001$, ***: $p < 0.0001$

Supplementary Data 2 : Coefficient of variations with standard deviation of all the texture features

Texture Features	Changes in temporal resolution				Changes in scan duration	
	AUC		Ktrans		Ktrans	
	CV-dt	SD	CV-dt	SD	CV-T	SD
FOS_Energy	0.2908	0.2318	1.108	0.5278	0.9572	0.3552
FOS_Entropy	0.2455	0.3807	0.4601	0.3781	0.2832	0.2421
FOS_Inter-quartile Range	0.1050	0.1544	0.6555	0.3906	0.5671	0.2505
FOS_Kurtosis	0.4055	0.4336	0.648	0.5613	0.4800	0.3940
FOS_Average	0.1110	0.1434	0.6374	0.3750	0.5495	0.2472
FOS_skewness	0.2231	0.2515	0.5436	0.9945	0.3226	0.2292
FOS_Standard deviation	0.1752	0.1546	0.6718	0.3632	0.5814	0.2611
GLCM_Cluster prominence	0.5036	0.6206	0.6195	0.4919	0.3647	0.3050
GLCM_Cluster shade	0.2717	1.1540	0.6354	0.6317	0.3370	0.3028
GLCM_ClusterTendency	0.3289	0.4432	0.6052	0.4936	0.3552	0.2895
GLCM_Contrast	0.3119	0.4381	0.4429	0.3240	0.3910	0.4643
GLCM_correlation	0.4435	0.6429	0.5834	0.5057	0.3461	0.2879
GLCM_Inverse difference moment	0.0071	0.0069	0.0182	0.0253	0.0076	0.0081
GLCM_JointAverage	0.0484	0.0924	0.0239	0.0472	0.0054	0.0063
GLCM_JointEnergy	0.0332	0.0323	0.0834	0.1076	0.0368	0.0388
GLCM_JointEntropy	0.2429	0.3834	0.5005	0.4486	0.2989	0.2594
GLRLM_GL non-uniformity	0.0276	0.0251	0.0592	0.0896	0.0313	0.0371
GLRLM_GL variance	0.2623	0.3950	0.3379	0.3735	0.2231	0.2444
GLRLM_High GL run emphasis	0.0986	0.1700	0.0434	0.0691	0.0176	0.0143
GLRLM_Low GL run emphasis	0.1156	0.1935	0.0998	0.1633	0.0419	0.0369
GLRLM_Long run emphasis	0.1054	0.0910	0.2155	0.1855	0.1541	0.1081
GLRLM_Run length non-uniformity	0.0968	0.0873	0.2158	0.2664	0.1144	0.1330
GLRLM_Run entropy	0.0115	0.0097	0.0267	0.0326	0.0165	0.0150
GLRLM_Run variance	0.0911	0.0811	0.1758	0.1690	0.1230	0.1052
GLRLM_Short run emphasis	0.0636	0.0969	0.1187	0.1266	0.0776	0.0812
GLSZM_GL non-uniformity	0.2447	0.2635	0.3658	0.2589	0.3153	0.2432
GLSZM_GL variance	0.4287	0.3884	0.4387	0.3409	0.3192	0.1916
GLSZM_Large area emphasis	0.2186	0.1291	0.2899	0.1705	0.2964	0.1873
GLSZM_Small area emphasis	0.1510	0.2329	0.1447	0.1465	0.1641	0.2821
GLSZM_Size zone non-uniformity	0.2525	0.2201	0.3187	0.1765	0.2978	0.1463
GLSZM_Zone variance	0.2240	0.1781	0.2606	0.1451	0.2658	0.1570
GLSZM_Zone entropy	0.1305	0.1897	0.1359	0.1106	0.7271	0.08471

NOTE. Abbreviations : AUC: area under the 'time-concentration curves' at 90s after arrival of the contrast agent bolus in the acquisition volume ; Ktrans: influx volume transfer constant ; CV: coefficient of variation; CV-dt, CV-T: CV as a function of changes in temporal resolution, or changes in scan duration, for a given texture feature; SD: standard deviation, FOS: first order statistics ; GL: grey level ; GLCM: grey level co-occurrence matrix, GLRLM: grey level run length matrix ; GLSZM: grey level size zone matrix.

Supplementary Data 3 : Univariate analysis of the associations between texture features and response to chemotherapy for each model (depending on temporal resolution and scan duration).

Texture Features	p-value at univariate analysis									
	Temporal Resolution							Scan Duration §		
	dt2	dt4	dt6	dt8	dt10	dt12	dt20	T3'00	T4'00	T5'00
AUC_FOS_Energy	0.275	0.275	0.458	0.239	0.407	0.315	0.631	*0.458*	*0.458*	*0.458*
AUC_FOS_Entropy	0.407	0.827	0.965	0.458	0.827	0.896	0.631	*0.965*	*0.965*	*0.965*
AUC_FOS_Inter-quartile Range	0.275	0.570	0.760	0.896	0.631	0.694	0.827	*0.760*	*0.760*	*0.760*
AUC_FOS_Kurtosis	0.896	0.513	0.965	0.760	0.965	0.896	0.458	*0.965*	*0.965*	*0.965*
AUC_FOS_Average	0.513	1.000	1.000	0.896	0.827	0.760	0.827	*1.000*	*1.000*	*1.000*
AUC_FOS_skewness	0.896	0.694	0.965	0.694	0.827	0.965	0.359	*0.965*	*0.965*	*0.965*
AUC_FOS_Standard deviation	0.359	0.359	0.570	0.631	0.694	0.407	0.631	*0.570*	*0.570*	*0.570*
AUC_GLCM_Cluster prominence	0.239	0.407	0.76	0.239	0.513	0.513	0.694	*0.760*	*0.760*	*0.760*
AUC_GLCM_Cluster shade	0.896	0.631	0.631	0.631	0.694	0.513	0.694	*0.631*	*0.631*	*0.631*
AUC_GLCM_Cluster tendency	0.359	0.694	0.827	0.407	0.694	0.760	0.513	*0.827*	*0.827*	*0.827*
AUC_GLCM_Contrast	0.407	0.631	0.827	0.407	0.760	0.827	0.631	*0.827*	*0.827*	*0.827*
AUC_GLCM_Correlation	0.275	0.061	0.089	**0.032***	0.127	**0.040***	0.458	*0.089*	*0.089*	*0.089*
AUC_GLCM_Inverse difference moment	0.407	0.694	0.896	0.407	0.760	0.827	0.570	*0.896*	*0.896*	*0.896*
AUC_GLCM_Joint average	0.965	0.513	0.458	0.965	0.631	0.513	0.513	*0.458*	*0.458*	*0.458*
AUC_GLCM_Joint energy	0.359	0.631	0.827	0.359	0.760	0.760	0.570	*0.827*	*0.827*	*0.827*
AUC_GLCM_Joint entropy	0.359	0.694	0.896	0.407	0.760	0.896	0.570	*0.896*	*0.896*	*0.896*
AUC_GLRLM_GL non-uniformity	0.407	0.407	0.458	0.458	0.407	0.407	0.458	*0.458*	*0.458*	*0.458*
AUC_GLRLM_GL variance	0.275	0.570	0.76	0.275	0.631	0.513	0.458	*0.760*	*0.760*	*0.760*
AUC_GLRLM_High GL run emphasis	0.570	0.458	0.458	0.827	0.570	0.458	0.631	*0.458*	*0.458*	*0.458*
AUC_GLRLM_Low GL run emphasis	0.760	0.570	0.631	0.965	0.631	0.570	0.513	*0.631*	*0.631*	*0.631*
AUC_GLRLM_Long run emphasis	0.570	0.359	0.275	0.694	0.513	0.275	0.458	*0.275*	*0.275*	*0.275*
AUC_GLRLM_Run length non-uniformity	0.407	0.407	0.458	0.275	0.458	0.407	0.513	*0.458*	*0.458*	*0.458*
AUC_GLRLM_Run entropy	**0.050***	**0.050***	**0.050***	0.127	0.074	**0.032***	**0.040***	***0.050****	***0.050****	***0.050****
AUC_GLRLM_Run variance	0.359	0.275	0.206	0.176	0.694	0.275	0.458	*0.206*	*0.206*	*0.206*
AUC_GLRLM_Short run emphasis	0.570	0.631	0.965	0.359	0.760	0.570	0.896	*0.965*	*0.965*	*0.965*
AUC_GLSZM_GL non-uniformity	0.458	0.513	0.359	0.432	0.458	0.513	0.359	*0.359*	*0.359*	*0.359*
AUC_GLSZM_GL variance	0.570	0.238	0.965	0.337	0.827	0.631	0.315	*0.965*	*0.965*	*0.965*
AUC_GLSZM_Large area emphasis	0.570	0.513	0.359	0.570	0.206	0.513	0.407	*0.359*	*0.359*	*0.359*
AUC_GLSZM_Small area emphasis	0.239	0.315	0.570	0.896	0.239	0.127	0.061	*0.570*	*0.570*	*0.570*
AUC_GLSZM_Size zone non-uniformity	0.238	0.275	0.359	0.239	0.407	0.359	0.275	*0.359*	*0.359*	*0.359*
AUC_GLSZM_Zone variance	0.513	0.513	0.315	0.570	0.275	0.513	0.458	*0.315*	*0.315*	*0.315*
AUC_GLSZM_Zone entropy	0.150	0.176	0.513	0.275	0.896	0.206	0.407	*0.513*	*0.513*	*0.513*
Ktrans_FOS_Energy	0.359	0.150	0.359	0.896	0.458	0.896	0.359	0.194	0.484	0.359
Ktrans_FOS_Entropy	0.570	0.074	0.458	0.359	0.176	0.407	0.176	0.271	0.317	0.458
Ktrans_FOS_Inter-quartile range	0.458	**0.026***	0.206	0.359	0.127	0.570	0.407	0.162	0.549	0.206
Ktrans_FOS_Kurtosis	0.694	0.359	0.631	0.631	0.631	0.359	0.206	1.000	0.689	0.631
Ktrans_FOS_Average	0.315	0.760	0.150	0.359	0.176	0.631	0.315	0.271	0.617	0.150

Ktrans_FOS_Skewness	0.570	0.407	0.513	0.206	0.827	0.106	0.896	0.841	0.841	0.513
Ktrans_FOS_Standard deviation	0.827	**0.013***	0.206	0.359	0.127	0.694	0.760	0.110	0.110	0.206
Ktrans_GLCM_Cluster prominence	0.694	0.150	0.513	0.458	0.359	0.631	0.176	0.317	0.424	0.513
Ktrans_GLCM_Cluster shade	0.760	0.176	0.570	0.513	0.359	0.694	0.176	0.368	0.424	0.570
Ktrans_GLCM_Cluster tendency	0.760	0.127	0.513	0.458	0.275	0.570	0.150	0.317	0.368	0.513
Ktrans_GLCM_Contrast	0.694	0.074	0.359	0.458	0.315	0.458	0.150	0.317	0.368	0.359
Ktrans_GLCM_Correlation	0.176	0.239	0.359	0.827	0.965	0.206	0.827	1.000	0.841	0.359
Ktrans_GLCM_Inverse difference moment	0.694	0.074	0.359	0.458	0.315	0.458	0.150	0.317	0.368	0.359
Ktrans_GLCM_Joint average	0.760	0.106	0.513	0.760	0.275	0.458	0.150	0.317	0.368	0.513
Ktrans_GLCM_Joint energy	0.760	0.089	0.513	0.458	0.275	0.458	0.127	0.317	0.368	0.513
Ktrans_GLCM_Joint entropy	0.760	0.106	0.513	0.458	0.275	0.458	0.150	0.317	0.368	0.513
Ktrans_GLRLM_GL non-uniformity	0.570	0.631	0.458	0.513	0.570	0.407	0.570	0.617	0.617	0.458
Ktrans_GLRLM_GL variance	0.631	0.176	0.570	0.631	0.570	0.513	0.458	0.317	0.424	0.570
Ktrans_GLRLM_High GL run emphasis	0.631	0.176	0.570	0.965	0.631	0.513	0.513	0.317	0.424	0.570
Ktrans_GLRLM_Low GL run emphasis	0.631	0.176	0.570	0.965	0.631	0.513	0.513	0.317	0.424	0.570
Ktrans_GLRLM_Long run entropy	0.127	**0.021***	0.074	0.074	**0.032***	0.089	0.074	0.057	0.110	0.074
Ktrans_GLRLM_Run length non-uniformity	0.896	0.965	0.631	0.827	0.827	0.760	0.896	0.841	1.000	0.631
Ktrans_GLRLM_Run entropy	**0.021***	**0.032***	0.074	0.061	0.176	0.074	0.089	0.162	0.194	0.074
Ktrans_GLRLM_Run variance	0.074	0.061	0.150	0.106	0.061	0.239	0.089	0.110	0.134	0.150
Ktrans_GLRLM_Short run emphasis	0.513	0.127	0.827	0.760	0.407	0.760	0.359	0.549	0.764	0.827
Ktrans_GLSZM_GL non-uniformity	0.631	0.694	0.407	0.458	0.570	0.407	0.513	0.549	0.617	0.407
Ktrans_GLSZM_GL variance	0.760	0.760	0.315	0.106	0.205	0.359	0.176	0.484	0.317	0.315
Ktrans_GLSZM_Large area emphasis	0.239	0.150	0.407	0.239	0.275	0.275	0.275	0.617	0.549	0.407
Ktrans_GLSZM_Small area emphasis	0.513	0.275	0.407	0.965	0.074	**0.050***	0.106	**0.021***	**0.016***	0.407
Ktrans_GLSZM_Size zone non-uniformity	0.694	0.694	0.458	0.458	0.570	0.513	0.727	0.689	0.689	0.458
Ktrans_GLSZM_Zone variance	0.275	0.150	0.458	0.407	0.239	0.315	0.315	0.617	0.549	0.458
Ktrans_GLSZM_Zone entropy	0.570	0.407	0.827	0.827	0.827	0.275	0.315	0.230	0.109	0.827

NOTE. Abbreviations : AUC: area under the 'time-concentration curves' at 90s after arrival of the contrast agent bolus in the acquisition volume ; Ktrans: influx volume transfer constant ; FOS: first order statistics ; GL: grey level ; GLCM: grey level co-occurrence matrix, GLRLM: grey level run length matrix ; GLSZM: grey level size zone matrix. 'dtx' corresponds to a sampling of 'x' seconds; 'Ty'00' corresponds to a scan duration of 'y' minutes.

§: regarding the changes in scan duration, the temporal resolution was fixed at dt=6s. The statistical analyses based on AUC parametric maps were not influenced because the scan duration was >> to 90s.

* : $p \leq 0.050$

Supplementary Data 4: Comparisons of the AUROC of the different models, depending on scan duration and temporal resolution

Comparisons		p-value
Temporal resolution		
Model_dt2 versus:	Model_dt4	0.9277
	Model_dt6	0.5832
	Model_dt8	0.8537
	Model_dt10	0.8537
	Model_dt12	0.8637
	Model_dt20	0.7884
Model_dt4 versus:	Model_dt6	0.6817
	Model_dt8	0.9349
	Model_dt10	0.9439
	Model_dt12	0.9385
	Model_dt20	0.8700
Model_dt6 versus:	Model_dt8	0.7378
	Model_dt10	0.7378
	Model_dt12	0.7514
	Model_dt20	0.8122
Model_dt8 versus:	Model_dt10	1.000
	Model_dt12	1.000
	Model_dt20	0.9321
Model_dt10 versus:	Model_dt12	1.000
	Model_dt20	0.9321
Model_dt12 versus:	Model_dt20	0.9356
Scan duration		
Model_T3'00 versus:	Model_T4'00	0.2788
	Model_T5'00	0.3521
Model_T4'00 versus:	Model_T5'00	0.8552

NOTE. Results are the p-value of the pair-wise comparisons according to the Delong method, without correction for multiple comparisons.
'Model_dti' corresponds to a prediction model based on texture features extracted from parametric maps built with a sampling of 'dt=i' seconds. 'Modele_Tj'00' corresponds to a prediction model based on texture features extracted from parametric maps built with a scan duration of 'T=j'00' minutes.

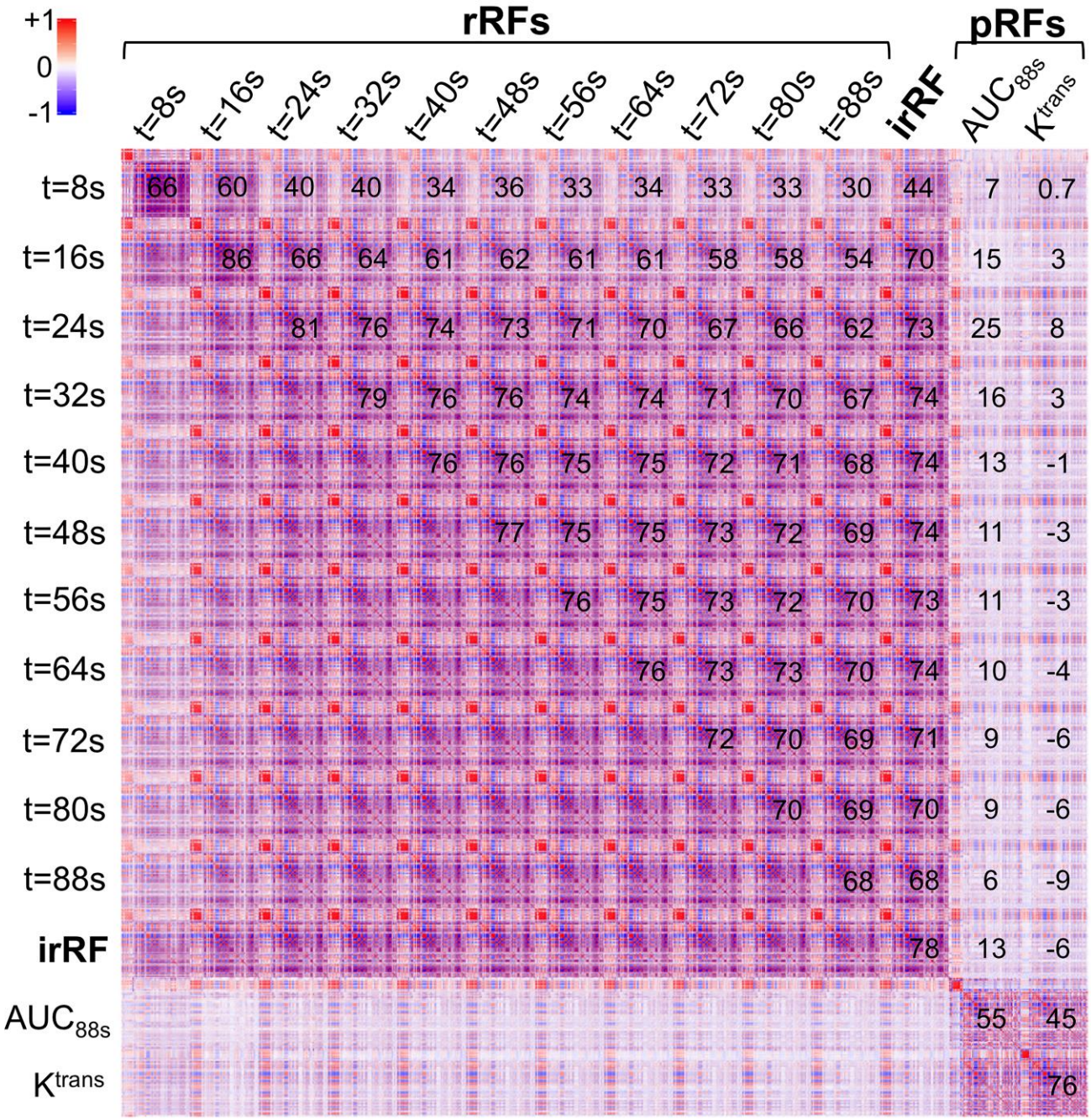
Annexe 7 : Supplementary Data Article 5

Supplementary Data 1: Inter- and intra-observer agreements for the semantic radiological analysis.

Variables	Inter-observer agreement[§]						Intra-observer agreement[§§]	
	R1 vs. R2	p-value	R1 vs. R3	p-value	R2 vs. R3	p-value	R1 vs. R1-2	p-value
Longest diameter[1]	0.990	<0.00001***	0.994	<0.00001***	0.998	<0.00001***	0.994	<0.00001***
Heterogeneous SI on T1-WI[2]	0.673	<0.00001***	0.843	<0.00001***	0.829	<0.00001***	0.862	<0.00001***
Heterogeneous SI on T2-WI[2]	0.642	<0.00001***	0.804	<0.00001***	0.833	<0.00001***	0.970	<0.00001***
Heterogeneous SI on T1-WI[2]	0.490	<0.00001***	0.735	<0.00001***	0.743	<0.00001***	0.862	<0.00001***
Necrotic signal[2]	0.580	<0.00001***	0.922	<0.00001***	0.641	<0.00001***	0.922	<0.00001***
Haemorragic signal[2]	0.613	<0.00001***	0.68	<0.00001***	0.886	<0.00001***	0.807	<0.00001***
Peritumoral edema[2]	0.424	0.00002***	0.714	<0.00001***	0.56	<0.00001***	0.808	<0.00001***
Peritumoral enhancement[2]	0.576	<0.00001***	1	<0.00001***	0.576	<0.00001***	0.855	<0.00001***
MRI growth pattern[2]	0.567	<0.00001***	0.969	<0.00001***	0.598	<0.00001***	0.938	<0.00001***
Tail sign[2]	0.823	<0.00001***	0.648	<0.00001***	0.653	<0.00001***	0.477	0.00002***
Vessel and/or nerve invasion[2]	0.955	<0.00001***	0.884	<0.00001***	0.932	<0.00001***	0.788	<0.00001***
Bone invasion[2]	0.814	<0.00001***	0.873	<0.00001***	0.939	<0.00001***	0.900	<0.00001***

NOTE. The following tests were performed: [1]: intra-class correlation coefficients; [2]: weighted Kappa.
[§]: Inter-observer agreement between the 3 radiologists: R1 (: A.C., a senior radiologist with 8 years of experience in MRI including 3 years in a sarcoma reference center), R2 (: D.F., a fellow with 3 years of experience including a 6-month internship in a sarcoma reference center with) and R3 (: M.K., a senior radiologist with 27 years of experience in sarcoma imaging).
[§§]: Intra-observer agreement for one of the senior radiologists (R1/A.C.) with a delay of 3 months between the 2 readings (:R1-2).
Abbreviations: SI: signal intensity, WI: weighted imaging.
*: $p<0.05$, **: $p<0.005$, ***: $p<0.001$

Supplementary Data 2. Correlation plot of all the radiomics features (RFs), relative RFs (rRFs), integrated rRF (irRFs) and parametric RFs (pRFs).
In total, 11 x 92 rRFs were extracted from each phase of the DCE-MRI acquisition (t=8s to t = 88s), 92 irRFs were calculated (as $\int$ rRF(t).dt), and 2 x 92 pRFs were extracted from Ktrans and AUC88s (semi)-parametric maps.
The color of each pixel correspond the value of the spearman correlation coefficient (: ρ). The number inside each larger square correspond to the average ρ for the 92 x 92 correlations between 2 (ir/r/p)RFs categories (x10^{-3}). A large number indicates that the (ir/r/p)RFs categories tend to be correlated.



Supplementary Data 3. Comparisons of the performances of the predictive models measured with concordance-index (c-index) and integrated are under the curve (iAUC).

	Model-1	Model-2	Model-3	Model-4	Model-5	Model-6	Model-7
Model-1	0 (0; 0)	**-0.08* (-0.294; -0.02)**	-0.026 (-0.172; 0.068)	0.004 (-0.111; 0.126)	0.004 (-0.111; 0.126)	-0.035 (-0.188; 0.05)	0.018 (-0.017; 0.089)
Model-2	0.1 (-0.05; 0.208)	0 (0; 0)	0.054 (-0.001; 0.21)	**0.085* (0.062; 0.275)**	**0.085* (0.062; 0.275)**	0.046 (-0.009; 0.175)	**0.099* (0.06; 0.33)**
Model-3	0.033 (-0.164; 0.134)	-0.067 (-0.186; 0.037)	0 (0; 0)	0.031 (-0.01; 0.117)	0.031 (-0.01; 0.117)	-0.008 (-0.114; 0.085)	0.044 (-0.008; 0.184)
Model-4	0.014 (-0.185; 0.117)	-0.086 (-0.209; 0.022)	-0.019 (-0.059; 0.043)	0 (0; 0)	0 (0; 0)	-0.039 (-0.176; 0.023)	0.014 (-0.059; 0.115)
Model-5	0.014 (-0.185; 0.117)	-0.086 (-0.209; 0.022)	-0.019 (-0.059; 0.043)	0 (0; 0)	0 (0; 0)	-0.039 (-0.176; 0.023)	0.014 (-0.059; 0.115)
Model-6	0.054 (-0.125; 0.152)	-0.046 (-0.135; 0.039)	0.021 (-0.056; 0.103)	0.04 (-0.045; 0.119)	0.04 (-0.045; 0.119)	0 (0; 0)	0.053 (-0.015; 0.218)
Model-7	-0.015 (-0.113; 0.044)	**-0.115* (-0.22; -0.003)**	-0.048 (-0.115; 0.081)	-0.029 (-0.092; 0.098)	-0.029 (-0.092; 0.098)	-0.068 (-0.152; 0.05)	0 (0; 0)

NOTE. The upper right part of the table (in light blue) shows the bootstrapped differences between c-indices for each pair of models for 10000 replicates.
The lower left part of the table (in light orange) shows the bootstrapped differences between iAUCs for each pair of models for 10000 replicates.
Data in parentheses are 95% confidence intervals (95%CI). A comparison was considered significant if the 95%CI did not include 0. *: $p<0.05$

Annexe 8 : Revue des approches radiomics dédiées aux sarcomes

Critical Review of Sarcomas Radiomics Studies: Quantifying Quality to Improve Reproducibility

ABSTRACT

Objectives: Sarcomas are a model for intra- and inter-tumoral heterogeneities making them particularly suitable for radiomics analyses. Our purposes were to review the aims, methods and results of radiomics studies involving sarcomas

Methods: Pubmed and Web of Sciences databases were searched for radiomics or textural studies involving bone, soft-tissues and visceral sarcomas until February 2020. Two radiologists evaluated their objectives, results and quality of their methods, imaging pre-processing and machine-learning workflow helped by the items of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), Image Biomarker Standardization Initiative (IBSI) and 'Radiomics Quality Score' (RQS).

Statistical analyses included inter-reader agreements, correlations between methodological assessments, scientometrics indices, and their changes over years, and between RQS, number of patients and models performance.

Results: Forty-seven studies were included involving: soft-tissue sarcomas (27/47, 57.4%), bone sarcomas (12/47, 25.5%), gynecological sarcomas (6/47, 12.8%) and mixed sarcomas (2/47, 4.3%), mostly imaged with MRI (29/47, 61.7%), for a total of 3657 distinct patients. Median RQS was 4 (27.3% of the maximum, range: -7 – 17). Performances of predictive models and number of patients negatively correlated (p=0.026). None of the studies detailed all the items from the IBSI guidelines. There was a significant increase in studies' impact factors since the establishing of the RQS in 2017 (p=0.029).

Conclusion: Although showing promising results, further efforts are needed to make sarcoma radiomics studies reproducible with an acceptable level of evidence. A better knowledge of the RQS and IBSI reporting guidelines could improve the quality of sarcoma radiomics studies and accelerate clinical applications.

HIGHLIGHTS

1. Preliminary retrospective studies suggest that radiomics could improve the diagnostic management and prognostic predictions for sarcoma patients.
2. Several methodological items depicting the pre and post-processing of medical images and the machine-learning pipeline are commonly missing in current sarcoma radiomics studies
3. The IBSI reporting guidelines and the Radiomics Quality Score (RQS) can help structuring and improving the reproducibility of radiomics studies

KEY-WORDS

Radiomics;

Meta-analysis;
Sarcomas;
Machine-learning;
Quality improvement

ABBREVIATIONS

^{18}F-FDG-PET: ^{18}F-fluorodeoxyglucose positron emission tomography
AUROC: area under the receiver operating characteristics curve
C-index: concordance index
DCE-MRI: dynamic contrast enhanced MRI
DWI: diffusion weighted imaging
IBSI: image biomarkers standardization initiative
PDI: proportion of detailed items
QUADAS: quality assessment of diagnostic accuracy studies
RF: radiomics features
ROI: region of interest
RQS: radiomics quality score
STS: soft tissue sarcoma
TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis
VOI: volume of interest
WI: weighted imaging

INTRODUCTION

Radiomics has now become one of the largest fields of research in medical imaging. In oncologic imaging, radiomics consists in the extraction and mining of hundreds of numeric variables that non-invasively quantify the radiological phenotype (i.e. their shape, texture and heterogeneity) of whole tumor, *in situ* and *in vivo*, from any imaging modality [1,2]. Therefore, radiomics appears complementary to histopathological and molecular analyses, which, by definition, rely on biopsy samples at risk of sampling bias. The underlying hypotheses of radiomics are that it reflects the molecular features of tumor and that integrating radiomics data with non-radiological and '*-omics'* data could enhance predictive models. Hence, screening all these tumors features would enable to identify more relevant subgroups of patients, to better tailor their treatments and, consequently, to improve their survivals compared with traditional '*one-size-fits-all'* approaches.

Sarcomas are malignant mesenchymal ubiquitous tumors that can affect bone, soft tissues or the viscera [3]. Sarcomas are characterized by their heterogeneity at all

levels, from molecular subgroups (from simple amplification to highly complex genomic profiles) to radiological presentations (from uniform superficial nodules to multicompartmental heterogeneous masses) [4,5]. Thus, although rather rare tumors, sarcomas are a model for radiomics approaches with promising results. Indeed, radiomics has demonstrated improvements in the prediction of malignancy [6–14], histological grading of sarcoma [15-19], patient's prognosis [15,19–25], and response evaluation [26–30] over traditional radiological approaches.

However, to our knowledge, no oncologic radiomics studies have translated to clinical applications and clinicians could question their validity, as for other imaging biomarkers [31]. Referring to the imaging biomarker roadmap by O'Connor *et al.* can help understanding why [32]. Several properties defining biomarkers are missing and none of the translational gaps has been overcome. Indeed, there is a lack of evidence of intra- and multi-centric repeatability, reproducibility, specificity consistency, temporality or cost-effectiveness of the radiomics signatures that have been developed so far. These failures are partly due to the technical and complex pipelines necessary to the extraction and computation of radiomics features and to their integration into machine-learning algorithms. Besides general tools to assess the quality of diagnostic accuracy studies (such as TRIPOD - for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis – or QUADAS-2 – for QUality Assessment of Diagnostic Accuracy Studies), two main initiatives have emerged to clarify the methodology and to improve the quality of radiomics studies [33,34]. In 2017, Lambin *et al.* have defined the Radiomics Quality Score (RQS) that lists the crucial aspects of radiomics studies and estimates their quality [1]. Second, the Image Biomarkers Standardization Initiative (IBSI), an independent international collaboration, have provided a comprehensive and detailed review of each mandatory step for radiomics analyses including nomenclature of the radiomics features, general schemes and datasets for calibration [35–37] (https://ibsi.readthedocs.io/en/latest/). Hence, the next challenge for radiomics is to overcome the turning point between proofs of concept and real-life application.

Therefore, considering sarcomas as a paragon of this challenge, our aims were to review the sub-fields and objectives of sarcoma radiomics studies, to evaluate their results and quality in the light of the RQS and IBSI reporting guidelines in order to stress their limits and discuss further efforts and solutions.

MATERIALS AND METHODS

Study design

We used the PubMed and Web of Science electronic database to identify full research articles investigating radiomics approaches dedicated to sarcomas from 01/01/2000 to 02/01/2020. The search items consisted in: ("histogram" or "texture" or "textural" or "radiomics") and ("sarcoma" or "soft-tissue tumor" or "osteosarcoma" or "chondrosarcoma or "leiomyosarcoma"). This search identified: 497 results with Web of Science and 480 results with PubMed. After exclusion of duplicates, non-English speaking articles, single abstract, congress publications, non-medical imaging studies, and studies with less than 5 radiomics features, a total of 47 articles were finally included. Full-text selection and review were performed by two radiologists, alone and in consensus: one senior radiologist from a sarcoma reference center and PhD student in applied mathematics (A.C.) and one fellow with 3.5 years of experience in radiology including one year in a sarcoma reference center (D.F.).

Assessment of the quality of radiomics studies

Descriptive analysis of the studies. The two reviewers reported: (1) the year of publication, (2) scientometrics of the publication journal (impact factor, h-index, citescore in 2018), (3) specialty of the journal, (4) if the study had a methodological/technical intention, (5) the specialty and (6) gender of the first author, (7) if a biostatistician, data scientist or applied mathematician participated in the study, (8) the proportion of radiologists among the authors, (9) the study first objectives, (10) the imaging modality, (11) the number of patients, (12) the use of public datasets and (13) if semantic radiological features were added to radiomics-based models.

Estimation of the RQS. The RQS is made of 16 items reflecting 6 different methodological domains of the radiomics studies that are summarized in Table 1. A note was attributed to each item by the two radiologists (blinded from the reading to the other radiologist) and the sum of these notes provided the RQS, which can range from -8 to 36.

Estimation of the quality of image processing and RFs extraction. Based on the IBSI guidelines, the radiologists consensually evaluated if the following 'processing' items were detailed in the methods and how: (1) use of denoising algorithm; (2) use of bias field correction algorithm for MRI studies; (3) voxel-size standardization; (4) gray-

levels discretization; (5) use of signal intensity (SI) harmonization technique for MRI studies; (6) segmentation method; (7) multiple segmentation; (8) segmentation of volume or region of interest (VOI and ROI, respectively); (9) meshing method for shape radiomics features (RFs); (10) aggregation method for 2nd order texture features; (11) details and number of RFs per RFs categories; and (12) software. A percentage of detailed items (PDI) for reproducibility of the image processing was calculated as follows: 100 × the number of detailed items divided the total number of processing items assessable (i.e. excepted denoising; bias-field correction; SI normalization if MRI was not used, aggregation if not needed; mesh in the absence of shape features).

Estimation of the quality of the statistical / machine-learning pipeline. The radiologists reported if the following items of a predictive model were given, and, if present, their details, namely: (1) multivariate analysis; (2) RFs selection before multivariate analysis; (3) resampling of the cohort used to train model; (4) machine-learning hyperparameters tuning (if the study did not rely on a single univariate analysis and used algorithms with hyperparameters); (5) metrics to assess models performance; (6) evaluation on an independent validation cohort; (7) calibration curve; (8) cut-off assessment of the radiomics score or models predictions' probabilities; and (9) analysis of the best model outliers. Similarly, we calculated a PDI to estimate the quality of the statistical / machine-learning pipeline.

Estimation of the quality of the general methodology. The radiologists consensually assessed if the following information were detailed in the methods: (1) study design: (1) retrospective / prospective; (2) single / multi-centric; (3) inclusion method; (4) flow-chart; (5) depiction of the cohort(s); (6) depiction of the imaging protocols; (7) depiction of the patients' management following imaging; (8) comparison of the best model to the best radiological model; (9) comparison of the best model to the best clinical-pathological model. Similarly, we calculated a PDI for methodological quality.

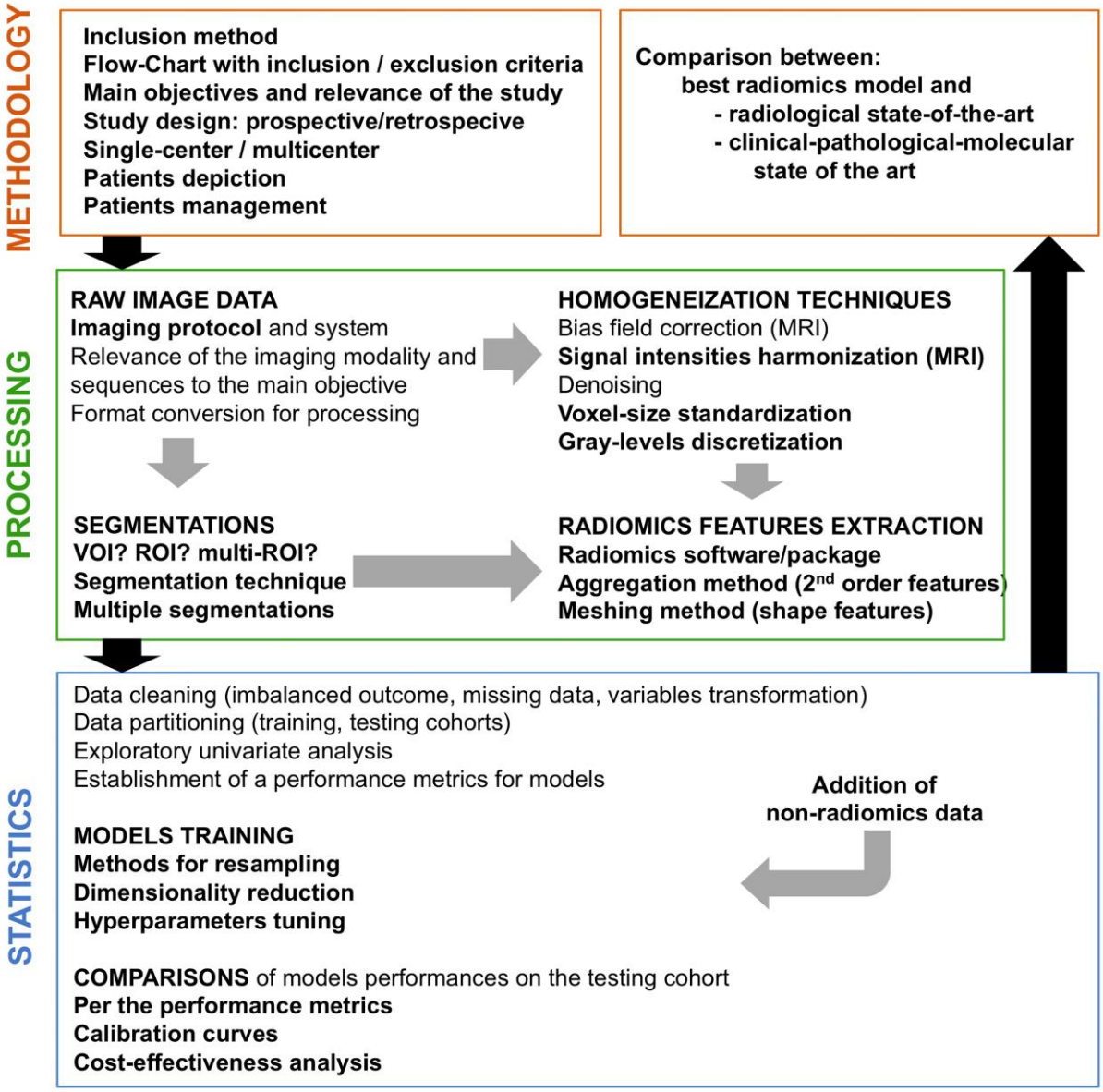
Results of the studies. When given, the performance metrics of the final radiomics-based model of the article was reported, namely: accuracy, area under the ROC curve (AUROC) or concordance-index (c-index). For univariate analysis alone was done, we reported the best univariate metrics.

Figure 1 summarizes the entire radiomics pipeline and all the items we reviewed.

Table 1. Radiomics Quality Score (RQS) domains, items and levels for each item.

RQS domain	RQS item	RQS levels
1. Protocol quality and stability in image an segmentation (0 to +5)	Image protocol	(1) for well documented protocol (2) for publicly available protocol
	Multiple segmentations	(1) if performed multiple times, with feature robustness assessment
	Phantom study	(1) if texture phantoms were used for feature robustness assessment
	Test-retest	(1) for multiple time points for feature robustness assessment
2. Feature selection and validation (-8 to +8)	Feature reduction	(-3) if nothing is done (+3) if either feature reduction or correction for multiple testing
	Validation	(-5) if no validation cohort (2) for internal validation (3) for external validation (4) two external validations or validation of previously published signature (5) validation on ≥ 3 datasets from > 1 institution
3. Model performance index (0 to +5)	Discrimination statistics	(1) for discrimination statistic and statistical significance (2) if resampling is applied
	Calibration statistics	(1) for discrimination statistic and statistical significance (2) if resampling is applied
	Cutoff analysis	(1) if cutoff either pre-defined or at median or continuous risk variable reported
4. Biologic / clinical validation and utility (0 to +6)	Non radiomics features	(1) if multivariable analysis with non radiomics features
	Biologic correlates	(1) if present
	Comparison to gold standard	(2) if comparison to gold standard
	Potential clinical utility	(2) for reporting clinical utility
5. Level of evidence	Prospective study	(7) for prospective analysis
	Cost-effective analysis	(1) for cost effective analysis
6. Open science and data	-	(1) for open source scans (2) + open source segmentations (3) + open source codes (4) + open source representative segmentations and features

Figure 1. Schematic view of the study assessments. Items in bold correspond to those specifically investigated in the present study.



Statistical analyses

Statistical analyses were performed with R (version 3.5.3; R Foundation for Statistical Computing) using the 'tidyverse' and 'irr' packages [38]. Categorical variables were given as number and percentages. Numeric variables were given as mean and standard deviation or median and range as appropriate. Correlations between numeric variables were evaluated with linear regressions. Associations between numeric and categorical variables were assessed with unpaired Student t-test of Mann-Whitney tests, as appropriate. Inter-observer agreements were assessed with: (1) interclass correlation coefficients (ICC) for RQS using a two-way random effect model determining absolute agreement between the radiologists; (2) classical or weighted Kappa (κ) for the categorical ordinal items of the RQS, as appropriate. All tests were two-tailed. A p-value of less than 0.05 was deemed significant.

RESULTS

Summary of the radiomics studies

Table 2 depicts the general features of the radiomics studies. They were mostly published in radiological journals (31/47 – 66%). The first authors were radiologists (or nuclear physicians) in 29/47 (61.7%), mostly men (18 out of 43 analyzable names, 36.4%) and the average percentage of radiologists among the authors was of 53.3% (± 33).

In total, 27/47 (57.4%) involved soft-tissues sarcomas, 12/47 (25.5%) osseous sarcomas, 6/47 (12.8%) gynecologic sarcomas and 2/47 (4.3%) studies mixed soft-tissues and osseous sarcomas. The most frequent imaging modality was MRI (31/47, 66% - including combinations with PET, structural MRI, diffusion-weighted imaging [DWI] and dynamic-contrast-enhanced [DCE] MRI).

The most frequent first aim of the study was the distinction between benign and malignant tumors (18/47, 38.3%), followed by the prediction of patients' prognosis (11/47, 23.4%). Overall, 3657 non-redundant patients were included with a median of 58 patients per study (range: 11 – 226).

Table 2. Characteristics of the sarcoma radiomics studies.

Characteristics	No. of studies (%)
Journal Speciality	
Imaging*	31/47 (66%)
Clinical	8/47 (17%)
Generalist	6/47 (12.8%)
Medical physics	2/47 (4.3%)
Methodological paper	
No	42/47 (89.4%)
Yes	5/47 (10.6%)
Specialty of the 1st author	
Radiologist[1]	29/47 (61.7%)
Radiotherapist	4/47 (8.5%)
Clinician[2]	2/47 (4.3%)
Physicist, Computer scientist	12/47 (25.5%)
Gender of the 1st author[3]	
Women	16/43 (36.4%)
Men	18/43 (63.6%)
Percentage of radiologists among authors[4]	58.3 (0 - 100)
Biostatician, data/computer scientists, applied mathematicians among authors	
No	21/47 (44.7%)
Yes	26/47 (55.3%)
Geographical origin of the study	
Asia	20/47 (42.6%)
Europe	19/47 (40.4%)
North America	8/47 (17%)
Imaging modality	
Ultrasound	1/47 (2.1%)
CT-scan	8/47 (17%)
18F-FDG-PET/CT	7/47 (14.9%)
Structural MRI	19/47 (40.4%)
Advanced' MRI	5/47 (10.6%)
Structural and advanced MRI	5/47 (10.6%)
Multi-modalities (PET and MRI)	2/47 (%)
Types of sarcoma	
Bone sarcoma	12/47 (25.5%)
Gynecologic sarcoma	6/47 (12.8%)
Soft-tissues sarcoma	27/47 (57.4%)
Mixed	2/47 (4.3%)
Total number of patients[5]	3657
Number of patients per study[4]	58 (11 - 226)
Public data sets	2/47 (4.3%)
Study objectives	
Discrimination benign/malignant tumor	18/47 (38.3%)
Prognosis prediction	11/47 (23.4%)
Response to treatment prediction	7/47 (14.9%)
Correlation with histological or molecular features	1/47 (2.1%)
Local relapse prediction	2/47 (4.3%)
Grading prediction	4/47 (8.5%)
Grading and prognosis prediction	2/47 (4.3%)

Other	2/47 (4.3%)
Radiological inputs to boost radiomics model	
No	39/47 (83%)
Yes	8/47 (17%)

NOTE. Data are number (: no.) of patients with percentage in parentheses.
1. Radiologist and nuclear physician
2. Clinician are oncologic surgeons and medical oncologists
3. Gender was not assessable for 4 studies
4. Results are given as median and range in parentheses

RQS

Table 3 shows the notes attributed to each item of the RQS and its final values with inter-rater agreements. None of the study did a test-retest analysis of the RFs, nor a phantom study. Only one (methodological) study was prospectively designed. Thirty-four out of 47 (72.3%) studies did not validate their results on an independent cohort. Biological correlations and cost-effectiveness analysis were available in 2/47 (4.3%) studies, respectively. Correlations between raters were all at least moderate (i.e. > 0.40) except for cut-off analysis (weighted-κ = 0.28), comparison to gold standard (κ = 0.33), the use of non-RF variables (κ = 0.35) and assessment of potential clinical utility (κ = 0.23). The median RQS was of 4 (range: -7 to +17) with an excellent ICC (0.90, $p < 0.0001$).

Quality of the image processing and RFs extraction

Figure 2 shows the proportions of studies that detailed each processing item. The median number of extracted RFs was 44 (range: 9 – 210105). First-order feature were extracted in 45/47 (95.7%) studies, 2nd order texture features in 42/47 (89.4%) studies, shape features in 30/47 (63.8%) and other features in 8/47 (17%) studies. VOIs were used in 35/47 (74.5%) studies, single ROI in 8/47 (17%) studies and multiple-ROI in 3/47 studies (6.4%) – with one study not giving this information. Seven out of 47 (14.9%) used a semi-automatic segmentation with a manual correction; all the others used a manual segmentation. Information was missing: in 8/47 (17%) studies regarding which software (or package) was used to extract RFs; in 17/31 (54.8%) regarding the use of SI harmonization techniques; in 29/47 (61.7%) regarding eventual voxel-size standardization. The aggregation method was indicated in 3 out of the 42 (7.1%) studies with 2nd order RFs. The median PDI for image processing and RFs extraction was 55.6% (range: 30 - 88.9).

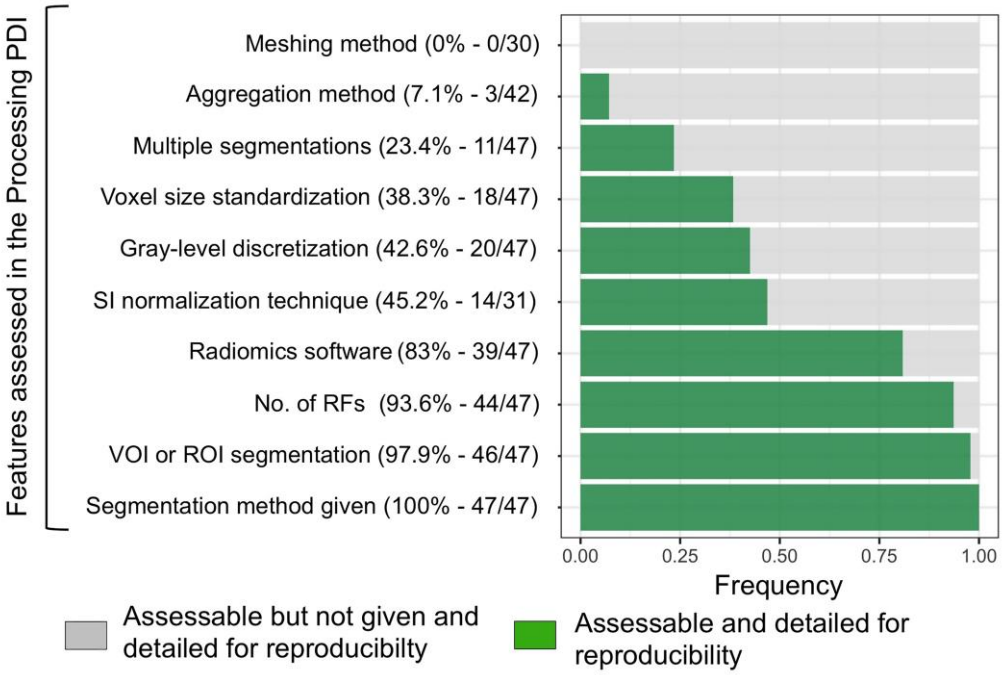
Table 3. Results of Radiomics Quality Score (RQS) analysis.

RQS items	No. of studies (%)	Inter-observer agreements (p-value)
Protocol quality		
0	16/47 (34%)	0.448 (<0.0001***)
1	29/47 (61.7%)	
2	2/47 (4.3%)	
Multiple segmentations		
0	36/47 (76.6%)	0.601 (<0.0001***)
1	11/47 (23.4%)	
Phantom study		
0	47/47 (100%)	1. (-)
1	0/47 (0%)	
Test retest		
0	47/47 (100%)	1. (-)
1	0/47 (0%)	
Features reduction		
3	15/47 (31.9%%)	0.928 (<0.0001***)
-3	32/47 (68.1%%)	
Validation		
-5	34/47 (72.3%)	0.918 (<0.0001***)
2	10/47 (21.3%)	
3	2/47 (4.3%)	
4	1/47 (2.1%)	
Discrimination statistics		
0	4/47 (8.5%)	0.588 (<0.0001***)
1	16/47 (34%)	
2	27/47 (57.4%)	
Calibration statistics		
0	43/47 (91.5%)	0.513 (<0.0001***)
1	3/47 (6.4%)	
2	1/47 (2.1%)	
Cutoff analysis		
0	42/47 (89.4%)	0.28 (0.0536)
1	5/47 (10.6%)	
Non radiomics features		
0	33/47 (70.2%)	0.352 (0.0157*)
1	14/47 (29.8%)	
Biological correlates		
0	45/47 (95.7%)	0.647 (<0.0001***)
1	2/47 (4.3%)	
Comparison to gold standard		
0	24/47 (51.1%)	0.334 (0.0058*)
1	23/47 (48.9%)	
Prospective study		
0	46/47 (97.9%)	1 (<0.0001***)
7	1/47 (2.1%)	
Cost effectiveness analysis		
0	45/47 (95.7%)	1 (<0.0001***)
1	2/47 (4.3%)	
Potential clinical utility		
0	12/47 (25.5%)	0.234 (0.0521)

2	35/47 (74.5%)	
Open science		
0	45/47 (95.7%)	1 (<0.0001***)
3	2/47 (4.3%)	
RQS§	**4 (-7 - +17)**	**0.899 (<0.0001***)**

NOTE. Results are number of patients with percentage in parentheses and inter-observer agreements (Cohen's Kappa, weighted Kappa and interclass correlation coefficients) with p-value in parentheses, except for RQS (§) given as median and range in parentheses. The first column of results corresponds to the reading of the senior radiologist.
*: $p < 0.05$; **: $p < 0.005$; ***: $p < 0.001$.

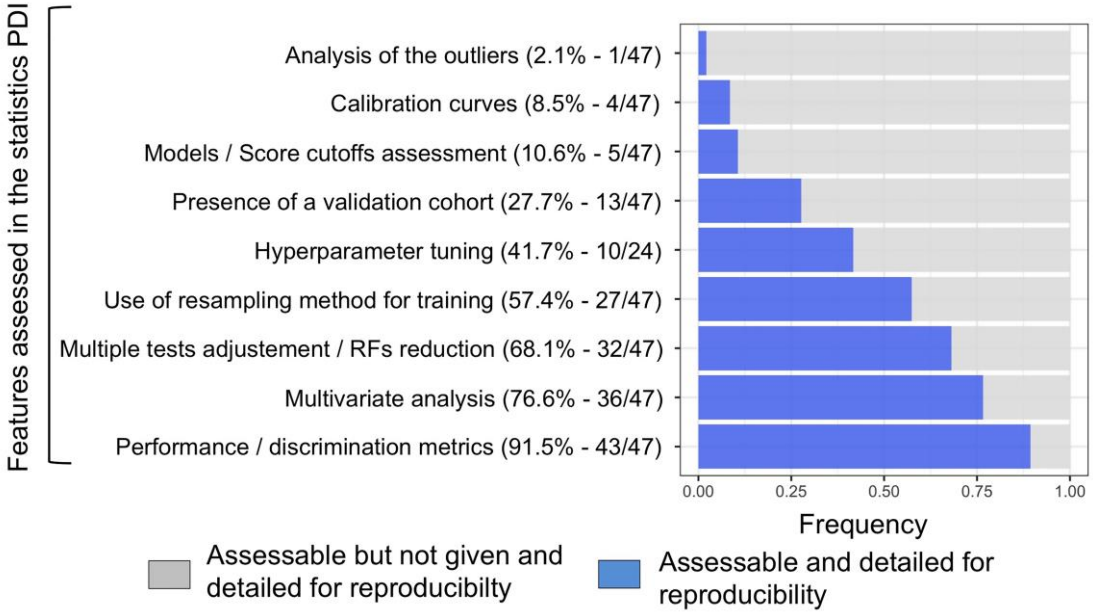
Figure 2. Proportions of studies providing the data necessary to the reproducibility of the results relative to image processing and radiomics features computation.
Abbreviations: No.: number; PDI: proportion of detailed items; RF: radiomics features



Quality of the statistical / machine-learning pipeline

Figure 3 shows the proportions of studies that detailed each statistical items. Eleven of the 47 (23.4%) studies only performed an exploratory univariate analysis. Careful analysis of the failures (i.e. outliers) of the best final model was found in only one study. The median PDI for the statistical / machine-learning pipeline was 44.4% (range: 0 – 77.8).

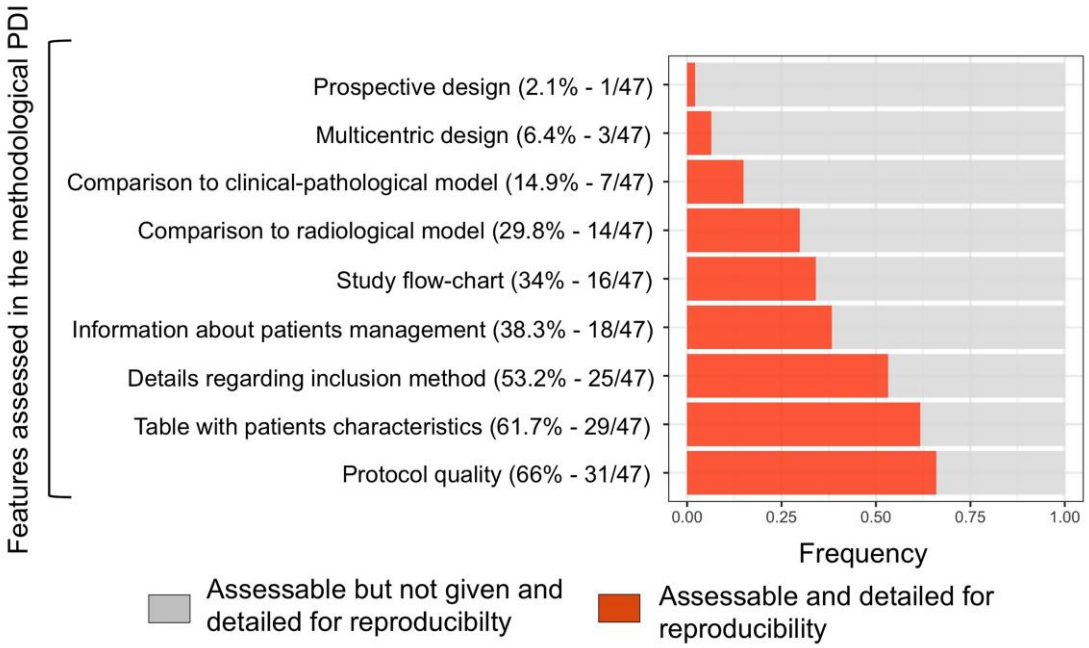
Figure 3. Proportions of studies providing the data necessary to the reproducibility of results relative to statistical analysis. Abbreviations: PDI: proportion of detailed items; RF: radiomics features.



Quality of the method

Figure 4 provides the proportion of studies that detailed each methodological item. Three out of the 47 (6.4%) were multi-centric. The flow-chart (written or in a figure was missing in 31/47 (66%) studies. The depiction of the patients' epidemiological and clinical characteristics from each cohort was lacking in 18/47 (38.3%) studies. The inclusion method was not given in 22/47 (46.8%) studies – and when given, consisted in case-control in 1/47 (2.1%) study and consecutive inclusion in 24/47 (46.8%) studies. The diagnostic and therapeutic managements of patients after imaging were provided in 18/47 (38.3%) studies and proved to be homogeneous in only 11 of them. A comparison with best clinical-pathological models was performed in 7/47 (14.9%) studies and with best radiological models in 14/47 (29.8%). The median method PDI was 33.3% (range: 0 – 77.8).

Figure 4. Proportions of studies providing information relative to the general method.
Abbreviations: PDI: proportion of detailed items

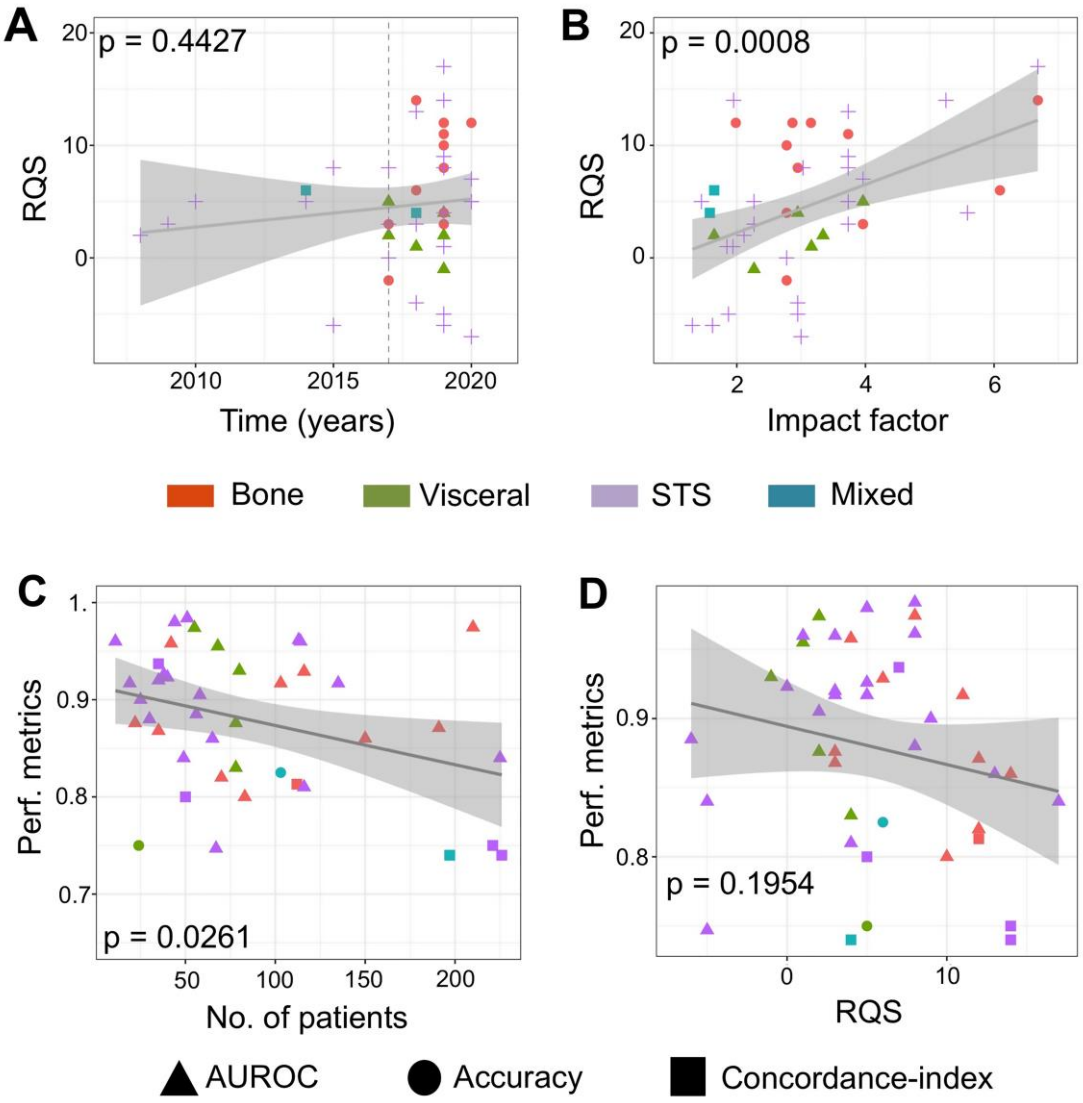


Correlations between performances metrics, quality of the radiomics studies and time (Figure 5)

There was a significant positive correlation between RQS and impact factor of journals (adjusted-R^2 = 0.205, p = 0.0008), and RQS and citescore (adjusted-R^2 = 0.138, p = 0.006) but not with h-index (p = 0.3). No correlation was found between time and RQS (p = 0.4), even when dichotomized as < 2017 and ≥ 2017 (: year of the publication of the RQS) (3. ± 3.7 versus 5.3 ± 6.5, p = 0.1). RQS was significantly higher when a co-author had academic degrees in statistics and/or data science (2.4 ± 6.1 versus 6.5 ± 5.1, p = 0.02). RQS positively correlated with the average of the 3 PDIs (adjusted-R^2 = 0.61, p < 0.0001).

The performance metrics of the studies were missing in 5/47 (10.6%) studies, and, when present, corresponded to accuracy in 2/47 (4.3%) studies, AUROC in 34/47 (72. %) studies and to c-index in 6/47 (12.8%) studies. In their best models, these values ranged from 0.74 to 0.98 (for a maximum of 1 in a perfect model). The performance metrics negatively correlated with the number of patients (adjusted-R^2 = 0.118, p = 0.03) but not with the RQS (p = 0.2) though a tendency towards a negative correlation could be seen.

Figure 5. Correlations between radiomics quality score (RQS), performance metrics, time and impact factor. (**A**) Correlation between RQS and time: the vertical dashed line corresponds to the year of publication of the RQS. (**B**) Correlation between RQS and impact factor. (**C**) Correlation between performance (perf.) metrics and number (No.) of patients. (**D**) Correlation between performance metrics and RQS. Each point corresponds to a study. The regression line and its 95% confidence interval are shown in grey with p-value.



DISCUSSION

The interest in radiomics in oncologic studies has exponentially increased since its first definition by Gillies *et al.* [39]. Yet, no clinical application has been developed so far. To overcome the translational gaps, international initiatives have proposed quality score dedicated to radiomics and reference standards for computing radiomics features, which are complementary to prior methodological guidelines, such as

TRIPOD or QUADAS-2. Our present study focused on sarcoma, a model for tumor heterogeneity, to draw an inventory of the research in radiomics, to illustrate its numerous potential applications and the efforts to make to put it into clinical practice. Overall, our results show that RQS of sarcoma studies are low on average, with several missing items limiting their reproducibility.

Research in radiomics is particularly attractive because it enables to explore new quantitative biomarkers by combining trendy multidisciplinary domains including computer science, data science and the medical '-omics' without needing investment in new imaging systems and tracers. However, each step of radiomics study can influence its results, as highlighted by several methodological reports focusing of imaging systems, image acquisition parameters, image correction, SIs harmonization (for MRI), voxel size, number of gray levels, or software... Herein, we do not mean that one method to tackle one of these steps is better than the others, but that this method should be given and detailed in order to ensure the reproducibility of the study and its comparability with other studies.

Some studies have already investigated the overall quality of radiomics studies with RQS in other fields. Regarding studies looking at correlations with tumor biology, the average RQS was of less than 50% [40]. In 51 neuro-oncologic studies, the median RQS was 11 [41]. Its median was 4.5 for studies involving renal cell carcinomas [42]. Our median RQS (: 4) proved to be lower than those obtained in these studies. This could be explained by our slightly different inclusion criteria. We choose not exclude studies because of a low impact factor of the journal in which they were published, while Park *et al.*, and Ursprung *et al* did [41–43]. Indeed, we wanted to have a realistic view. Moreover, it should be noted that research in sarcoma generally does not reach the same audience than research in uro-oncology and neuro-oncology. Nevertheless, the RQS items with the lowest notes were globally the same. Few studies performed multiple segmentations to assess the stability of RFs. None presented test-retest analysis because of their retrospectivity as well as organizational and ethical difficulties to perform a contrast-enhanced examination twice. None showed a phantom analysis either maybe because of the lack of consensual radiomics phantom, validated by international imaging societies. Calibration statistics were frequently missing, which may be due to the lack of knowledge regarding these analyses in the radiological community. Cost-effectiveness was almost never at least

discussed. Non-RFs variables and comparisons to gold standard were also lacking in more than a third of studies, although several clinical-pathological models (‘sarculator’), molecular signature (i.e. CINSARC) or radiological models have shown good performances [4,44–46]. For instance, two studies have highlighted that combining visual heterogeneity on T2 weighted-imaging (-WI), peritumoral enhancement on contrast-enhanced T1-WI and estimation of the amount of necrosis was strongly correlated to histological grade, overall survival and metastasis-free survivals, but none of the radiomics studies aiming at finding signature for grading and/or survivals compared its best model with a model based on these conventional radiological variables, or considered combining RFs with these non-RFs variables to improve their results [48,49]. Finally, only 2 complementary studies, based on public datasets provided open-source codes [20,21].

Analyzing the RQS items with a low inter-observer reproducibility (with one observer having experience in imaging and data analysis and the other one only in imaging) underlines lacks of clarity in their definition. For instance, protocol quality had only a moderate κ because what defines well-defined protocol for each imaging modality, or DWI and DCE-MRI is missing. The κ for cutoff analysis was only fair because this analysis may be implicit (for instance dichotomizing predictions with a cutoff of 0.5, median, tertiles, keeping continuous risk scores) and, when explicit, may be the opposite of the meaning of the given definition (for instance: looking for optimal cutoff, at risk of overfitting). Finally, the definition for potential clinical utility appears unclear, subjective and qualitative resulting in a fair κ.

In addition to previous studies, we tried to deepen the 3 main aspects of a radiomics study through PDIs for image processing, RFs extraction, statistical learning and general methods. Although partly correlated with RQS, this complementary approach highlights which information was the most frequently missing, that-is-to-say in ascending order: processing, statistics and methods. Furthermore, we investigated the relationships between the RQS and original variables. First, we did not find a significant correlation between RQS and time, though the average RQS tended to increase with time, but it should be noted that the spread of the RQS strongly increased after 2017 with more publications with higher RQS. Second, as expected, publications with higher RQS were published in journal with higher impact factor, although high RQS were also found among journal with low impact factor – reflecting the limits of this scientometrics, especially for rare diseases. Third, though not

significant, the performance metrics tended to decrease with higher RQS, which could be explained by over-optimistic and overfitted results in study with low quality. In the same way, we found a significant negative relationship between performance metrics and number of included patients. It should be noted that men remained over-represented as first or last authors with similar proportions as reported proportions in non-interventional radiology journals [50]. Finally, no study explicitly reported negative results, as already commented by Buvat and Orlhac, which questions the number of false-positive findings and the need for reconsider designs, controls and reviewing of radiomics studies in order to reach clinical applications instead of accumulating proofs of concept [51].

Our study has limits. The number of studies was relatively small but it was already sufficient to highlight the needs for improving the quality of radiomics approaches. We included studies with various first objectives and subgroups of sarcoma with a rather small number of studies per objective and subgroups. This prevented us to perform an exhaustive quantitative meta-analysis. We did not exhaustively investigate all the items of TRIPOD and IBSI but we tried to focus on some of them among the PDIs analyses.

To conclude, radiomics may improve current clinical-pathological and molecular predictive models involving sarcomas. However, our study stresses the urgent need for improving the quality and reproducibility of radiomics studies. Increasing education about the particular radiomics' methodology and quality metrics – in addition to multidisciplinarity and multi-centric prospective collaborations - are necessary to fasten clinical applications.

REFERENCES

[1] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 14 (2017) 749–62.

[2] Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuzé S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Ann. Oncol. 28 (2017) 1191–206.

[3] Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F (2013) WHO classification of tumours of soft tissue and bone, vol 5, 4th edn. IARC Press, Lyon

[4] Chibon F, Lagarde P, Salas S, Pérot G, Brouste V, Tirode F, et al. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. Nat. Med. 16 (2010) 781–7.
[5] Lucchesi C, Khalifa E, Laizet Y, Soubeyran I, Mathoulin-Pelissier S, Chomienne C, et al. Targetable Alterations in Adult Patients With Soft-Tissue Sarcomas: Insights for Personalized Therapy. JAMA Oncol. 4 (2018) 1398–404.
[6] Malinauskaite I, Hofmeister J, Burgermeister S, Neroladaki A, Hamard M, Montet X, et al. Radiomics and Machine Learning Differentiate Soft-Tissue Lipoma and Liposarcoma Better than Musculoskeletal Radiologists. Sarcoma (2020) 7163453.
[7] Vos M, Starmans MPA, Timbergen MJM, van der Voort SR, Padmos GA, Kessels W, et al. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. Br. J. Surg.106 (2019) 1800–9.
[8] Thornhill RE, Golfam M, Sheikh A, Cron GO, White EA, Werier J, et al. Differentiation of lipoma from liposarcoma on MRI using texture and shape analysis. Acad. Radiol. 21 (2014) 1185–94.
[9] Kim HS, Kim J-H, Yoon YC, Choe BK. Tumor spatial heterogeneity in myxoid-containing soft tissue using texture analysis of diffusion-weighted MRI. PLoS ONE 12: (2017) e0181339.
[10] Martin-Carreras T, Li H, Cooper K, Fan Y, Sebro R. Radiomic features from MRI distinguish myxomas from myxofibrosarcomas. BMC Med Imaging 2019;19:67.
[11] Mayerhoefer ME, Breitenseher M, Amann G, Dominkus M. Are signal intensity and homogeneity useful parameters for distinguishing between benign and malignant soft tissue masses on MR images? Objective evaluation by means of texture analysis. Magn. Reson. Imaging 26 (2008) 1316–22.
[12] Wang H, Chen H, Duan S, Hao D, Liu J. Radiomics and Machine Learning With Multiparametric Preoperative MRI May Accurately Predict the Histopathological Grades of Soft Tissue Sarcomas. J. Magn. Reson. Imaging 51 (2020) 791–7.
[13] Juntu J, Sijbers J, De Backer S, Rajan J, Van Dyck D. Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. J. Magn. Reson. Imaging 31 (2010) 680–9.
[14] Chen C-Y, Chiou H-J, Chou S-Y, Chiou S-Y, Wang H-K, Chou Y-H, et al. Computer-aided diagnosis of soft-tissue tumors using sonographic morphologic and texture features. Acad Radiol 16 (2009) 1531–8.
[15] Peeken JC, Bernhofer M, Spraker MB, Pfeiffer D, Devecka M, Thamer A, et al. CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. Radiother. Oncol. 135 (2019) 187–96.
[16] Zhang Y, Zhu Y, Shi X, Tao J, Cui J, Dai Y, et al. Soft Tissue Sarcomas: Preoperative Predictive Histopathological Grading Based on Radiomics of MRI. Acad. Radiol. 26 (2019) 1262–8.
[17] Xiang P, Zhang X, Liu D, Wang C, Ding L, Wang F, et al. Distinguishing soft tissue sarcomas of different histologic grades based on quantitative MR assessment of intratumoral heterogeneity. Eur. J. Radiol. 118 (2019) 194–9.
[18] Corino VDA, Montin E, Messina A, Casali PG, Gronchi A, Marchianò A, et al. Radiomic analysis of soft tissues sarcomas can distinguish intermediate from high-grade lesions. J. Magn. Reson. Imaging 47 (2018) 829–40.
[19] Peeken JC, Spraker MB, Knebel C, Dapper H, Pfeiffer D, Devecka M, et al. Tumor grading of soft tissue sarcomas using MRI-based radiomics. EBioMedicine 48 (2019) 332–40.
[20] Vallières M, Freeman CR, Skamene SR, 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 (2015) 5471–96.
[21] 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. 62 (2017) 8536–65.

[22] Crombé A, Le Loarer F, Sitbon M, Italiano A, Stoeckle E, Buy X, et al. Can radiomics improve the prediction of metastatic relapse of myxoid/round cell liposarcomas? Eur. Radiol. 30 (2020) 2413–24.
[23] Crombé A, Fadli D, Buy X, Italiano A, Saut O, Kind M. High-Grade Soft-Tissue Sarcomas: Can Optimizing Dynamic Contrast-Enhanced MRI Postprocessing Improve Prognostic Radiomics Models? J. Magn. Reson. Imaging (2020) https://doi.org/10.1002/jmri.27040.
[24] Hayano K, Tian F, Kambadakone AR, Yoon SS, Duda DG, Ganeshan B, et al. Texture Analysis of Non-Contrast-Enhanced Computed Tomography for Assessing Angiogenesis and Survival of Soft Tissue Sarcoma. J. Comput. Assist. Tomogr. 39 (2015) 607–12.
[25] Spraker MB, Wootton LS, Hippe DS, Ball KC, Peeken JC, Macomber MW, et al. MRI Radiomic Features Are Independently Associated With Overall Survival in Soft Tissue Sarcoma. Adv. Radiat. Oncol. 4 (2019) 413–21.
[26] Crombé A, Saut O, Guigui J, Italiano A, Buy X, Kind M. Influence of temporal parameters of DCE-MRI on the quantification of heterogeneity in tumor vascularization. J. Magn. Reson. Imaging 50 (2019) 1773–88.
[27] Crombé A, Périer C, Kind M, De Senneville BD, Le Loarer F, Italiano A, et al. T2 - based MRI Delta-radiomics improve response prediction in soft-tissue sarcomas treated by neoadjuvant chemotherapy. J. Magn. Reson. Imaging 50 (2019) 497–510.
[28] Esser M, Kloth C, Thaiss WM, Reinert CP, Kraus MS, Gast GC, et al. CT-morphologic and CT-textural patterns of response in inoperable soft tissue sarcomas treated with pazopanib-a preliminary retrospective cohort study. Br J Radiol. 92 (2019) 20190158.
[29] Esser M, Kloth C, Thaiss WM, Reinert CP, Fritz J, Kopp H-G, et al. CT-response patterns and the role of CT-textural features in inoperable abdominal/retroperitoneal soft tissue sarcomas treated with trabectedin. Eur. J. Radiol. 107 (2018) 175–82.
[30] Tian F, Hayano K, Kambadakone AR, Sahani DV. Response assessment to neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion. Abdom. Imaging 40 (2015) 1705–12.
[31] Burke HB. Independent imaging biomarkers do not exist. Nat. Rev. Clin. Oncol. 14 (2017) 452.
[32] O'Connor JPB, Aboagye EO, Adams JE, Aerts HJWL, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat. Rev. Clin. Oncol .14 (2017)169–86.
[33] Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155 (2011) 529–36.
[34] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med.162 (2015) W1-73.
[35] Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J. Nucl. Med. 59 (2018) 189–93.
[36] Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur. J. Nucl. Med. Mol. Imaging 46 (2019) 2638–55.
[37] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology (2020):191145.
[38] Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software 2019;4:1686.
[39] Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. Clin. Radiol. 65 (2010) 517–21.

[40] Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother. Oncol. 127 (2018) 349–60.
[41] Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, et al. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer (2020) 20-29.
[42] Ursprung S, Beer L, Bruining A, Woitek R, Stewart GD, Gallagher FA, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. Eur. Radiol. (2020). https://doi.org/10.1007/s00330-020-06666-3.
[43] Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur. Radiol. 30 (2020) 523–36.
[44] Coindre JM, Terrier P, Bui NB, Bonichon F, Collin F, Le Doussal V, et al. Prognostic factors in adult patients with locally controlled soft tissue sarcoma. A study of 546 patients from the French Federation of Cancer Centers Sarcoma Group. J. Clin. Oncol. 14 (1996) 869–77.
[45] Pasquali S, Colombo C, Pizzamiglio S, Verderio P, Callegaro D, Stacchiotti S, et al. High-risk soft tissue sarcomas treated with perioperative chemotherapy: Improving prognostic classification in a randomised clinical trial. Eur. J. Cancer 93 (2018) 28–36.
[46] Pasquali S, Pizzamiglio S, Touati N, Litiere S, Marreaud S, Kasper B, et al. The impact of chemotherapy on survival of patients with extremity and trunk wall soft tissue sarcoma: revisiting the results of the EORTC-STBSG 62931 randomised trial. Eur. J. Cancer 109 (2019) 51–60.
[48] .Zhao F, Ahlawat S, Farahani SJ, Weber KL, et al. Can MR Imaging Be Used to Predict Tumor Grade in Soft-Tissue Sarcoma? Radiology 272 (2014) 192-201.
[49] Crombé A, Marcellin PJ, Buy X, Stoeckle E, Brouste V, et al. Soft-Tissue Sarcomas: Assessment of MRI Features Correlating With Histologic Grade and Patient Outcome. Radiology 291 (2019) 710-721.
[50] Bernard C, Pommier R, Vilgrain V, Ronot M. Gender gap in articles published in European Radiology and CardioVascular and Interventional Radiology: evolution between 2002 and 2016. Eur. Radiol. 30 (2020) 1011–9.
[51] Buvat I, Orlhac F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. J. Nucl. Med. 60 (2019) 1543–4.