Whole-body / total-body biomarkers in PET imaging

Louis Rebaud

université PARIS-SACLAY

Whole-body / total-body biomarkers in PET imaging

Biomarqueurs corps entier en imagerie par Tomographie d'Emission de Positons (TEP)

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°575 : electrical, optical, bio : physics and engineering (EOBE)
Spécialité de doctorat : Physique et imagerie médicale
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire d'Imagerie Translationnelle en Oncologie** (Institut Curie, Inserm), sous la direction de **Irène BUVAT**, Directrice de recherche, et la co-supervision de **Bruce SPOTTISWOODE**, PhD, Director

Thèse soutenue à Paris-Saclay, le 07 mai 2024, par

Louis REBAUD

Composition du Jury

Membres du jury avec voix délibérative

Adrien DEPEURSINGE Professeur des universités, HES-SO Valais-Wallis	Président
Sally BARRINGTON Professeure des universités, Médecin nucléaire, King's College London	Rapporteur & Examinatrice
Ronald BOELLAARD Professeur des universités, University Medical Center Groningen, VU University Medical Center Amsterdam	Rapporteur & Examinateur
Anne-Ségolène COTTEREAU Médecin nucléaire, PhD, Hôpital Cochin APHP	Examinatrice

THESE DE DOCTORAT

NNT : 2024UPAST047

université PARIS-SACLAY | DOCTORAL SCHOOL Electrical, Optical, Bio: Physics and Engineering (EOBE)

Titre : Biomarqueurs corps entier en imagerie par Tomographie d'Emission de Positons (TEP)

Mots clés : Radiomique, Imagerie moléculaire, Intelligence artificielle, Oncologie, Imagerie médicale, Apprentissage automatique

Résumé : Cette thèse, réalisée en partenariat avec l'Institut Curie et Siemens Healthineers, explore l'utilisation de l'imagerie par tomographie par émission de positrons (TEP) pour le pronostic du cancer, en se concentrant sur les lymphomes non hodgkiniens, en particulier le lymphome folliculaire (FL) et le lymphome diffus à grandes cellules B (DLBCL). Partant de l'hypothèse que les biomarqueurs actuels calculés dans les images TEP sous-utilisent leur richesse en informations, ce travail se concentre sur la recherche de nouveaux biomarqueurs en imagerie TEP corps entier. Une première approche manuelle a permis de valider une caractéristique précédemment identifiée (fragmentation de la tumeur) et d'explorer l'importance pronostique de l'atteinte splénique dans les DLBCL, en constatant que le volume de l'atteinte splénique ne permet pas de stratifier davantage les patients présentant une telle atteinte. Pour dépasser les limites empiriques de la recherche manuelle, une méthode d'identification semi-automatique des caractéristiques a été mise au point. Elle consiste à extraire automatiquement des milliers de biomarqueurs candidats et à les tester à l'aide d'un pipeline de sélection conçu pour trouver des caractéristiques quantifiant de nouvelles informations pronostiques.

Les biomarqueurs sélectionnés ont ensuite été analysés et recodés de manière plus simple et plus intuitive. Cette approche a permis d'identifier 22 nouveaux biomarqueurs basés sur l'image, qui reflètent des informations biologiques sur les tumeurs, mais aussi l'état de santé général du patient. Parmi eux, 10 caractéristiques se sont avérées pronostiques à la fois pour les patients atteints de FL que pour ceux souffrant de DLBCL. La thèse aborde également le défi que représente l'utilisation de ces caractéristiques dans la pratique clinique, en proposant le modèle ICARE (Individual Coefficient Approximation for Risk Estimation). Ce modèle d'apprentissage automatique, conçu pour réduire le surapprentissage et améliorer la généralisation, a démontré son efficacité dans le cadre du challenge HECKTOR 2022 visant à prédire le risque de rechute de patients atteints de cancer des voies aérodigestives supérieures à partir de leurs images TEP. Ce modèle s'est également avéré plus résistant au surapprentissage que d'autres méthodes d'apprentissage automatique lors d'une comparaison exhaustive sur un benchmark de 71 jeux de données médicales. Ces développements ont été implémentés dans une extension logicielle d'un prototype développé par Siemens Healthineers.

Title : Whole-body / total-body biomarkers in PET imaging

Keywords : Radiomics, Molecular Imaging, Artificial Intelligence, Oncology, Medical imaging, Machine Learning

Abstract : This thesis in partnership with Institut Curie and Siemens Healthineers explores the use of Positron Emission Tomography (PET) for cancer prognosis, focusing on non-Hodgkin lymphomas, especially follicular lymphoma (FL) and diffuse large B cell lymphoma (DLBCL). Assuming that current biomarkers computed in PET images overlook significant information, this work focuses on the search for new biomarkers in whole-body PET imaging. An initial manual approach validated a previously identified feature (tumor fragmentation) and explored the prognostic significance of splenic involvement in DLBCL, finding that the volume of splenic involvement does not further stratify patients with such an involvement. To overcome the empirical limitations of the manual search, a semi-automatic feature identification method was developed.
It consisted in the automatic extraction of thousands of candidate biomarkers and there subsequent testing by a selection pipeline design to identify features quantifying new prognostic information.

The selected biomarkers were then analysed and re-encoded in simpler and more intuitive ways. Using this approach, 22 new image-based biomarkers were identified, reflecting biological information about the tumours, but also the overall health status of the patient. Among them, 10 features were found prognostic of both FL and DLBCL patient outcome. The thesis also addresses the challenge of using these features in clinical practice, proposing the Individual Coefficient Approximation for Risk Estimation (ICARE) model. This machine learning model, designed to reduce overfitting and improve generalizability, demonstrated effectiveness in the HECKTOR 2022 challenge for predicting outcomes from head and neck cancer patients [18F]-PET/CT scans. This model was also found to overfit less than other machine learning methods on an exhaustive comparison using a benchmark of 71 medical datasets. All these developments were implemented in a software extension of a prototype developed by Siemens Healthineers.

Acknowledgments

I want to begin by expressing my heartfelt gratitude to my supervisor, Irène Buvat. Her guidance, trust, and the freedom she provided during my PhD were invaluable. Her enthusiasm, rigor, curiosity, energy, and remarkable ability to always ask the right questions truly inspired me. I learned a lot working with her.

I would like to express my sincere gratitude to all the individuals at Siemens Healthineers with whom I had the privilege to work. Special appreciation goes to Bruce Spottiswoode for his unwavering support, insightful guidance, and kindness throughout this journey. Particular thanks are extended to Nicolò Capobianco and Ludovic Sibille for their regular engagement, valuable advice, and wisdom. Their time and insights have been immensely beneficial and greatly appreciated, and I have learned a lot from our discussions. I am also profoundly grateful to the rest of the team for their warm welcome when I arrived in Knoxville. Their hospitality made my experience truly enriching. Lastly, I would like to acknowledge Matthieu Lepetit-Coiffe for ensuring that all necessary resources were provided for the successful completion of this project. His assistance and availability were crucial and greatly appreciated.

I also thank Pr. Adrien Depeursinge et Pr. John Prior for accepting to be members of my follow up committee and making sure that everything was on track during my PhD.

I extend my warm gratitude to all the members of the LITO lab for their invaluable help, advice, and support. Your collective experience and the countless debates and discussions we shared have been a source of knowledge and inspiration, sparking numerous new ideas and profoundly enriching my research journey.

I acknowledge Dr. Clémentine Sarkozy, Dr. Anne-Ségolène Cottereau and Pr. Franck Morschhauser, who helped me tremendously by bringing their knowledge and experience to analyze my results and give me the medical background that I was lacking, being a computer scientist by trade.

I would also like to give a special thanks to all the peoples who helped collect, prepare, and provide the data I have been working with. Pr. Michel Meignan, Dr. Anne-Ségolène Cottereau, Pr. Franck Morschhauser, Pr. Laetitia Vercellino, Dr. Olivier Casasnovas, Pr. Catherine Thieblemont, Loic Chartier and Cédric Portugues and the LYSARC. Large quantity of high-quality data is a key factor in the success of such studies, and without this colossal work nothing would have been possible.

Je voudrais remercier mes amis pour m'avoir aidé à penser à autre chose qu'à ma thèse. Vous avez su me faire oublier la pression et m'apporter une véritable bouffée d'air frais. Je suis profondément reconnaissant de vous avoir à mes côtés et j'attends avec impatience tous les moments que nous partagerons encore.

Je remercie chaleureusement toute ma famille, pour leur amour, leur soutien, leur présence et leurs conseils, et pour m'avoir aidé à devenir qui je suis. J'étends ces remerciements à ma belle-famille qui, bien que présente depuis moins longtemps, m'inspire, me construit, et me soutient tout autant.

Je remercie tout particulièrement mes parents et mon frère pour tout ce qu'ils m'ont apporté et pour leur soutien inconditionnel. Merci de m'avoir emmené jusqu'ici, vous avez été ma force et mon inspiration. Cette thèse est aussi la vôtre. Plus spécifiquement, je remercie mon père pour m'avoir donné le gout de la science, ma mère, celui de l'effort et du travail bien fait, et mon frère, celui d'aller faire autre chose que travailler.

Enfin, je terminerai en remerciant Lola, mon soleil. Pour tout ce qu'elle apporte dans ma vie et celle des autres. Ta présence et ta bienveillance ont été des piliers dans cette aventure. Merci de rendre chaque jour plus lumineux et chaque moment plus précieux.

Synthèse en français

Le cancer est une des principales causes de mortalité dans le monde, étant la première cause de décès prématuré dans 57 pays en 2019. Cette maladie complexe, caractérisée par une croissance cellulaire anormale et incontrôlée, comprend aujourd'hui plus de 200 types de cancer. En 2018, il a été estimé que 18 millions nouveaux cas de cancer ont été détectés, entraînant 9 millions de décès. L'augmentation prévue du nombre de cas de cancer est notamment liée au vieillissement de la population mondiale.

Heureusement, la médecine a réalisé des progrès considérables au cours du dernier siècle. Par exemple, le taux de survie à cinq ans pour les patients atteints de cancer du sein est passé de 40% il y a cent ans à plus de 90% aujourd'hui, grâce aux traitements modernes. Des résultats encore plus impressionnants ont été obtenus dans le domaine des cancers pédiatriques, avec un taux de survie passant de 10% à près de 80%.

Ces progrès sont dus à une meilleure compréhension de la maladie et au développement d'une large gamme de stratégies thérapeutiques, comme la chirurgie, la chimiothérapie, la radiothérapie, l'immunothérapie, la thérapie ciblée et la thérapie hormonale. Le traitement efficace du cancer ne réside pas seulement dans la disponibilité de ces options, mais dans leur application personnalisée à chaque patient, nécessitant une compréhension fine de la maladie pour optimiser les résultats.

L'imagerie médicale, et notamment la combinaison de la Tomographie par Émission de Positons (TEP) avec la Tomodensitométrie (TDM) ou l'Imagerie par Résonance Magnétique (IRM), jouent un rôle central dans le diagnostic et le traitement du cancer. Cependant, le potentiel de l'imagerie TEP est probablement sous-exploité. En effet, les approches traditionnelles d'interprétation des images TEP ont tendance à simplifier l'information présente dans l'image, se concentrant sur les tumeurs primaires ou un sous-ensemble limité de lésions, et négligeant souvent les autres informations reflétées par l'image, y compris dans les tissus apparemment sains. Cette lacune dans l'utilisation des données d'imagerie TEP que cette thèse vise à combler en identifiant de nouvelles informations pronostiques, est une nouvelle opportunité pour améliorer la gestion des cancers.

La première approche que j'ai explorée était l'identification et la construction de caractéristiques construites à la main, basées sur l'analyse visuelle des images et l'aide de la littérature scientifique. Bien que cette approche ait été insuffisante pour la découverte de nouvelles caractéristiques, elle a aidé à valider une caractéristique déjà identifiée (la fragmentation tumorale estimée par le rapport volume/surface) par Decazes et al. (2018) et nous a donné l'occasion de mieux comprendre l'impact de l'envahissement splénique pour les patients atteints de lymphomes diffus à grandes cellules B (DLBCL).

Les chapitres 2 à 4 de ce manuscrit introduisent les concepts nécessaires en imagerie médicale, en apprentissage automatique, et sur les lymphomes. La suite de cette synthèse résume chaque chapitre présentant les travaux originaux réalisés pendant cette thèse.

Chapitre 5 Étude du rôle de l'envahissement splénique dans le pronostic du DLBCL

Dans ce chapitre, nous avons exploré le rôle de l'envahissement splénique dans le pronostic des patients atteints de lymphome diffus à grandes cellules B (DLBCL). L'étude se concentre sur une analyse détaillée de plusieurs marqueurs, notamment le Volume Tumoral Métabolique Total (TMTV), l'envahissement splénique (SI), la taille de la rate, le Volume Tumoral Métabolique à l'Intérieur de la Rate (MTVIS), et le Volume Tumoral Métabolique à l'Extérieur de la Rate (MTVOS), sur une cohorte de 377 patients atteints de DLBCL. Les indicateurs évalués pour mesurer le devenir des patients étaient la survie sans progression (PFS) et la survie globale (OS).

Les résultats obtenus dans cette étude nous éclairent sur l'impact de l'envahissement splénique sur le pronostic des patients. Il a été constaté que les patients avec SI présentent une PFS et une OS significativement plus faibles ($p < 0.03$ pour la PFS et $p < 0.04$ pour l'OS) par rapport à ceux sans envahissement. De plus, ces patients avaient un TMTV nettement plus élevé ($p < 0.001$), ce qui souligne le lien entre l'envahissement splénique et la charge tumorale globale. La capacité de prédire l'envahissement splénique à partir du TMTV avec une précision moyenne de 0.62 ($p < 0.001$) indique une corrélation forte entre ces deux paramètres, et l'envahissement splénique n'offre pas d'informations pronostiques supplémentaires au-delà de celles fournies par le TMTV.

L'analyse du Volume Tumoral Métabolique à l'Intérieur de la Rate (MTVIS) a révélé que ce marqueur n'était pas prédictif du devenir du patient. Les patients avec un volume élevé de tumeur à l'intérieur de la rate ne présentaient pas un risque plus élevé comparé à ceux avec un volume faible, suggérant que le volume tumoral à l'intérieur de la rate n'influence pas le pronostic de manière significative.

Par ailleurs, le Volume Tumoral Métabolique à l'Extérieur de la Rate (MTVOS) s'est avéré aussi prédictif que le TMTV. Cette découverte suggère que l'essentiel de l'information pronostique liée au TMTV se situe en dehors de la rate.

La présence de splénomégalie, bien qu'associée à l'envahissement splénique, n'a pas fourni d'informations pronostiques additionnelles. Les patients avec une rate hypertrophiée en raison de l'envahissement splénique n'ont pas montré de différence significative en termes de risque par rapport à ceux avec un envahissement splénique sans augmentation du volume de la rate, ce qui souligne que la taille de la rate seule n'est pas un indicateur pronostique dans le DLBCL.

Ces résultats, bien qu'intéressants, nécessitent une validation dans d'autres cohortes de patients atteints de DLBCL pour être confirmés. De plus, l'exploration du rôle de l'envahissement splénique dans d'autres sous-types de lymphomes pourrait fournir des éclairages supplémentaires sur son impact sur le pronostic, étant donné que la pathophysiologie et la réponse au traitement peuvent varier considérablement entre les différents types de lymphomes.

En conclusion, cette étude met en lumière le fait que, bien que l'envahissement splénique soit un facteur pronostique significatif dans le DLBCL, sa valeur est fortement liée au TMTV et n'offre pas d'avantages supplémentaires pour la stratification des risques des patients au-delà de ce que le TMTV fournit déjà.

Chapitre 6 Développement d'un outil de sélection de biomarqueurs (ROBI)

Bien que des résultats intéressants aient été obtenus au moyen de cette recherche manuelle, je n'ai pas découvert de nouvelles informations pronostiques avec cette approche. Pour surmonter les limitations de l'approche, j'ai développé une méthode semi-automatique pour identifier de nouveaux biomarqueurs pronostiques. Les principaux avantages de la méthode proposée sont une exploration étendue de l'espace de recherche et un contrôle précis des fausses découvertes possiblement causées par des tests multiples. Cette recherche semi-automatique consiste en la construction automatisée d'un grand nombre de biomarqueurs candidats basés sur l'image, calculés dans les lésions, leur environnement immédiat et les organes segmentés. Les milliers de biomarqueurs candidats résultants sont ensuite analysés par un pipeline de sélection automatisé pour identifier les biomarqueurs codant de nouvelles informations pronostiques non déjà mesurées par les biomarqueurs actuels utilisés par les médecins en clinique. En raison du nombre important de candidats testés, ce pipeline doit être robuste aux faux positifs (FP, caractéristiques non pronostiques sélectionnés par hasard) et optimisé pour maximiser les découvertes.

Ce chapitre détaille le développement d'un outil de sélection de biomarqueurs, ROBI (Robust Biomarker Identifier), conçu pour identifier les biomarqueurs radiomiques basés sur l'image les plus susceptibles de refléter de nouvelles informations pronostiques tout en minimisant et contrôlant le nombre de faux positifs. L'outil combine plusieurs techniques de sélection de caractéristiques pour choisir des biomarqueurs qui codent des informations pertinentes non déjà quantifiées par des biomarqueurs établis et qui sont les plus susceptibles de prédire le devenir des patients dans l'ensemble de données utilisé pour la sélection. Le pipeline minimise la sélection de FP et estime leur nombre, avec une rigueur de sélection ajustable.

Pour valider cet outil, 500 ensembles de données synthétiques et des données rétrospectives issues de d'images TEP/TDM au 18F-FDG de 378 patients DLBCL ont été ana-

lysés. Sur ces données DLBCL, deux biomarqueurs radiomiques établis, le volume tumoral total (TMTV) et la plus grande distance entre deux lésions (Dmax), ont été mesurés à partir de la segmentation des images TEP/TDM au 18F-FDG, et 10 000 biomarqueurs aléatoires ont été générés. La sélection a été effectuée et vérifiée sur chaque ensemble de données, avec une signification statistique évaluée par des tests de Wilcoxon. L'efficacité de ROBI a été comparée à des méthodes contrôlant les tests multiples et un modèle de Cox avec pénalité Elasticnet.

Dans les ensembles de données synthétiques, le pipeline a sélectionné significativement plus de vrais positifs (TP) que de FP ($p < 0,001$). Pour 99,3 % des ensembles de données synthétiques, le nombre de FP était dans l'intervalle de confiance à 95 % estimé par le pipeline. Ce pipeline a significativement augmenté le nombre de TP par rapport aux méthodes habituelles de sélection de caractéristiques ($p < 0,001$). Dans l'ensemble de données réel, ROBI a sélectionné les deux biomarqueurs établis et un biomarqueur aléatoire, estimant à 95 % le risque de sélectionner 0 ou 1 FP et une probabilité de 0,0014 de sélectionner uniquement des FP. La correction de Bonferroni n'a sélectionné aucune caractéristique, tandis que l'Elasticnet a sélectionné 73 caractéristiques aléatoires et manqué l'un des deux biomarqueurs établis.

ROBI contrôle donc efficacement le nombre de faux positifs tout en augmentant le nombre de biomarqueurs sélectionnés par rapport à la méthode standard pour contrôler le taux de fausses découvertes. L'outil trouve des biomarqueurs pertinents parmi des milliers de candidats, tandis que d'autres méthodes standard échouent avec un si grand nombre de candidats potentiels. Il n'est pas un substitut à une validation externe mais identifie les candidats les plus prometteurs parmi un grand nombre. Le pipeline est agnostique au domaine d'application, ce qui le rend utile pour une large gamme de disciplines traitant des grands nombres de paramètres, telles que la génomique.

Une implémentation Python est disponible sur https ://github.com/Lrebaud/robi

Chapitre 7 Découverte de nouveaux biomarqueurs pronostiques pour les lymphomes non hodgkiniens

Dans ce chapitre, nous présentons l'identification de nouveaux biomarqueurs pronostiques pour les lymphomes non hodgkiniens (NHL) en exploitant les capacités de ROBI pour analyser de vastes ensembles de biomarqueurs candidats issus d'images TEP/TDM au 18F-FDG. Deux cohortes de patients atteints de NHL, l'une composée de patients avec un Lymphome Diffus à Grandes Cellules B (DLBCL) et l'autre de patients présentant un Lymphome Folliculaire (FL), ont servi à cette exploration. Ces cohortes, issues de trois essais cliniques et comptant environ 350 patients chacune, ont fourni la puissance statistique nécessaire pour tester des milliers de biomarqueurs candidats au moyen de notre pipeline ROBI.

En analysant ces images TEP/TDM réalisées avant le traitement, nous avons automatiquement construit et testé un grand nombre de caractéristiques issues des images. ROBI a ainsi sélectionné 28 nouveaux biomarqueurs pronostiques extraits des images TEP/TDM pour les patients FL et 28 autres pour les patients DLBCL. En analysant ces biomarqueurs, nous avons manuellement identifié 22 informations biologiques intuitives et pronostiques que nous avons ré-encodées en 22 biomarqueurs toujours calculés dans l'image, mais plus simples. Parmi ces derniers, 10 se sont révélés pronostiques à la fois pour les patients atteints de DLBCL et de FL, suggérant une applicabilité clinique étendue et de meilleures chances de reproduction des résultats. Parmi ceux-ci, on trouve des biomarqueurs qui se concentrent sur les lésions, tel que le nombre de lésions, l'envahissement de la trachée ou la présence de lésions occultes (petites lésions avec une faible activité métabolique). D'autres en revanche concernent plutôt l'état de santé global du patient. Ainsi, un faible volume de graisse sous-cutanée ou une plus forte densité des tissus bronchiques étaient associés à un risque plus élevé. Les informations d'activité métabolique mesurée par la TEP, de densité de tissu par la TDM et de forme définie par la segmentation sont toutes représentées par au moins un de ces nouveaux biomarqueurs.

Ces découvertes, prometteuses pour l'amélioration de l'évaluation pronostique des patients atteints de DLBCL et de FL, nécessitent une validation externe pour confirmer leur utilité et applicabilité en pratique clinique. Le code pour le calcul et le test de ces biomarqueurs a été rendu disponible ici :
https ://github.com/Lrebaud/exhaustive_radiomics

Chapitre 8 Développement d'un nouveau modèle d'apprentissage automatique (ICARE) lors d'une compétition (HECKTOR)

Dans le chapitre précédent, nous avons identifié des dizaines de nouveaux biomarqueurs radiomiques présentant une valeur pronostique. Si la valeur pronostique de certains d'entre eux est confirmé par d'autres équipes, se posera alors la question de la meilleure façon de les utiliser en clinique. Dans ce chapitre, nous avons abordé le défi rencontré par les oncologues face à l'abondance de biomarqueurs pronostiques issus de diverses modalités (examens cliniques, biopsies, analyses sanguines, génétiques, etc.). Pour simplifier cette complexité, les modèles agrègent souvent de multiples caractéristiques en un seul score de risque, comme l'IPI et AnnArbor chez les patients atteints de lymphomes, élaborés à travers des concertations de spécialistes et de nombreuses analyses. Une approche alternative utilise des modèles d'apprentissage automatique pour apprendre à partir des données un score (signature) basé sur les valeurs des caractéristiques et les résultats observés chez de nombreux patients. Typiquement, un modèle de Cox est entraîné pour prédire le risque des patients à partir d'un ensemble de caractéristiques, et ce modèle peut être transformé en un nomogramme pour une utilisation et un déploiement facile.

Cependant, définir quelle caractéristique est plus pronostique qu'une autre s'avère extrêmement difficile en raison du bruit affectant les valeurs des caractéristiques et la valeur à prédire. Par exemple, la survie globale, fréquemment utilisée est très bruitée par nature, et souvent censurée. Nous avons alors développé l'idée qu'il pourrait être préférable de ne pas apprendre de poids spécifique pour chaque caractéristique, mais simplement un signe, permettant ainsi à chaque caractéristique de contribuer équitablement à la prédiction. Cette intuition a conduit au développement du modèle ICARE, détaillé dans ce chapitre.

Nous avons testé cette idée lors du challenge HECKTOR présenté durant la conférence MICCAI 2022, où différentes équipes du monde entier devaient concevoir un modèle pour segmenter automatiquement les tumeurs et les ganglions lymphatiques envahis sur les images TEP/TDM au 18F–FDG de patients atteints de cancer de la tête et du cou provenant de plusieurs hôpitaux. Pour cette tâche, nous avons utilisé un nnUNet simple et nous nous sommes classés 4e parmi les 36 équipes participantes. La deuxième tâche de ce challenge consistait à entrainer un modèle à prédire le risque de rechute des patients. Pour cette tâche, nous avons utilisé le modèle ICARE et nous avons été classés 1er parmi les 18 équipes qui ont participé.

Notre classement pour la tâche de prédiction de survie renforce l'idée qu'une stratégie d'apprentissage minimaliste est adaptée au contexte de la prédiction de survie, confirmant notre hypothèse : dans certaines situations, il est préférable de ne pas apprendre de poids. Une propriété intéressante de ICARE découverte pendant le défi est sa capacité à gérer un grand nombre de caractéristiques, semblant échapper au fléau de la dimension. Nous suggérons qu'une analogie avec la sagesse des foules pourrait expliquer ce phénomène, où l'erreur de chaque caractéristique est annulée par l'erreur des autres, menant à une bonne estimation du risque.

Cependant, ne pas apprendre de poids pourrait être sous-optimal dans des scénarios disposant de données suffisantes ou avec un bruit limité, où un modèle attribuant un poids plus important aux caractéristiques les plus prédictives serait plus efficace que ICARE. Nous avons donc cherché à déterminer dans quelles situations il est préférable d'utiliser le modèle ICARE.

Chapitre 9 Comparaison d'ICARE à d'autres modèles d'apprentissage automatique

Dans ce chapitre, nous avons comparé le modèle ICARE à d'autres modèles d'apprentissage automatique pour comprendre dans quelles conditions il est préférable d'utiliser le modèle ICARE, qui n'attribue pas de poids aux caractéristiques, plutôt qu'un modèle d'apprentissage automatique traditionnel. Pour cela, 71 ensembles de données médicales réelles issus de deux collections, SurvSet et TCGA, ont été collectés. Ces grands ensembles de données diversifiés permettent une comparaison exhaustive

d'ICARE avec d'autres modèles d'apprentissage automatique. Les données comprenaient plusieurs caractéristiques et une cible censurée à prédire (par exemple, la prédiction de survie). Neuf modèles ont été évalués sur ces ensembles de données, et les hyperparamètres de chacun des modèles ont été optimisés, afin de rendre la comparaison plus équitable et plus proche de leur utilisation réelle.

L'évaluation a révélé une performance relativement uniforme des modèles à travers la majorité des ensembles de données. En particulier, dans la moitié des ensembles de données analysés, la différence de score entre les modèles les plus et les moins performants était faible, à l'exception du modèle d'arbre de décision qui avait des performances systématiquement inférieures. De la même façon, l'optimisation des hyperparamètres et de la sélection et le prétraitement des caractéristiques n'ont pas apporté d'améliorations significatives, mettant en lumière un impact limité du choix du modèle et de son optimisation sur la performance prédictive globale.

ICARE, en apprenant de manière univariée seulement le signe de chaque caractéristique au lieu d'un poids, a démontré des performances similaires aux autres modèles dans la plupart des ensembles de données tout en présentant moins de surapprentissage, en particulier dans les ensembles de données de haute dimension. Ces résultats suggèrent que l'utilisation du modèle ICARE, réduisant le surapprentissage, pourrait améliorer la généralisation des modèles entre centres.

Les résultats de l'étude invitent également à reconsidérer l'accent communément mis sur les tests et l'optimisation extensifs des modèles. Ils suggèrent qu'il pourrait être avantageux d'utiliser des modèles plus simples comme ICARE et de se focaliser en parallèle sur la recherche de nouvelles informations biologiques plutôt que de passer beaucoup de temps à optimiser la combinaison d'informations préalablement identifiées comme pronostiques.

Content

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

Cancer is a leading cause of mortality worldwide and was the first cause of premature death in 57 countries in 2019 [1]. This multifaceted disease, characterized by abnormal and uncontrolled cell growth, stands out by its diversity and complexity, as we know today more than 200 types of cancer. It was estimated that 18 million new cancer cases were identified, and 9 million fatalities were due to cancer in 2018 [2]. As global population is aging, it is expected that the overall number of cancer cases will rise in the future [3].

Fortunately, medicine made tremendous progress over the last century. While breast cancer patients had a five-year survival rate of 40% one hundred year ago, this rate is now over 90% thanks to modern treatments and understanding of the disease [4], [5]. Even more impressive results were achieved in pediatric cancers as the survival rate went from 10% to almost 80% today [6].

These remarkable progresses are partly due to our better understanding of the disease, but also to the wide range of therapeutic strategies that have been developed in the last decades. Combinations of surgery, chemotherapy, radiotherapy, immunotherapy, targeted therapy, and hormone therapy are commonly used to treat cancer patients. However, effective cancer treatment lies not just in the availability of these options, but in their tailored application to individual patients. This requires a nuanced understanding of how to predict the severity of the disease and its evolution, and monitor the efficacy of selected treatments, a process crucial for optimizing patient outcomes while minimizing unnecessary toxicity.

At the heart of patient management in cancer care is medical imaging, a key tool that provides a wealth of anatomical, functional, and molecular insights. The integration of Positron Emission Tomography (PET) with Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) has become a cornerstone in cancer diagnosis and treatment. These modalities offer comprehensive overviews of a patient's condition, from disease profiling at diagnosis to monitoring during and after treatment. As most tumors can be easily spotted on whole-body PET scans, as well as their metabolic activity, this imaging modality has become pivotal in characterizing disease stage,

metabolic heterogeneity, and systemic responses to therapy, thereby informing treatment decisions.

Despite the critical role of PET imaging, its full potential is likely underexploited. Traditional approaches to PET image interpretation tend to oversimplify, focusing on primary tumors or a limited subset of lesions, and often overlooking the rich information present in the entire scan, including in apparently healthy tissues. This gap in the utilization of PET imaging data presents a significant opportunity for enhancing cancer management.

1.2 Contribution of the PhD

Given that context, the focus of this thesis was to find new prognostic information present in whole-body PET images. We hypothesis that more information prognostic of the outcome of cancer patients is present in PET images. We also hypothesize that this information can be identified through data mining approaches and encoded in meaningful and practical biomarkers.

A strong focus was put on the interpretability and reproducibility of the identified features, as well as their potential use for patient management. While the methodological developments were not bound to a specific disease, the actual biomarker search was conducted on two diseases: follicular lymphoma (FL) and diffuse large B cell lymphoma (DLBCL).

A semi-automated method was designed to identify new prognostic image-based features and several tools were developed to make this approach feasible. With this method, I identified dozens of new features prognostic of the outcome of FL and DLBCL patients. We were able to provide an intuitive explanation and definition to many of them, and several of the identified features appeared to be prognostic in two cohorts of patients, suggesting some commonalities between these two subtypes of non-Hodgkin lymphoma.

A new machine learning model (ICARE) designed for outcome prediction was also developed to better aggregate the identified biomarkers in a single prognostic signature. With this model, we won an international challenge. I demonstrated its utility for signature design.

1.3 Summary of Chapters

The manuscript is organized as follows:

Section I Introduction of the concepts

Chapter 2 covers the basics of medical imaging and focuses specifically on the two modalities used during the thesis: PET and CT imaging. This chapter also

presents the principles of image segmentation and introduces the concept of radiomics.

Chapter 3 is focused on machine learning and covers all the statistical concepts used extensively in our work.

Chapter 4 presents general information about cancers with a focus on lymphomas. It specifically dives into Follicular Lymphoma (FL) and Diffuse Large B Cell Lymphoma (DLBCL), the two diseases studied during this research.

Section II Original developments

All original developments are presented as articles (submitted, in preparation, or already accepted).

Chapter 5 reports the analysis of the impact of spleen tumor involvement for the prognosis of DLBCL patients.

Chapter 6 presents the biomarker selection tool (ROBI) developed for the semi-automated search of biomarkers.

Chapter 7 illustrates the use of the ROBI tool to identify new biomarkers prognostic of the outcome of FL and DLBCL patients and includes a tentative interpretation of those biomarkers.

Chapter 8 presents the ICARE model developed in the context of the HECKTOR challenge held during the MICCAI 2022 conference. It introduces the model and explains how it was used for the challenge.

Chapter 9 extensively compares the ICARE model to other traditional machine learning models on a large collection of medical datasets, to clarify its asset.

Chapter 10 summarizes the achievements and lessons learned from the work and draw some perspectives.

Section I

Introduction of the concepts

Chapter 2

Medical Imaging and Radiomics

Medical imaging is a key technology in modern healthcare, enabling the visualization of the internal structure of the body for diagnostic, monitoring, and treatment management. Being extensively used across various medical fields, it provides clinicians with critical insights that aid in the accurate diagnosis and effective treatment selection for numerous conditions. Billions of medical imaging studies are performed each year [7]. Computed Tomography (CT) and Positron Emission Tomography (PET) are two examples of imaging modalities. CT imaging uses X-rays to produce detailed cross-sectional images, or slices, of the body, offering valuable information about the body's anatomical structures with precise measurement of tissue density. PET, on the other hand, employs a radioactive tracer to visualize metabolic or biochemical activity within the body, thus providing functional insights, for instance, how much glucose is metabolized in each region of the body.

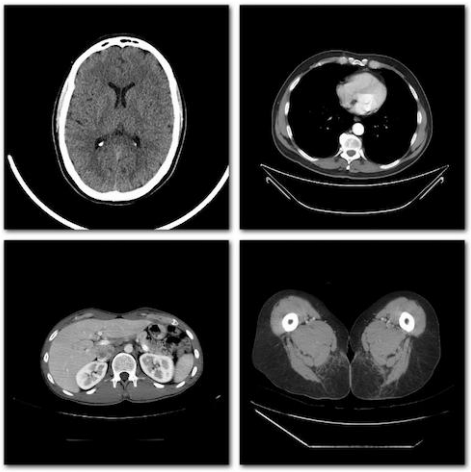
The quantitative interpretation of medical images sometimes requires a segmentation step. This consists in partitioning an image into one or multiple segments to simplify or change the representation of an image into something more meaningful and easier to analyze. Segmentation is particularly useful in medical imaging for the delineation of anatomical structures and areas of interest, supporting accurate diagnoses and treatment planning. For instance, in oncology, delineating the tumors allows to estimate the total metabolic tumor volume, a powerful prognostic factor in multiple cancers.

Finally, radiomics involves the extraction of many quantitative features from medical images, transforming the unstructured information of images into a structured table, allowing for detailed data analysis.

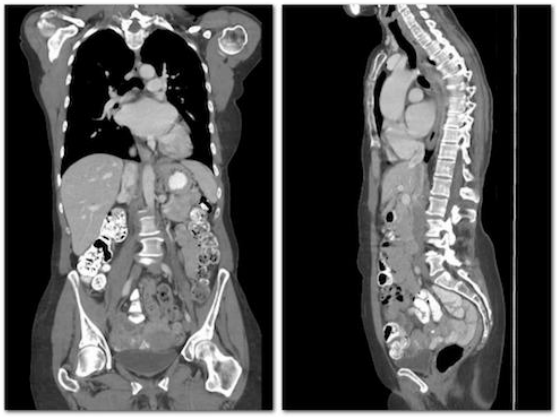
In this chapter, we will cover the basics concepts of PET and CT imaging, the two modalities used during the PhD, as well as image segmentation and radiomics.

2.1 Computed Tomography

Computed Tomography (CT) is a medical imaging modality offering precise mapping of tissue density in three dimensions [8]. This technique differentiates tissue types by measuring their variation in density, which can then be coded with colors, typically using grey scales with white representing denser tissues such as bones and black for less dense materials like air. Figure 1 displays various examples of CT images.



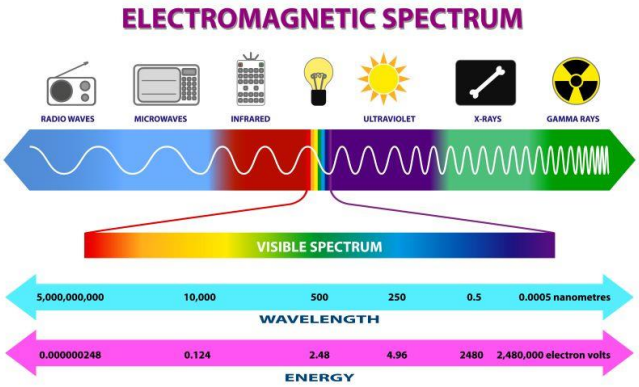
CT images of head, chest, abdomen, and thighs (left to right and top to bottom). Taken from [9].



Coronal reconstruction (left) and Sagittal reconstruction (right) of CT scan. Taken from [9].

Figure 1: Examples of Computed Tomography (CT) images illustrating how organs can be identified. The skeleton (white) stands out clearly and air (black) in the lung can also be easily detected.

The origin of CT can be traced back to Wilhelm Conrad Roentgen's discovery of X-rays in 1895 [10]. Initially met with skepticism, this discovery quickly became pivotal in medical diagnosis. X-rays, a form of high-energy electromagnetic radiation, penetrate diverse materials, including human tissues. The first X-ray image, which depicted Roentgen's wife's hand, not only proved the potential of this technology but also revolutionized the concept of non-invasive internal examinations. Figure 2 shows where X-rays are located on the electromagnetic spectrum and the first X-ray image.



The electromagnetic spectrum with X-rays marked. Taken from [11].

The first X-ray image. Taken from [12].

Figure 2: The electromagnetic spectrum with X-rays marked and the first X-ray image taken in 1895 depicting Roentgen's wife's hand. The ring and skeleton are clearly visible.

Early adoption of X-ray technology in medical diagnosis encountered significant challenges, including concerns about radiation exposure and the rudimentary nature of the initial equipment. Despite these hurdles, the intrinsic diagnostic value of X-rays gradually gained recognition, catalyzing enhancements in imaging technologies. This evolution culminated in the early 1970s with the development of Computed Tomography (CT) by Godfrey Hounsfield and Allan Cormack [13]. CT scans, offering three-dimensional imaging, yielded a substantial improvement over traditional X-rays, facilitating more comprehensive analysis of internal structures and pathologies. The transition from single-slice to multi-slice CT scanners, incorporating sophisticated techniques like helical scanning, marked a pivotal progression in medical imaging, significantly improving both the speed and resolution of scans [14], [15]. Figure 3 illustrates the difference between standard X-ray scans and CT images.

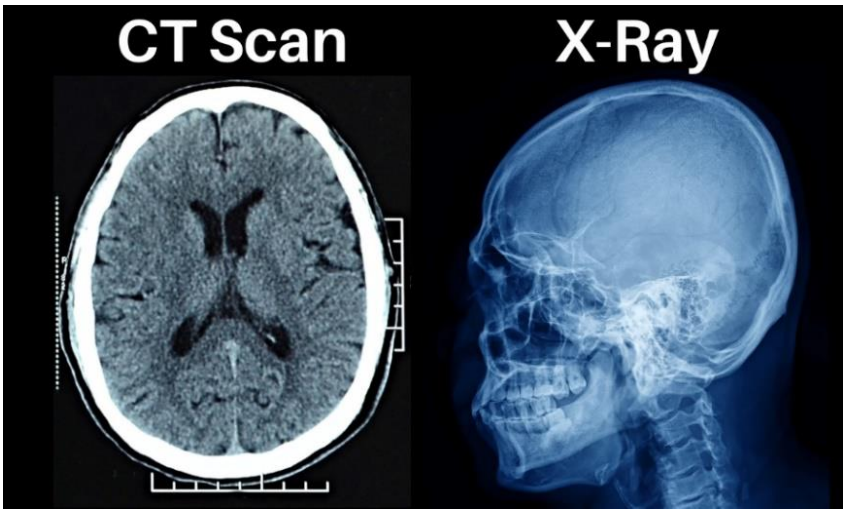


Figure 3: Difference between CT scan and X-ray scan. While X-ray only offers a unique 2D shadow projection of X-rays passing through the whole body, CT scan gives multiple 2D cross sections (slices) of the whole scanned region. Taken from [16].

CT scanners operate by rotating an X-ray source around the patient, capturing images from multiple angles, as shown in Figure 4.

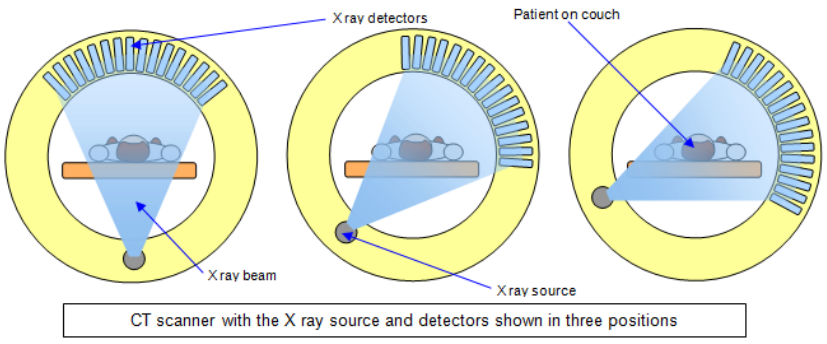


Figure 4: Basic principle of the CT scan: multiple images of the patients are taken at different angles. These images are then combined to estimate the signal in slices. Taken from [17].

These images are then reconstructed into cross-sectional slices using algorithms such as filtered back projection and iterative reconstruction. The refinement of these algorithms, propelled by advances in computational power, has profoundly impacted image quality and diagnostic accuracy [18].

A critical aspect of CT scan interpretation is the use of Hounsfield units (HU), which provide a standardized scale for measuring tissue density, ranging from -1000 HU for air to +2000 HU for dense bone [19]. This scale enables differentiating different tissue types, as it quantifies variations in tissue density. The Hounsfield unit is defined by the equation:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}$$

where μ, μ_{water}, and μ_{air} represent the attenuation coefficients in cm^{-1} of the voxel, water, and air, respectively. The attenuation coefficient represents the degree to which different tissues in the body reduce the intensity of X-ray beams.

The resolution of CT images, referring to the minimum discernible detail, is a key feature characterizing the image quality. Modern CT scanners have remarkably high resolutions, often to fractions of a millimeter, enabling the detection of small lesions, offering more precise diagnosis and treatment planning.

Contrast agents, typically iodine or barium-based, are employed to enhance the visibility of internal structures in CT imaging, such as vascular structures [20]. Figure 5 presents an example of the benefit of contrast agent. While these agents are widely used, they are some risks associated with their injection, such as allergic reaction, leading to the development of safer alternatives and protocols[21].

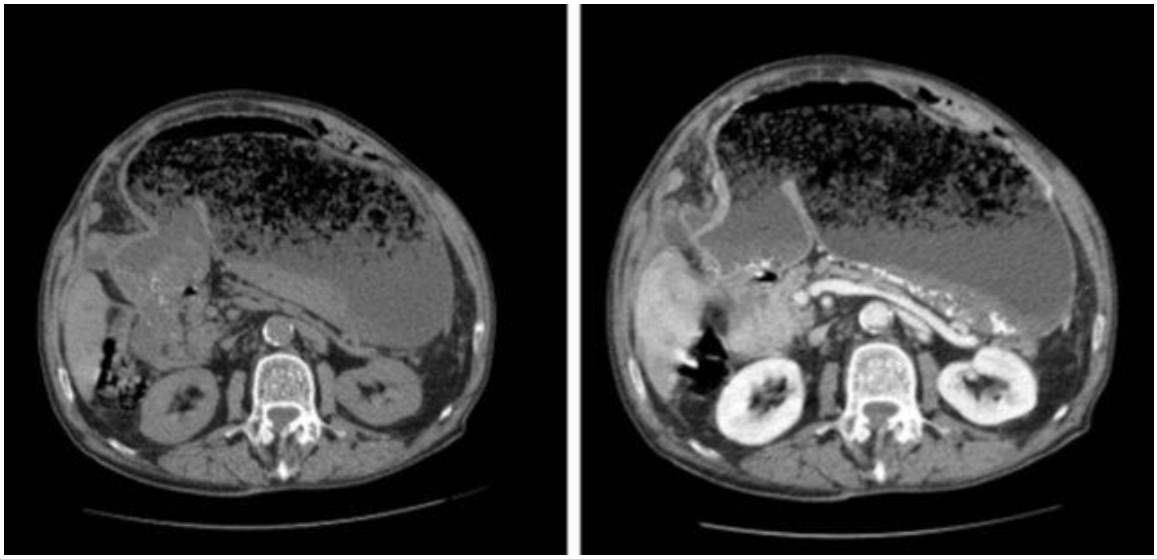


Figure 5: CT scan before (left) and after (right) injection of contrast agent. The kidneys stand out more clearly with the contrast agent. Taken from [22].

While CT is known for its rapid, high-resolution imaging capabilities, alternative modalities like MRI and ultrasound offer benefits such as the absence of radiation exposure. Ensuring patient and personnel safety in CT imaging involves minimizing radiation doses through specific technologies and protocols. Current research endeavors are directed towards mitigating CT's inherent limitations, such as radiation risks and challenges in imaging certain soft tissues, by developing lower-dose techniques and enhancing image quality [23].

2.2 Positron Emission Tomography

Positron Emission Tomography (PET) stands as a pivotal imaging modality in medical diagnosis and research, owing to its unique capability to visualize physiological processes within the human body. This technique finds extensive application in various fields such as oncology, neurology, and cardiology, and plays an important role in disease detection and treatment monitoring [24].

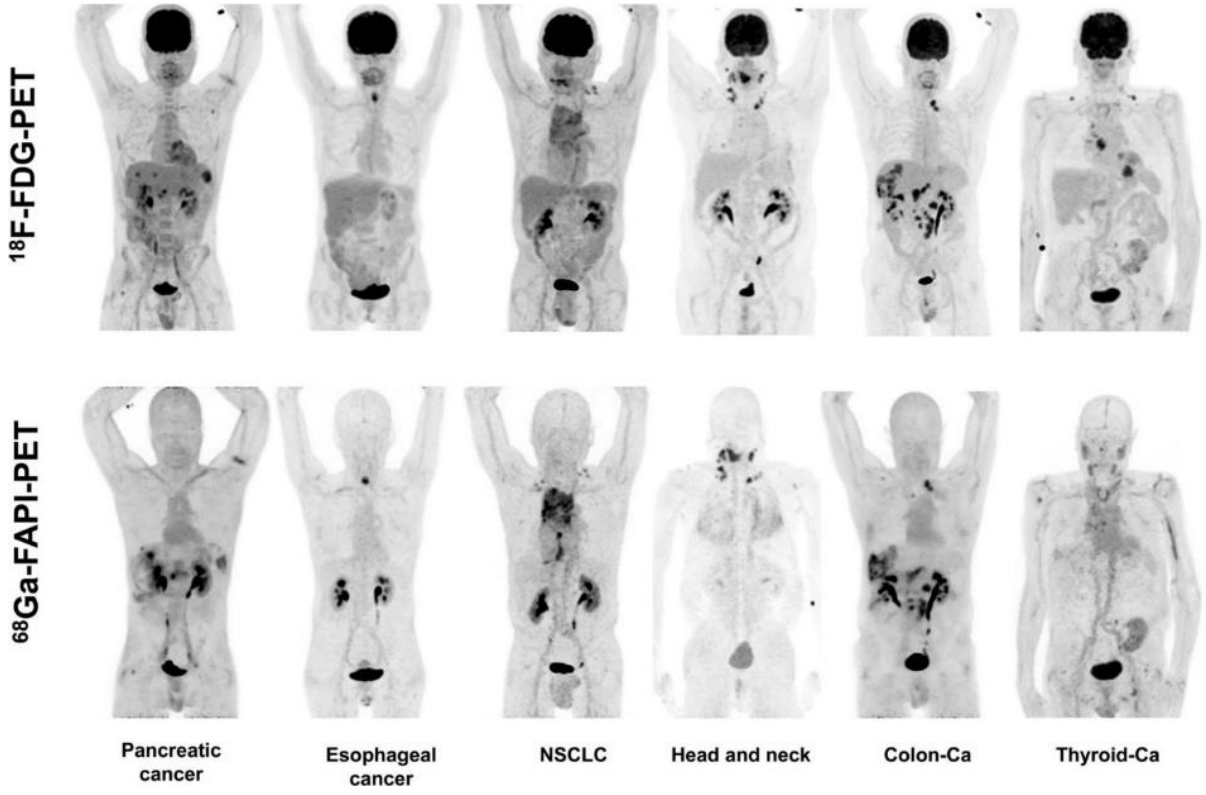


Figure 6: Maximum intensity projection of PET scans of different patients with two radiotracers: ^{18}F-FDG (top row), which reflects glucose consumption, and ^{68}Ga-FAPI (bottom row), which maps fibroblast activation protein (FAP) that is often over expressed in cancerous tissues. The darker the image, the higher the concentration of radiotracer. Each column corresponds to one patient. While brain and liver are clearly visible with ^{18}F-FDG, they are not with ^{68}Ga-FAPI, while cancerous lesions can still be observed, showing that ^{68}Ga-FAPI might be more specific to cancerous areas in certain anatomical regions. Taken from [25].

The process of PET imaging requires the administration of a radioactive tracer, such as fluorodeoxyglucose (FDG), which makes it possible to map glucose metabolism. A large variety of tracers can target many different specific physiological processes such as oxygen consumption, blood flow, or receptor binding. This diversity of radiotracers underscores PET's adaptability across different medical and research applications. Figure 6 illustrates this by showing the PET images of the same patients scanned with two different radiotracers.

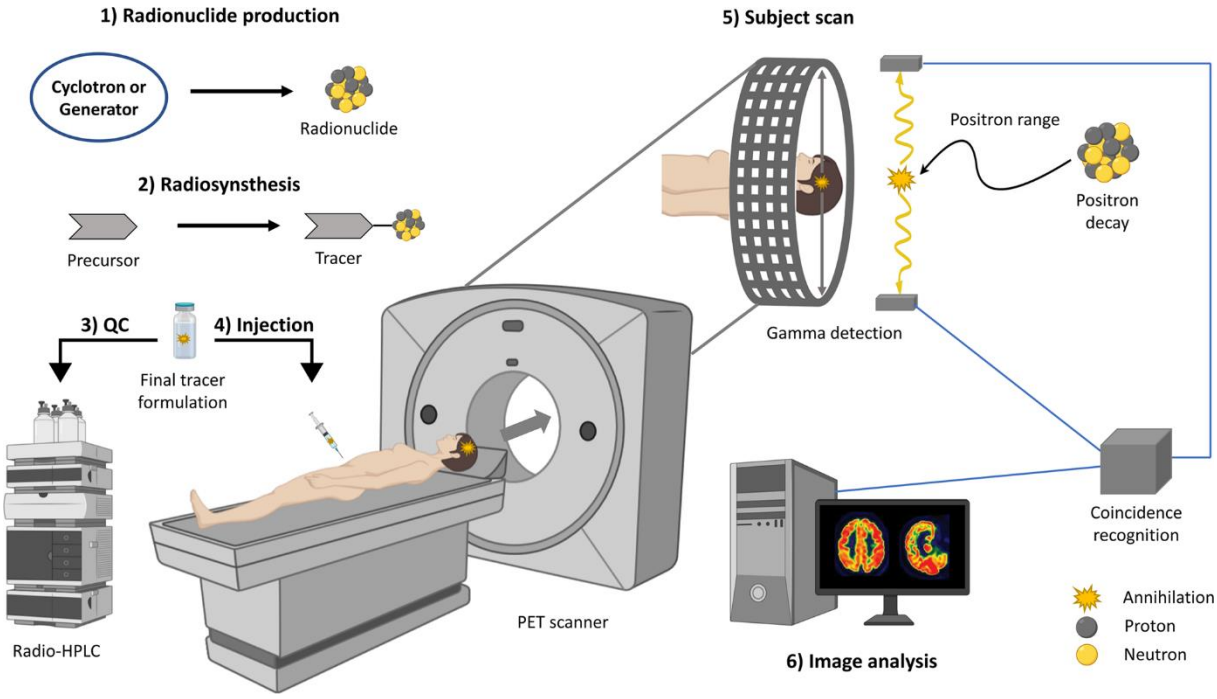


Figure 7: Overview of the whole process of PET imaging, from the production of the radiotracer to the reconstruction of the image. Taken from [26].

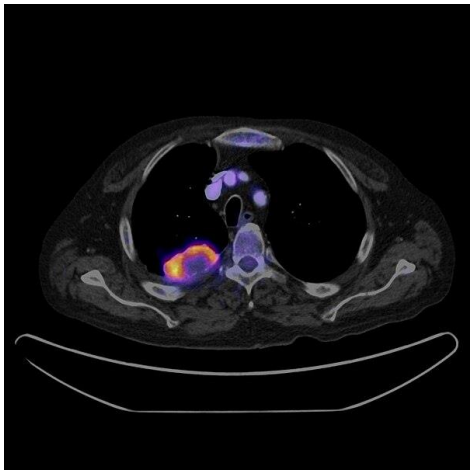
Upon administration, the tracer, containing a positron-emitting radionuclide like fluorine-18 for FDG, engages in the body's biochemical pathways. In cancer detection, for example, the elevated glucose consumption of cancer cells results in higher FDG accumulation, leading to an increased emission of positrons in tumor regions [27]. The annihilation of these positrons with electrons generates two gamma photons traveling in opposite directions. The concurrent detection of these photons allows the three-dimensional reconstruction of the image, accurately locating the tracer within the body [28]. Algorithms such as Ordered Subset Expectation Maximization are employed to transform these signals into a three-dimensional image, delineating the tracer distribution and providing insights into the metabolic activity of tissues. Figure 7 summarizes the whole process.

The radioactive nature of the radiotracer implies that its activity diminishes over time. For instance, ^{18}F-FDG has a half-life of 110 minutes. The positron emission rate is affected by this decay and other factors as well such as patient weight and machine parameters. To address this variability, the Standardized Uptake Value (SUV) has been introduced. The SUV, a dimensionless ratio, reflects the tracer concentration in a

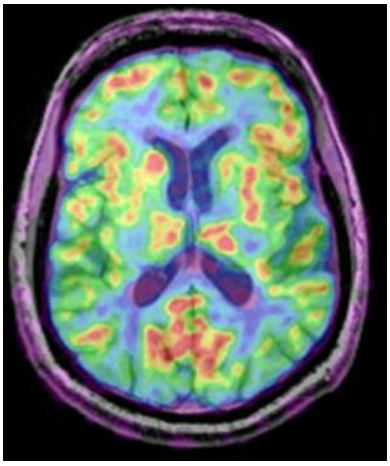
specific region, normalized against the administered dose and the patient's body weight. It offers a standardized quantitative measure of radiotracer concentration, enabling comparisons across different patients. SUV is usually defined as:

$$SUV = \frac{activity\ concentration}{injected\ dose} \times patient\ weight$$

with the $injected\ dose$ in megabecquerel (MBq), $patient\ weight$ in gram, and $activity\ concentration$ in MBq/milliliter. PET excels in detecting metabolic alterations before structural changes become evident. This is particularly valuable in early cancer detection and in identifying initial signs of neurodegenerative diseases, such as the early detection of amyloid plaques in Alzheimer's disease [29].



PET/CT scan of lungs taken from [30].



PET/MRI of the brain taken from [31].

Figure 8: Examples of PET combined with CT and MRI. The PET image is shown in color while the information of CT and MRI are shown in grey scale.

In clinical settings, PET is often combined with computed tomography (CT) or magnetic resonance imaging (MRI) to create hybrid modalities like PET/CT or PET/MRI. Examples are provided in Figure 8. This fusion enhances diagnostic precision by integrating PET's functional insights with the anatomical details provided by CT or MRI. However, it is noteworthy that PET's spatial resolution is generally inferior to that of CT and MRI, with modern PET systems achieving resolutions around 4 mm, compared to the submillimeter resolution of CT or MRI.

Despite these technological advances, PET imaging faces certain limitations. Radiotracers, for example, do not exclusively accumulate in target regions. While ^{18}F-FDG concentrates in tumors, it also accumulates in areas naturally exhibiting high glucose metabolism, such as the brain, and in inflammatory sites [32]. This can complicate image interpretation, driving ongoing research towards more condition-specific tracers. Moreover, the short lifespan of radiotracers introduces logistical challenges. Some must be produced in a cyclotron and promptly transported to the imaging site, limiting PET's availability in areas remote from production facilities. This

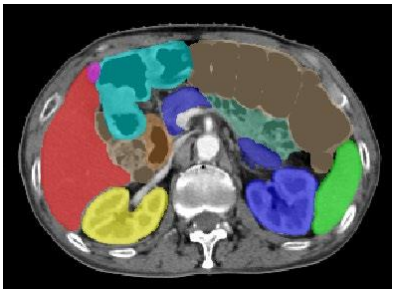
also complicates scheduling, as delays in tracer production or transportation can result in tracer wastage and require rescheduling patient appointments.

PET imaging also involves radiation exposure from the radioactive tracers. Although the radiation levels are low and within accepted safety margins, there remains an inherent risk, particularly for repeated scans or sensitive groups such as children or pregnant women. Nonetheless, the significant diagnostic and therapeutic benefits of PET imaging justify this exposure but require a balance of risks and benefits [33]. The decision to employ PET imaging takes into account the patient's overall radiation exposure history, especially in scenarios involving multiple imaging procedures or longitudinal studies with repeated scans.

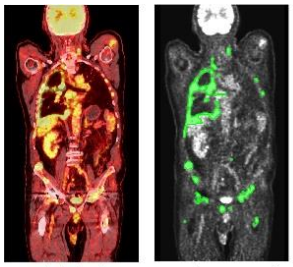
2.3 Image segmentation

Medical image segmentation is a useful tool for the quantitative analysis and interpretation of diverse imaging modalities [34], [35], [36]. This involves dividing a digital image into segments or sets of voxels, thereby streamlining the image's structure and augmenting its suitability for analytical purposes. In clinical practice, segmentation can help differentiating between various anatomical or functional structures or regions of clinical interest, such as organs, tumors, or tissue types.

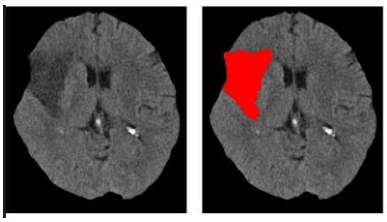
Figure 9 shows several examples of segmentation of PET and CT images.



Organ segmentation in CT images. Taken from [37].



Tumor segmentation example. Taken from [38].



Example of stroke lesions segmentation on CT images. Taken from [39].

Figure 9: Examples of segmentation on CT and PET images of organs, tumors, and stroke lesion.

In specialties like radiology, oncology, and neurology, segmentation can play a key role [40], [41], [42]. Radiologists use segmentation to accurately delineate organ boundaries in CT or MRI scans, to guide treatment in radiotherapy for instance, or extract prognostic parameters from the segmented areas. In the field of oncology, segmentation help characterize tumors, thus informing cancer treatment strategies. For neurologists, segmentation is essential for identifying cerebral abnormalities, including lesions or areas impacted by stroke, significantly influencing patient management strategies [43].

Segmentation techniques vary, ranging from manual delineation by experts to semi-automatic and fully automatic methods. A prevalent semi-automatic approach involves initially running an automatic threshold-based segmentation (e.g., using a 2.5 SUV threshold for segmenting tumors in PET images of lymphoma patients [44]), followed by manual adjustment and classification of the segmented regions by a physician (e.g., metabolic or tumoral in oncologic PET imaging). Although straightforward, this method has several drawbacks. Determining appropriate cut-off values is challenging, and this approach requires time and effort by the physician. Consequently, there has been a growing interest in fully automatic segmentation, propelled by advances in deep learning models like U-Net and V-Net. The architecture of U-Net, characterized by its contracting and expansive pathways, enables precise localization, rendering it exceptionally suitable for medical image segmentation. These Convolutional Neural Networks (CNNs) are trained on extensive datasets, allowing them to distinguish between various tissues and structures effectively. An advanced iteration, nnU-Net, has further enhanced this approach, extending its applicability to a wider range of scenarios [45].

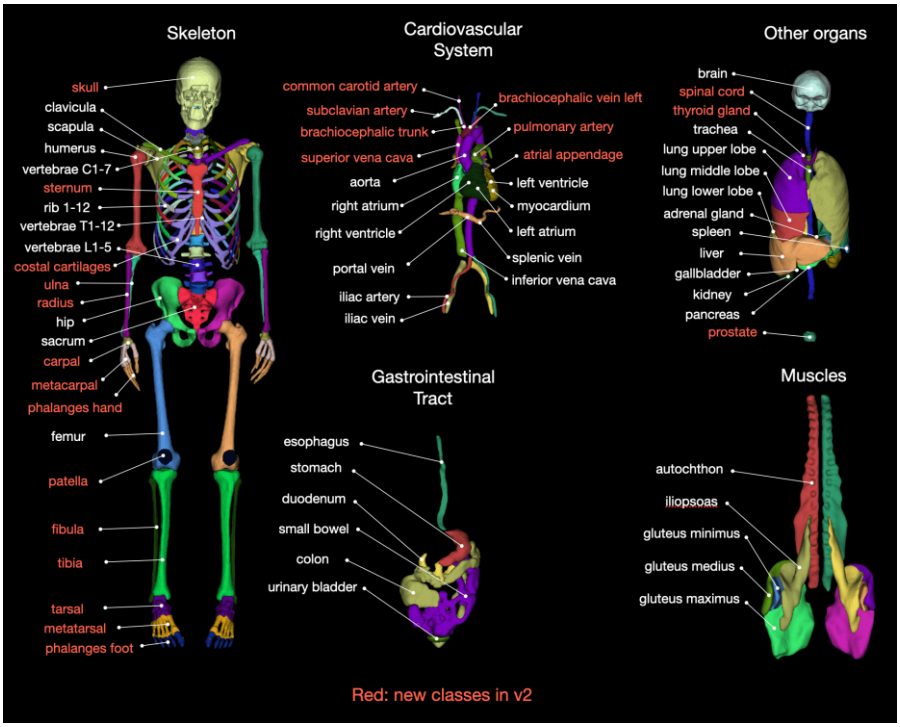


Figure 10: List of organs automatically segmented by TotalSegmentator. Taken from [46].

Although fully automated tumor segmentation remains challenging, organ segmentation using deep learning models has become remarkably precise and robust. User-friendly implementations, such as TotalSegmentator [46] and MOOSE [47], are now available through open-source software. These models, trained on numerous examples, achieve high levels of segmentation accuracy. This automated approach paves the way for comprehensive analyses of whole-body information in large patient cohorts, a task that would have been prohibitively resource intensive in the past. Figure 10 lists all the organs segmented by TotalSegmentator.

However, these technological advances also introduce ethical and regulatory considerations [48]. For instance, algorithmic bias in segmentation can have clinical consequences, emphasizing the need for diverse training datasets to ensure equitable and accurate diagnoses across various patient demographics. Moreover, addressing privacy concerns related to the use of patient data for training these algorithms is a critical ethical issue. The integration of deep learning models into clinical settings brings unique challenges, primarily due to the complexity and unfamiliarity of AI-based systems. Physicians, traditionally relying on established diagnostic methods, may struggle with interpreting and trusting AI-generated results, particularly when the AI's decision-making process is opaque or inexplicable. This "black box" nature of AI can impede effective collaboration between human clinicians and AI systems, potentially leading to resistance against adopting such technologies.

Another challenge is the efficient use of these models. It is unrealistic to expect to remove human expertise from medical image analysis. The next natural question is therefore how to get the best synergy from human-AI collaboration. For instance, numerous works are being conducted on how to point to the physician the automatically segmented regions that need the most attention and review [49].

Additionally, establishing clear legal responsibilities, especially in cases of misdiagnosis or treatment errors, remains a contentious issue [50]. The determination of liability, whether it rests with the physician or the AI model (and, by extension, its developers), is a complex legal and ethical matter. This ambiguity may cause reluctance among physicians in relying on AI tools, despite their evident potential benefits.

2.4 Radiomics

Radiomics represents an innovative intersection of medical imaging, computer vision, and data science, transforming imaging data into a high-dimensional space well suited for advanced statistical analysis [51], [52]. Leveraging common imaging modalities like PET, CT, and MRI, radiomics aims to transform unstructured image data into structured data amenable to advanced statistical examination and automated image analysis.

The radiomic workflow starts with image acquisition, followed by segmentation to identify areas of interest, such as tumors or specific organs. This segmentation stage varies in approach, manual, semi-automatic, or fully automatic. Preprocessing then normalizes variances due to different imaging equipment, patient positioning, population characteristics, and inherent noise. This step, which is key for maintaining consistency and reproducibility across datasets [53], [54], includes processes like interpolation for uniform voxel spacing, outlier elimination, and value discretization via binning, thereby mitigating noise-related feature distortion (binning is also required for some radiomic features) [55]. Subsequent stages involve automatic extraction of diverse features (shape, intensity, texture), employing tools like the open-source PyRadiomics package [56], which we used throughout this thesis. These features

effectively convert qualitative visual data into quantifiable metrics [57]. Figure 11 summarizes the whole process.

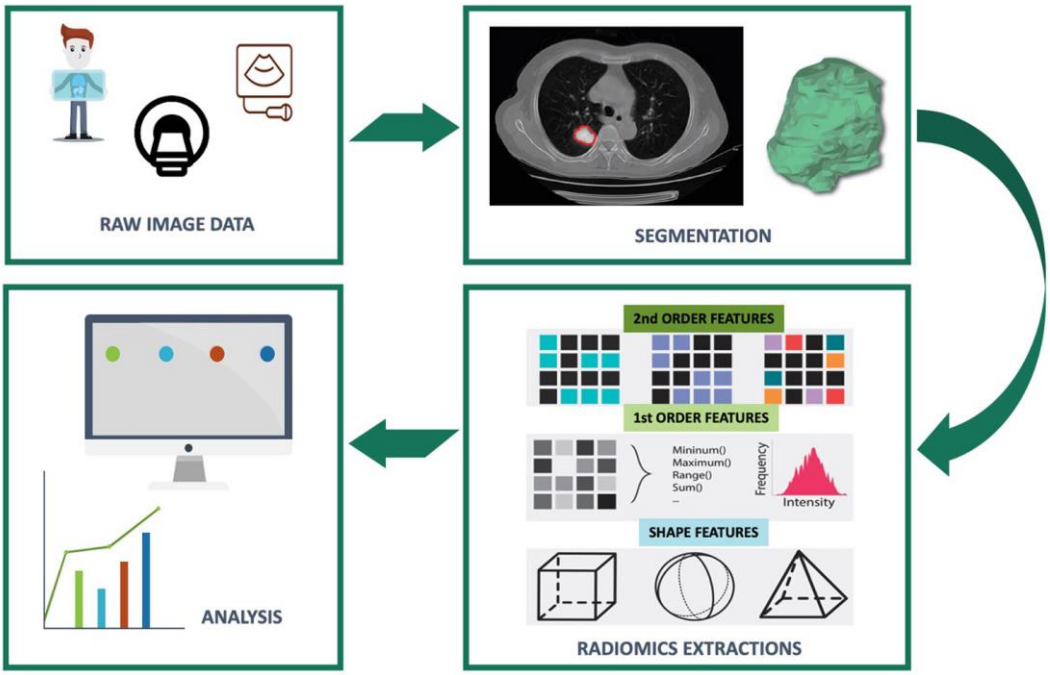


Figure 11: Overview of the full radiomic analysis pipeline. Taken from [58].

Radiomics' potential is extensive, offering pathways to new hypotheses, image-based biomarker discovery, and partial automation of image analysis – prospects that could revolutionize diagnosis, prognosis, and treatment strategies. Studies have repeatedly established significant correlations between radiomic features and relevant clinical endpoints, such as disease phenotypes and treatment responses [59], [60], [61], [62], [63].

However, the translation of radiomics into clinical practice remains limited. Despite extensive literature referencing radiomics, few results have reached clinical application [64], [65]. Challenges impeding broader adoption include small cohort sizes, issues with reproducibility and standardization, complex model interpretation, and methodological concerns. Initiatives like the Imaging Biomarker Standardization Initiative (IBSI) [66], [67] are pivotal in promoting standardization, thereby enhancing study reproducibility and clinical applicability. Tools like PyRadiomics [56] and LIFEx [68] provide consistent analysis platforms, while techniques like ComBat [69] address cohort variability, increasing radiomics' applicability. The radiomics community's commitment to developing best practice guidelines is also instrumental in improving study quality [70], [71], [72], [73].

A critical hurdle remains in the interpretability of radiomic data, essential for both scientific credibility and clinical application. The challenge lies in deciphering the intricate mathematical definitions of the radiomic features and ensuring that these features correlate meaningfully with biological or pathological processes. Without

clarity, there is a risk of radiomic models becoming opaque "black boxes", technically accurate but clinically elusive. This opacity can impede clinical integration and increase the likelihood of false discoveries. The challenge intensifies when integrating radiomic features with machine learning models. Efforts to explain radiomics findings remain rare [74], [75], [76]. An appealing scenario could be to find new relevant biological information through radiomic data mining, and then reencode this information in simpler and more interpretable ways. Enhancing interpretability not only improves confidence in radiomic findings but also bridges the communication gap between researchers and clinicians. Moreover, it is a necessary step to use radiomics for knowledge discovery. Understanding the biological mechanisms encoded by a feature is a promising way to improve our understanding of the disease.

An alternative approach to radiomic is deep radiomic. Instead of manually defining the features with mathematical formula, we let a model learn them directly from the images. More specifically, a neural network, often a Convolutional Neural Network (CNN), is trained on the images, in supervised or unsupervised manner, and creates abstractions (e.g., features) in its inner representation. Then, interpretation techniques are used to isolate and understand specific features learned by the model that could be useful. This is a promising but challenging approach. Its greater flexibility and abstraction capabilities make it more powerful than manual definition, potentially discovering more meaningful features, but it also makes the approach more prone to overfitting. Higher volume of training data is often needed to train neural networks and the datasets available for radiomic analysis rarely include more than a thousand patients. A second challenge is the identification and interpretation of the features learned by the model. While many interpretation techniques exist to decipher CNNs (e.g., GradCam, SHAP, …), it remains a difficult task. I explored this methodology during the PhD. While I was able to identify known biomarkers for DLBCL patients (TMTV and Dmax), I did not find new ones. Despite many efforts, I found this approach less suited than manual definition of the features for the task and data at hand.

Chapter 3

Machine learning

Machine learning sits at the intersection of statistics and computer science. Despite being an old discipline, it has gained unprecedented momentum in recent times, propelled by advancements in computational power, the availability of large datasets, and continuous innovations in algorithmic techniques. Unlike traditional programming methods, which rely on explicit instructions for specific tasks, machine learning enables computers to learn and make decisions from data, offering a more flexible and dynamic approach to problem-solving. This adaptability makes it particularly valuable in complex scenarios where predefined rules fall short [77].

Today's medicine produces a phenomenal quantity of data. Tools that can effectively explore, understand, and retrieve useful information from these large and complex pools of data have the potential to greatly improve our knowledge and the toolbox of the physicians [78]. For instance, successful applications of the method in radiology, dermatology and oncology have been reported, such as automated disease classification, organ segmentation and tumor segmentation [46], [79], [80], [81].

In this chapter, we will cover the basics of the discipline, present some algorithms that can learn from data, how they can be used for medical applications, and finally concepts and tools that are extensively used in this thesis to find new image-based biomarkers.

3.1 Basic principles

Data science is an interdisciplinary domain that uses scientific methods, statistics, and algorithms to extract knowledge and insights from data. Machine learning is one of the tools used by data scientists [82].

Traditional programming involves the explicit definition of a precise set of instructions defined by a human, that are then executed by a machine to achieve the desired goal. For instance, let's say we want to write a program that return tomorrow's temperature based on today's weather data. With the traditional approach, a programmer could produce this type of equation:

$$T_{tomorrow} = 0.85 \times T_{today} + 0.03 \times H_{today}$$

with T the temperature and H the humidity. But then, the programmer compares the output of its program to real weather reports and find discrepancies. He reads in

meteorology literature that atmospheric pressure can influence the temperature of the next day. Thus, he updates his previous program to account for this and update some coefficients based on the discrepancies he saw in the data:

$$T_{tomorrow} = 0.85 \times T_{today} + 0.03 \times H_{today} + 0.03 \times P_{today}$$

with P the atmospheric pressure. With additional trials and errors, intuition and expert knowledge, the programmer can keep updating his program to have the results closest to the weather reports.

Machine learning takes the opposite direction: learning the program directly from the data. For instance, a linear regression, a type of machine learning model, would define the temperature as:

$$T_{tomorrow} = \beta_0 + \beta_1 \times T_{today} + \beta_2 \times H_{today} + \beta_3 \times P_{today}$$

The β values are called parameters and are unknown by default. Only the available features (T, H and P) are specified to the model. Then based on the actual values in the weather report, the linear regression algorithm will define the best β values to produce a predicted temperature as realistic as possible.

Another intuitive machine learning algorithm is the K-nearest neighbor. The idea is straightforward: in the data, find the K examples that are the most similar to the sample for which we want to make a prediction. In our example, we will find the K days with T, H and P closest to the T, H and P of the day for which we want to make a prediction. We will then take the average temperature of the days following our K neighbors to get the final prediction. This approach is sometimes referred to as digital twins in medicine applications.

The machine learning approach described above is called "supervised" because we provide examples of the desired output to the algorithm. Another approach is called "unsupervised". In this scenario, we do not explicitly tell the algorithm what the desired output is. We rather let it find relevant patterns in the data. If we give the weather reports to an unsupervised machine learning, it could group the days by seasons or create cluster of days with and without storm for instance.

One important concept in modelling was captured by the statistician George Box: "All models are wrong, but some are useful". No model will ever predict the temperature of the next day to a thousandth of a degree, but this precision is not required for the model to be useful. For this reason, model evaluation plays a central role in machine learning. It consists in estimating the performance (e.g., accuracy) of the model on new unseen data. In other words, once the model is deployed and used in real world applications, how precise will its prediction be (the underlying question being, is the model precise enough to be useful)? This is called the testing phase. It is important to use different datasets for training and testing. Most of the time, a model will perform

better on the data used to train it than on new data the model never learned from. For this reason, we often used a "training set" and a "testing set".

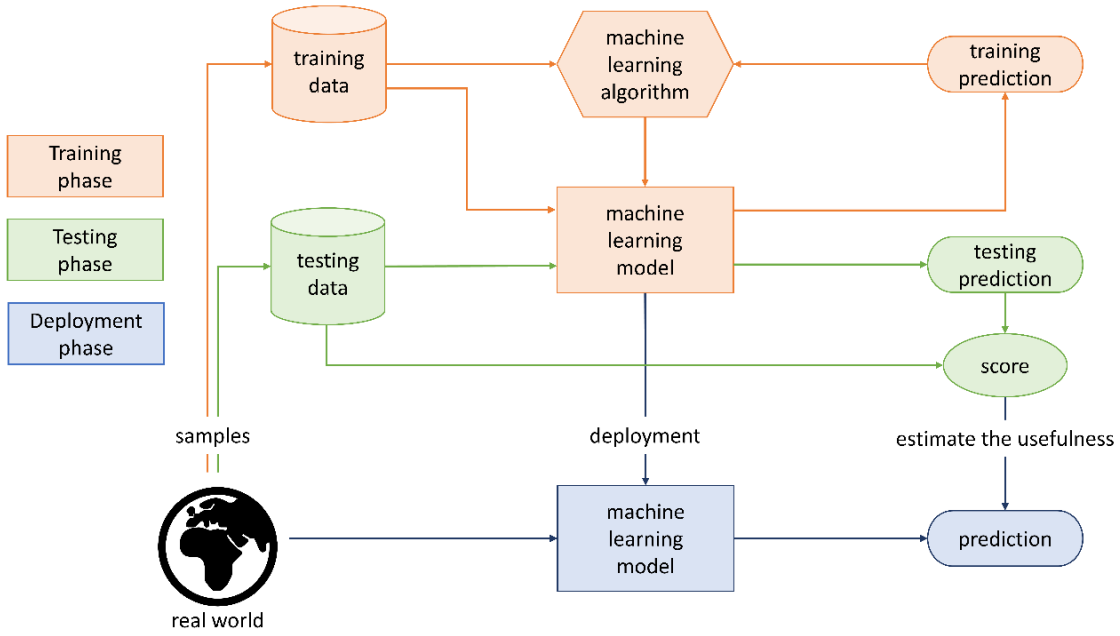


Figure 12: Overall diagram of the machine learning development and deployment process.

Samples of the real world are necessary to build (e.g., train) the model and evaluate it. It is necessary to have enough data for efficient training and testing. It is intuitive that the model will not be realistic if we only have three weather reports. A good rule of thumb is the more data the better. More specifically, it is important that the data available are representative of the data on which the model will be used. Having a million weather reports of London will probably not produce an accurate model for weather prediction in Rio de Janeiro. Figure 12 shows a diagram of the notions introduced above.

One common challenge in this learning process is overfitting and underfitting. Overfitting is like a student memorizing answers without understanding the underlying concepts, leading to poor performance in unfamiliar situations. This happens when a model learns too much from the training data, including the noise and inaccuracies, and fails to generalize to new data. Underfitting, on the other hand, is when the model has not learned enough from the training data, akin to a student who has not studied enough and thus cannot make accurate predictions or decisions.

Regularization addresses overfitting by adding a penalty to the learning algorithm. This penalty discourages the model from learning a more complex or flexible model, thus forcing it to learn only the most important patterns in the training data. Common forms of regularization include techniques like L1 and L2 regularization, which add different types of penalties to the cost function used by the learning algorithm.

Another key component of the machine learning process is feature preprocessing. It involves the techniques applied to raw data before feeding it into a machine learning algorithm. The purpose of this step is to convert or encode the data in a manner that enhances the algorithm's performance, making it easier for the model to learn and make predictions. This process can include scaling or normalizing features so that they are on a similar scale, handling missing values, encoding categorical variables into numerical values, and creating new features from the existing ones (feature engineering). Effective feature preprocessing can significantly improve the performance of a machine learning model, as it helps in reducing the complexity of the data and highlights the most important attributes for making predictions [83].

Classification and regression are two fundamental types of tasks in the field of machine learning and statistics, each serving different purposes. Classification involves categorizing data into predefined classes or groups. For instance, a classification model might be used to determine whether an email is "spam" or "not spam", or to identify the species of a plant based on its features. The key characteristic of classification is that the output is categorical, not numerical. On the other hand, regression deals with predicting a continuous, numerical output. The goal is to understand the relationship between variables and to predict a quantity. In our weather forecast example above, the task of predicting the temperature of the next day was a regression task.

3.2 Models

This section will briefly describe common machine learning models. The goal is not to give a complete understanding of each model, but rather an intuitive and simple explanation to grasp the core idea, the strengths, and weaknesses of each model.

3.2.1 Linear regression

Linear regression is a method used for modelling the relationship between a dependent variable and one or more independent variables [84]. The aim is to fit a linear equation to observed data. In simple linear regression, this equation takes the form:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots$$

where y is the dependent variable, x_i are the independent variables, β_0 is the y-intercept, and β_i are the slope of the line. This equation allows us to predict the value of y based on the values of x_i. Linear regression is widely used because of its simplicity and interpretability, and it finds applications in numerous fields, such as economics, biology, engineering, and social sciences. The key assumption in linear regression is that there is a linear relationship between the variables, and it requires careful examination of data for accuracy and validity of the model. While it might seem simplistic, in practice, this assumption often allows for good performance and robust generalization on to new data.

3.2.2 Logistic regression

Logistic regression is a model used for binary classification tasks, where the goal is to predict a binary outcome (e.g., yes/no, success/failure). Unlike linear regression, which predicts a continuous output, logistic regression predicts the probability of an event occurring [85]. It models this probability in relation to one or more variables using a logistic function. The logistic function ensures that the output probability is always between 0 and 1. The basic form of the logistic regression equation is:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots)}}$$

where $P(Y = 1)$ is the probability of the event occurring, β_0 the intercept, and β_i are the coefficients of the features x_i. It is particularly useful because it not only provides a classification but also quantifies the odds of the event occurring as a function of the independent variables.

3.2.3 Decision tree

A decision tree is a machine learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure, where each internal node represents a test on a feature (e.g., is the temperature above 20°C), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

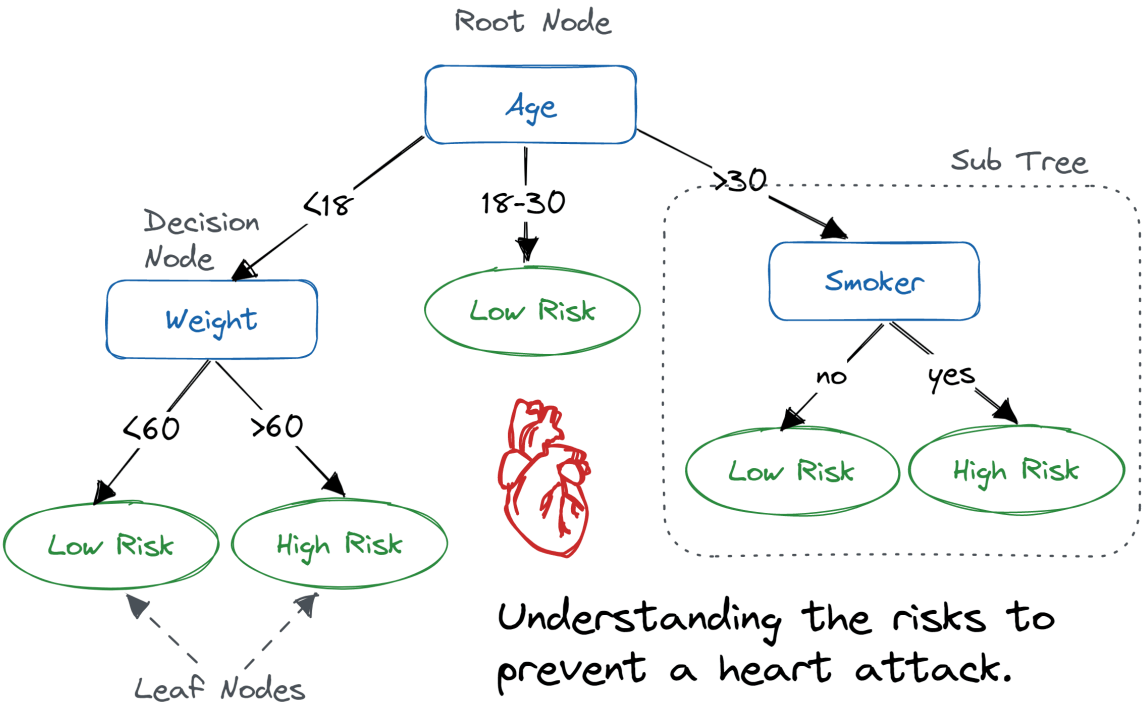


Figure 13: Example of a decision tree to classify patients as low or high risk of heart attack, based on their age, weight and tobacco consumption. Taken from [86].

When constructing a decision tree, the algorithm selects the attribute that best splits the data into groups with the most homogeneous (or similar) outcomes. This selection is often based on criteria like Gini impurity or information gain for classification tasks, and variance reduction for regression. The process is recursive, splitting each subset further until the algorithm reaches a predetermined stopping condition (like a maximum depth or minimum number of samples required to split a node). Figure 13 shows an example of a decision tree that classifies patients as low or high risk of heart attack.

Decision trees are popular due to their simplicity, ease of interpretation, and ability to handle both numerical and categorical data. They also do not require feature normalization. They visually represent the decision-making process, which can be easily understood by non-experts. However, they can be prone to overfitting, especially if they grow too deep or complex, thus failing to generalize well from the training data to unseen data. Techniques like pruning (removing parts of the tree that provide little power to classify instances) and limiting the depth are used to prevent this overfitting [87].

3.2.4 Ensemble models

Ensemble models in machine learning are advanced methods that combine multiple individual models to improve overall predictive performance, compared to using a single model. The underlying principle is that a group of "weak learners" can, when combined, form a "strong learner". There are various types of ensemble methods, with Bagging, Boosting, and Stacking being the most prominent [88]. Figure 14 illustrates these three methods.

In Bagging, multiple models are trained in parallel on different subsets of the data (sampling with replacement also known as bootstrapping), and their predictions are averaged (for regression) or voted (for classification) to produce the final output. Random Forest is a well-known example of bagging, where multiple decision trees are combined.

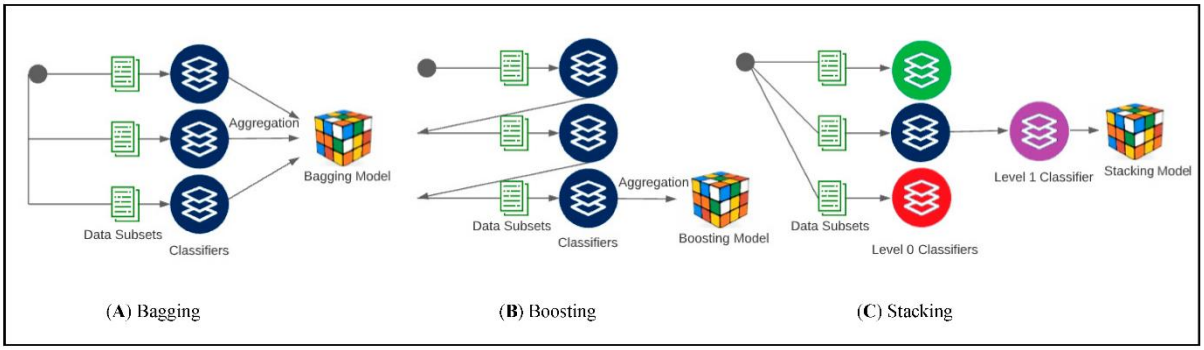


Figure 14: The three types of models ensembling methods. Taken from [89].

Boosting, on the other hand, trains models sequentially, each new model focusing on the mistakes of the previous ones. The idea is to gradually improve the model's performance.

Stacking involves training multiple models and then using another model to combine their predictions. The first-layer models are trained on the full dataset, and their predictions are used as inputs for the second-layer model to make the final prediction.

Ensemble methods are widely used because they often lead to more robust and accurate models, reducing the likelihood of overfitting and improving performance on diverse datasets. However, they can be computationally intensive and less interpretable than individual models.

3.2.5 Support Vector Machines

Support Vector Machines (SVMs) are a set of supervised learning methods used for classification and regression. The core idea behind SVM is to find the best hyperplane (e.g., a line in two dimensions, a plane in three dimensions) that separates different classes in the feature space. For binary classification, this hyperplane is chosen to maximize the margin between the two classes, where the margin is defined as the distance between the hyperplane and the nearest data points from each class, known as the support vectors [90]. Figure 15 illustrates this type of model in action.

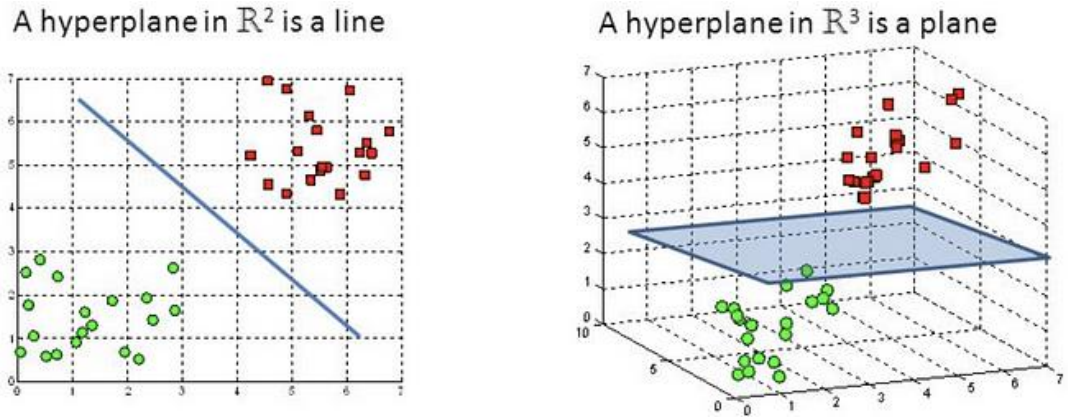


Figure 15: Visualization of SVMs separating the green and red points in a two- and three-dimensions spaces. Taken from [91].

SVMs are effective in high-dimensional spaces and are versatile, as different kernel functions can be specified for the decision function. Common kernels include linear, polynomial, and radial basis function (RBF). The choice of kernel and its hyperparameters can have a significant impact on the performance of the SVM. Figure 16 shows the impact of the choice of kernel on the model. SVMs are known for their accuracy and robustness, particularly in cases where the number of dimensions exceeds the number of samples. However, they can be computationally intensive, especially for large datasets, and may require careful tuning of hyperparameters and choice of kernel to achieve optimal performance.

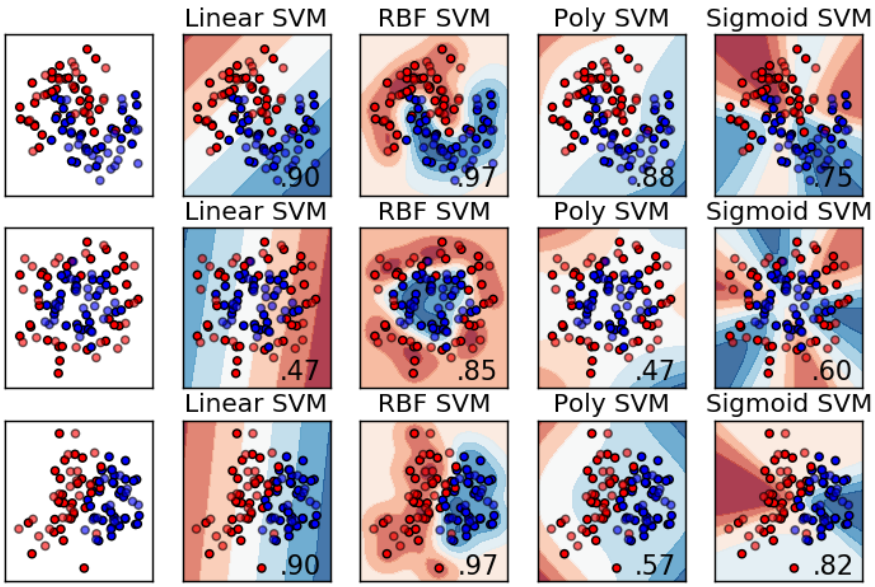


Figure 16: Visualization of the impact of the choice of kernel on the SVM decision boundaries. Accuracy is given in each plot. Taken from [92].

3.2.6 Neural Networks

Neural networks (NN) are machine learning algorithms inspired from an old understanding of the human brain. The key element of the NN are the artificial neurons. Each neuron performs a weighted sum of its inputs, add a bias, and pass it to an activation function. Common activation functions are the sigmoid and ReLU (0 if x < 0 ; x otherwise). The free parameters learned by the model are the weights and biases of the neurons [77].

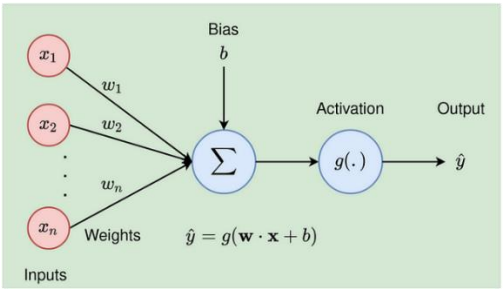
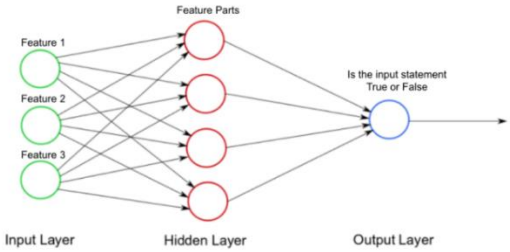


Diagram of a single neuron. Taken from [93].



A basic neural network. Taken from [94].

Figure 17: Diagrams of a node (e.g., neuron) of a neural network and a simple neural network with one hidden layer.

Stacked in layers of interconnected neurons, they transform input data through various stages, from the initial input layer, through one or more hidden layers, to the final output layer. Figure 17 shows diagrams that summarize this. A simple NN with only one hidden layer can theoretically code any function with enough neurons. In practice, it is often better to have multiple hidden layers (hence the term deep learning) to

improve performance. Each layer will use the output of the previous layer. This allows the model to create and handle more and more abstract and complex concepts as we go deeper in the model.

One of the key strengths of NN is their versatility in architecture, allowing them to be tailored for a wide range of applications. For instance, Convolutional Neural Networks (CNNs) are highly effective for image processing, while Transformers excel in handling sequential data like natural language. NN architectures are numerous, and it is a vast domain.

However, NN generally require large datasets to perform well; they are not ideally suited for small datasets as they can easily overfit, learning the noise in the training data rather than the intended patterns. Overfitting is a significant challenge with NN and regularization techniques, such as dropout, L1/L2 regularization, or early stopping, are employed to prevent this by penalizing overly complex models and promoting simpler, more generalizable models.

Despite their power and flexibility, NN have limitations. They demand substantial computational resources, especially deep networks with many layers. Also, their 'black box' nature can make it difficult to understand the exact reasoning behind their decisions. This complexity, combined with the need for large datasets and careful regularization to avoid overfitting, are important considerations in their application.

3.3 Survival analysis

Survival analysis is fundamentally designed to analyze and interpret data where the outcome of interest is the time until an event occurs. This event could be anything from the failure of a machine to the death of patients, or the time until a cancer under control starts developing again.

One unique aspect of survival data is the presence of censored data, particularly right-censored data. Right-censoring occurs when the event of interest has not happened for some subjects during the study period. For instance, if a study ends after five years, but some patients are still alive, their survival time is unknown beyond those five years. This incomplete data cannot be discarded, as it still provides valuable information about survival times. Figure 18 shows examples of censored data. Traditional machine learning models, which expect complete information for all cases, are inadequate for handling censored data. If classic machine learning algorithms were used as is on this type of data, they will be effectively trained to predict the date of censoring rather than the date to the timepoint of interest. This limitation requires the adaptation of models to accommodate the uncertainty introduced by censoring. The models must estimate survival functions while considering that some data points do not represent actual event occurrences but rather the minimum time until the event could occur [95]. Many machine learning models such as tree-based models, SVMs and boosted models have been adapted to this type of data.

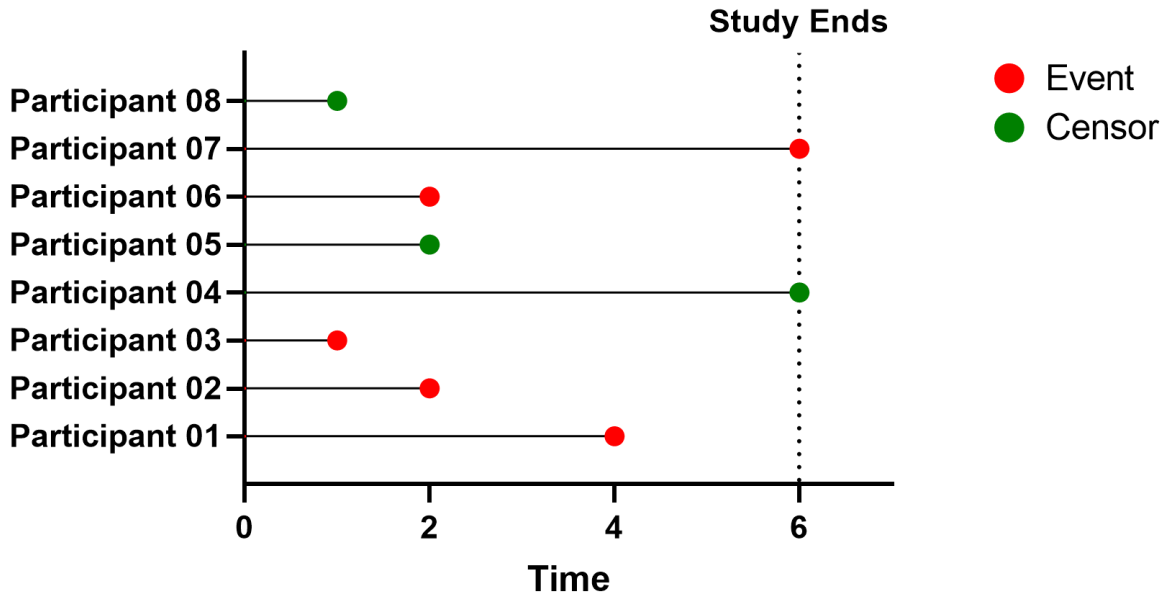


Figure 18: Diagram showing examples of censored data. Points in red represent patients who died during the study, and for which the death was observed. Patients in green are censored: we know they were alive until the time point, but we do not know if nor when they died. Taken from [96].

The Cox proportional hazards model, often referred to as Cox model, is a statistical technique used predominantly in medical research to investigate the relationship between the survival time of patients and one or more predictor variables. It models how each factor influences the outcome of the patients. A key feature of the Cox model is that it can handle multiple biomarkers (e.g., variables) at the same time, combining them into a unique score, easier to use in clinical practice, as well as insights into which factors are more significant. This model assumes that the hazard ratios (coefficients) of the features are constant over time, an assumption known as the proportional hazard assumption. The model is defined as:

$$p = \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)$$

$$h(t) = h_0(t) \times p$$

with p the partial hazard, t the timepoint for which the prediction is done, $h(t)$ the hazard of the patient at the time t estimated by the model from the m features $(x_1, x_2, \ldots, x_m)$, $h_0(t)$ the baseline hazard and β_i the coefficients of each covariate. $\mathrm{Exp}(\beta_i)$ is called the hazard ratio (HR) of feature x_i. HR equals to 1 mean no effect, higher than 1 means a positive correlation with the risk and vice versa for a HR below 1. Partial hazard p can be used as prediction if we are only interested in ranking the patients by hazard (e.g., risk).

Other models for survival analysis include parametric models like the exponential, Weibull, and log-normal models, which assume a specific form for the hazard function.

These models can provide more detailed estimates under certain conditions but require more assumptions about the data, which may not always be appropriate [97].

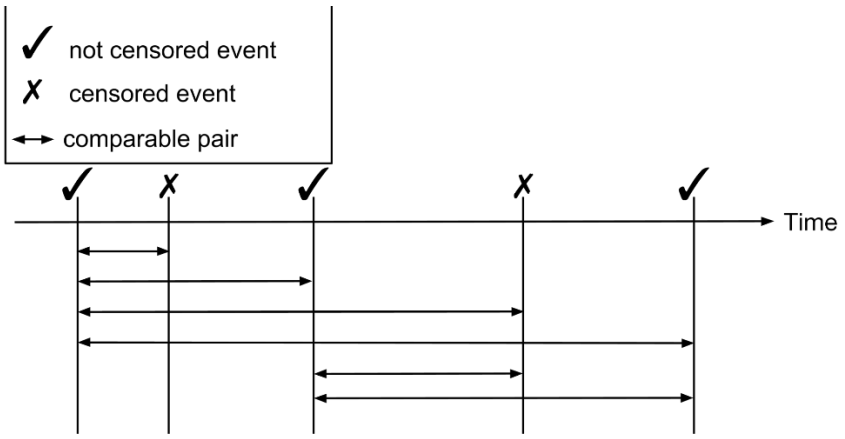


Figure 19: Diagram showing how comparable pairs are defined to compute a concordance index. Taken from [98].

Harrell's concordance index (e.g., concordance index or C-index) is an important measure in survival analysis. It gauges the model's predictive accuracy by quantifying how well the model predicts the ordering of subjects' event times. It compares the predicted and observed outcomes to see how often the model correctly predicts the order of events. For example, if a model predicts that one patient will experience an event (like a disease recurrence) before another, and this prediction is true in the actual data, it contributes positively to the C-index. A C-index of 0.5 suggests no predictive discrimination, akin to random guessing, while a C-index of 1 indicates perfect predictive accuracy. Generally, a C-index higher than 0.7 is considered acceptable in many clinical models. C-index used pairs of samples that are comparable (e.g., the sample with the lowest value should not be censored). Figure 19 illustrates this comparability of samples.

However, Harrell's C-index can become optimistically biased in scenarios with high censoring (overestimate the model's predictive ability). To address this limitation, Uno's C-index was developed and is less biased when censoring is high [99].

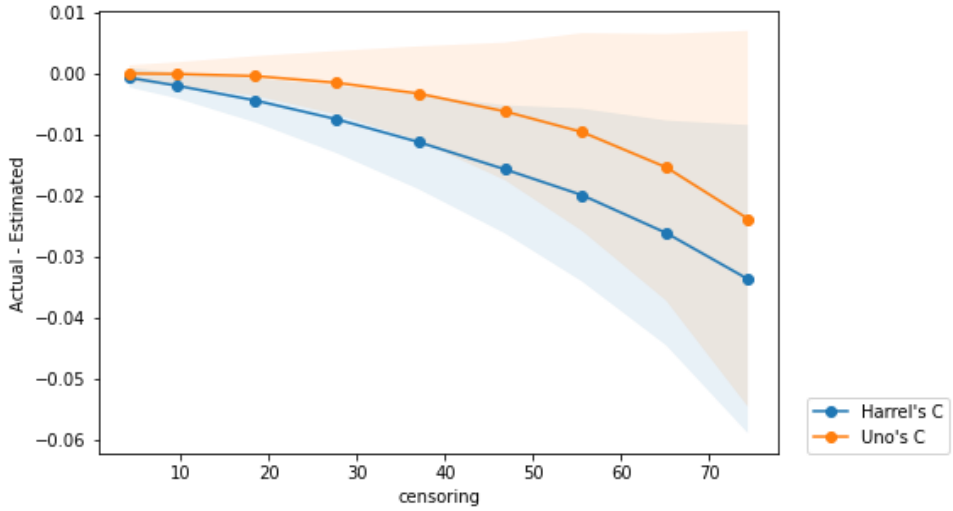


Figure 20: Estimation of the optimistic bias of Harrell and Uno C-index on synthetic data. The y axis is the C-index without any censoring minus the C-index with the censoring. The x axis is the proportion of censored samples. The higher the censoring, the higher the bias. Even if Uno handles it better, it is still biased when censoring increases.

Yet, it does not completely remove the optimistic bias. Figure 20 shows a measure of this bias on synthetic data. A better approach is to use the Receiver Operating Characteristic curve (ROC) and its associated Area Under the Curve (AUC) metric. They are used to evaluate the performance of a binary classification model, indicating its ability to distinguish between two classes by plotting the true positive rate against the false positive rate at various threshold settings. The cumulative/dynamic receiver operating characteristic curve (AUC) (e.g., time-dependent AUC, tAUC) extends the concept of AUC to time-to-event data by considering the probability of an event occurring at or before various time points [100].

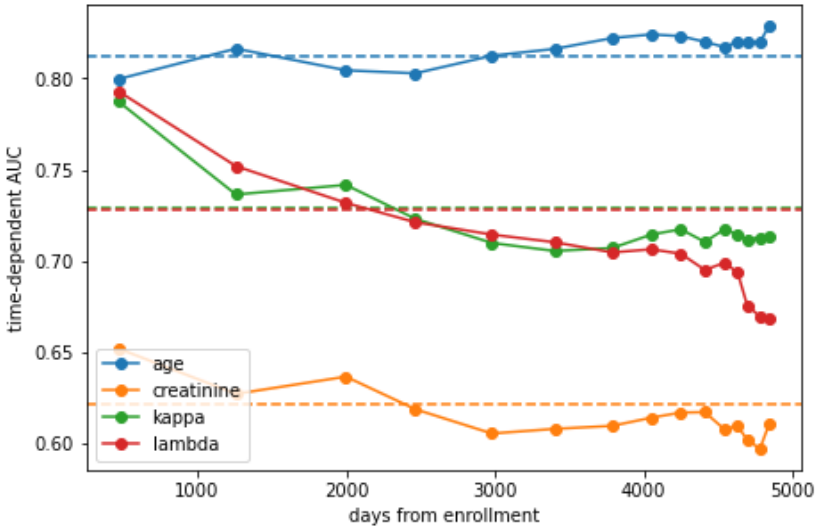


Figure 21: Time-dependent AUC (tAUC) of multiple clinical features to predict time to death. tAUC is given on the y axis, and time on the x axis. The average tAUC are given with the dashed lines. Taken from [101].

tAUC assesses the model's discriminatory power at different times, acknowledging that the ability to predict an event may change over time. For each time point, it calculates an AUC reflecting how well the model distinguishes between individuals who will experience the event before that time and those who will not. This approach allows for a more nuanced and temporally detailed assessment of model performance, particularly important in clinical settings where risk predictions are needed at different follow-up times. The measure can be averaged over time to have a unique value as a score. Figure 21 shows an example on real data.

Kaplan-Meier curves are another fundamental tool in survival analysis [102]. They provide a non-parametric way to estimate and visualize the survival function from the lifetime data. By plotting the proportion of subjects surviving against time, these curves offer an intuitive understanding of the survival experience of a group. It is often used in clinical trials to compare the survival of different groups of patients and assess if one group survives better than another. Figure 22 illustrates this.

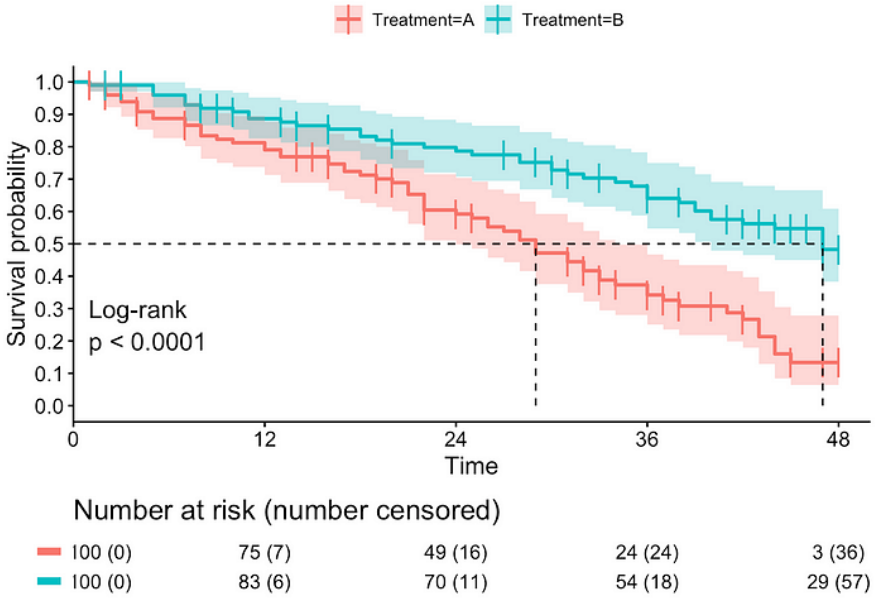


Figure 22: Example of Kaplan-Meier curves used to assess the efficacy of two treatments A and B. It is clearly visible that patients who received treatment B survived better than those who received treatment A. Taken from [103].

The logrank test is commonly used to compare the survival distributions of two or more groups. It is a non-parametric test that assesses whether there is a statistically significant difference in survival between groups, making it invaluable in analyzing clinical trial data where comparing different treatment groups is essential [104].

3.4 Feature selection

Feature selection is a critical process in the development of machine learning models [105]. It involves identifying and selecting those input variables that are most relevant

to the task at hand. This process is crucial because the quality and quantity of features directly influence the performance of machine learning models. By selecting the most pertinent features, the model can focus on the most relevant data, reducing the risk of overfitting and improving its ability to generalize to new, unseen data. Overfitting occurs when a model learns not only the underlying patterns in the training data but also its noise. This makes the model less effective at predicting outcomes for new data. Therefore, by reducing the number of irrelevant, noisy or redundant features, feature selection helps in building more robust models.

The concept of the "curse of dimensionality" is closely tied to feature selection. This term describes the phenomenon where the feature space increases exponentially with the addition of each new feature, which leads to a significant increase in the amount of data needed to ensure that the model can learn effectively. In high-dimensional spaces, data becomes sparse, and the model struggles to learn from it, necessitating an even larger dataset for training. This sparsity makes it difficult for the model to find and learn patterns in the data, which, in turn, can lead to poor performance. Numerous features in a model imply a higher requirement for training samples. As the number of features grows, the complexity of the model increases, requiring more data to capture the relationships between these features and the output variable. Without sufficient data, the model might fail to learn these relationships accurately, leading to poor predictions. This requirement of more data for more features can become a significant challenge, especially in scenarios where data collection is expensive or time-consuming. A common rule of thumb is that a model needs at least 10 examples for each feature to train effectively [106].

Several common techniques are used for feature selection. These include methods like filter, wrapper and embedded methods. Each of these methods has its strengths and is chosen based on the specific requirements and constraints of the problem at hand.

Filter methods, such as mutual information, chi-squared tests, and correlation coefficient rankings, prioritize features based on their statistical properties, independent of any machine learning model. These methods are computationally efficient and provide a straightforward means to eliminate irrelevant or redundant features based on statistical measures.

Wrapper methods, like recursive feature elimination (RFE), genetic algorithms, and sequential feature selection algorithms, assess subsets of features based on the performance of a specific machine learning model. They iteratively add or remove features and evaluate model performance, effectively ‘wrapping’ the model evaluation in the feature selection process. This approach is more computationally intensive but tends to yield features more tailored to the model’s performance.

Embedded methods, exemplified by LASSO (Least Absolute Shrinkage and Selection Operator), Elastic Net, and Decision Trees, incorporate feature selection as part of the model training process. These methods optimize the feature selection and model

training simultaneously, which can lead to more efficient and potentially more effective models, especially when dealing with high-dimensional data like medical images.

Feature engineering, another crucial aspect of preparing data for machine learning, involves creating new features from the existing ones. This process is driven by domain knowledge and is aimed at enhancing the model's performance by introducing new features that capture additional information, which might not be present explicitly in the raw data. Feature engineering can be as important as feature selection because it adds valuable information that can improve the model's ability to learn and make predictions.

Lastly, it is essential to discuss confounders. These are variables that can influence both the features and the target variable, leading to spurious associations [107]. In the context of feature selection, it is crucial to identify and appropriately handle confounders to ensure that the model captures true relationships and not those influenced by these confounding variables. The failure to account for confounders can lead to models that are biased or incorrect in their predictions. For instance, a feature confounded by the clinical center from which the patient is originating will not bring valuable information to a model that will be deployed in a unique center. One common method to identify confounders is through statistical techniques such as stratification or multivariable regression analysis, where variables are examined for their impact on both the outcome and the primary variables of interest, helping to isolate those that exert an undue influence on the relationship being studied.

3.5 Hyperparameter tuning

Hyperparameter tuning is a fundamental aspect of building and refining machine learning models. Hyperparameters are not actually part of the machine learning model. They are the parameters of the machine learning algorithm that will build the machine learning model from the data. Based on the values of the hyperparameter, the model will learn its parameters from the data. For instance, in a neural network, the weights are parameters learned during training. In contrast, the number of hidden layers in the network are hyperparameters set before training begins. Hyperparameters play a crucial role in controlling the behavior of the learning algorithm and can significantly impact the performance of the model.

The process of hyperparameter tuning involves finding the right combination of hyperparameters that results in the best performance of a model on a given task. This is not a trivial task as the space of possible hyperparameter values is often large and complex. Additionally, the optimal hyperparameters can vary significantly between different datasets and different types of models.

There are several techniques for hyperparameter tuning. The simplest one is manual tuning, where a practitioner uses their experience and intuition to choose hyperparameters. This method can be effective but is often time-consuming and relies

heavily on the practitioner's expertise.

A more systematic approach is grid search, where a predefined set of hyperparameter values is exhaustively tried. This method guarantees that the best combination in the predefined set will be found, but it can be computationally expensive, especially if the number of hyperparameters and their potential values are large.

Random search is another technique where hyperparameter values are randomly selected from a defined range. This method is often more efficient than grid search [108]. Figure 23 illustrates how this method can be more effective than the grid search.

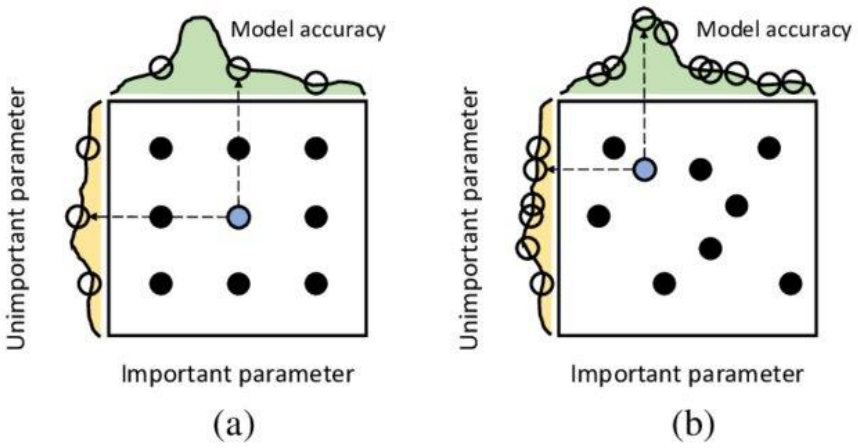


Figure 23: Comparison of grid search (a) and random search (b). With the same number of trials, random search explores the search space more effectively. Taken from [109].

An advanced technique is Bayesian optimization, which uses a probabilistic model to guide the search for the best hyperparameters. This method can be more efficient than random search as it learns from the results of previous iterations to improve the search [110].

Each of these methods has its strengths and weaknesses, and the choice of method depends on the specific problem, the computational resources available, and the experience of the practitioner. However, random search was found to be a safe go to method, that quickly finds good configurations. Since extensive hyperparameter testing can lead to overfitting (over adapting the hyperparameters to a specific set of data), random search is a good approach since it often finds decent combinations with a minimal number of trials.

3.6 Evaluation

The evaluation of models is a critical step to assess their performance and applicability in real-world scenarios. The necessity of evaluation stems from the fundamental goal of machine learning: to create models that not only learn from data but also generalize well to new, unseen data. This generalization capability is what makes a model truly useful, as it indicates the model's ability to make accurate predictions or decisions

beyond the specific examples it was trained on.

In classification tasks, where the goal is to assign each input to one of several categorical classes, metrics like accuracy, precision, recall, and the F1 score are commonly used. Accuracy measures the proportion of correctly predicted instances among the total instances, while precision and recall focus on the model's performance in predicting a specific class. The F1 score provides a balance between precision and recall, offering a single metric that considers both false positives and false negatives.

In regression tasks, which involve predicting continuous values, different metrics are used, like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. MSE measures the average squared difference between the predicted and actual values, emphasizing larger errors. MAE offers a more straightforward interpretation by calculating the average absolute difference, and R-squared provides a measure of how well the observed outcomes are replicated by the model.

To effectively evaluate these metrics, data is typically split into three sets: training, validation, and testing. The training set is used to train the model. Its performance is then assessed on the validation set. It is used to test different models, preprocessing and hyperparameters. The testing set, on the other hand, is used to assess the final model's performance, offering an unbiased evaluation of its generalization capability. The distinction between the validation and test sets is crucial. Repeated use of the validation set can inadvertently lead to overfitting, where the model becomes overly tailored to the validation data. To counteract this, the test set is used only once, providing a final, unbiased assessment of the model's performance on new data.

Data leakage is a critical issue in model evaluation, occurring when information from outside the training dataset is inadvertently used to create the model. This can lead to overly optimistic results during training and validation but poor performance in real-world application. Preventing data leakage involves careful data handling and bug control, ensuring that the model is never exposed to test data during training [111].

However, dividing the data available in three sets is sometimes not a good solution [112]. If the data available is limited, the train, validation and test sets will be too small to effectively train and test the model. In such cases, it is better to use cross-validation. It is a robust method for assessing a model's performance when dealing with limited data. It involves dividing the dataset in a train and a validation sets. These sets are used to train and test the model and its score is saved. Then, the whole dataset is split again in new train and validation sets, the model is retrained from scratch on the new train set, and its score on the new validation set is saved. This is repeated multiple times, and the average score on all the validation set is used as a final score.

There are several ways data can be split during cross-validation, each with its unique advantages. One of the most common methods is k-fold cross-validation, where the dataset is divided into 'k' equal parts, or folds. In each iteration, one fold is used for

validation while the others are used for training, ensuring that every data point gets to be in the validation set exactly once. This method is particularly effective in providing a robust estimate of the model's performance, especially when data is limited.

Another method is the Monte Carlo cross-validation, which randomly splits the dataset into training and validation sets multiple times. This approach differs from k-fold cross-validation in its randomness and the fact that data points may appear in the validation set multiple times or not at all across different iterations. This method provides a more optimistically biased estimate of the performance than K folds, but is more reliable when comparing models on the same task, since more folds can be created, reducing the variability of the final score.

Stratification is an approach often used in conjunction with these methods, especially in classification tasks where class imbalance might be an issue (one class of samples is more present than another). Stratified sampling ensures that each fold or split maintains the same proportion of classes as the original dataset. This is crucial for preventing biased estimates of the model's performance, especially in scenarios where one class significantly outnumbers the others.

Additionally, techniques like leave-one-out cross-validation, where each data point is used once as a single validation set while the rest of the dataset serves as the training set, can be beneficial for small datasets. However, this method can be computationally expensive for larger datasets.

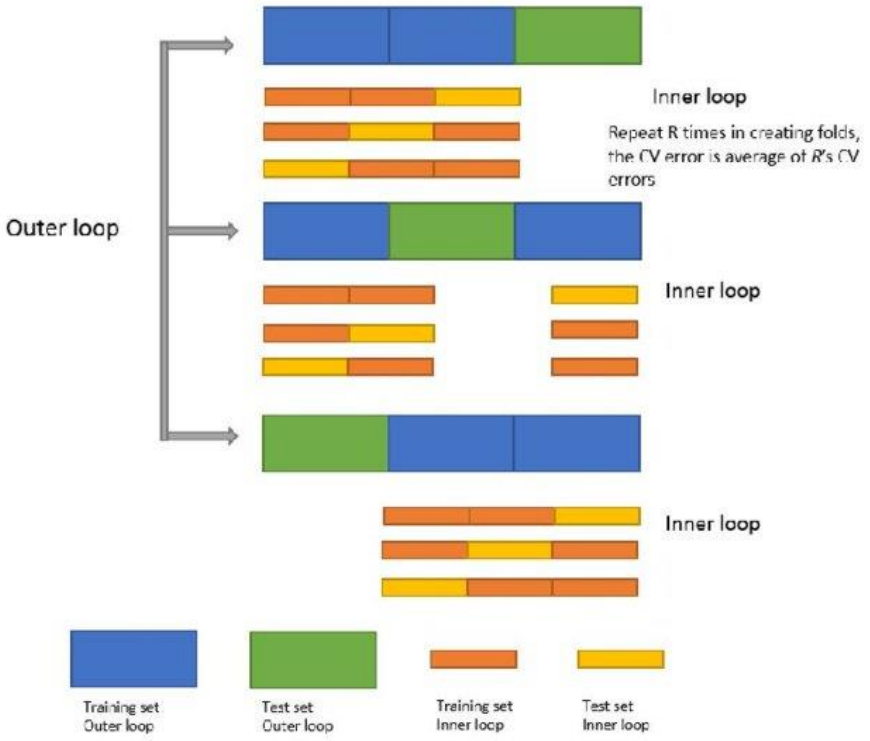


Figure 24: Diagram showing how nested cross validation works. Taken from [113].

Nested cross-validation takes this concept further to include hyperparameters tuning. In this approach, an inner cross-validation loop selects the best hyperparameters and/or model, while an outer cross-validation loop provides an unbiased evaluation of the model's performance. This nested structure ensures that the evaluation of the model's performance is not influenced by the specific selection of hyperparameters, leading to a more reliable assessment. While it is computationally intensive, this method is considered the gold standard approach when dealing with a limited amount of data. Figure 24 shows how nested cross validation can be implemented.

One important aspect of model score is the significance of the score. It is possible on some tasks and some datasets to achieve high score only by chance, without any real predictive power. For instance, on small survival datasets, it is common to have random variables with decent C-index [114]. To compensate for this, scores need to be tested. A common way to do this is to use a permutation test [115].

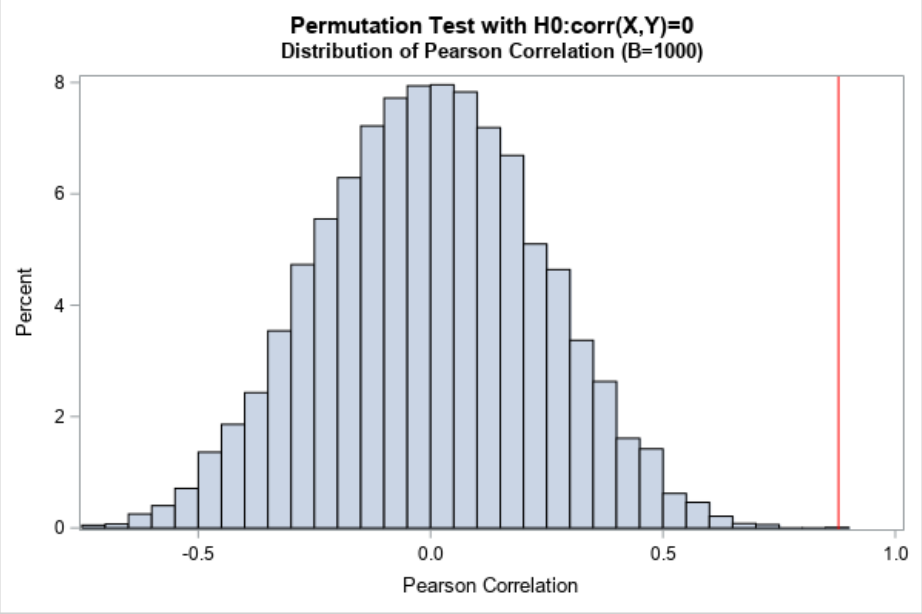


Figure 25: Example of a permutation test to assess the significance of a Pearson correlation. The distribution of the permuted features correlation is given in blue, and the correlation of the non-permuted one in red. This correlation is statistically significant as it is extremely unlikely to have a correlation this high only by chance. Taken from [116].

The goal of this test is to assess the likelihood of having a given score just by chance. In other word, we assess how exceptionally predictive a model, or a feature is compared to random predictions. If it is very unlikely to reach a given score with just chance, then it is reasonable to assume that the model or feature has actual predictive value. Permutation is a robust and reliable way to evaluate this probability. By randomly shuffling the target value, it breaks the relationship between the covariates and the outcome. These permuted values are evaluated, and the corresponding score is saved. Repeating this process hundreds of times effectively emulates the absence of predictive value. A distribution of scores of random predictions is thus created.

Comparing the score of the non-permuted model to this distribution allows for the estimation of the likelihood of not being predictive. This probability is defined has the proportion of permuted scores greater than or equal to the non-permuted score. Figure 25 shows an example of permutation test.

3.7 Multiple testing

In the realm of scientific research, particularly in statistical analysis, understanding the concept of false positives is crucial. A false positive occurs when a test incorrectly indicates the presence of a condition, such as a disease in medical testing, when it is actually not present. This can be especially problematic in studies where numerous hypotheses are tested simultaneously. For instance, a test saying that a feature is prognostic of a condition while it is not.

This leads us to the multiple testing problem, a challenge that arises when multiple statistical tests are made simultaneously. In such cases, the likelihood of encountering at least one false positive increases with the number of tests conducted. Imagine testing a hundred independent hypotheses, each at a 5% significance level. The chance of observing at least one false positive is no longer just 5% but substantially higher. This inflation of false positive rates can lead to misleading conclusions if not properly addressed.

To mitigate the risks associated with the multiple testing problem, correction methods are employed. One of the simplest and most widely used methods is the Bonferroni correction. This technique adjusts the significance threshold based on the number of tests performed. For instance, if ten hypotheses are tested, the Bonferroni correction would divide the standard significance level (usually 0.05) by ten. This stricter criterion helps control the rate of false positives but can be overly conservative, potentially leading to false negatives or missed discoveries.

In statistical terms, we often focus on controlling the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR) [117]. FWER is the probability of making one or more false discoveries among the rejected hypotheses, while FDR is the expected proportion of false discoveries among the rejected hypotheses. Controlling FWER, as in the Bonferroni correction, ensures a low probability of any false discovery, but can be too stringent in cases where many hypotheses are tested. FDR control, on the other hand, offers a balance, allowing for a controlled proportion of false discoveries, which can be more suitable for exploratory research where some false leads are acceptable [118]. Examples of techniques to control the FDR are q-values, the Benjamini-Hochberg procedure and its improved version called the two-stage linear step-up [119].

Chapter 4

Lymphoma

Cancer, a disease characterized by abnormal and uncontrolled cell growth, is a leading cause of mortality worldwide, affecting millions annually. Among its various forms, lymphoma, a cancer form originating in the lymphatic system, presents a significant public health challenge. Epidemiologically, lymphomas represent a large portion of hematologic cancers, with Non-Hodgkin Lymphoma (NHL) being more prevalent than Hodgkin Lymphoma. NHL represents nearly 3% of all cancer diagnoses and deaths worldwide [120]. This chapter focuses on two specific types of NHL: Follicular Lymphoma (FL) and Diffuse Large B-Cell Lymphoma (DLBCL), each exhibiting unique clinical and biological characteristics. Accurate staging of these cancers is crucial, as it determines the disease's extent and guides therapeutic decisions, impacting patient survival and quality of life. This introductory chapter will establish the basic concepts necessary to understand the work presented in this thesis. It will cover the broader context of cancer and delve into the specifics of lymphomas, particularly focusing on FL and DLBCL. Additionally, this chapter will detail various patient staging methods, highlighting their critical role in the diagnosis and management of these cancers.

4.1 Cancer general principles

Cancer, a complex and multifaceted disease, is marked by the uncontrolled proliferation of aberrant cells, driven by multiple cellular mechanisms [121]. The human body, comprising trillions of cells, maintains a meticulous equilibrium of cellular growth, division, and apoptosis (programmed cell death). Cancer disrupts this balance, primarily due to mutations in cellular DNA [122]. These mutations, frequently occurring in proto-oncogenes and tumor suppressor genes, are pivotal in the malignant transformation of normal cells [123]. These mutations can occur due to various factors, including genetic predisposition, environmental exposures (e.g., radiation, chemicals, viral infections, pollution, ...), and lifestyle [124]. As these mutated cells continue to divide uncontrollably, they can form a mass called a tumor. Tumors can be benign (non-cancerous) or malignant (cancerous). Malignant tumors can invade surrounding tissues and spread to other parts of the body through a process known as metastasis, whereas benign tumors stay localized and do not spread [125].

It was estimated that in 2018, there were 18.1 million new cancer cases and 9.6 million deaths from cancer worldwide [2]. In 2019, cancer was the first cause of premature death in 57 countries, and the second after cardiovascular disease in 70 [1]. Figure 26 is a map showing the rank of cancer in each country.

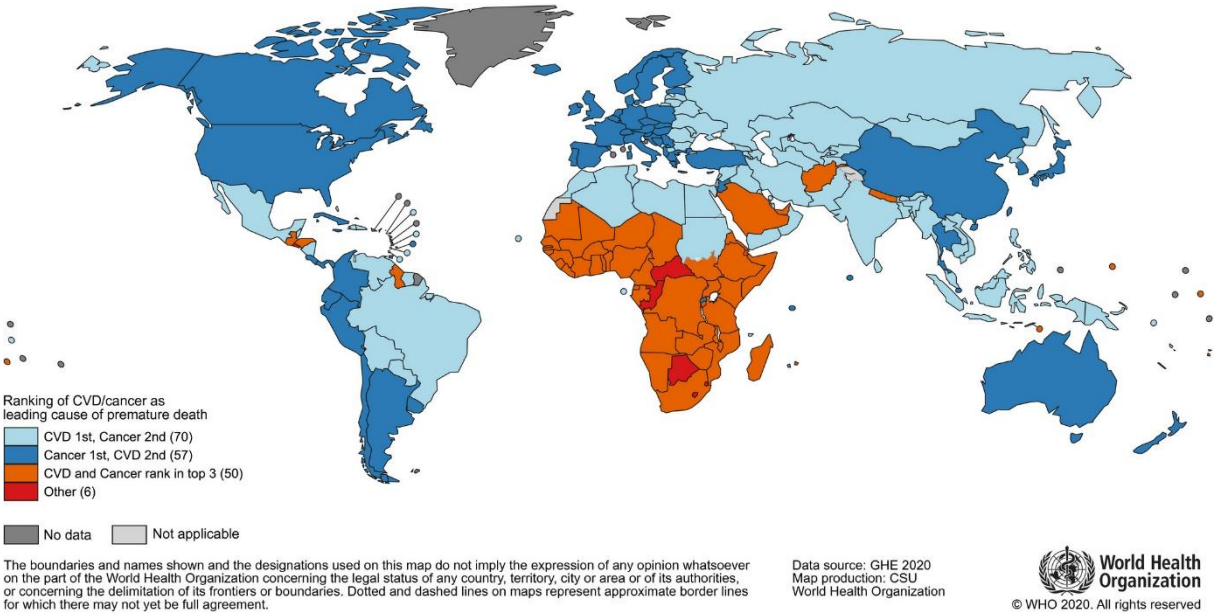
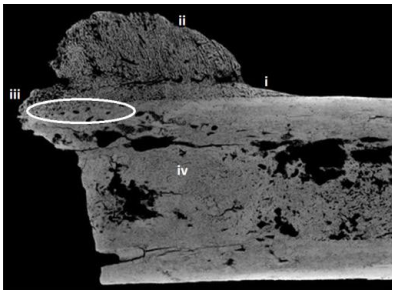


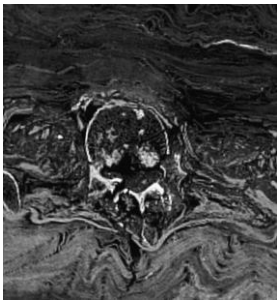
Figure 26: Global distribution of cardiovascular diseases and cancer as leading causes of death, with color-coded rankings by country, according to World Health Organization data from 2020. Taken from [1].

Cancer, commonly perceived as a contemporary ailment, has historical roots in both humans and other species [126]. Figure 27 shows several pieces of evidence of this. However, cancer incidence increased in the recent years [127]. This escalation is multifactorial, primarily caused by factors such as demographic shifts towards an aging population, advances in diagnostic techniques, changes in lifestyle patterns characteristic of modern society, and increased exposure to environmental pollutants [124].

Radiograph showing metastatic cancer traces in dinosaur fossil from the Jurassic era. Taken from [128].



micro-CT image of an osteosarcoma in a foot bone of a human relative who lived 1.7 million years ago. Taken from [129].



CT scan of vertebra of a human mummy dating around 2000 B.C showing bone lesions caused by prostate cancer metastasis. Taken from [130].

Figure 27: Examples of cancer traces found on fossil of dinosaur, human relative and a human mummy, testifying that cancer is an old disease not specific to humans.

The evolution of cancer research spans from ancient Egyptian and Greek texts to modern discoveries. Notably, the term 'cancer' itself has historical roots, with the first description likening tumors to a crab, an analogy drawn by the ancient Greek physician Hippocrates due to the crab-like spread of the disease [131]. Hippocrates' contributions remain significantly relevant in contemporary times. He reported two key insights: firstly, that cancer is a systemic ailment, impacting the entire body rather than just a single organ; and secondly, that effective cancer treatment requires restoring balance to the entire organism with a comprehensive, multidisciplinary strategy, beyond merely removing the tumor [132]. The identification of oncogenes and tumor suppressor genes stands as a cornerstone in the history of cancer research [133]. Groundbreaking experiments, like the identification of the Philadelphia chromosome in chronic myelogenous leukemia, have significantly shaped contemporary understanding of cancer biology [134].

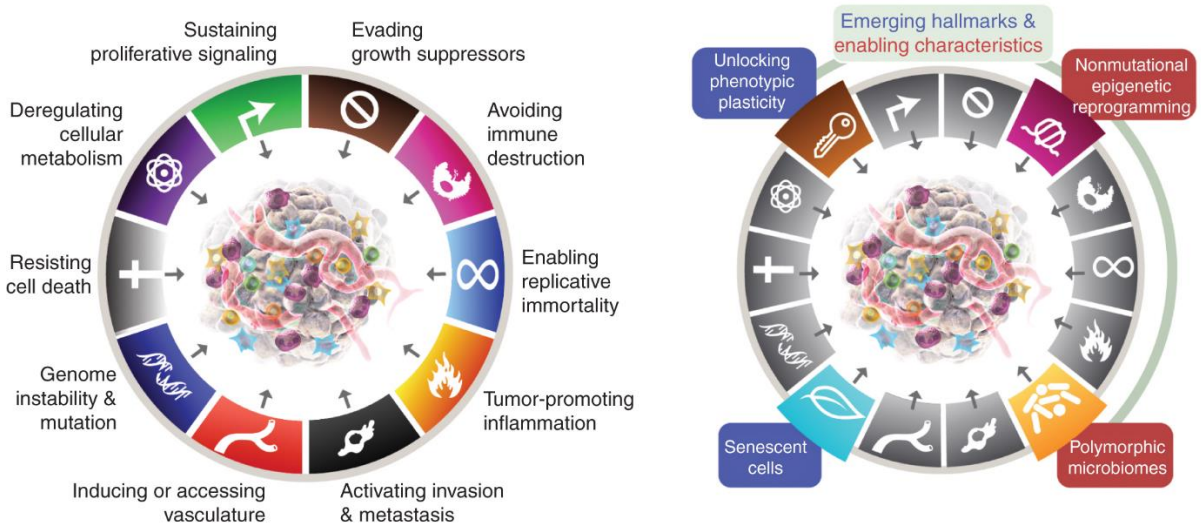


Figure 28: Latest version of the hallmark of cancer. Taken from [135].

The concept of "hallmarks of cancer," introduced by Hanahan and Weinberg [136], offers a superior understanding of cancer pathophysiology. It defines a set of characteristics that collectively define the transformative process that normal cells undergo to become cancerous. These hallmarks encompass key processes such as sustained proliferative signaling, evasion of growth suppressors, resistance to apoptosis, replicative immortality, angiogenesis, metastasis, metabolic reprogramming, and immune evasion. For example, sustained proliferative signaling enables incessant cellular division, while evasion of immune destruction allows cancer cells to circumvent immunological defense mechanisms. Figure 28 presents the latest version of the list of identified hallmarks of cancer. A cell does not necessarily have to exhibit all the hallmarks of cancer to become cancerous and different types of cancers may display different combinations of these hallmarks. Some of these characteristics are considered fundamental for the transformation of a normal cell into a cancerous one, like sustained proliferative signaling and evading growth suppressors, but the specific combination can vary.

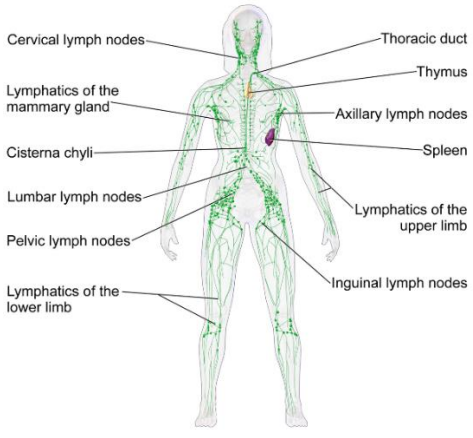
Cancer's lethality predominantly arises from its capacity to compromise vital organ function, either through direct invasion or distant metastasis. For instance, lung cancer can obstruct airways or disrupt pulmonary structure impairing breathing and can result in post-obstructive pneumonia, while brain cancer can directly impair neurological functions [137], [138], [139]. Beyond physical symptoms, the psychological and social impact of a cancer diagnosis is profound, often entailing considerable mental distress and uncertainty for the patient and its relatives [140], [141], [142].

Cancer diagnosis typically involves imaging modalities (e.g., PET, CT, MRIs) [143] and biopsies, where a small sample of tumor tissue is extracted and examined [144]. Recent strides in molecular diagnostics, such as genomic and proteomic analyses, have enabled more nuanced cancer characterizations, fostering personalized treatment strategies [145].

The treatment options for cancer are diverse, tailored to cancer type, stage, location, and the patient's health status [146], [147]. Key modalities include surgery [148], radiation therapy [149], chemotherapy [150], immunotherapy [151], and targeted therapy [152], each undergoing significant advances. For example, surgical procedures have evolved towards minimally invasive approaches, and targeted therapies now more precisely attack cancer cells based on genetic mutations. While surgery aims to eliminate tumors, radiation and chemotherapy focus on eradicating or stopping the proliferation of cancer cells, and immunotherapy strengthens the body's innate immune response against cancer.

4.2 Lymphomas

Lymphomas, a heterogeneous group of hematological malignancies, originate within the lymphatic system, an integral part of the body's immune defense mechanism [153].



Schematic of the lymphatic system. Taken from [154].

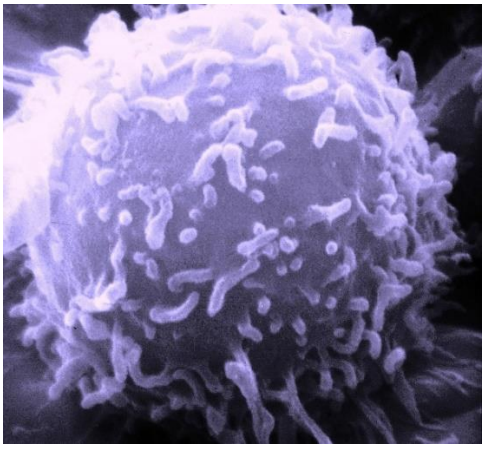


Image of a human lymphocyte. Taken from [155].

Figure 29: Diagram of the human lymphatic system and the image of a human lymphocyte taken via electron microscopy.

This system encompasses a vast network of lymphatic vessels, similar to blood vessels, responsible for circulating lymph, a clear fluid, throughout the body. Components of this system include lymph nodes, spleen, thymus, and bone marrow, all playing an essential role in mediating immune responses [156]. Figure 29 shows a diagram of the lymphatic system and the image of a lymphocyte.

Lymphocytes are a type of white blood cell. They are primarily involved in the body's adaptive immune response, responsible for recognizing and reacting to specific pathogens, such as bacteria, viruses, cancerous cells, and foreign substances. Lymphocytes are mainly divided into two subtypes: B cells and T cells. B cells are responsible for antibody production. They recognize pathogens and produce specific antibodies that bind to antigens, helping to neutralize and eliminate them. T cells, on the other hand, are crucial for cell-mediated immunity. They are further subdivided into helper T cells (which assist other immune cells), cytotoxic T cells (which kill infected or cancerous cells), and regulatory T cells (which help modulate the immune response and maintain tolerance) [157].

Dysregulation of these lymphocytes, particularly under the influence of genetic mutations and environmental factors, precipitates the development of various lymphomas. For instance, in follicular lymphoma, a transformation in B cells is frequently observed, stemming from aberrations in normal apoptotic pathways, resulting in their uncontrolled growth and proliferation [158].

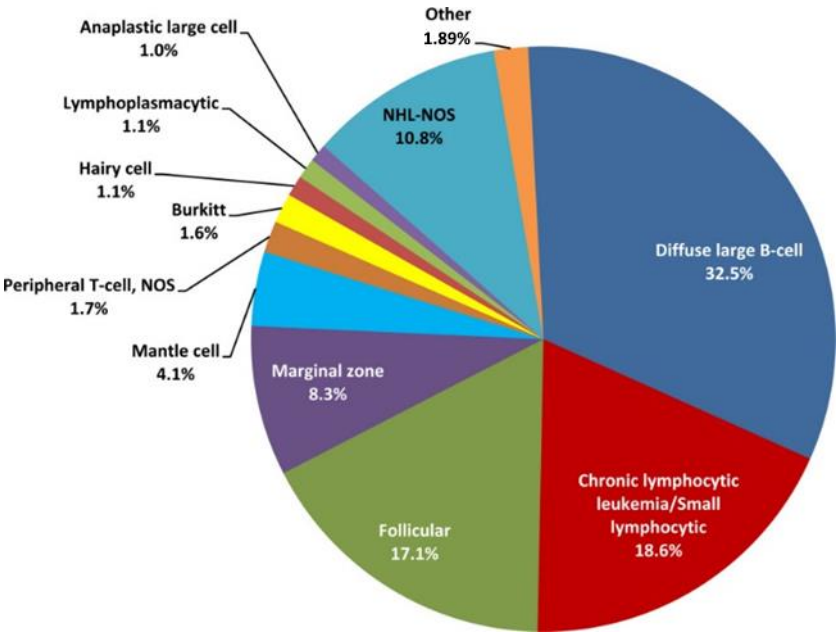


Figure 30: Repartition of non-Hodgkin lymphoma by subtypes. Adjusted from [159].

Lymphomas are primarily categorized into Hodgkin lymphoma (HL) and non-Hodgkin lymphoma (NHL), with the latter encompassing different subtypes [160]. This includes the aggressive diffuse large B-cell lymphoma (DLBCL) and the comparatively indolent follicular lymphoma (FL). The next section of this chapter describes these specific NHL

subtypes. Figure 30 presents the subtype repartition of NHL.

The exact cause of lymphoma is not completely understood, but factors such as genetic predisposition, exposure to certain chemicals or radiation, and some infections (such as the Epstein-Barr virus) are believed to increase the risk of developing lymphoma [161], [162], [163], [164].

The symptoms of lymphomas vary with the disease type and stage. Common manifestations include swelling of lymph nodes and night sweats, whereas advanced lymphomas might present with organ-specific symptoms due to compression. Notably, systemic 'B symptoms' (fever, weight loss, night sweats) are more pronounced in later stages [165].

Diagnostic approaches for lymphomas go beyond simple biopsy and include imaging modalities, blood analyses, and advanced techniques like immunophenotyping and genetic profiling [166], [167]. Positron emission tomography (PET) scans play an important role for staging, and the advent of liquid biopsies offers a minimally invasive alternative for monitoring treatment efficacy [168]. Figure 31 shows how response to treatment can be assessed with PET imaging.

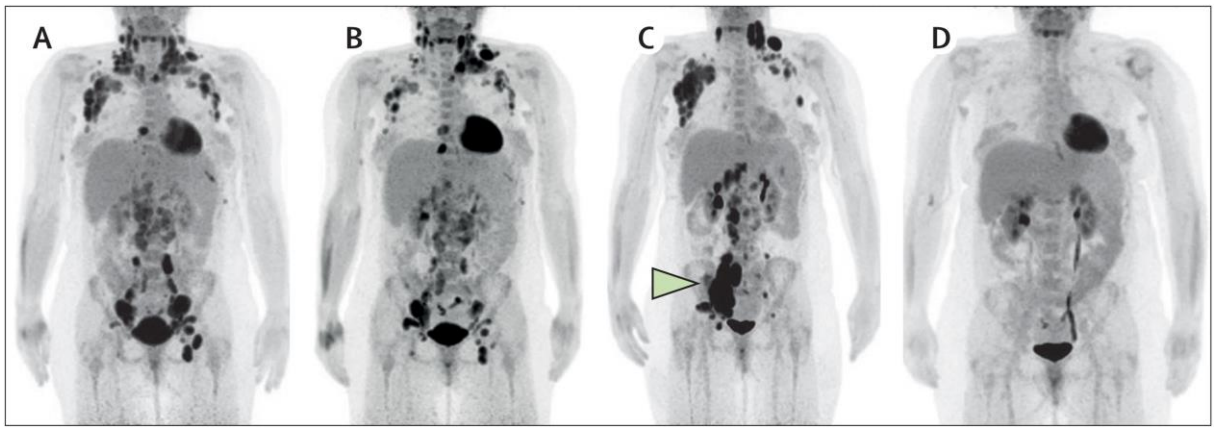


Figure 31: Maximum intensity projection of ^{18}F-FDG PET scans of DLBCL patients at different timepoint, showing complete metabolic response on the final image. Taken from [169].

Therapeutic strategies for lymphomas have evolved substantially. Traditional methods like chemotherapy and radiation therapy are now augmented by targeted and immunotherapies [170], [171]. Chimeric antigen receptor (CAR) T-cell therapy has notably transformed treatment paradigms, particularly for refractory lymphomas [172]. Recent clinical trials have brought to the forefront novel therapeutic agents, such as bispecific antibodies and immune checkpoint inhibitors, expanding patient-specific treatment options and heralding a new era of personalized oncology [173], [174].

4.3 Follicular & Diffuse Large B cell lymphomas

Follicular and Diffuse Large B Cell Lymphomas (FL and DLBCL) originate from B cells but display divergent pathophysiological and clinical characteristics. DLBCL is an aggressive subtype while FL is a slow-growing form and untreated FL patients or patients who have refractory or relapsed disease have a much longer overall survival than DLBCL patients [175]. Figure 32 displays histopathological slices of the two types of NHL.

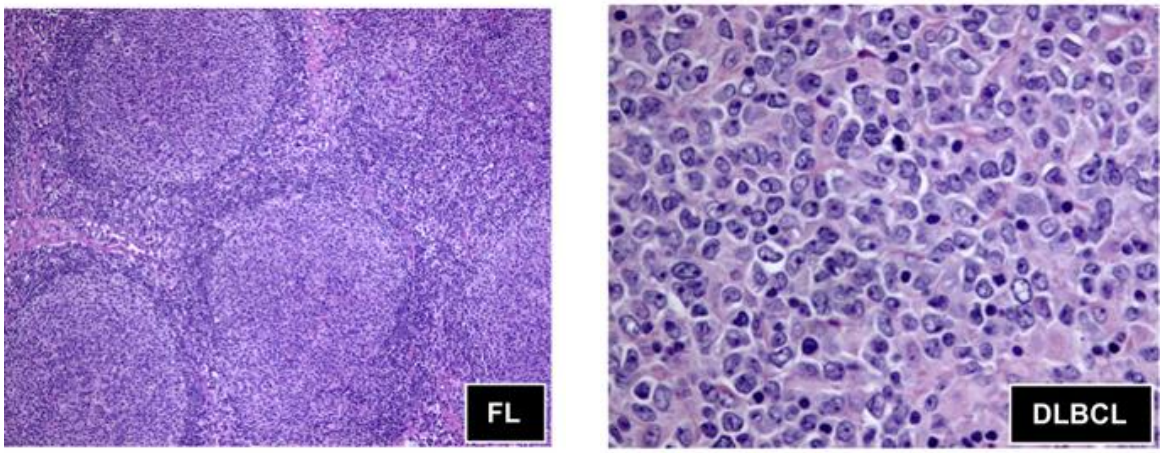


Figure 32: Histopathological slices of biopsies of Follicular Lymphoma (FL) (left) and Diffuse Large B Cell Lymphoma (DLBCL) (right). Taken from [176].

FL, an indolent subtype, emerges from the B cells of the follicular center in the lymph nodes [158]. Its growth pattern mimics that of normal lymph node follicles, which is reflected in its name. The hallmark genetic abnormality in FL is the chromosomal translocation t(14;18), which leads to the overexpression of the BCL2 gene, integral to the inhibition of apoptosis [177]. Dysregulation of this gene facilitates prolonged survival of B cells, thereby creating an environment conducive to further oncogenic transformations. A noteworthy trait of FL is its potential evolution into a more aggressive form, commonly into DLBCL [178].

DLBCL, in contrast, is characterized by the rapid proliferation of atypical large B cells across lymph nodes. Its genetic landscape is heterogeneous, marked by various genetic alterations that drive its aggressive nature. In particular, mutations in genes like MYC, BCL6, and EZH2 have been implicated, each playing a unique role in the growth and development of lymphoma [179].

Therapeutic strategies for these lymphomas, while overlapping in some aspects such as the employment of the monoclonal antibody Rituximab targeting the CD20 antigen on B cells, differ significantly. In DLBCL, Rituximab combined with the CHOP chemotherapy regimen (Cyclophosphamide, Hydroxydaunorubicin, Oncovin, and Prednisone) has significantly improved patient outcomes [180], [181]. Moreover, novel treatments like CAR T-cell therapy are gaining traction, especially in refractory DLBCL cases [172].

4.4 Patient prognosis assessment

In oncology, particularly for Follicular Lymphoma (FL) and Diffuse Large B Cell Lymphoma (DLBCL), patient staging is crucial for selecting effective treatment strategies and prognostic outcomes. Staging meticulously evaluates the cancer's spread within the body and is pivotal for guiding therapeutic decisions, prognosing, and comparing outcomes across treatment modalities [182]. The ^{18}F-FDG PET/CT scan is considered the gold standard for DLBCL staging and is recommended for FL [180], [181]. This technique enables tumor volumes to be measured, offering valuable insights into disease progression.

The concept of extranodal sites, namely locations outside the lymph nodes invaded by tumoral lymphoma cells, is critical. Their presence often indicates an advanced disease stage, directly influencing treatment decisions. Thorough identification and evaluation of their volume are useful for a comprehensive characterization of the disease's extent.

The Eastern Cooperative Oncology Group (ECOG) performance status is a scale used to assess how a patient's disease affects their ability to function on a daily basis. It ranges from 0, indicating fully active, to 5, denoting death [183]. Figure 33 lists the criteria of the different stages. This status is instrumental in understanding how well a patient can endure various treatments and is often used alongside other staging methods to personalize treatment plans.

Grade	ECOG
0	Fully active, able to carry on all predisease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature—for example, light house work, office work
2	Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours
3	Capable of only limited self-care, confined to bed or chair more than 50% of waking hours
4	Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair
5	Dead

ECOG, Eastern Cooperative Oncology Group.

Figure 33: Eastern Cooperative Oncology Group (ECOG) performance status. Taken from [184].

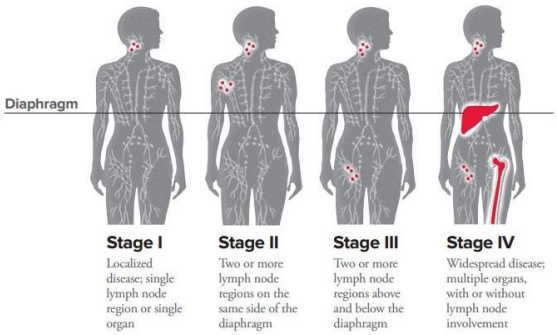


Figure 34: Ann Arbor staging. Taken from [185].

The Ann Arbor staging system, the standard method used in lymphoma staging, categorizes the disease into four stages. It considers factors such as the number of lymph nodes involved, whether the lymphoma has spread to both sides of the diaphragm and if organs are invaded [180], [181]. Figure 34 shows the staging rules based on the location of the tumor sites.

The Lugano classification, which is an evolution of the Ann Arbor system, incorporates the use of modern imaging techniques like CT and PET scans. This classification provides a more detailed assessment of the disease, particularly regarding the identification of extranodal involvement [186].

The Deauville score grades the response to treatment in lymphoma patients [187]. It ranges from 1 to 5, with lower scores indicating a better response. This score has become a key tool in evaluating how well a patient responds to therapy and in making decisions about continuing, changing, or stopping treatment.

For prognosis, the International Prognostic Index (IPI) and its variations – age-adjusted IPI (aaIPI), National Comprehensive Cancer Network (NCCN) IPI, and FLIPI (Follicular Lymphoma International Prognostic Index) – are essential tools. The IPI accounts for factors like age, stage, ECOG status, extranodal involvement, and serum lactate dehydrogenase (LDH) levels to predict outcomes. The aaIPI is tailored for younger patients, while the NCCN IPI offers a more nuanced approach for Diffuse Large B Cell Lymphoma [188], [189], [190]. FLIPI, specifically designed for Follicular Lymphoma, incorporates different parameters relevant to this subtype [191].

Total Metabolic Tumor Volume (TMTV) is an emerging metric in lymphoma staging. It quantifies the total volume of metabolically active tumor using PET scans, providing a more comprehensive view of the tumor burden in the body. It was found prognostic of the outcome of DLBCL patients in multiple studies [192], [193], [194]. In FL, its prognostic value was also recently highlighted [195], [196], [197].

More recently, another measure, Dmax, or the maximum distance between two lesions was introduced [198]. It helps in assessing the spread of lymphoma. It is particularly useful in understanding the anatomical extent of the disease. Multiple studies have found that it was a prognostic factor in DLBCL, but its prognostic capabilities remain to be confirmed in FL [199]. Figure 35 shows two DLBCL patients with similar TMTV but different Dmax.

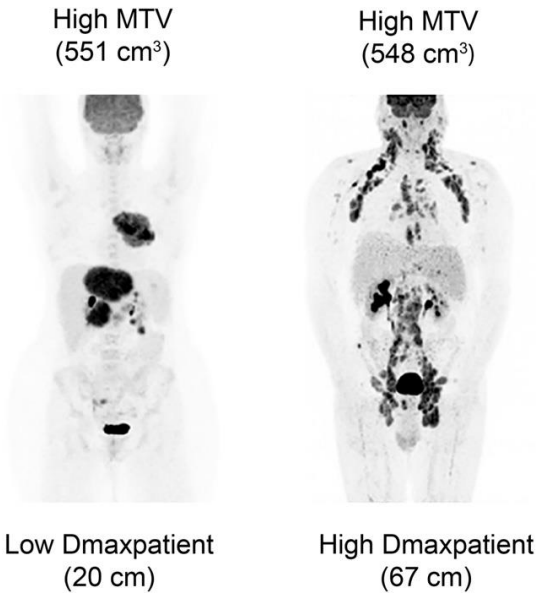


Figure 35: Maximum intensity projection of ^{18}F-FDG PET scans of DLBCL patients with similar TMTV but different Dmax [198].

PET/CT scans play a pivotal role in staging FL and DLBCL patients, predominantly by delineating the location and volume of lesions. However, this approach might overlook additional prognostic information embedded within the scans. Beyond the lesions and their immediate surrounding, PET/CT images could contain valuable information about the disease's characteristics and the patient's overall health status, including signs of comorbidities. In particular, metabolic activity measured by the 18F-FDG scans all over the body and tissue density obtained from the CT systematically associated with the PET scan are only used to detect lesions and might be underutilized in clinical practice.

A critical challenge in this context is the effective integration of this detailed information from the lesion level to a holistic patient-level perspective. Moreover, patients with similar Total Metabolic Tumor Volume (TMTV) and maximum tumor diameter (Dmax) still demonstrate visual differences in tumor burden and varying clinical outcomes, as shown in Figure 36. These observations suggest that factors beyond the quantifiable volumes and location of tumors, potentially captured in the imaging data, might significantly impact patient prognosis. Developing methodologies to detect and interpret this complex information is essential for advancing staging and prognostic accuracy in FL and DLBCL.

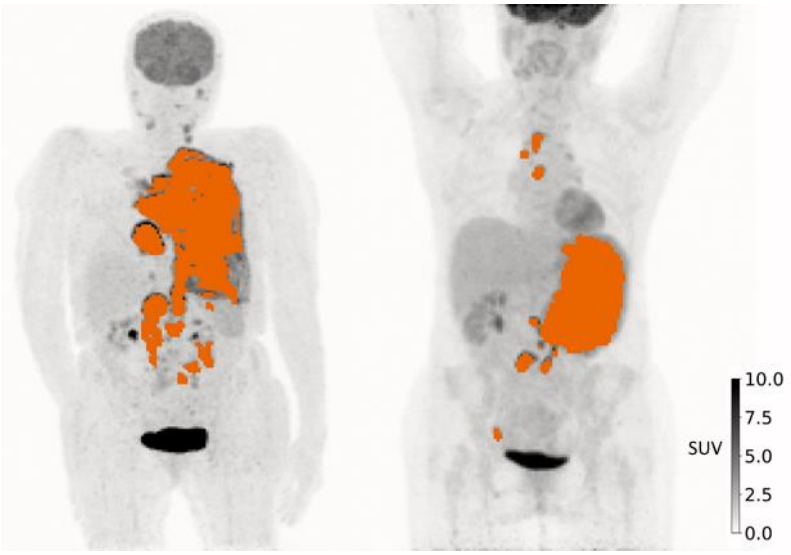


Figure 36: Maximum intensity projection of ^{18}F-FDG PET scans of DLBCL patients with similar TMTV (left: 761 cm^3, right: 819 cm^3) and Dmax (left: 0.30 m, right: 0.28 m). The lesions are represented in orange. Taken from [200].

Different studies have thus been conducted to find and quantify new prognostic information in the PET images of lymphoma patients. Zanoni et. al. [201] published a review on how PET/CT is used in NHL. They showed that the search for new image biomarkers and radiomic signatures is an active field. For instance, they reported multiple studies that found radiomic features to be correlated with patient outcome and response to treatment. However, they concluded that more data, standardization, and external validation are needed before these findings could be used in the clinic. It is recognized that small dataset size is frequent in radiomic and that studies with less

than 100 patients are common and subject to a significant risk of bias [202], [203]. Moreover, most of these findings remains unexplained and the prognostic biological information encoded by somehow cryptic features is not understood, as mentioned in chapter 2 [74]. Furthermore, despite numerous published studies, few radiomic models have been translated into the clinic and no radiomic features other than TMTV are being used to stage DLBCL or FL patients [64], [65].

For these reasons, a strong emphasis was put on evaluating the robustness and interpretability of the results in this thesis. Two cohorts of NHL patients, one of DLBCL and one of FL patients, with almost 400 patients each were used for analysis and validation of the findings. A significant part of the time was spent developing and validating a reliable biomarker selection pipeline, and substantial efforts were made to simplify and understand the image biomarkers found to be of prognostic value.

Section II

Original developments

Chapter 5

Investigating the role of spleen involvement in DLBCL prognosis

5.1 Introduction

A first straightforward way to find new relevant image-based biomarkers in PET/CT images is to identify new features visually and/or based on some intuition and define them empirically. This is actually the way Dmax, which characterizes the dissemination of the disease, has been discovered. By looking at multiple examples of patients that are similar according to currently used biomarkers, but with different outcomes, we hoped to capture some differences visually, the human visual system being excellent at finding patterns. This is the first methodology I adopted during the PhD, which also allowed me to get familiar with the PET/CT images of lymphoma patients.

One of the many handcrafted features I tested is the tumor burden fragmentation. As shown at the very end of chapter 4, DLBCL patients with almost identical TMTV and Dmax can still have significant differences in their tumor burden spatial distribution. Part of this difference can be captured by the surface to volume ratio of the total tumor burden. This feature was found prognostic in a cohort of 215 DLBCL patients by Decazes et al. [204]. I found that for PFS prediction on another cohort of DLBCL patients, this feature had a univariate C-index of 0.58 ($p < 0.034$) and was significantly improving a multivariate model with TMTV and Dmax ($p < 0.04$). But it was not prognostic of the OS. I found that the 6 mm cutoff identified by Decazes et al. was not significantly separating DLBCL patients as a function of their survival ($p < 0.076$ for PFS and $p < 0.116$ for OS). Therefore, the feature had moderate predictive power but was complementary of TMTV and Dmax. This result was presented as a poster during SNMMI 2022 conference [200].

Apart from visual differences, another strategy would be to use the scientific knowledge of the disease studied to build new biomarkers. For instance, we know that the spleen plays a critical role in the lymphatic system. It functions as a site for lymphocyte proliferation and immune surveillance. The spleen filters blood, removing old or damaged red blood cells and platelets. It supports the immune system by producing lymphocytes, particularly B and T cells. Additionally, the spleen's macrophages destroy pathogens and cellular debris, contributing to the body's defense mechanism against infections. Lymphoma originating from the lymphatic system, it is not rare to see patients with splenic tumor involvement. This can lead to

splenomegaly (enlargement of the spleen), which may cause discomfort or pain and increase the risk of spleen rupture. Moreover, tumor infiltration of the spleen can impair its functions, vital to the immune system, which can make the body more susceptible to infections.

Based on these observations, we investigated the impact of splenic tumor involvement in lymphoma. In particular, we studied the survival of DLBCL patients with various degrees of splenic involvement.

5.2 Article in review

Risk Stratification in Diffuse Large B-Cell Lymphoma: Evaluating Splenic Tumor Volume Impact on Patient Prognosis

IN REVIEW

Louis Rebaud[1,2], Nicolò Capobianco[3], Anne-Ségolène Cottereau[2,4], Clémentine Sarkozy[2,5], Laetitia Vercellino[6,7], Olivier Casasnovas[8], Franck Morschhauser[9], Catherine Thieblemont[10-12], Bruce Spottiswoode[13], Irène Buvat[2]

[1]Siemens Healthcare SAS, Saint Denis, France; [2]LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France; [3]Siemens Healthcare GmbH, Germany; [4]Department of Nuclear Medicine, Cochin Hospital, AP-HP, Université Paris Cité, Paris, France; [5]Institut Curie, Saint Cloud, Paris, France; [6]Department of Nuclear Medicine, Saint-Louis Hospital, AP-HP, Paris, France; [7]Inserm, UMR_S942 MASCOT, Université Paris Cité, F-75006, Paris, France; [8]Department of Hematology, University Hospital of Dijon, INSERM 1231, Dijon, France; [9]Department of Hematology, University of Lille, CHU Lille, ULR 7365; [10]Université Paris Cité, 85 boulevard St Germain, F-75006 Paris, France; [11]Assistance Publique – Hôpitaux de Paris, Saint Louis Hospital, Hemato-oncology, Paris, France; [12]Inserm U1153, Hôpital Saint Louis, 1 avenue Claude Vellefaux, F-75010 Paris, France; [13]Siemens Medical Solutions USA, Inc., Knoxville, Tennessee, United States;

Keywords

DLBCL, Spleen involvement, Splenic involvement, Prognosis, TMTV

Abstract

This study investigates the prognostic information of splenic tumor involvement (SI) in patients with Diffuse Large B-Cell Lymphoma (DLBCL).

Methods: Prognostic value of Total Metabolic Tumor Volume (TMTV), SI, the size of the spleen, Metabolic Tumor Volume Inside the Spleen (MTVIS) and Metabolic Tumor Volume Inside the Spleen (MTVOS) was assessed on a cohort of 377 DLBCL patients. Progression-free survival (PFS) and overall survival (OS) were used as endpoints.

Results: SI patients showed poorer PFS ($p < 0.03$) and OS ($p < 0.04$) than non-SI patients and had higher TMTV ($p < 0.001$). SI was predictable from TMTV with an average precision of 0.62 ($p < 0.001$). SI did not provide additional prognostic information beyond TMTV. MTVIS was not prognostic of the outcome. Patients with SI and elevated MTVIS were not at higher risk than patients with SI and low MTVIS. The same observation was made with splenomegaly. MTVOS was as predictive as TMTV.

Conclusion: The prognostic value of TMTV in DLBCL predominantly resides outside the spleen. Splenic involvement does not give additional prognostic information. Further validation in different patient cohorts is needed.

Introduction

Non-Hodgkin lymphomas (NHL) are the most common hematological malignancy worldwide, accounting for 3% of all cancer diagnoses and deaths [120]. Diffuse Large B-Cell Lymphoma (DLBCL) is the most common subtype of NHL cancers (33% of NHL) with an aggressive behavior, and only 65% of patients alive 5 years after diagnosis [205].

18F-FDG PET/CT scans are routinely used to stage DLBCL patients and assess response to treatment, based on international criteria (Lugano 2014), with response adapted strategies that are now standard of care. By segmenting the lesions on the PET scan, the Total Metabolic Tumor Volume (TMTV) can be calculated. This biomarker has been found to be prognostic in multiple studies and is increasingly being used to stage DLBCL patients [206], [207]. The images are also used to assess tumor location, organ involvement and response to treatment.

Splenic Involvement (SI) is also routinely evaluated in the International Prognostic Index (IPI) through the quantification of the number of extra-nodal sites, and in the Ann Arbor staging. SI is recognized as a poor prognostic factor for the outcome of DLBCL patients. Multiple studies found that patients with SI were at significantly higher risk than patients without SI [208], [209]. Yet, the prognostic value of the volume of tumor in the spleen is unknown.

In this study, we investigated a cohort of 377 DLBCL patients to evaluate the prognostic value of SI and its volume.

Materials and Methods

A total of 377 DLBCL patients from the REMARC (NCT01122472) and LNH073B (NCT00498043) cohorts were analyzed. The detailed compositions of the cohorts have been described elsewhere [210], [211]. Baseline 18F-FDG PET/CT scans in the form of anonymized DICOM files, Progression Free Survival (PFS) and Overall Survival (OS) were available for all patients. The treatment received was available for all patients.

All lesions were segmented by expert medicine nuclear physicians (ASC, LV, MM) in the PET images, as already described [198], [207], [212].

For each patient, the spleen was automatically segmented on the CT image of the PET/CT scans using a deep-learning model called TotalSegmentator [46]. With this model, a Dice score of 0.983 was reported for spleen segmentation when using the segmentation by expert radiologists as the ground truth [46].

All automatic spleen segmentations were visually checked and automatically adjusted whenever needed. Adjustments consisted in the automated removal of small Region of Interest (ROI) splenic segments located outside the spleen and spatially disconnected from the main spleen region. These small ROIs produced by the segmentation algorithm being orders of magnitude smaller than the actual spleen region, their removal was performed automatically by keeping only the largest ROI produced by TotalSegmentator.

Because the PET images and the CT images had different voxel size, all CT-based segmentation masks and PET images were resampled to 1x1x1mm voxel size using nearest neighbor interpolation so that they could be further subtracted, intersected, and aligned using SimpleITK [213], [214], [215].

For each patient, Total Metabolic Tumor Volume (TMTV) was calculated by adding the volumes of all segmented lesions. Metabolic Tumor Volume Outside the Spleen (MTVOS) was calculated by measuring the total volume of lesion voxels not intersecting with the spleen mask. The volume defined as the intersection of the spleen mask and the lesion masks was measured and referred to as the Metabolic Tumor Volume Inside the Spleen (MTVIS). The spleen volume (SV) was also calculated from the CT-derived spleen mask (regardless of the metabolic activity of the spleen). For a more precise estimation of the volumes, the shapes of the masks were refined as triangle meshes as recommended by the image biomarker standardization initiative [66]. These meshes were calculated from the resampled segmentation masks using a marching cube algorithm implemented in the PyRadiomics package [56]. Spleen involvement (SI) was defined as a binary variable equal to 1 in the presence of segmented lesions in the spleen and 0 otherwise.

The group of patients with SI was divided into two groups: patients with less than half of the volume of their spleen segmented as lesion by physicians (so called “focal” SI

group), and patients with more than half of their splenic volume segmented as lesion (called "extensive" SI hereafter).

The prognostic values of SI, MTVIS, SV and MTVOS for PFS and OS were characterized based on their hazard ratio (HR) values computed using a Cox proportional hazard model. HR were controlled for treatment and TMTV. Kaplan-Meier (KM) curves were computed for each patient group and logrank tests were used to assess the significance of the difference for both PFS and OS. The cutoff of 220 mL was used when binarizing TMTV, as reported and validated in multiple studies [216], [217], [218]. Difference of distribution of TMTV and SV between patient groups was assessed with Mann-Whitney U rank test. Accuracy of the prediction of SI from TMTV, MTVOS or SV was assessed using the average precision defined as the area under the precision-recall curve and its significance was estimated with a permutation test of 10,000 permutations.

Results

The cohort was composed of 377 DLBCL patients, with 107 PFS event and 63 deaths. Median age was 65 years. The majority of patients had an Ann Arbor status of 4 (n=284), and other either 3 (n=64) or 2 (n=28). One patient had a status of 1. Most patients had an age-adjusted IPI of 2 (n=201) or 1 (n=117), and other either 3 (n=52) or not provided (n=7). Concerning treatment, 331 patients received R-CHOP, some of them also receiving lenalidomide (n=143) or a placebo (n=136). The 46 remaining patients received R-ACVBP.

A total of 130 patients (34%) had SI, among which 79 were focal SI (less than half of the volume of spleen segmented as lesion) and 51 were extensive SI (more than half of the splenic volume was tumoral).

0. Splenic Involvement (SI) prognostic power

Figure 37 shows the KM curves for PFS and OS of patients with and without SI.

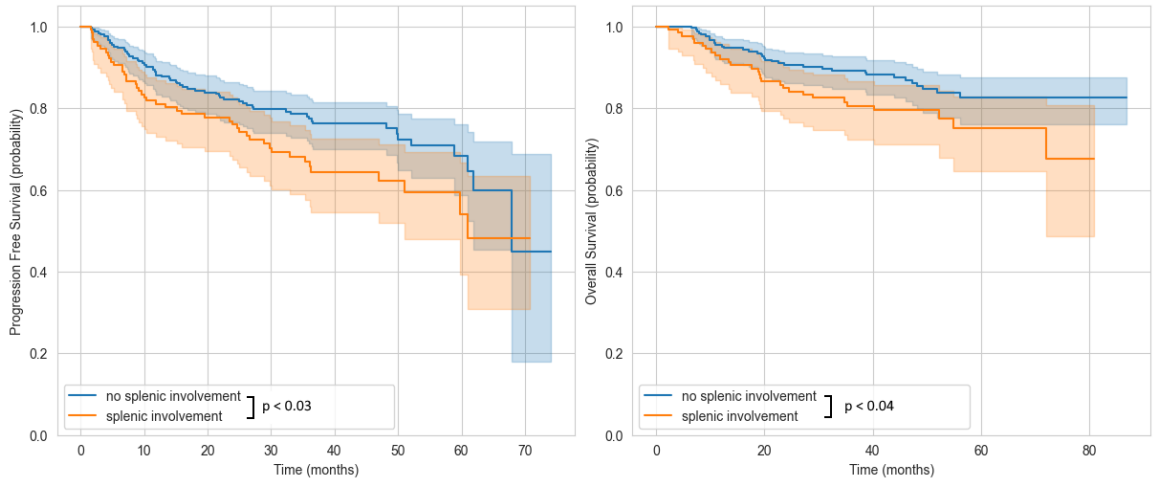


Figure 37: Kaplan-Meier curves of patients with and without splenic involvement, for Progression Free Survival and Overall Survival.

When controlling for treatment, hazard ratios (HR) of SI were 1.24 ($p < 0.02$) and 1.29 ($p < 0.03$), for PFS and OS respectively. When HR was controlled for both treatment and TMTV, HRs were 1.16 ($p < 0.13$) and 1.17 ($p < 0.22$), for PFS and OS respectively, suggesting that SI was confounded by TMTV.

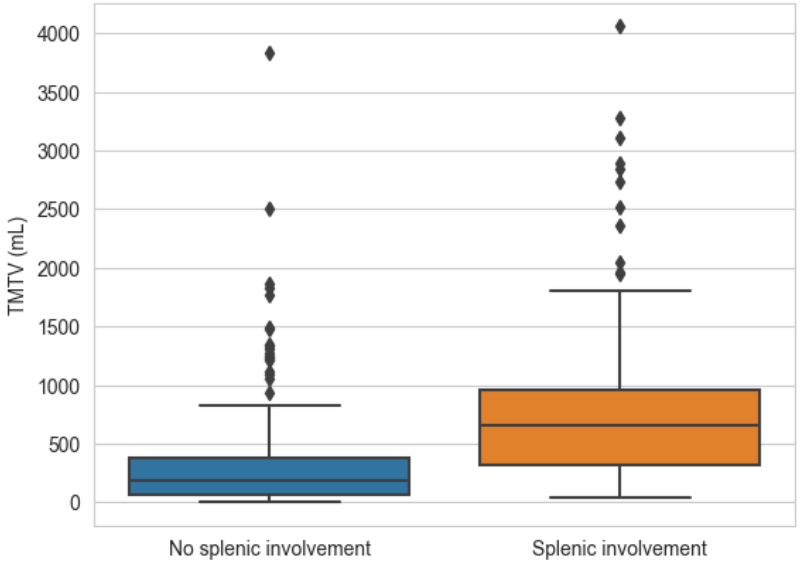


Figure 38: Distribution of Total Metabolic Tumor Volume (TMTV) for patients with and without splenic involvement.

Figure 38 shows the distribution of TMTV for patients with and without SI. Patients with SI had significantly higher TMTV ($p < 0.001$) than patients without SI. TMTV could predict SI with an average precision of 0.62 ($p < 0.001$).

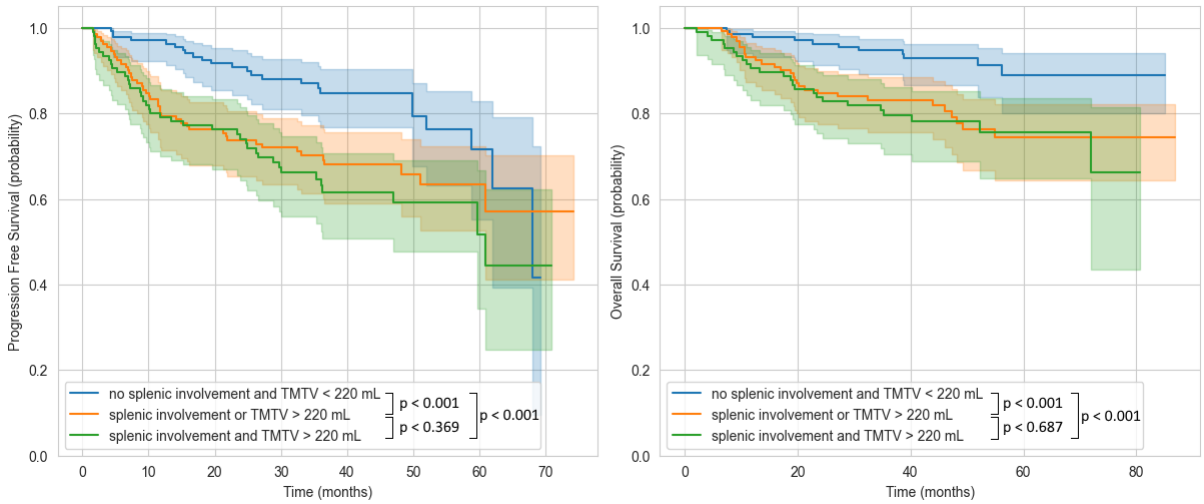


Figure 39: Kaplan-Meiers curves of three groups of patients: patients without splenic involvement and low TMTV (blue), patients with splenic involvement or high TMTV (orange), and patients with splenic involvement and high TMTV (green).

Figure 39 shows KM curves for the combination of SI and TMTV. Patients were grouped as a function of their number of risks factor. SI did not further stratified patients with elevated TMTV.

2. Metabolic Tumor Volume Inside the Spleen (MTVIS) prognostic value

Table 1 shows the HR for PFS and OS of MTVIS controlled for treatment and controlled for treatment and TMTV in different patient groups, suggesting an absence of prognostic value.

	Number of patients	Hazard Ratio controlled for treatment		Hazard Ratio controlled for treatment and TMTV	
		PFS	OS	PFS	OS
All	377	1.10 (p < 0.28)	1.21 (p < 0.06)	0.97 (p < 0.73)	1.04 (p < 0.72)
SI	130	1.02 (p < 0.91)	1.23 (p < 0.29)	0.82 (p < 0.31)	1.06 (p < 0.79)
SI – Focal	79	1.12 (p < 0.53)	0.96 (p < 0.88)	1.06 (p < 0.73)	0.92 (p < 0.73)
SI – Extensive	51	1.07 (p < 0.83)	1.59 (p < 0.16)	1.03 (p < 0.94)	1.57 (p < 0.32)

Table 1: Hazard Ratio of the Metabolic Tumor Volume Inside the Spleen (MTVIS) controlled for treatment and controlled for treatment and TMTV, for Progression Free Survival (PFS) and Overall Survival (OS), computed in different patients groups: entire cohort (All), patients with splenic involvement (SI), patients with less than half of the spleen involved (SI – Focal) and patients with more than half of the spleen involved (SI – Extensive).

Figure 40 displays the KM curves for the combination of SI and MTVIS. Patients with SI were not further stratified by having a MTVIS below or above the median (102 mL).

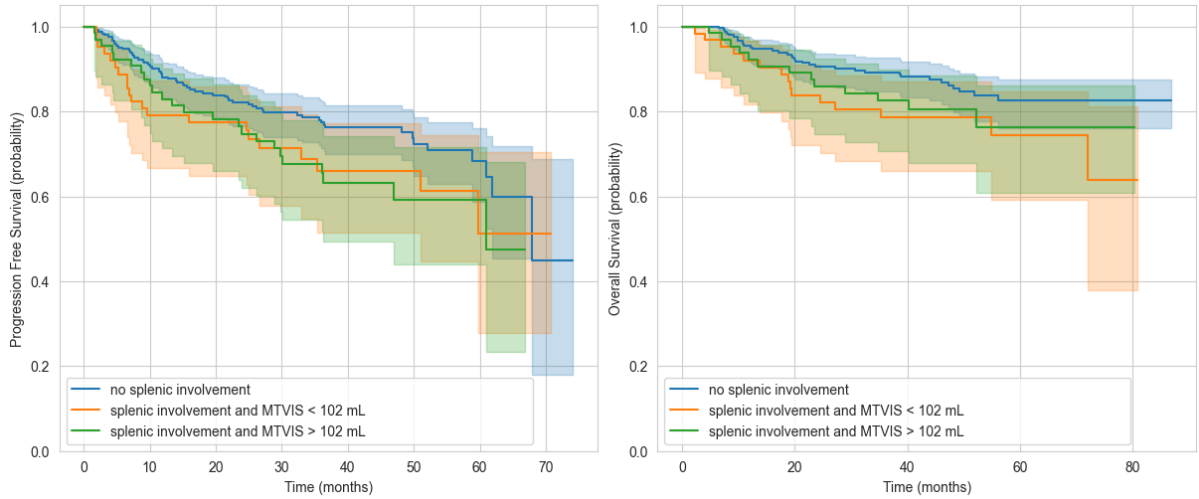


Figure 40: Kaplan-Meiers curves of three groups of patients: patients without splenic involvement (blue), patients with splenic involvement and Metabolic Tumor Volume Inside the Spleen (MTVIS) below the median (orange), and patients with splenic involvement and MTVIS above the median (green).

Various cut-offs values were tested to define low and high MTVIS groups. Figure 41 shows the results of a logrank test for both PFS and OS for these values. Patients were never significantly separated in low and high-risk groups based on the MTVIS values (Figure 41A). The same analysis was conducted for the cut-off used to separate the patients into "focal" or "extensive" SI. None of the cut-off values significantly yielded two SI groups with significantly different outcomes (Figure 41B).

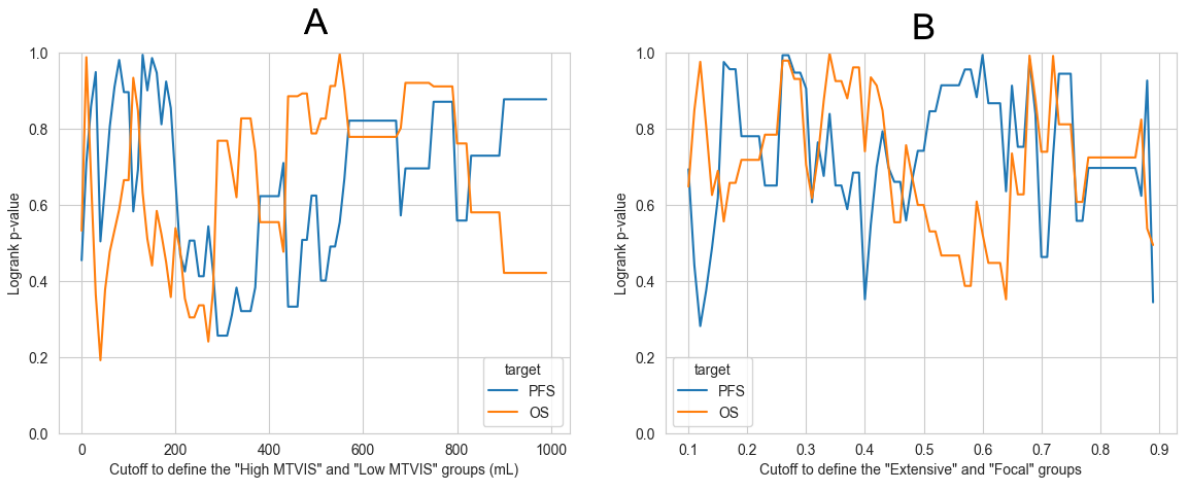


Figure 41: P-values of logrank tests testing for significance of difference in Progression Free Survival (PFS) and Overall Survival (OS) for patients with splenic involvement grouped as a function of their Metabolic Tumor Volume Inside the Spleen (MTVIS) (A) or as a function of the proportion of splenic volume involved to assign them in "focal" or "extensive" group (B). For each criterion, multiple cut-off values were tested (x-axes).

3. Spleen Volume (SV) prognostic value

Table 2 shows the HR for PFS and OS of SV controlled for treatment and controlled for treatment and TMTV in different patient groups. SV was significantly associated with a shorter OS on the entire cohort, but when controlled for TMTV the significance was lost.

	Number of patients	Hazard Ratio controlled for treatment		Hazard Ratio controlled for treatment and TMTV	
		PFS	OS	PFS	OS
All	377	1.15 (p < 0.11)	**1.28 (p < 0.02)**	1.04 (p < 0.72)	1.13 (p < 0.27)
SI	130	0.99 (p < 0.96)	1.14 (p < 0.51)	0.87 (p < 0.43)	1.03 (p < 0.90)
SI – Focal	79	1.08 (p < 0.67)	0.93 (p < 0.76)	1.06 (p < 0.74)	0.92 (p < 0.71)
SI – Extensive	51	0.90 (p < 0.75)	1.44 (p < 0.29)	0.83 (p < 0.63)	1.31 (p < 0.49)

Table 2: Hazard Ratio of Spleen Volume (SV) controlled for treatment and controlled for treatment and TMTV, for Progression Free Survival (PFS) and Overall Survival (OS), calculated in different patients groups: entire cohort (All), patients with splenic involvement (SI), patients with less than half of the spleen involved (SI – Focal) and patients with more than half of the spleen involved (SI – Extensive).

Figure 42 shows the distribution of SV for patients without SI, with focal SI and with extensive SI. Patients with focal SI had a significantly higher SV than patients without SI ($p < 0.001$) but significantly lower than patients with extensive SI ($p < 0.001$). SV could predict SI with an average precision of 0.72 ($p < 0.001$).

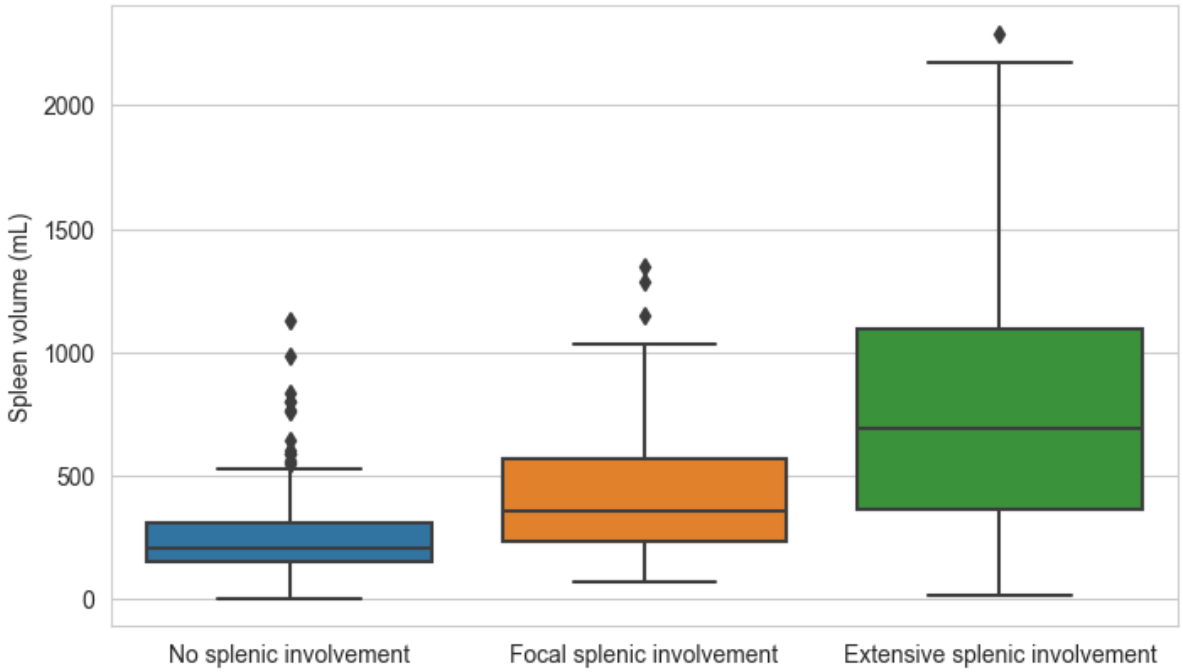


Figure 42: Distribution of Spleen Volume (SV) for patients without splenic involvement, patients with focal splenic involvement (less than half of the spleen involved) and with extensive splenic involvement (more than half of the spleen involved).

4. Metabolic Tumor Volume Outside the Spleen (MTVOS) prognostic value

Table 3 shows the HR of TMTV and MTVOS controlled for treatment. When TMTV was prognostic, MTVOS was also prognostic with almost identical HR. The table also features MTVOS HR controlled for treatment and TMTV. In this case, MTVOS was never prognostic. The table also shows that TMTV and MTVOS were highly correlated, even in groups composed of patients with SI. MTVOS could predict SI with an average precision of 0.51 ($p < 0.001$).

	TMTV Hazard Ratio controlled for treatment		MTVOS Hazard Ratio controlled for treatment		MTVOS Hazard Ratio controlled for treatment and TMTV		TMTV and MTVOS Spearman's correlation
	PFS	OS	PFS	OS	PFS	OS	
All	**1.26 (p < 0.01)**	**1.35 (p < 0.01)**	**1.24 (p < 0.01)**	**1.30 (p < 0.01)**	1.09 (p < 0.73)	0.90 (p < 0.72)	0.94
SI	**1.30 (p < 0.05)**	1.33 (p < 0.10)	**1.34 (p < 0.02)**	1.27 (p < 0.16)	1.45 (p < 0.31)	0.89 (p < 0.79)	0.87
SI – Focal	**1.63 (p < 0.01)**	1.41 (p < 0.10)	**1.62 (p < 0.01)**	1.43 (p < 0.09)	0.64 (p < 0.73)	1.79 (p < 0.73)	0.96
SI – Extensive	1.08 (p < 0.79)	1.35 (p < 0.36)	1.06 (p < 0.84)	1.11 (p < 0.75)	0.96 (p < 0.94)	0.50 (p < 0.32)	0.85

Table 3: Hazard Ratios of Total Metabolic Tumor Volume (TMTV) and Metabolic Tumor Volume Outside the Spleen (MTVOS) controlled for treatment and TMTV, and Spearman's correlation between TMTV and MTVOS for different groups of patients: entire cohort (All), patients with splenic involvement (SI), patients with less than half of the spleen involved (SI – Focal) and patients with more than half of the spleen involved (SI – Extensive).

Discussion

In this study, we characterized the different features of splenic involvement and analyzed their impact on the outcome of a cohort of DLBCL patients.

Patients with splenic involvement were found to have shorter PFS and OS. Yet, splenic involvement was strongly associated with TMTV: patients with splenic involvement had a significantly higher TMTV, to the extent that it was possible to predict splenic involvement from TMTV alone with high accuracy. Therefore, splenic involvement did not improve the stratification based on TMTV only.

When we evaluated the Metabolic Tumor Volume Inside the Spleen (MTVIS), we did not find any additional prognostic information compared to splenic involvement or TMTV only. Patients with a large volume of tumor inside their spleen were not at higher risk than patients with a small splenic tumor volume. The Metabolic Tumor Volume Outside the Spleen (MTVOS) was highly correlated with and as predictive as TMTV, meaning that most of the prognostic information of TMTV is located outside the splenic region. This finding is similar with previous results from Guerra et al. as they found MTVOS to be as predictive as TMTV for PFS prediction of follicular lymphoma patients [219].

To the best of our knowledge, no study reported that patients with large splenic tumor volume were at higher risk than patients with a low volume of tumor in the spleen. Yamanaka et al. [220] found no statistical differences between these two groups, but they concluded that it might be due to a too small population of 108 patients.

Splenomegaly was associated with splenic involvement but was not adding additional prognostic information. Patients with splenic involvement and a hypertrophic spleen were not at higher risk than patients with splenic involvement and a normal spleen volume.

While these observations are statistically sound, these findings need to be confirmed in other cohorts of DLBCL patients. The role of splenic involvement in other lymphoma subtypes also warrants further investigation.

Conclusion

While splenic involvement was significantly prognostic of the outcome in a cohort of 377 DLBCL patients, it was confounded by TMTV and did not improve patient stratification. Furthermore, accounting for the volume of splenic involvement did not improve patient stratification. Therefore, DLBCL patients with high splenic tumor volume were not at higher risk than those with lower splenic tumor volume.

5.3 Discussion

In this study, the impact of splenic tumor involvement on the prognosis of DLBCL patients was assessed. While patients with splenic involvement had significantly lower PFS and OS, they also had higher TMTV. It is therefore unknown if their higher risk was due to the splenic involvement or their high TMTV. We demonstrated that the metabolic tumor volume inside the spleen did not offer additional prognostic value beyond splenic involvement or TMTV. It was also found that splenomegaly was correlated with splenic involvement but did not further impact prognosis.

While this study's findings are relevant and help us understand better how to diagnose DLBCL patients, it illustrates an issue I encountered for many image-based features I tested: many of them were confounded by TMTV. This is a common issue in radiomics when promising findings were later found to be surrogate of TMTV's prognostic information [221]. This was an important problem since the main goal of the project was to find new information in the images.

Another problem with the empirical search is the multiple testing issue. By testing one by one new radiomic features, the high number of tests is not controlled. This can result in many false positives as the number of tested features increases.

Lastly, the intuition-based definition of new features leads to a narrow exploration of the image information. By looking at whole body images, only differences visible on the displayed images (e.g., MIPs) will be explored. It is easy to miss small differences in lesion shape or PET or CT values. While it is unlikely that radiomic features can capture information invisible to our eyes if we know where to look at, they can be a formidable tool to mine image data. By measuring numerous radiomic features in all the images, in multiples regions such as lesions, lesion surroundings and organs, a large fraction of the image information can be automatically tested for its prognostic value and the novelty of the information it provides. All these candidate biomarkers would be tested one by one on a cohort of patients and the selected ones would then be analyzed to understand the encoded biological information, allowing for a data-driven approach to the task of biomarker discovery. Once the biological information is understood, we would ideally try to reencode it in simpler and more direct ways. This would allow for a more efficient, reliable, and interpretable feature definition. These new features would then be tested on new cohorts of patients to ensure the validity of the discovery and their interest for clinical applications.

This automated approach is not without limitations. First, the biomarkers found prognostic might not always be easy to interpret and linking them to biological interpretation can be challenging. Secondly, being a univariate approach, biomarkers prognostic only when associated with other biomarkers will not be identified. While we could miss some interesting information by not exploring this search space, it is preferable to not try to find such combination as the number of possibilities will be too

high for the number of samples available, and the risk of overfitting would have been too great. Lastly, testing thousands of features at once requires a powerful selection process. Without it, the risk of false positives (selecting features that are good prognostic only by chance) and false negatives (missing relevant features) becomes high. For this reason, a significant amount of work was dedicated to the construction of robust selection tool that ensures that the selected biomarkers bring new prognostic information, not already quantified by known features.

Chapter 6

Development of a biomarker selection tool (ROBI)

6.1 Introduction

The selection tool used to screen the candidate image-based radiomic biomarkers had several objectives. First it should test the candidates for the novelty of the information they provided compared to already known prognostic information. A feature reencoding the TMTV or Dmax will not help improve our understanding of the disease and the diagnosis of DLBCL patients. We could try to find new features re-encoding TMTV and Dmax information, but with a higher prognostic value. However, even if we find such features, it would be impossible to prove that they are better than TMTV and Dmax with the limited number of patients available, especially since TMTV and Dmax were validated on multiple cohorts and numerous patients.

The pipeline must also control for confounders such as the clinical center from which the patient was enrolled. A radiomic feature able to differentiate the machine used to scan the patient will not be useful in clinic.

The second goal of the pipeline is to test the prognostic value of the candidates. We found that trying to have a precise estimate of the prognostic value of a feature is extremely challenging because the data and the outcome are noisy. We conjectured that for survival analysis tasks, we often do not have enough data to estimate the predictive power of a feature reliably and accurately. Yet, this is actually not our goal when looking for new biomarkers. We do not want to know how much a candidate is prognostic, we just want to know if it is prognostic. Therefore, in the pipeline, we estimate the likelihood of the feature of not having any prognostic value with a permutation test. If it is unlikely that a feature does not have any prognostic value, we select it.

However, if this test is performed on thousands of features, multiple testing should be accounted for. We used false discovery rate (FDR) estimation techniques for their flexibility. Since this is an exploratory phase, it is not too much of a problem to select some false positives (FP) if it allows for the discovery of many new relevant biomarkers. Given the selected candidates will ultimately be tested on other cohorts, FP will be discarded in these external validations. Yet, these techniques to control FDR have a limited statistical power, and if too many candidates are tested at once, they will not

be able to reliably identify features related to the target from those that are not. If too many candidates are tested, none or all candidates will be selected. Therefore, we designed a pipeline to minimize the number of tested candidates in the FDR estimation to maximize the number of discoveries. This led us to the biomarker selection pipeline described in this chapter.

6.2 Article in review

ROBI: a Robust and Optimized Biomarker Identifier to increase the likelihood of discovering relevant radiomic features.

IN REVIEW

Louis Rebaud[1,2], Nicolò Capobianco[3], Clémentine Sarkozy[2,4], Anne-Ségolène Cottereau[2,5], Laetitia Vercellino[6,7], Olivier Casasnovas[8], Catherine Thieblemont[9-11], Bruce Spottiswoode[12], Irène Buvat[2]

[1]Siemens Healthcare SAS, Saint Denis, France; [2]LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France; [3]Siemens Healthcare GmbH, Germany; [4]Institut Curie, Saint Cloud, Paris, France; [5]Department of Nuclear Medicine, Cochin Hospital, AP-HP, Université Paris Cité, Paris, France; [6]Department of Nuclear Medicine, Saint-Louis Hospital, AP-HP, Paris, France; [7]Inserm, UMR_S942 MASCOT, Université Paris Cité, F-75006, Paris, France; [8]Department of Hematology, University Hospital of Dijon, INSERM 1231, Dijon, France; [9]Université Paris Cité, 85 boulevard St Germain, F-75006 Paris, France; [10]Assistance Publique – Hôpitaux de Paris, Saint Louis Hospital, Hemato-oncology, Paris, France; [11]Inserm U1153, Hôpital Saint Louis, 1 avenue Claude Vellefaux, F-75010 Paris, France; [12]Siemens Medical Solutions USA, Inc., Knoxville, Tennessee, United States;

Abstract:

Purpose: To design and validate a feature selection tool that selects biomarkers most likely to reflect new prognostic information while minimizing and controlling the number of false positives (FP).

Materials and Methods: The ROBI feature selection pipeline is a software combining several feature selection techniques to select biomarkers that encode relevant information not already quantified by established biomarkers and that are most likely to predict patient outcome in the dataset used for selection. The pipeline minimizes the selection of FP and estimates their number. Selection stringency can be adjusted.

A total of 500 synthetic datasets and retrospective data from 18F-FDG PET/CT scans of 378 Diffuse Large B Cell Lymphoma (DLBCL) patients were analyzed to validate the tool. On the DLBCL data, two established radiomic biomarkers, TMTV and Dmax, were measured from the segmentation of the 18F-FDG PET/CT scans, and 10,000 random ones were generated. Selection was performed and verified on each dataset. Statistical significance was evaluated with Wilcoxon signed-rank tests. The efficacy of ROBI has been compared to methods controlling for multiple testing and a Cox model with Elasticnet penalty.

Results: In the synthetic datasets, the pipeline selected significantly more true positives (TP) than FP ($p < 0.001$). For 99.3% of the synthetic datasets, the number of FP was within the 95% confidence interval estimated by the pipeline. The proposed pipeline significantly increased the number of TP compared to usual feature selection methods ($p < 0.001$). In the real dataset, ROBI selected the two established biomarkers and one random biomarker and estimated 95% chance of selecting 0 or 1 FP and a probability of 0.0014 of selecting only FP. The Bonferroni correction selected no feature, and the Elasticnet selected 73 spurious features and missed one of the two established biomarkers.

Conclusion: The ROBI pipeline effectively selected relevant biomarkers while controlling FP, demonstrating robust performance on both synthetic and real datasets.

Keywords: Biomarker, biomarker discovery, feature selection, multiple testing, false positive

Abbreviations:

ROBI: Robust and Optimized Biomarker Identifier

FP: False Positive

TP: True Positive

CB: Candidate biomarker

TST: two-stage linear step-up procedure

FDR: False discovery rate

CCO: Correlation Clustering Optimization

DLBCL: Diffuse Large B Cell Lymphoma

TMTV: Total Metabolic Tumor Volume

Dmax: maximum distance between two lesions

Introduction

Radiomics involves the extraction and analysis of quantitative medical image features [51], [52]. By converting images into mineable data, radiomics may reveal disease characteristics that are currently overlooked, improving diagnosis, prognosis, and treatment planning. A great number of scientific publications have mentioned radiomics since its introduction, but reproducibility, standardization, interpretability, and methodological issues limit its potential, and few radiomics results have been translated into the clinic [64], [65].

Standards for radiomic feature definition and calculation, and guidelines for best practices are being developed to accelerate clinical translation [66], [70], [71], [72]. Lack of external validation and methodological flaws in assessing biomarker novelty and prognostic power partly explain why radiomics has not been adopted in the clinics yet. Statistical methods, such as robust feature selection algorithms, cross-validation techniques for model evaluation, and statistical tests for assessing the significance of prognostic biomarkers, can address some of these challenges by ensuring the reliability and generalizability of radiomics studies. On the other hand, improper use of these techniques—including overfitting models to specific datasets, not controlling for C-index inflation, ignoring multiple testing corrections, data leakage in the machine learning pipeline and failing to validate findings externally—can lead to misleading results, characterized by either overly optimistic or pessimistic evaluations of radiomic features and models.

In this context, we introduce the Robust and Optimized Biomarker Identifier (ROBI), not as a novel feature selection method, but as a software solution designed to combine a range of established techniques in a simple yet efficient manner. ROBI is a streamlined Python package designed to facilitate the selection of radiomic features, thereby mitigating the risk of selecting features that either mirror existing biomarkers (Orlhac et al., 2014 [221]) or lack prognostic relevance. By implementing current best practices within an optimized framework, ROBI aims to minimize false positives—erroneously selected non-relevant features—while enhancing the detection of true positives—genuinely relevant features. It employs time-efficient permutation tests to precisely estimate the number of false positives, offering users the flexibility to tailor selection stringency according to their research objectives.

ROBI's efficacy is demonstrated through validation on synthetic datasets with established truths, and on a cohort of Diffuse Large B Cell Lymphoma (DLBCL) patients, where it successfully identified two known biomarkers out of many random ones. This underscores ROBI's utility as a practical tool that leverages existing methodologies to overcome some of the current barriers in radiomics, paving the way for more reliable and clinically applicable radiomic research outcomes.

Material and methods

1. Pipeline

Candidate biomarkers (CB) are assessed for their predictive potential by ROBI, based on their values in a patient cohort and their association with the outcome (e.g., time before relapse, response to treatment). To avoid selecting candidates that replicate known predictive information, previously known predictive biomarkers must be identified. Figure 43 presents the overall pipeline. More details on the choice of the parameter values are provided in the supplemental data.

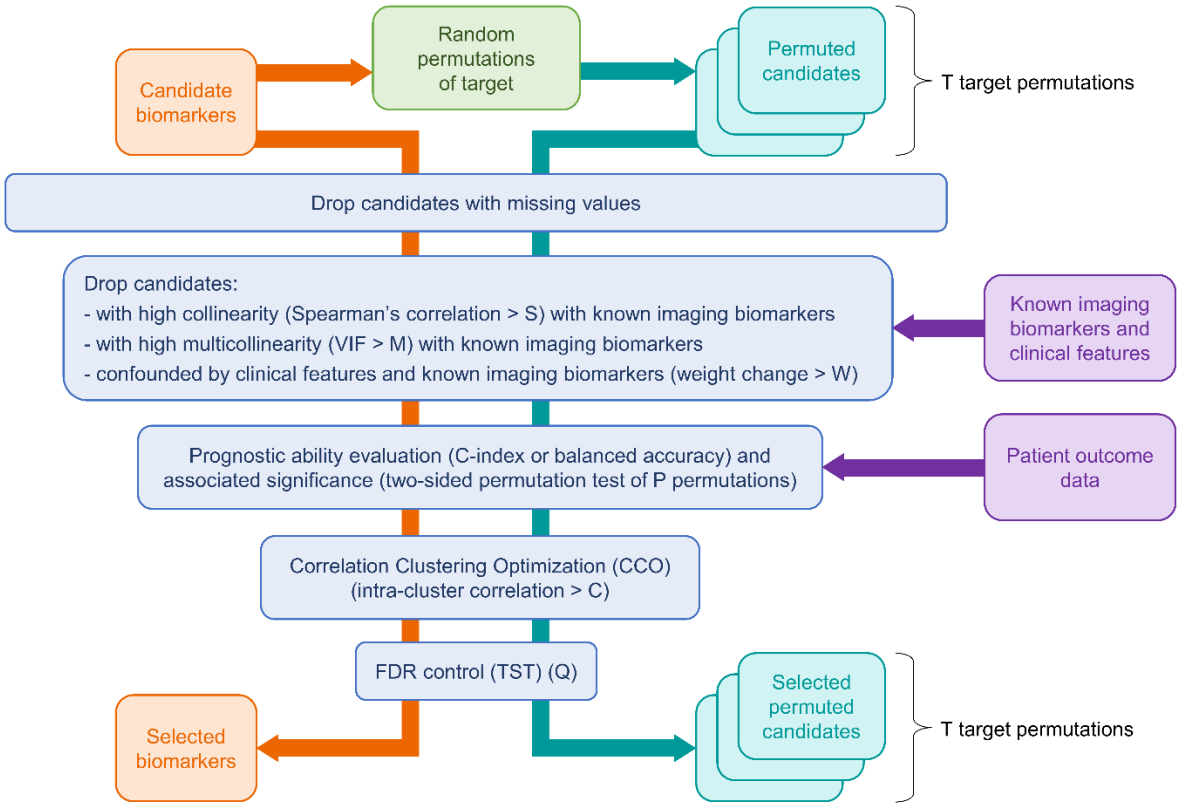


Figure 43: Diagram of the ROBI selection pipeline. Each free tuning parameter is denoted by a capital letter (S, M, W, P, C, Q and T). Intuitive explanation and range of values of these parameters are provided in supplemental materials. "VIF" is the Variance Inflation Factor. "weight change" is the relative change in weight when confounders are introduced. "FDR" is the False Discovery Rate and "TST" stands for two-stage linear step-up procedure, the technique used to control for FDR. Filtering candidates reproducing known information and CCO are optional.

a. Discarding missing values

A low number of samples and high numbers of censored samples artificially increase biomarker prognostic value [99], [114]. Any CB with missing values is thus discarded to avoid favoring CBs unavailable to all patients.

b. Discarding already known information

CBs with an absolute Spearman correlation coefficient greater than a tunable cut-off S (0.5 by default) with a known imaging or clinical biomarker are discarded to ensure that the selected CB capture new information. In case of multiple established biomarkers, multicollinearity is assessed using the Variance Inflation Factor (VIF). CBs exceeding a certain tunable multicollinearity threshold M (5 by default) are discarded. A linear model (Cox for survival and logistic regression for classification) controls for confounders (e.g., age, treatment) [222]. A univariate model with only the evaluated CB is trained first and assigns a weight W_{uni} to the CB. Then, a multivariate model with the evaluated CB and known covariates is trained and the new weight W_{multi} is assigned to the CB. The relative change in weight is defined as:

$$W_{shift} = \frac{|W_{uni} - W_{multi}|}{W_{multi}} \times 100$$

Any CB with W_{shift} above a threshold W (10% by default) is discarded.

c. Assessment of CBs performance

Each CB's prognostic ability is assessed using Harrell's Concordance Index (C-index) against patient outcome data such as time of death or relapse, accommodating censored outcomes, or balanced accuracy for classification task, accommodating imbalanced datasets. These scores are tested for significance using a two-sided permutation test of P permutations (1,000 by default). A two-stage linear step-up procedure (TST) is used to control the false discovery rate (FDR, the proportion of false positive in selected biomarkers) and address multiple testing [119]. This statistical method uses a conservative threshold to identify potential selection and adjusts this threshold in a second stage based on the initial results to increase power while controlling the overall FDR. Adjusting TST's Q parameter allows flexibility in balancing numbers of FPs and selected CBs. To increase the yield, ROBI performs the TST last in the selection process when the number of tested CBs has already been substantially reduced through the previous selection steps.

d. Optimization of the number of selected biomarkers

To optimize the selection of biomarkers, we employ a correlation clustering optimization (CCO) strategy, where CBs conveying similar information are grouped based on their absolute Spearman's correlation. Within each cluster, only the biomarker demonstrating the highest predictive accuracy is retained. This approach is informed by methods previously utilized in genomics, notably the weighted gene correlation network analysis (WGCNA) technique, which clusters genes based on similarity in expression patterns to identify modules of highly correlated genes, thereby facilitating the interpretation of complex biological phenomena [223]. By adopting a similar methodology, we adjust the maximum allowable correlation between two clusters, C (0.5 by default), to fine-tune the granularity of the clustering and thus the number of

biomarkers selected. This method not only enhances the specificity of our biomarker selection process but also ensures that the biomarkers retained offer of more unique predictive value, thereby avoiding redundancy.

e. False positive estimation

Because it is selecting the CBs with the best p-values, CCO may optimistically bias TST's FDR. To correct and improve the number of FP estimation, ROBI randomly permutes outcome data during selection. This preserves the relationships among CBs but breaks their association with patient outcomes. The features selected using the permuted outcome are thus FP. After repeating this process T times (by default 1,000 times), ROBI calculates the average number of FPs and its 95% confidence interval. The probability of only selecting FPs is assessed by the proportion of permutated datasets with as many as or more selected CBs than the non-permuted selection.

2. Synthetic data evaluation

A total of 500 synthetic datasets were generated with scikit-learn [224] and scikit-survival [225] Python packages to evaluate ROBI. These datasets varied in the number of samples, number of genuine (associated with the outcome) and spurious (not associated with the outcome) biomarkers, censoring, correlation between biomarkers, and target noise. Table 4 shows the parameter distributions and ranges. Details about the generation of the datasets are provided in the documentation of scikit-learn [224].

	Average (and std)	**Min**	**Max**
Number of samples	423 (±260)	10	1000
Proportion of censored samples	0.57 (±0.27)	0.1	0.9
Noise	4.92 (±3.04)	0	10
Number of predictive biomarkers	515 (±280)	1	1000
Number of non-predictive biomarkers	3164 (±2268)	10	50000
Proportion of predictive biomarkers	0.196 (±0.153)	0.004	0.812
Average correlation between candidates	0.13 (±0.08)	0	0.6

Table 4: Average, standard deviations, and range of the synthetic dataset features.

A linear regression with random weights on genuine biomarkers defined the target using Scikit-learn's "make_regression" function. The target was not built using spurious biomarkers. ROBI processed each dataset with CCO with $P = 10^6$, and $T = 10^3$. Genuine biomarkers selected by ROBI were defined as TPs and selected spurious biomarker as FPs. The "effective_rank" parameter within "make_regression" allowed for the simulation of correlations among features (biomarkers) by controlling the rank of the covariance matrix used to generate the features. A lower "effective_rank" implies a higher correlation among a subset of features, thereby simulating real-world scenarios where biomarkers might exhibit interdependencies. The same datasets were also processed with the two-stage linear step-up procedure (TST) alone to compare its results to ROBI's selections and verify that ROBI's optimization improves the number of biomarkers rightly selected.

The average, standard deviation and 95% confidence interval of the number of selected CBs, TPs and FPs were calculated as well as the percentage of datasets with more TPs than FPs, for different values of Q. Wilcoxon signed-rank tests were used to determine whether ROBI selected more TPs than FPs and if using ROBI increased the number of rightly selected biomarkers compared to TST alone. The distribution of the difference of TPs for the ROBI and TST selection for the same number of FPs was plotted.

3. Real data evaluation

DLBCL patients from REMARC (NCT01122472) and LNH073B (NCT00498043) cohorts were analyzed. Detailed cohort compositions have been described elsewhere [210], [211]. All patients had baseline anonymized ^{18}F-FDG PET/CT scans, with 5 years Progression Free Survival (PFS) and 5 years Overall Survival (OS) available. Lesions were segmented by expert nuclear medicine physicians (ASC, LV, MM) in the PET images [198], [212].

In DLBCL, Total Metabolic Tumor Volume (TMTV) and maximum distance between two lesions (Dmax) are known to be prognostic of PFS and OS [198], [226]. These two biomarkers were calculated on the segmented PET images using PyRadiomics [56].

10,000 spurious biomarkers were randomly generated for all patients and input to ROBI in addition to TMTV and Dmax. ROBI parameter settings were S = 0.5, M = 5, W = 10%, P = 10^7 and T = 10^4. Q was set to have at least one selected CB. No CCO was used because spurious biomarkers are random and have low correlation. Biomarkers were tested to not replicate the information of ECOG, age adjusted International Prognostic Index [189], treatment, and sex. We then checked whether TMTV and Dmax were selected by ROBI and whether the number of selected spurious biomarkers was within the 95% confidence interval of ROBI's estimated number of FPs. Selection was performed for progression (PFS) or death from any cause (OS) prediction.

Selection was also performed with other feature selection techniques: TST with a Q value chosen to have less than one false positive, the Bonferroni procedure with a probability of 0.05 of having 1 or more false positive, and a Cox model with Elasticnet penalty.

Results

1. Synthetic data evaluation

A total of 99.3% of datasets had the number of FPs within ROBI's 95% confidence interval. Table 5 shows ROBI's selection results on synthetic datasets, and Figure 44 shows the average number of selected features and FPs, with their 95% confidence intervals, as a function of Q, for both ROBI's and TST's selection. More CBs were selected with higher Q. ROBI selected significantly more TPs than FPs ($p < 0.001$). For the same Q, ROBI significantly increased numbers of TPs, FPs, and the difference between them compared to TST ($p < 0.001$).

Q	Method	Number of selected CB	Number of TP	Number of FP	Percentage of datasets with more TP than FP
0.01	TST	4 (±20)	4 (±20)	0 (±0)	100.0 %
	ROBI	6 (±28)	6 (±27)	0 (±0)	99.8 %
0.10	TST	9 (±33)	9 (±32)	0 (±0)	99.8 %
	ROBI	15 (±47)	14 (±44)	1 (±12)	99.2 %
0.25	TST	13 (±43)	13 (±42)	0 (±2)	99.4 %
	ROBI	28 (±81)	21 (±56)	6 (±48)	97.1 %
0.50	TST	19 (±57)	17 (±52)	1 (±10)	99.2 %
	ROBI	52 (±141)	32 (±71)	20 (±99)	94.6 %

Table 5: Average values and standard deviation of the number of selected candidate biomarkers (CB), number of true positives (TP), false positives (FP), and percentage of datasets with more FP than TP, for different levels of Q, for the ROBI pipeline and the TST procedure alone.

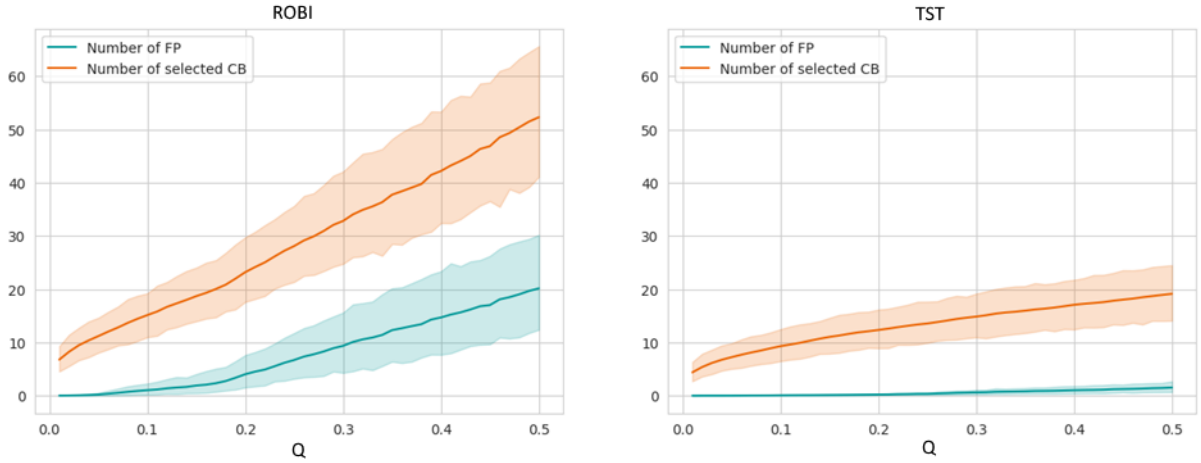


Figure 44: Average number of selected candidate biomarkers (CB) and average number of false positives (FP) among the selected CB, with the associated 95% confidence interval, for the ROBI pipeline and the TST procedure alone, at various levels of Q.

Figure 45 plots the difference between numbers of TPs of ROBI and the number of TPs of TST for samples in which the same numbers of FPs were selected. For the same number of FPs, ROBI selected significantly more TPs than TST alone ($p < 0.001$).

The probability of having only FPs in the selection estimated by ROBI was strongly correlated with the number of TPs ($\rho = -0.96$, $p < 0.001$). For 60% of cases with at least one TP, this probability was below 0.05. For the cases with only FPs selected (3.3% of all cases), 0.6% of them had a probability below 0.05.

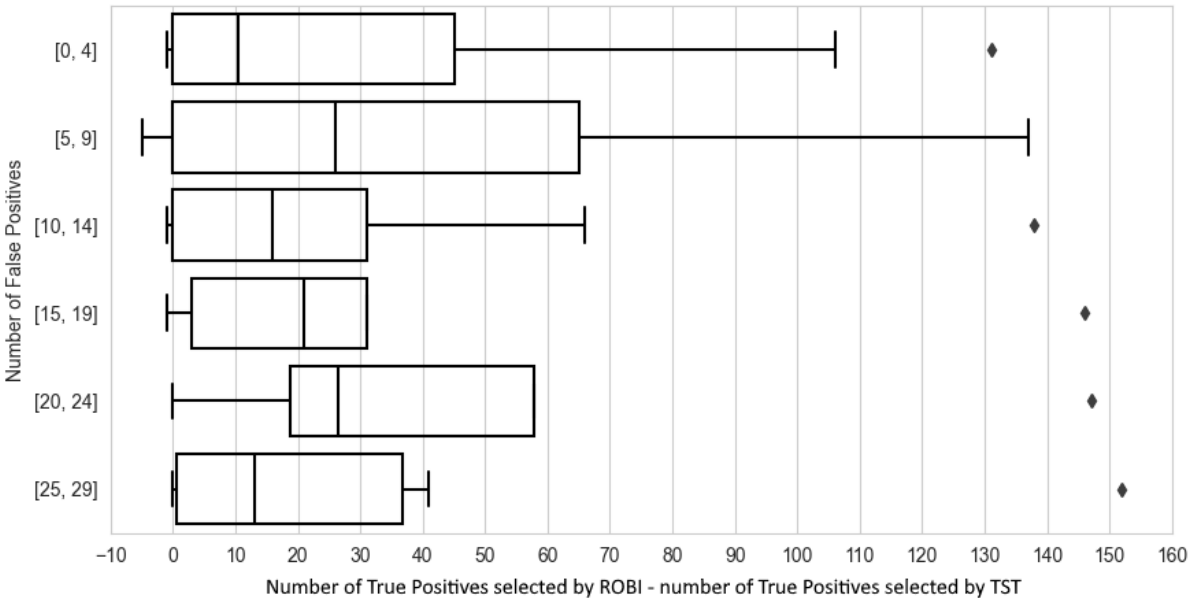


Figure 45: Difference between the number of True Positives (TP) selected by ROBI and the number of TPs selected by TST alone when the two methods had the same number of False Positives (FP). The difference is positive most of the time, meaning that ROBI effectively improved the rate of TP selection.

2. Real data evaluation

The DLBCL cohort included 378 patients, among whom 96 had progressive disease and 55 died.

For PFS prediction, TMTV and Dmax both yielded a C-index of 0.63, while 105 spurious features had a C-index > 0.58 and 16 had a C-index > 0.60. The significance of the spurious features was $p < 0.01$ for 103 of them, and $p < 0.001$ for 13 of them. The Bonferroni selection did not select any feature. TST selected both TMTV and Dmax and one spurious feature. An Elasticnet selected Dmax and ranked it first, but it did not select TMTV and selected 101 spurious features. ROBI selected TMTV and Dmax, and one spurious feature. ROBI predicted a 95% chance of having 0 or 1 FP with an average of 0.1 FP. ROBI estimated the probability of having only FP to be 0.0014.

For OS prediction, TMTV and Dmax had respectively a C-index of 0.63 and 0.60, and 137 spurious features had a C-index > 0.60. The significance of the spurious features was $p < 0.01$ for 110 features, and $p < 0.001$ for 8 of them. The Bonferroni selection did not select any feature. TST did not select any spurious features, nor TMTV nor Dmax. An Elasticnet selected TMTV and ranked it 47. It did not select Dmax and selected 73 spurious features. ROBI did not select any feature.

Discussion and Conclusion

This study introduced ROBI, the Robust and Optimized Biomarker Identifier. We called it "Robust" because false discoveries are controlled and "Optimized" because multiple strategies increase the number of rightly selected biomarkers. We showed that this selection tool efficiently controls the false positive numbers while increasing the number of selected biomarkers compared to the standard two-stage linear step-up procedure (TST) alone. ROBI's 95% confidence interval estimating the number of false positives was correct for 99.3% of the synthetic datasets, small difference between these two numbers being probably explained by statistical fluxes. It can find relevant biomarkers among thousands of candidates with enough data (96 events for PFS prediction in our real dataset), while other standard methods fail with such a high number of potential candidates. For instance, in the test performed on the DLBCL, enough patients had PFS observed to select TMTV and Dmax, but not enough events were observed in OS to selected them.

As shown by the evaluation on the synthetic datasets, ROBI's utility transcends the radiomic domains, making it a versatile tool for biomarker selection across various fields (e.g., genomics).

ROBI has limitations. Only biomarker screening is addressed. Validating a new biomarker requires definition, measurement, standardization, modeling, and interpretation. More importantly, ROBI does not replace external validations. It only increases the chance of replicating the findings by controlling the risk of false positive selection.

Limitations include dropping candidate biomarkers with missing data. This step may eliminate promising biomarkers by reducing the number of candidates. Removing a few patients (preferably those with a censored target value) to avoid discarding too many candidate biomarkers can mitigate this limitation.

ROBI is more time consuming than other selection methods. However, thousands of biomarkers can be accurately evaluated in a reasonable time. On a PC an Intel Core i7-11800H (2.30 GHz), NVIDIA GeForce RTX 3070 (8 GB), and 16 GB RAM, 5,000 candidates could be evaluated in less than 9 min, with $T = 10^3$ et $P = 10^7$.

ROBI may not identify all relevant biomarkers among candidates, and the number of false negatives (predictive biomarkers that are not selected) remains unknown. In addition, because it uses univariate tests, ROBI may not always choose biomarkers that improve multivariate models. Furthermore, while the pipeline can estimate the number of false positives and the probability of selecting only false positives, it cannot tell which feature is more likely to be a true positive, and external validation remains required to validate the selected features.

Because ROBI uses a multivariate model to address confounders, only a finite number

of them can be handled. For survival prediction, the general guideline is 10 non-censored samples per confounder [106].

An important future work needed is a more thorough comparison to other feature selection techniques on non-synthetic datasets.

In conclusion, ROBI selects biomarkers that best predict patient outcomes in a cohort, by discarding candidates that do not measure any new predictive information. ROBI identifies the most promising candidates, which will then have to be tested on external cohorts to confirm their predictive value. ROBI might facilitate feature selection in radiomics and beyond, and to support this effort, we provide a user-friendly Python implementation at https://github.com/Lrebaud/robi.

6.3 Discussion

In this article, we introduced the ROBI pipeline. This tool can select relevant biomarkers from large pools of candidate biomarkers by testing them on a cohort of patients. The candidates are selected for the novelty of their information with respect to existing prognostic factors and their prognostic value. While ROBI does not guarantee that the selected biomarkers are true ones for the corresponding disease, it minimizes the chances of selecting non relevant candidates for the training cohort. It does not replace an external validation but identify the most promising candidates among a large number. ROBI is optimized for screening of large ensembles of candidates. We showed in the paper that it can find relevant candidates among thousands of spurious ones. It is domain agnostic, making it useful for a wide range of domains dealing with abundant features, such as genomic. We released this tool as a Python package for an easy and reliable use of the pipeline.

With this tool, we can explore numerous image-based features in whole-body PET/CT scans, to extensively screen the content of these images.

Chapter 7

Discovery of new prognostic biomarkers for Non-Hodgkin Lymphomas

7.1 Introduction

Once I was able to deal with large numbers of candidates thanks to the ROBI tool, I could start extensively exploring the whole PET/CT scans of cancer patients. I had two cohorts of Non-Hodgkin Lymphoma patients available: one composed of DLBCL patients, and another of FL patients. These two groups of patients came from three clinical trials. Both had approximately 350 patients, providing enough statistical power to test thousands of candidate biomarkers.

In this chapter, we detailed how a large pool of candidate image-based features was automatically constructed and tested, covering a significant amount of the information present in the PET/CT scans performed before treatment. We present the result of their selection with the ROBI tool and the following analysis to understand and reencode the information conveyed by the selected features.

7.2 Article in preparation for submission

Discovery of new prognostic biomarkers in Diffuse Large B Cell Lymphoma and Follicular Lymphoma using comprehensive 18F-FDG PET/CT mining

IN PREPARATION FOR SUBMISSION

Louis Rebaud[1,2], Nicolò Capobianco[3], Clémentine Sarkozy[2,4], Anne-Ségolène Cottereau[2,5], Laetitia Vercellino[6,7], Olivier Casasnovas[8], Franck Morschhauser[9], Catherine Thieblemont[10-12], Bruce Spottiswoode[13], Irène Buvat[2]

[1]Siemens Healthcare SAS, Saint Denis, France; [2]LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France; [3]Siemens Healthcare GmbH, Germany; [4]Institut Curie, Saint Cloud, Paris, France; [5]Department of Nuclear Medicine, Cochin Hospital, AP-HP, Université Paris Cité, Paris, France; [6]Department of Nuclear

Medicine, Saint-Louis Hospital, AP-HP, Paris, France; [7]Inserm, UMR_S942 MASCOT, Université Paris Cité, F-75006, Paris, France; [8]Department of Hematology, University Hospital of Dijon, INSERM 1231, Dijon, France; [9]Department of Hematology, University of Lille, CHU Lille, ULR 7365; [10]Université Paris Cité, 85 boulevard St Germain, F-75006 Paris, France; [11]Assistance Publique – Hôpitaux de Paris, Saint Louis Hospital, Hemato-oncology, Paris, France; [12]Inserm U1153, Hôpital Saint Louis, 1 avenue Claude Vellefaux, F-75010 Paris, France; [13]Siemens Medical Solutions USA, Inc., Knoxville, Tennessee, United States;

Abstract

In this study, we explored new prognostic biomarkers in Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) through an extensive analysis of ^{18}F-FDG PET/CT scans, focusing on Progression Free Survival as a primary endpoint. Utilizing the open-source Robust and Optimized Biomarker Identifier (ROBI) pipeline, we assessed thousands of radiomic features for their prognostic values and the novelty of their information, leading to the identification of 28 new prognostic image-based biomarkers for FL patients, and 28 others for DLBCL patients. Through careful analysis of the relationship between feature values and visual appearance of the image signal , we manually identified 22 biological information prognostic of the outcome that we re-encoded into 22 more interpretable image-based biomarkers. Among these 22 features, 10 demonstrated prognostic significance across both DLBCL and FL, suggesting a higher likelihood of replication, as well as clinical applicability to a larger number of patients. While several surrogate biomarkers were calculated inside or close to the lesions, others reflected the patient's general state of health and comorbidities, calling for a novel approach to image-based patient stratification. Using these newly-identified prognostic biomarkers enabled a more accurate prediction of patient outcomes, highlighting the potential of these biomarkers to refine therapeutic strategies. Our findings show promise for enhancing the prognostic assessment of DLBCL and FL patients, warranting further validation in external cohorts to confirm their clinical utility. The code for computation and test of these biomarkers is freely available.

Introduction

Non-Hodgkin lymphomas (NHL) rank as the fifth most prevalent cancer with 72,035 new cases in Western Europe in 2020, and is the most common hematological malignancy worldwide, accounting for 3% of all cancer diagnoses and deaths [120], [175]. Diffuse Large B-Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL) are the two most common subtypes of NHL, as DLBCL represents 30%-58% of NHL cases, and FL 20%-25% of them [158], [180]. DLBCL is an aggressive but curable disease, with 70% of the patients considered as cured after the standard of care RCHOP; FL is characterized by an important heterogeneity in disease presentation and outcome: some patients remaining asymptomatic without treatment for decades and other presenting a refractory disease with a particularly poor outcome [227]. Importantly, FL

can transform into DLBCL forms, with an annual incidence of 2-3% and a poor outcome.

The standard practice in managing DLBCL and FL involves ^{18}F-FDG PET/CT scanning, at baseline for staging and end of treatment for response assessment. These scans assist doctors in determining cancer's stage and monitoring its progression and response to treatments. Specifically, by locating tumors in the images and detecting organ involvement, medical experts can evaluate prognostic scoring system, relying on staging and other biological parameters, such as the Ann Arbor staging and the International Prognostic Index (IPI) and its variants like the age-adjusted IPI (aaIPI) and Follicular Lymphoma International Prognostic Index (FLIPI) specifically designed for FL. Furthermore, delineating lesions on the PET scan allows for the calculation of the Total Metabolic Tumor Volume (TMTV), a prognostic factor validated in various series in DLBCL, and more recent in FL. Multiple studies found that TMTV was prognostic of DLBCL patient outcomes [206], [207], and could be used to guide treatment strategies, as in the CAR-T cell setting. While it is not a standard of care, it is used more and more to stage those patients. Recent work also suggests that TMTV has potential prognostic value in FL [197]. Other PET derived metrics, such as the maximum distance between two lesions (Dmax) was found prognostic on several cohorts of DLBCL patients [198], [199], and promising results were found on a cohort of 126 FL patients [228].

Beyond these features, an obvious follow-up question would be: is there additional useful information still overlooked in these ^{18}F-FDG PET/CT images? Non-medical experts can easily spot evident imaging disparities in PET scans of patients with similar stage, TMTV and Dmax but whether these differences impact the outcome is an open question. We can also wonder whether there is prognostic information outside the tumor regions reflecting patient's specific conditions and state, and possibly in the CT images that are always associated with the PET scan.

The goal of this study was therefore to search for potential prognostic biomarkers present in baseline ^{18}F-FDG PET/CT scans of FL and DLBCL patients, using Progression Free Survival (PFS) as an endpoint, PFS also being a surrogate of the Overall Survival (OS) [229]. We aimed to identify biomarkers that do not merely reiterate information already conveyed by TMTV, Dmax or the existing clinical staging features.

Our strategy consisted in generating plethora of potential biomarkers, aiming to capture a wide range of information within and outside tumor lesions, in both PET and CT images. Then, the Robust and Optimized Biomarker Identifier (ROBI) selection pipeline was used to select biomarkers identified as prognostic of PFS without replicating known information [230]. This pipeline was designed to maximize the probability of identifying true prognostic biomarkers by controlling the false discovery rate and optimizing the order of the selection steps. We then tried to understand the biological meaning of the selected biomarkers and re-encoded these hypotheses into simpler features. To try to identify new prognostic biomarkers that would be robust and not specific to one subtype of NHL, these features were searched for on one cohort

of DLBCL patients and another of FL patients independently. Features found prognostic on the two cohorts would be more valuable as they will be applicable to more patients but would also have higher chances to be confirmed in external validation studies.

Material and methods

1. Data description

Two cohorts of patients were analyzed in this study. The first one is composed of 347 DLBCL patients from the REMARC (NCT01122472) and LNH073B (NCT00498043) trials. The second cohort consists of 350 FL patients from the multicentric RELEVANCE clinical trial (NCT01476787). The detailed compositions of these cohorts have been described elsewhere [210], [211].

Baseline ^{18}F-FDG PET/CT scans were available for all patients as pseudonymized DICOM files. In addition, clinical and biological data at baseline (ECOG score, IPI index with FLIPI for FL and age-adjusted IPI (aaIPI) for DLBCL, treatment received and sex), treatment, outcome (PFS and OS) and recruitment center were available. For the FL cohort, comorbidities were also available.

2. Data preparation

All lesions were segmented by expert medicine nuclear physicians (ASC, LV, MM) in the PET images, applying the approach detailed by Cottereau et al. [212].

For every patient, 24 organs or group of organs were segmented from the CT images of the PET/CT scans using the TotalSegmentator deep-learning model [46]. Another deep-learning model, MOOSE, was employed to segment muscle and fat, which were not segmented by version 1.5.2 of TotalSegmentator [46]. The segmented organs were left and right adrenal glands, brain, blood vessels (aorta, pulmonary, iliac, portal, splenic, inferior vena cava), colon, duodenum, esophagus, gallbladder, heart, a merge of gluteus, autochthon, and iliopsoas, left and right kidneys, liver, left and right lungs, pancreas, skeleton, small bowel, spleen, stomach, trachea, urinary bladder, all skeletal muscles and fat (subcutaneous and visceral).

Because the PET and the CT images did not always share the same voxel spacing, all images and regions of interest were resampled to 1x1x1mm voxel size using SimpleITK [213], [214], [215]. Nearest neighbor interpolation was used for resampling segmentation masks (to keep the binary values of the segmentation masks) and spline interpolation was used for resampling images.

3. Calculation of candidate biomarkers

Numerous features were created from the regions of interest (ROI) corresponding to the tumor and organ segmentation masks. In the PET and CT images, all features of the PyRadiomics package were computed for each specific organ [56]. The metabolic

tumor volume within the organ was calculated by intersecting the tumor and organ segmentation masks. This volume was also normalized by the volume of the organ and the TMTV. A binary biomarker that indicated whether an organ contained tumor was created. Finally, the shortest and largest distances between the organ and the tumor were determined using all voxels of the regions (and not only their respective center of mass). The number of involved organs was measured, as well as total muscle volume, total fat volumes, and the TMTV divided by these volumes.

In addition, in the PET and CT images, all features of the PyRadiomics package were computed for each spatially disconnected segmented tumor lesion. The same features were also computed in the 8 millimeter thick shell that surrounded each lesion. At the patient level, each feature was aggregated using 6 approaches: minimum, mean, median, maximum, standard-deviation, and range. Last, the PyRadiomics features were computed considering all the lesions as a single unconnected ROI, and also in the 8 millimeter thick shells of tissues surrounding each unconnected ROI.

4. Selection of candidate biomarkers

The Robust and Optimized Biomarker Identifier (ROBI) selection pipeline [230] was used to select candidates that were the most likely to be predictive of the outcome (e.g. risk) of patients. PFS was used as endpoint as it was less censored than OS. Candidates that re-encoded already known prognostic information (like TMTV or Ann Arbor stage) were automatically discarded by the ROBI pipeline. For the selection using the DLBCL cohort, candidates were controlled for confounding effect with TMTV, Dmax, ECOG, aaIPI, treatment, sex, age, and Ann Arbor status. For the FL cohort, confounders were TMTV, ECOG, FLIPI, treatment, sex, age, and Ann Arbor status. A Kruskal-Wallis test was also used to discard candidates confounded by the recruitment center of the patients. ROBI's parameters were set to S = 0.5, W = 10% and M = 5 to avoid selecting biomarkers reproducing known information [230]. The C = 0.5 value was used to reduce the number of candidate biomarkers and force the selected ones to reflect a wide diversity of information. In the pipeline, candidates were grouped into cluster of similar information based on their correlation, and only the most prognostic candidate in each cluster was used in the rest of analysis. Values of $P = 10^7$, $T = 10^4$ were used to ensure a precise estimate of the significance of the prognostic value of the candidates and a reliable estimate of the number of false positives (FP). The Q parameter, which controls the selectivity of the ROBI pipeline, was chosen to ensure that fewer than one FP could be expected, thereby reducing the chance of selecting non-reproducible biomarkers. We reported the total number of selected biomarkers, averaged estimated number of FPs, the 95% confidence interval (CI) of the estimated number of FPs, and the probability of selecting only FPs.

The selection was performed on each cohort independently.

5. Interpretation of selected biomarkers

Our goal was to obtain new explainable biomarkers related to patient prognosis. To understand the key information reflected by the selected candidate biomarkers, we first determined if the selected candidates were positively or negatively correlated with the PFS. Then, using visual samples and correlations with more intuitive features and with other selected biomarkers, we formulated hypotheses that might explain the biological information reflected by each candidate biomarker and how this might relate to the patient's outcome. When possible, we re-formulated the hypothesis in simpler terms using a surrogate biomarker supposed to capture the same information as the selected candidate biomarker. This step allowed us to test our interpretation assumption and to make the biomarker easier to understand. This analysis was performed for the biomarkers selected on the FL cohort and on the DLBCL cohort independently. Each surrogate biomarker identified in one cohort was systematically tested for its prognostic power in the other cohort.

For each selected biomarker, we reported the C-index, its significance, and the sign of the correlation with the risk of PFS event. We also reported the prognostic performance of the surrogate biomarker, visual examples of high and low risk patients according to the surrogate biomarker, as well as Kaplan-Meier curves of the cohort as split based on the surrogate biomarker and TMTV values. Prognostic performances of the surrogate biomarkers were controlled for multiple testing with a two-stage linear step-up procedure [119]. All surrogate biomarkers were evaluated on both FL and DLBCL cohorts.

6. Response to treatment

Response to treatment at 120 weeks defined by Cheson et al 1999 [231] and assessed by the independent review committee of the RELEVANCE trial was available for all FL patients. FL patients were assigned to either the “responding” group encompassing all FL patients with complete, unconfirmed complete or partial response to treatment, or to the “progressive” group when progressive disease was identified at the 120 weekds timepoint. Only one FL patient was identified as presenting a stable disease and was discarded from the analysis.

Selected biomarkers and surrogate biomarkers identified based on the PFS outcome were tested for their ability to distinguish between patients responding to treatment vs patients with progressive disease in the FL cohort. They were evaluated through a univariate logistic regression model using the balanced accuracy calculated from a 10-fold stratified cross-validation. Significance of the balanced accuracies was assessed with a permutation test of 10,000 permutations.

This analysis was not performed on the DLBCL patients as the response to treatment was not available.

7. Multivariate model

On each cohort independently, a multivariate ICARE model [232] was trained to predict the PFS of patients from different groups of features: the baseline feature group (TMTV for FL; TMTV and Dmax for DLBCL), one group composed of the baseline features and the surrogate biomarkers identified as prognostic factors in both FL and DLBCL, and one group with all the surrogate biomarkers identified as prognostic factors in the tested cohort. Models were evaluated with a C-index calculated from a 10,000-folds Leave-Pair-Out cross validation. Average C-index were reported. The significance of the change of C-index when new features were added was assessed with Nadeau and Bengio correction to the paired Student-t test [233]. The same analysis was done with the selected biomarkers. The building of a multivariate model served as a sanity check and was not an attempt to estimate the actual performance of a model that would use the selected biomarkers. The model training and evaluation being performed on the same datasets used for the feature selection, data leakage was present, and the performance was likely over-optimistic. The goal of this multivariate model building was to verify that performance of a multivariate model was increasing when new selected biomarkers were added, as if the performance was decreasing, it would reveal a problem in the biomarker selection.

Results

FL patient data came from 39 centers and were acquired using 28 different machines. Median follow-up duration for the RELEVANCE, REMARC and LNH073B trials were respectively 72, 52 and 45 months. Out of the 350 included patients, 130 patients (37%) had disease progression and 36 patients (10%) died during the trial. For DLBCL patients, data came from 40 centers and images were acquired with 25 different machines. Of the 347 DLBCL patients, 96 patients (28%) had disease progression and 55 patients (16%) died during the trial.

A total of 6834 candidate biomarkers were computed in each cohort. 4231 were computed inside organ regions, 1116 inside distinct lesions, 1116 inside the shell of tissues around these lesions, 185 inside all lesions grouped as one unconnected ROI and 186 in the shell of tissues around this unique ROI. 3096 biomarkers were computed from the PET scans, 3096 from the CT and 503 from the shape of the organ or lesion ROIs.

Figure 46 shows the number of candidate biomarkers remaining after each step of the ROBI selection pipeline for each cohort. Most of the discarded candidates were dropped because they could not be computed on all patients (missing values) or because they were correlated with confounders (see Methods).

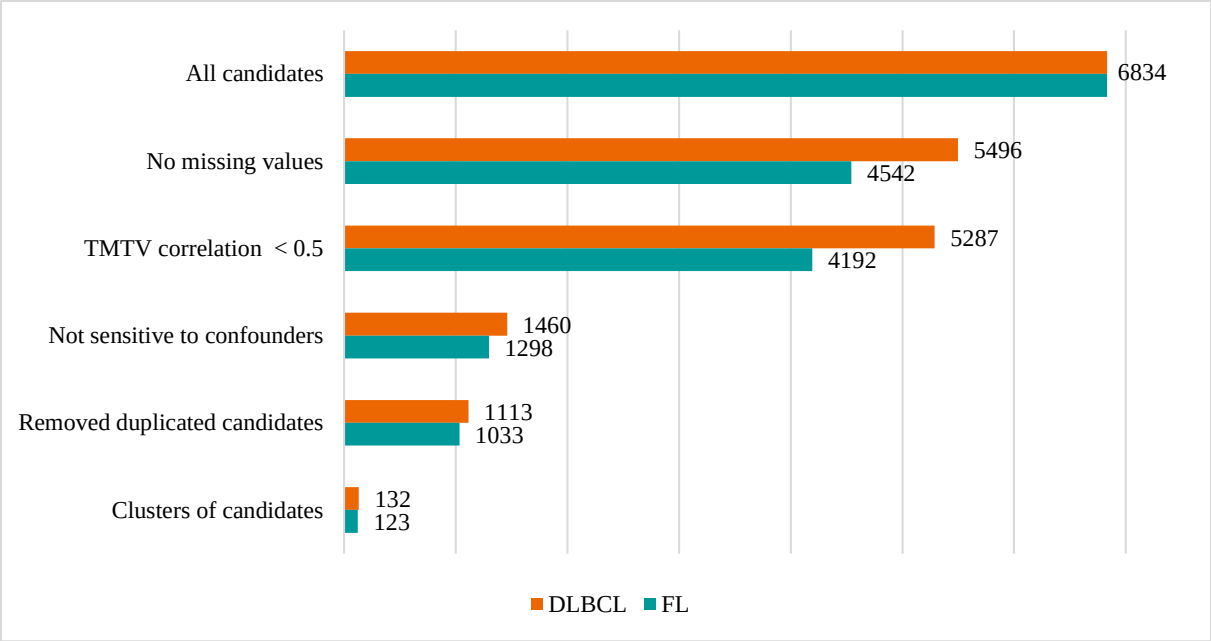


Figure 46: Number of candidate biomarkers selected at each step of the ROBI selection pipeline, for selection on the DLBCL cohort (orange) and on the FL cohort (green). Details about each step are provided in [230].

Figure 47 shows, for every value of Q, the number of candidates being selected, and the number of FPs estimated by ROBI (average value and 95% CI), for both cohorts. The number of selected candidates was always outside the 95% CI of the estimated number of FPs. Q = 0.07 was chosen for FL and Q = 0.08 for DLBCL, ensuring less than one FP can be expected. ROBI selected 28 features for the FL cohort, and 28 for the DLBCL cohort. Among the selected features, none were common to both diseases. ROBI estimated the probability of having only FPs in the selected biomarkers to be 0.001 for FL and 0.002 for DLBCL. ROBI estimated that 0.9 FP could be expected for FL and 0.9 FP for DLBCL, and the 95% confidence interval for the number of FPs was [0, 10] for both diseases.

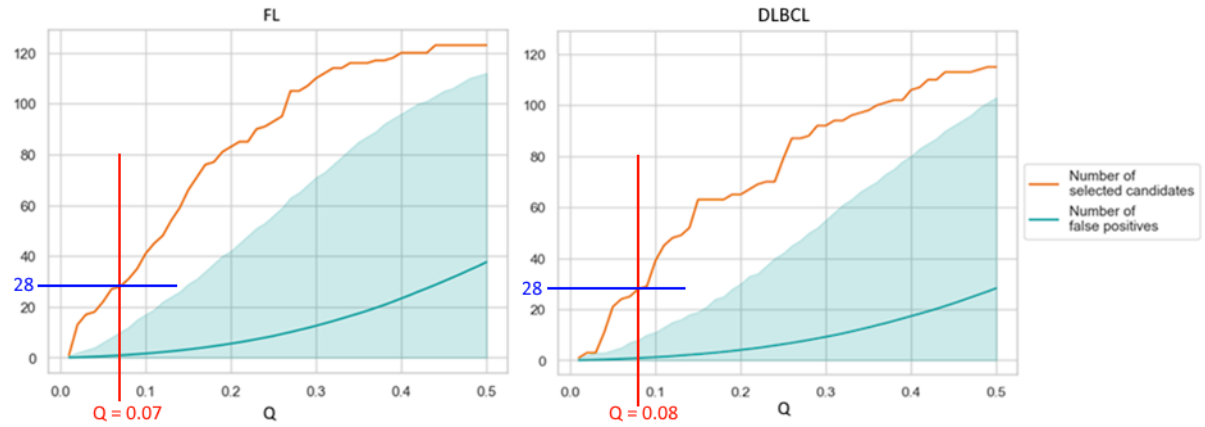


Figure 47: Number of selected candidates (orange) and average number of false positives and its 95% confidence interval (green) for all tested values of Q, for both selections. Q values chosen for the selection are depicted in red.

The complete lists of selected candidates are provided as supplemental data (Table S14 and Table S15), as well as their C-index and their significance, the sign of their correlations with the risk and their correlogram. In FL, most selected candidates were computed inside or near the tumoral ROIs (n=19) and 9 inside organ ROIs. In DLBCL, 14 features were related to organs, and 14 were related to lesions or their close surroundings. For FL, 12 selected candidate biomarkers were calculated from the PET, 9 from the CT, and 7 from the shape of the ROI only (regardless of the signal inside). For DLBCL 7 selected candidate biomarkers were calculated from the PET, 14 from the CT, and 7 were shape-related. Of the 28 candidate biomarkers selected on the DLBCL cohort, 9 were also prognostic for the PFS in the FL patients, although they were not selected by ROBI on the FL cohort. Of the 28 candidates selected on the FL cohort, 14 were prognostic for the PFS of the DLBCL patients but were not selected by ROBI on the DLBCL cohort. A correlogram showing the correlation between the 28 features selected on the FL cohort and the 28 features selected on the DLBCL cohort is provided as supplemental data (Figure S70).

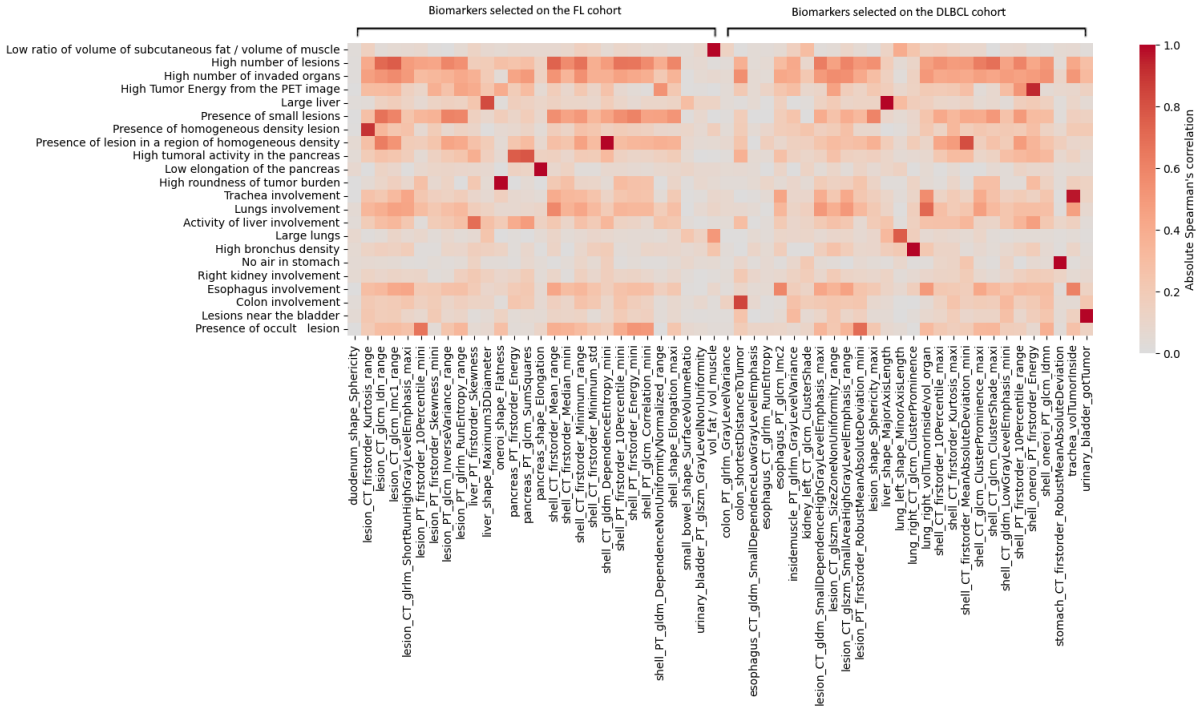


Figure 48: Absolute Spearman's correlations between the 22 surrogate biomarkers and the 2x28 biomarkers selected by the ROBI pipeline. The correlations were calculated by merging the two cohorts into one.

Table 6 list the 22 surrogate biomarkers manually created to interpret the information reflected by the biomarkers selected by ROBI and provides their C-index and significance for each cohort. Among these 22 biomarkers, 10 were prognostic in both FL and DLBCL cohorts, 5 were only prognostic in FL, and 7 were only prognostic in DLBCL. Correlograms of these biomarkers are provided as supplemental data (Figure S71 and Figure S72).

What increases the risk	ROI type	Computed from	Biomarker definition	C-index (p-value) DLBCL	C-index (p-value) FL
Low ratio of volume of subcutaneous fat / volume of muscle	Organ	Shape	Volume of subcutaneous fat / volume of muscle	**0.57 (p < 0.03)**	**0.58 (p < 0.01)**
High number of lesions	Lesion	Shape	Number of spatially disconnected segmented lesions	**0.63 (p < 0.01)**	**0.57 (p < 0.01)**
High Tumor Energy from the PET image	Lesion	PET	Sum of the squared SUV values of all the voxels segmented as lesion	**0.57 (p < 0.03)**	**0.60 (p < 0.01)**
Large liver	Organ	Shape	MajorAxisLength PyRadiomics feature of the segmented liver	**0.58 (p < 0.01)**	**0.60 (p < 0.01)**
Presence of small lesions	Lesion	Shape	Volume of the smallest lesion	**0.57 (p < 0.03)**	**0.56 (p < 0.03)**
Presence of homogeneous density lesion	Lesion	CT	The kurtosis of CT values in each lesion is computed. The highest kurtosis across all lesions is used.	**0.56 (p < 0.04)**	**0.57 (p < 0.01)**
Presence of lesion in a region of homogeneous density	Lesion	CT	The GLDM DependenceEntropy of the CT values in the shell of tissue around each lesion is computed. The lowest value across all lesions is used.	**0.60 (p < 0.01)**	**0.60 (p < 0.01)**
Trachea involvement	Lesion and organ	Shape	Whether or not the segmented trachea has at least one voxel segmented as lesion	**0.58 (p < 0.01)**	**0.55 (p < 0.01)**
High bronchus density	Lesion and organ	CT	The GLCM ClusterProminence of the CT values in the right lung	**0.59 (p < 0.01)**	**0.56 (p < 0.01)**
Presence of occult lesion	Lesion	PET and shape	Whether or not at least one lesion is smaller than 20 mL and with a SUVmax < 5	**0.57 (p < 0.01)**	**0.54 (p < 0.04)**
Large lungs	Organ	Shape	Volume of the lungs	**0.57 (p < 0.03)**	0.54 (p < 0.09)
High number of invaded organs	Lesion and organ	Shape	Number of segmented organs with at least one voxel segmented as lesion	**0.63 (p < 0.01)**	0.55 (p < 0.06)
No air in stomach	Organ	CT	RobustMeanAbsoluteDeviation of the CT values of the voxels in the stomach	**0.62 (p < 0.01)**	0.50 (p < 0.86)
Right kidney involvement	Lesion and organ	Shape	Whether or not the segmented right kidney has at least one voxel segmented as lesion	**0.57 (p < 0.01)**	0.51 (p < 0.37)
Esophagus involvement	Lesion and organ	Shape	Whether or not the segmented esophagus has at least one voxel segmented as lesion	**0.57 (p < 0.03)**	0.52 (p < 0.37)
Colon involvement	Lesion and organ	Shape	Whether or not the segmented colon has at least one voxel segmented as lesion	**0.58 (p < 0.01)**	0.51 (p < 0.25)
Lesions near the bladder	Lesion and organ	Shape	Whether or not the segmented bladder has at least one voxel segmented as lesion	**0.58 (p < 0.01)**	0.51 (p < 0.70)
High tumoral activity in the pancreas	Lesion and organ	PET	Sum of the squared SUV values of all voxels segmented as pancreas if pancreas is invaded.	0.55 (p < 0.13)	**0.57 (p < 0.01)**
Low elongation of the pancreas	Organ	Shape	Inverse of the Elongation PyRadiomics feature of the segmented pancreas	0.51 (p < 0.76)	**0.58 (p < 0.01)**
High roundness of tumor burden	Lesion	Shape	Inverse of the Flatness PyRadiomics feature of all voxels segmented as lesion	0.51 (p < 0.72)	**0.58 (p < 0.01)**
Lung involvement	Lesion and organ	Shape	Whether or not the segmented lungs have at least one voxel segmented as lesion	0.55 (p < 0.25)	**0.54 (p < 0.03)**
Activity of liver involvement	Lesion and organ	PET and shape	Maximum SUV value of all voxels segmented as both liver and lesion	0.55 (p < 0.15)	**0.58 (p < 0.01)**

Table 6: All manually created surrogate biomarkers with their respective C-index and p-values for each cohort. If the feature was binary, the significance was assessed with a long-rank test, otherwise with a permutation test. The significant C-index are highlighted in bold. The 10 surrogate biomarkers prognostic on the two cohorts are listed first, then the 7 surrogate biomarkers prognostic on the DLBCL cohort only, and then the 5 surrogate biomarkers only prognostic on the FL cohort.

Figure 48 shows the correlation between these 22 surrogate biomarkers and the 2×28 biomarkers selected by the ROBI pipeline. In this figure, it is observable that several selected biomarkers are correlated with several surrogate biomarkers, while some surrogate biomarkers encode very well a selected feature, while being much easier to interpret. Most but not all selected biomarkers were correlated with at least one surrogate biomarker: 79% had a correlation above 0.50 with at least one surrogate biomarker, and 40 % had a correlation greater than 0.70 with at least one surrogate biomarker. We were not able to understand the information reflected by three selected biomarkers as they were not linked to any surrogate biomarker (duodenum_shape_Sphericity, urinary_bladder_PT_glszm_GrayLevelNonUniformity, small_bowel_shape_SurfaceVolumeRatio).

A model predicting PFS with baseline features alone (TMTV for FL; TMTV & Dmax for DLBCL) had a C-index of 0.59 in FL, and 0.65 in DLBCL. When the 10 surrogate biomarkers found to be prognostic in the two diseases were added to the models, the C-index increased up to 0.64 in FL ($p < 0.19$) and 0.68 ($p < 0.42$) in DLBCL. When further introducing in the model the surrogate biomarkers that were prognostic only in the disease of interest (7 extra biomarkers for DLBCL and 5 for FL), the C-index went up at 0.65 in FL ($p < 0.13$ compared to TMTV only) and up to 0.69 in DLBCL ($p < 0.22$ compared to a model involving TMTV and Dmax only). A multivariate model with the baseline features along the selected biomarkers (and without any surrogate biomarkers) achieved similar performance of 0.66 ($p < 0.07$) in FL and 0.68 ($p < 0.39$) in DLBCL.

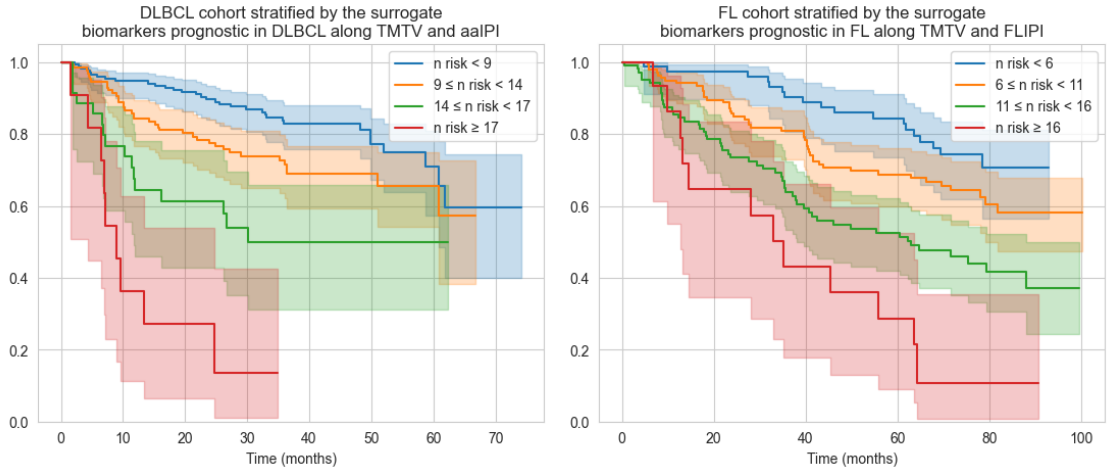


Figure 49: For each cohort, Kaplan-Meier curves showing the PFS of patients stratified by number of risk factors among TMTV, IPI (aaIPI for DLBCL, FLIPI for FL) and the surrogate biomarkers prognostic in each disease. All features were dichotomized. A IPI score > 2 was considered high risk. A TMTV above the median (328 cm^3 in FL, 292 cm^3 in DLBCL) was considered high risk. For other features, a value higher than the median was considered high risk for features positively correlated with the risk, and a value below the median was considered high risk for features negatively correlated with the risk. All risk groups had significantly different outcomes according to logrank tests ($p < 0.05$).

Figure 49 shows the Kaplan-Meier curves for each cohort separated by the number of risk factors among TMTV, IPI (aaIPI for DLBCL, FLIPI for FL) and the surrogate biomarkers that are prognostic in each disease. It shows that the surrogate biomarkers improved on TMTV and IPI and further stratify patients into more precise risk groups. The categories depicted in Figure 49 were able to predict PFS with a C-index of 0.64 in FL, and 0.66 in DLBCL.

In FL, 77% of patients responded to treatment (either complete, unconfirmed complete or partial response) and 23% had progressive disease at 120 weeks. When tested for their ability to predict the response to treatment (with a cross-validated univariate logistic regression), 13 of the 28 biomarkers selected in FL significantly discriminated patients with response to treament vs patients with progressive diseases. The list of these 13 biomarkers and the corresponding balanced accuracy are given as supplemental data (Table S16). When testing the 16 surrogate biomarkers that were prognostic in FL, 7 of them discriminated responding FL patients from non-responding patients. The corresponding balanced accuracy and associated p-values are given in Table 7.

Biomarker	ROI type	Computed from	Balanced accuracy (p-value)
High Tumor Energy from the PET image	Lesion	PET	0.60 ($p < 0.01$)
High tumoral activity in the pancreas	Lesion and organ	PET	0.60 ($p < 0.01$)
Presence of homogeneous density lesion	Lesion	CT	0.58 ($p < 0.03$)
Presence of lesion in a region of homogeneous density	Lesion	CT	0.64 ($p < 0.01$)
Trachea involvement	Lesion and organ	Shape	0.57 ($p < 0.04$)
Large liver	Organ	Shape	0.59 ($p < 0.01$)
Low elongation of the pancreas	Organ	Shape	0.60 ($p < 0.01$)

Table 7: Surrogate biomarkers that significantly discriminated FL patients responding to treatment from FL patients with progressive diseases.

The following section presents some of the surrogate biomarkers with examples. Only the 10 surrogate biomarkers that were predictive for the two diseases are presented.

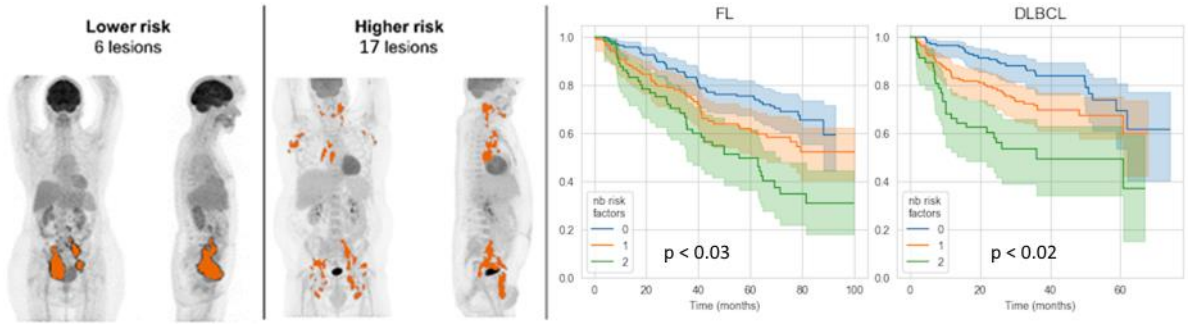


Figure 50: Maximum Intensity Projections (MIPs) of the PET images of low and high risks examples of the "High number of lesions" surrogate biomarker. Tumor segmentation is depicted in orange. The FL patient on the left had a TMTV of 355 cm^3 and the FL patient on the right had a TMTV of 371 cm^3. Kaplan-Meier curves of the PFS of the FL and DLBCL cohorts stratified with the biomarker and TMTV are displayed. Patient groups had significantly different outcomes. TMTV cutoff was 299 cm^3 for FL and 237 cm^3 for DLBCL, and biomarker cutoff was 23.

Some of the surrogate biomarkers identified describe the activity and the invasiveness of the tumor burden. Involvement of the trachea was found to be prognostic of the outcome. In addition, the tumor energy calculated from the PET image, which is defined as the sum of the squared SUV values inside all tumor regions, was found positively correlated with the risk. It was also positively correlated with TMTV (ρ = 0.69) and SUVmean inside all the tumor ROIs (ρ = 0.62). It had a C-index similar to TMTV on FL (0.59 and 0.60 for TMTV and Energy respectively), but lower than TMTV on DLBCL (0.63 and 0.57 for TMTV and Energy respectively). SUVmean was never significantly prognostic. The number of lesions was also a prognostic factor, as a higher number of lesions was correlated with higher risk and lower PFS. This feature is computed by counting the number of spatially disconnected regions of interest. This feature was highly correlated with Dmax (ρ = 0.80). Figure 50 shows examples of patients with low and high risks according to the number of lesions as well as Kaplan-Meier curves of the cohorts stratified based on this feature and TMTV.

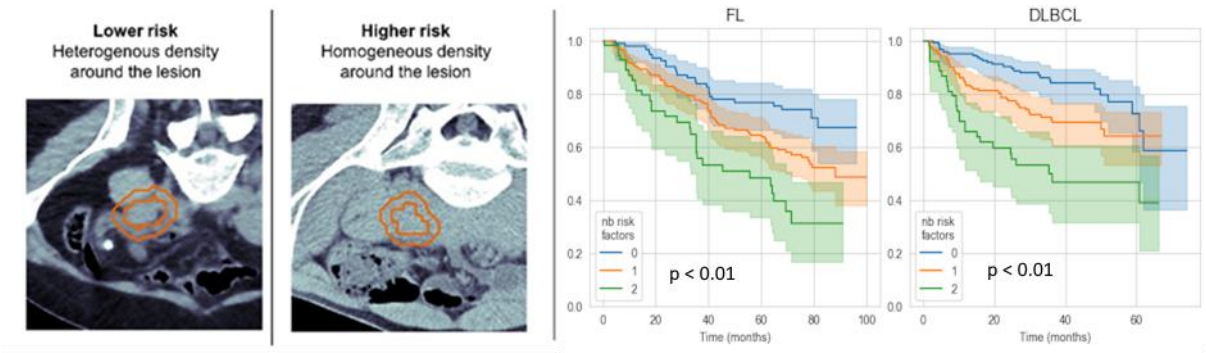


Figure 51: CT slices of low and high risks examples of the "Presence of lesion in a region of homogeneous density" surrogate biomarker. Density is given by a grey scale with the highest density being represented by white pixels. Tumor and shell surrounding lesion segmentations are depicted in orange. Kaplan-Meier curves of the PFS of the FL and DLBCL cohorts stratified with the biomarker and TMTV are displayed. Patient groups had significantly different outcomes ($p < 0.01$). TMTV cutoff was 849 cm^3 for FL and 364 cm^3 for DLBCL, and biomarker cutoff was 7.3 for FL and 6.9 for DLBCL.

Other surrogate biomarkers were focusing on a specific lesion. For instance, patients with at least one occult lesion, defined as being smaller than 20 mL and with a SUVmax < 5 were at higher risk than other patients. Similarly, it was found that the smaller the smallest lesion, the higher the risk. Other surrogate biomarkers were quantifying the homogeneity in density of lesions. Inside the lesions, the kurtosis of the CT values was calculated in each spatially disconnected lesion. The highest value among the patient lesions was used to stratify the patient. A higher maximum kurtosis (e.g., higher homogeneity) of the tumor density was associated with higher risk. Similarly, in the shell surrounding the lesions, the Dependence Entropy of the GLDM matrix was calculated in the 8 mm thick shell of tissues surrounding each lesion. The lowest value among all lesions was used to stratify the patient. A lower value (e.g., higher homogenity) was associated with an increased risk. Figure 51 shows examples of lesions of patients with low and high risks according to this feature, as well as Kaplan-Meier curves of the cohorts stratified using this feature and TMTV. While this last feature is related to the volume of the smallest lesion ($\rho = 0.44$), it is not entirely explained by it and carries additional prognostic information. We found no strong correlation between these last two features and the tumor location or host organ.

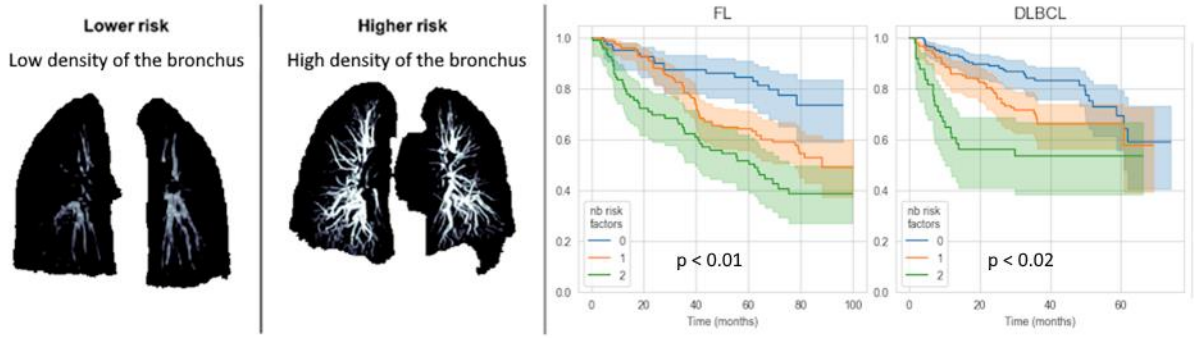


Figure 52: Maximum Intensity Projections (MIPs) of the CT images of low and high risks examples of the "High bronchus density" surrogate biomarker. Density is given by a grey scale with the highest density being represented by white pixels. Kaplan-Meier curves of the PFS of the FL and DLBCL cohorts stratified with the biomarker and TMTV are displayed. Patient groups had significantly different outcomes. TMTV cutoff was 437 cm^3 for FL and 326 cm^3 for DLBCL, and biomarker cutoff was 5132139 in FL and 8441570 in DLBCL.

Several surrogate biomarkers were not directly related to the tumor burden, nor to the features of a specific lesion, but potentially to the overall health state of the patient. In particular, the volume of subcutaneous fat divided by the volume of muscle, both segmented by MOOSE on the CT images, had a significant prognostic value (first row of Table 6). A low ratio (e.g., low volume of subcutaneous fat) was associated with shorter PFS. Such association was not observed for visceral fat. Likewise, patients with larger liver were at higher risk than other patients. This did not reflect fatty liver since only one FL patient was reported to have such condition. Another surrogate biomarker in this category reflected the density of the bronchus. Higher density was associated with higher risk. While we did not find any strong link to a specific commorbidity, patients with higher ECOG had a significantly higher density of the bronchus ($p < 0.05$). It was not correlated with lung involvement ($\rho = 0.15$). This feature was not correlated with the injection of contrast agent ($\rho = 0.07$ in FL, $\rho = 0.06$ in DLBCL). There were 24 FL patients and 13 DLBCL patients who received contrast agent, and no significant difference in the density of the bronchus was observed between patients with and without contrast agent ($p < 0.59$ in FL and $p < 0.58$ in DLBCL). Figure 52 shows examples of patients with low and high risks according to this feature, as well as Kaplan-Meier curves of the cohorts stratified using this feature and TMTV.

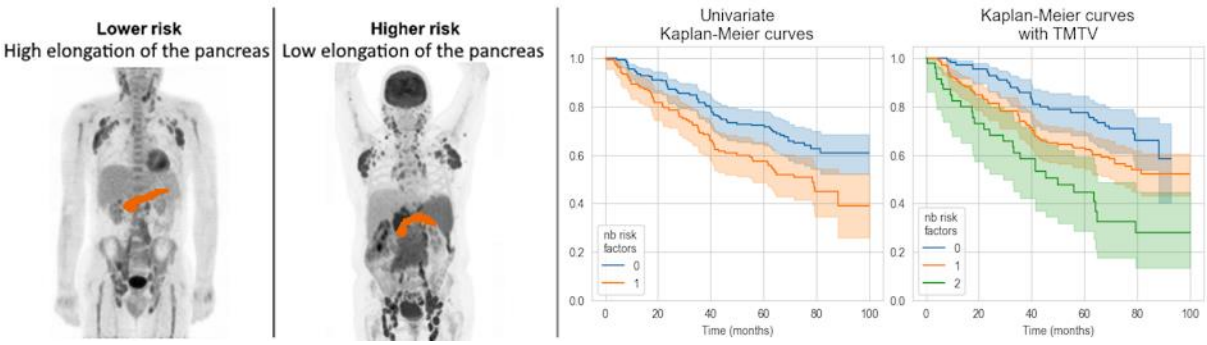


Figure 53: Low and high risks FL patients, univariate and multivariate PFS Kaplan-Meier curves of the FL cohort stratified using the "Low elongation of the pancreas" surrogate biomarker. Pancreas segmentation is depicted in orange. Patient groups had significantly different outcomes ($p < 0.02$). TMTV cutoff was 263 cm^3 and biomarker cutoff was -0.35 in univariable and -0.40 in multivariable analyses.

A last feature worth mentioning was computed from the pancreas ROI. In FL, the shape of the pancreas was found to be correlated with the PFS, with FL patients with more elongated pancreas having longer PFS. This feature was not prognostic in DLBCL. Elongation measures how stretched out a ROI is. We found that this feature could identify diabetic FL patients with a balanced accuracy of 0.69 ($p < 0.001$), with diabetic FL patients having a lower elongation. Diabetic FL patients had a lower PFS but this difference was not significant ($p = 0.71$). 27 FL patients were reported to have diabetes. Figure 53 shows examples of patients with low and high risks according to this features, as well as Kaplan-Meier curves of the cohort stratified using this feature alone or associated with TMTV.

Discussion and Conclusion

In this work, we extensively tested thousands of candidate biomarkers to discover new prognostic ones. This comprehensive approach yielded 28 new radiomic features prognostic of the PFS in FL patients, and 28 other radiomic features prognostic of the PFS in DLBCL patients. Through manual examination of these features, we identified 22 new prognostic biomarkers, most of them being easy to understand and compute. Among them, 10 features were more likely to be true prognostic biomarkers, as they were found prognostic of both DLBCL and FL patients. These 10 features were also moderately correlated with one another, offering a diverse set of biological information.

The candidate biomarkers were identified using the open-source ROBI pipeline [230]. This tool allowed us to select biomarkers that add new information with respect to TMTV and clinical features. It also reduced the risk of selecting biomarkers that would be prognostic of the PFS only by chance. By controlling the probability of false prognostic biomarker discovery, ROBI focuses on the most promising ones. This makes the selection process non exhaustive, with real prognostic biomarkers that may not be identified (false negatives). Yet, the goal was not to discover all biomarkers that might

be prognostic for patients but identify some strong ones and try to understand what they reflect.

Because of the low number of patients who died during the clinical trials, evaluating the candidate biomarkers on the prediction of OS was not an option since the statistical power was too low. We thus rather focused on the progression free survival endpoint, which is a proxy of the OS.

The selection process for biomarkers was carried out with rigorous and comprehensive testing. However, the interpretation of the selected biomarkers was more qualitative, subjective, and based on hypotheses. While certain biomarkers are straightforward and self-explanatory, such as the volume of subcutaneous fat and trachea involvement, many were far less obvious and were broken down into simpler features. This breakdown was achieved through visual examination and comparison with more straightforward candidate biomarkers. Despite the effort to simplify these features, it is important to note that this manual process may not fully capture all information contained within the biomarkers, potentially overlooking key details that are crucial for patient classification. Moreover, the simplified biomarkers are likely individually less powerful than the selected ones since they do not encompass multiple prognostic information such as many selected biomarkers. It is worth noting that in both FL and DLBLC, a multivariate model with all surrogate features was as predictive than one with all selected biomarkers, meaning that it is likely that a great fraction of the prognostic information coded in the selected biomarkers was successfully re-encoded in the surrogate biomarkers.

Interestingly, while multiple lesion-based biomarkers were selected, a significant number of the selected features do not seem to be directly linked to the tumors but rather the patient overall health status and comorbidities. If these findings are confirmed, this would demonstrate the interest of image-based radiomics to discover non image-specific information that could be measured without imaging equipment. As we observed, diabetes is a good example of this, as the shape of the pancreas on the CT image is a good predictor of both the diabetic status and the PFS of the FL patients.

If some of the multiple biomarkers presented in this study are validated through external cohorts, the challenge then becomes how to utilize them effectively. We demonstrated that these features enable a more accurate stratification of patients compared to using only TMTV and the IPI. They have the potential to identify patients with either extremely good or poor prognoses in both DLBCL and FL. Their greatest value, however, may lie in stratifying patients who appear similar when assessed using currently used features. For example, a group of FL patients with comparable TMTV and identical FLIPI scores may still not all have close outcome. Introducing a new biomarker for this homogeneous group could enable the identification of high and low-risk patients, thereby refining therapeutic strategies. As an illustration, we analyzed

18 FL patients from the RELEVANCE trial who had similar TMTV values (ranging from 291 cm^3 to 393 cm^3, corresponding to ± 5 percentiles around the median) and identical FLIPI scores of 3. FLIPI being constant, it cannot differentiate patients, and TMTV being similar, the C-index for PFS prediction in that group was 0.54 ($p < 0.73$) . However, some surrogate biomarkers were different between these 18 patients. For instance, the "number of lesions" biomarker achieved a C-index of 0.71 ($p < 0.04$) in predicting PFS for this population. Additionally, this specific group could be subdivided based on the size of the smallest lesion in each patient, resulting in significantly different PFS outcomes ($p < 0.02$). Another approach to effectively implement these biomarkers in clinical practice involves constructing predictive models or signatures, similar to the FLIPI, which integrates multiple features to provide a risk score. More precise signatures could be developed by incorporating several of the identified biomarkers. This research represents an initial step in that direction. If the biomarkers that we have discovered here are further validated in other DLBCL and FL patient cohorts, then we believe they can be integrated into a single model to enhance NHL patient stratification.

One limitation of this study is the definition of organ involvement. We used an automated tool to segment organs, producing approximate segmentations. We also used a simple rule to define organ involvement (at least one voxel shared between the segmentation of the organ and the segmentation of the lesions). This makes the definition of involvement prone to error, and in some cases, it might reflect the involvement of neighboring tissues rather than the organ itself.

While we minimized the risk of selecting false positive biomarkers from the DLBCL and FL cohorts through the ROBI settings and also by re-engineering the discovered prognostic features into simpler biomarkers whose prognostic value was confirmed in both DLBCL and FL cohorts, whether the biomarkers presented in this study will be prognostic in other cohorts of DLBCL or FL patients warrants further investigation. Since some biomarkers identified in this study were prognostic in the two diseases, we can also wonder if they may have prognostic value in other lymphoma subtypes or even other cancers, similar to TMTV which appears to be a good prognosticator for a variety of cancer types [234], [235], [236], [237]. We can also wonder if some of the identified biomarkers could help predict if a FL will transform in DLBCL. Thus, external validation on multiple external cohorts is of utmost importance to confirm our discoveries. To help in this endeavor, we made the code to compute and test the biomarkers freely available on GitHub at https://github.com/Lrebaud/exhaustive_radiomics.

7.3 Discussion

In this study, we used the previously developed ROBI pipeline to find new relevant biomarkers for prognosis of non-Hodgkin lymphoma patients. Through the extensive screening of thousands of candidate biomarkers measured in the images, we were able to identify 28 new radiomic features predictive of the PFS in the FL patients and 28 other features prognostic of the PFS in patients with DLBCL. The analysis of these radiomic features led us to develop several intuitive hypotheses to explain the biological information reflected by the selected features. Reencoding these hypotheses into simpler features allowed us to test our interpretation assumptions on the cohorts. Doing so, we could identify 22 simple features bearing distinct prognostic information, and 10 of them appeared to have a significant prognostic value on the two cohorts. While some of these features are related to lymphoma itself, such as trachea involvement or the presence of lesions with homogeneous density, some rather reflected the overall state of health of the patients, like the size of the liver or the amount of subcutaneous fat. We even found links between some of the selected features and diabetes or ECOG status.

The fact that some of these features were significantly prognostic in two cohorts, that is in two diseases, increases the likelihood of them being truly prognostic biomarkers. Of course, external validation is still required to rule out false positive findings.

One important aspect of this work is the interpretation of the features. Being able to identify key biological information encoded in the selected biomarkers and re-encode them in simpler terms increased both the interpretability but also the robustness. Despite our efforts, we were not able to provide an intuitive explanation for every feature. The prognostic power of the homogeneity in density of the surrounding of the lesions for instance, is still unexplained despite extensive research and numerous hypotheses tested. Furthermore, convoluted feature definitions do not inspire trust nor give physicians intuitive insights. It is worth noting that this specific feature, despite its cryptic nature, was one of the features with the lowest p-values on both DLBCL and FL cohort.

Beyond the identification of novel prognostic biomarkers, they will be of no use if we do not know how to use them for patient management.

Chapter 8

Development of a new machine learning model (ICARE) during a competition (HECKTOR)

8.1 Introduction

In the previous chapter, we identified multiple new prognostic biomarkers in both FL and DLBCL. These image-based features bring additional knowledge about the diseases and have the potential to improve patient prognosis. However, it is extremely challenging for oncologists to deal with too many biomarkers of prognostic values. They must already consider many parameters coming from numerous modalities (e.g., images, clinical exam, biopsy, blood tests, genetic). If the number of features in each modality is too high, it can become impossible to efficiently combine all the information coming from patient data. For this reason, models are often used to aggregate multiple features in one risk score that is easier to handle. IPI and AnnArbor are examples of such scores. They are created based on concertation between experts and numerous analyses and publications.

An alternative approach consists in the use of machine learning models to learn from the data a score (e.g., signature) based on feature values and outcome observed in many patients. Often, a Cox model is trained to predict outcomes of patients based on a set of features. It can be converted into a nomogram for easy use and deployment. This model works by assigning a signed weight (e.g., hazard ratio) to each feature.

However, during the search for new biomarkers, we found that defining which feature is more prognostic than another is extremely challenging, given the noise affecting the feature values and the outcome to be predicted. The weights in a model being a way to rank features as a function of their impact on the outcome, we emit the hypothesis that in certain situations, it might be preferable to not learn any weight, but simply a sign for each feature. This way, each feature contributes equally to the prediction. By not weighting the features, a model would minimize the risk to learn non generalizable information. This intuition led to the development of the ICARE model detailed in this chapter.

We tested this idea during the HECKTOR challenge held during the MICCAI 2022 conference. In this competition, different teams worldwide could design a model to

automatically segment head and neck tumor and invaded lymph nodes on 18F-FDG PET/CT image of head and neck cancer patients coming from many hospitals. A second task was to build a model that could predict the risk of relapse for each patient based on the PET/CT images and associated clinical data. For the segmentation, we used a simple nnUNet and ranked 4th among the 36 participating teams. For the outcome prediction, we ranked 1st among the 18 participating teams with the ICARE model. In this article, we present the details of our solutions to the challenge.

8.2 Article published

Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT

PUBLISHED IN LECTURE NOTES IN COMPUTER SCIENCE

*Louis Rebaud[1,2], *Thibault Escobar[1,3], Fahad Khalid[1], Kibrom Girum[1], and Irène Buvat[1]

[1]Siemens Healthcare SAS, Saint Denis, France; [2]LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France; [3]DOSIsoft SA, Cachan, France.

*co-first authors

Abstract

Automated lesion detection and segmentation might assist radiation therapy planning and contribute to the identification of prognostic image-based biomarkers towards personalized medicine. In this paper, we propose a pipeline to segment the primary and metastatic lymph nodes from fluorodeoxyglucose (FDG) positron emission tomography and computed tomography (PET/CT) head and neck (H &N) images and then predict recurrence free survival (RFS) based on the segmentation results. For segmentation, an out-of-the-box nnUNet-based deep learning method was trained and labelled the two lesion types as primary gross tumor volume (GTVp) and metastatic nodes (GTVn). For RFS prediction, 2421 radiomic features were extracted from the merged GTVp and GTVn using the pyradiomics package. The ability of each feature to predict RFS was measured using the C-index. Only the features with a C-index greater than C_{min}, hyperparameter of the model, were selected and assigned a +1 or –1 weight as a function of how they varied with the recurrence time. The final RFS probability was calculated as the mean across all selected feature z-scores weighted by their +/–1 weight. The fully automated pipeline was applied to the data provided through the

HECKTOR 2022 MICCAI challenge. On the test data, the fully automated segmentation model achieved 0.777 and 0.763 Dice scores on the primary tumor and lymph nodes respectively (0.770 on average). The binary-weighted radiomic model yielded a 0.682 C-index. These results allowed us to rank first for outcome prediction and fourth for segmentation in the challenge. We conclude that the proposed fully-automated pipeline from segmentation to outcome prediction using a binary-weighted radiomic model competes well with more complicated models. Team: LITO.

Keywords

Medical imaging, Survival prediction, Segmentation, FDG PET/CT, Head and neck, Machine learning

1. Introduction

Quantitative medical image analysis assists in patient staging, treatment planning and monitoring, and overall patient management. In head and neck (H &N) cancer, fluorodeoxyglucose (FDG) positron emission tomography combined with computed tomography (PET/CT) is a modality of choice for initial staging and patient follow-up and contributes to radiation therapy planning. Indeed, H &N cancer primary treatment mostly relies on radiotherapy and requires target volume delineation of the gross primary tumor volume (GTVp) and cancer node volumes (GTVn) on PET/CT images, which is time-consuming and prone to intra/inter-observer variabilities. Automated segmentation might allow radiation oncologists to optimize the treatment plan in a shorter time while improving reproducibility. In addition, the prediction of the risk of relapse based on medical images could help identify patients for whom treatment intensification and close monitoring might be needed.

In the recent years, machine learning (ML) and radiomics have been instrumental in advancing automated image segmentation and building predictive models. Yet, the diversity of datasets on which methods are designed and tested makes it difficult to compare their performance and determine which one is best suited in a particular context. Given the possible sensitivity of automated segmentation and predictive models to image quality, multi-center evaluation of these methods is absolutely needed before considering clinical deployment.

Challenges offer unique opportunities for testing and comparing the performance of different methods on a common database using large multi-center datasets. The HEad and neCK TumOR (HECKTOR) challenges organized as part of MICCAI aims at establishing best-performing methods for segmentation and prediction tasks [238], [239]. In 2022, the HECKTOR challenge first task was to automatically segment the H &N GTVp and GTVn from FDG PET/CT images. The second task consisted in automatically predicting patient outcomes from a PET/CT image, with or without clinical information, with PET/CT images and clinical information collected from nine

different centers.

Several contributions to the automated segmentation in the context of H &N cancer have already been published. Guo et al. proposed a modified U-net approach using dense blocks and reached 0.71 average Dice score on a public multi-center dataset of 250 PET/CT H &N patients [240]. Their study also showed that combining PET and CT in two channels substantially increased the segmentation performance compared to using PET (0.64 average Dice score) or CT (0.31 average Dice score) alone. Ren et al. compared several modality combinations including PET, CT, and magnetic resonance imaging (MRI) on a multi-center dataset of 153 patients for deep learning tumor segmentation using a U-net approach [241]. All combinations including PET provided similar results (0.72 to 0.74 Dice score), while the anatomic-only combination (CT and MRI) led to a lower score (0.58). More generally, automated medical image segmentation is currently dominated by deep convolutional neural networks (CNN) [242], [243], [244]. Most methods rely on U-net based approaches with several context-specific changes in model architecture, training scheme, and data pre- or post-processing. In HECKTOR 2021 challenge, the best-performing segmentation method used a tuned nnUNet with squeeze and excitation (SE) layers on fused PET and CT images, yielding a 0.779 Dice score on primary tumor [244], [245].

Similarly, models have been proposed to predict patient outcome from PET/CT images in H &N cancer (e.g., [246], [247]). In HECKTOR 2021, two different methods performed best at predicting the progression free survival [248], [249]. Both were based on a CNN trained on unsegmented images using large bounding boxes, and achieved 0.720 and 0.694 C-index on the test data respectively. A logistic model based on radiomic features calculated from the segmented tumor region also performed well with a 0.683 C-index [250].

This paper presents our simple and efficient pipeline for fully automatic segmentation and outcome prediction method and its performance on the HECKTOR 2022 challenge data. For the segmentation task, we adapted the publicly available nnUNet deep learning framework to detect and segment the H &N primary tumor (GTVp) and nodal gross tumor volumes (GTVn) [244]. For the prediction task, we introduce a novel binary-weighted model operating on radiomic features calculated from the tumor regions automatically segmented in the previous step. The evaluation was conducted on the HECKTOR 2022 challenge data and the models are publicly available.

2. Materials and methods

Here, we describe our proposed fully-automatic end-to-end framework to segment lesions and predict outcome from 18F-FDG PET/CT images (Figure 54). First, a well established out-of-the-box nnUNet deep learning method was trained to segment and label the GTVp and GTVn [244]. From the segmented GTVp and GTVn regions, we extracted radiomic features. We then applied the binary-weighted model to rank the patients as a function of their recurrence free survival.

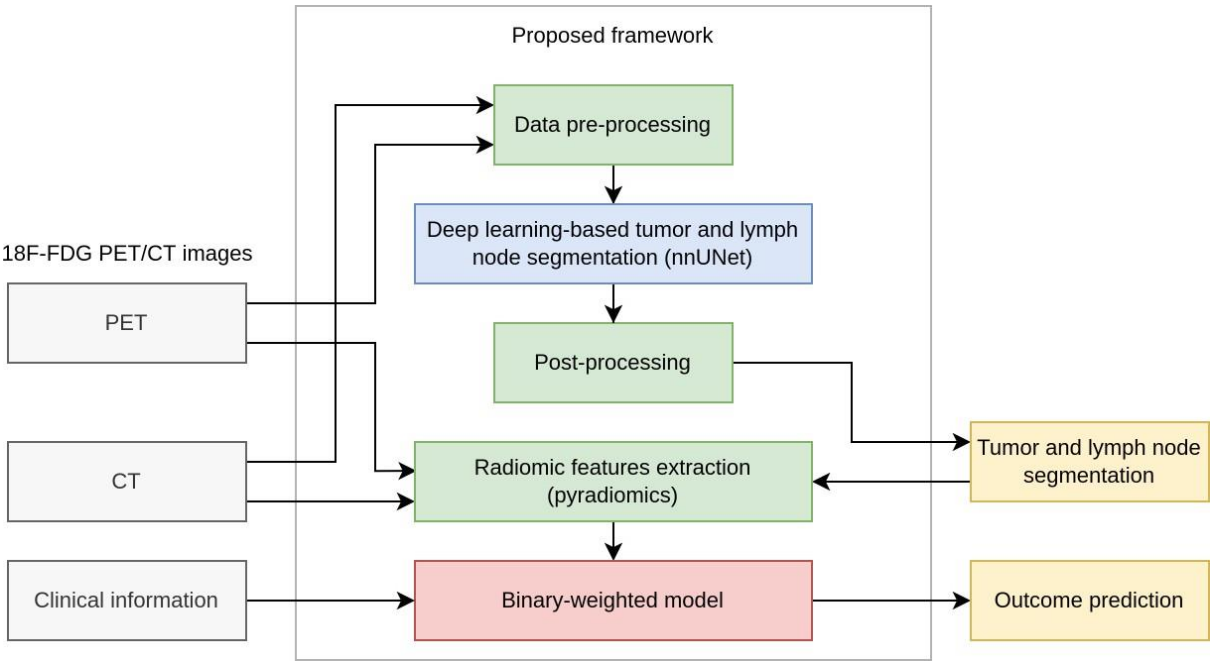


Figure 54: Proposed framework: schematic representation of the fully-automatic pipeline from segmentation to outcome prediction.

2.1. Data

To develop and evaluate the proposed method, we used the HECKTOR 2022 data that included FDG PET/CT images, clinical and survival data of 524 patients from 7 centers for training and PET/CT and clinical data only of 359 patients from 3 centers for blind testing of the models [238], [239]. In the training data, reference segmentations of the primary tumor (GTVp) and metastatic nodes (GTVn) were provided. Train and test PET/CT scans were provided with 9 clinical features with some missing values: gender, age, weight (1.23% missing values), tobacco (0 = no, 1 = yes) (61.1% missing), alcohol (0 = no, 1 = yes) (68.5% missing), performance status (56.0% missing), human papillomavirus (HPV) status (0 = no, 1 = yes) (35.2% missing), surgery (0 = no, 1 = yes) (38.7% missing), and chemotherapy (0 = no, 1 = yes). RFS was provided for 488 patients in the train set, and 339 patients of the test set for whom the outcome was known were concerned by the outcome prediction (task 2).

Data Pre-processing: The training CT images had an original median voxel-size of $0.976 \times 0.976 \times 2.798$ mm^3 and the PET images had median voxel-size of $4.000 \times 4.000 \times 3.270$ mm^3. All PET/CT images and corresponding segmentations were resampled to $2.0 \times 2.0 \times 2.0$ mm^3. CT and PET images were resampled using a third-order spline. The segmentation mask was resampled using nearest neighbor interpolation.

2.2. Tumor and lymph node segmentation

Deep Learning Model: All CT images were clipped between 0.5th and 99.5th percentile of the Hounsfield Units (HU) intensity values and normalized using z-score based on all training images. To favor contrast-based features in PET, PET standardized uptake values (SUV) were normalized using z-score patient-wise on the whole image. We used a nnUNet in "3D full resolution" mode to detect and segment the tumor and lymph nodes [244]. The pre-processed PET/CT images were given to the model as two-channel input images (PET and CT). Each PET/CT image was decomposed in random patches of 160×160×96×2 voxels before input into model training. The architecture of the 3D model was not modified except for the output channel. The output was a 1×1×1 convolution of size 160×160×96×2, where 2 corresponds to the tumor and lymph nodes channels. A softmax non-linear activation was used at the output layer of the 3D nnUNet model.

Training Scheme: The train set consisting of 524 patients was randomly divided into training and validation subsets using a five-fold cross-validation technique. Each fold contained data from 104 or 105 validation patients and 420 or 419 training patients. The nnUNet model was trained using the sum of Dice and cross-entropy losses. The initial number of feature maps in the architecture was 32. Performance assessment and post-processing strategy were determined based on the five-fold cross-validation with 1000 epochs training, with an initial learning rate of 0.01 and a scheduler weight decay of $3e^{-5}$. We selected a batch size of two. Other hyper-parameter settings, including data augmentation techniques, were the default settings of nnUNet. Implementation was done in Pytorch and training was performed using four GPUs: three NVIDIA Quadro RTX 5000 with 16 GB and one NVIDIA RTX A6000 with 49 GB GPU memory. On average, the training time was 141 s per epoch on NVIDIA Quadro RTX 5000 and 82 s on NVIDIA RTX A6000.

Post-processing: The segmentation output of the deep learning model had a 2×2×2 mm voxel spacing. It was then resampled into the corresponding original CT spacing. Then, a median filter with a 3×3×3 voxel kernel size was applied to smooth out the staircase effect.

Prediction on the Test Set: For predictions on the test set, three strategies were used. First we ensembled the five models trained during cross-validation. Second, a bagging strategy was adopted to increase the number of ensembled models to nine. Nine models were trained on random samples of size equal to the whole dataset drawn with replacement (i.e. bootstrap samples). The predictions from the models were then aggregated using majority voting. Nine was the maximum number of models we could train on our GPUs for this strategy within the allotted time of the challenge. Finally, we increased the number of epochs to 1500 and trained only one model on the whole dataset.

2.3 Outcome Prediction

Our prediction model was based on engineered radiomic features extracted from the tumor regions segmented using the automated approach described in Sect. 2.2. These features were then analyzed using an original approach yielding what we call a binary-weighted model.

Radiomic Features Extraction: We used the segmentation mask produced by the deep learning model described in Sect. 2.2. Primary tumor and lymph node regions were merged as a single "lesion" mask. To make the model less sensitive to potential segmentation errors, multiple masks were created from this binary lesion mask:

- Original lesion mask
- Smallest bounding box enclosing all the lesions
- Lesion mask refined by removing all voxels in which SUV was less than 2.5
- Lesion mask refined by removing all voxels in which SUV was less than 4
- Lesion mask re-segmented with a threshold of 40% of global SUVmax
- Lesion mask dilated by 1mm (resp 2, 4, 8 and 16 mm)
- A 2mm (resp 4, 8 mm) thick shell surrounding each connected component of the lesion mask

For each of these 13 masks, 93 radiomic features were computed on the PET image and 93 on the CT image with pyradiomics [56]. These features were the default features from pyradiomics, composed of features reflecting the ROI shape, and the signal intensity and texture. A fixed-bin size of 0.3 SUV units was used for PET images and 10 HU for the CT. Three handcrafted features were added: the number of tumor masses, the number of lymph nodes, and a binary variable indicating whether the scan was a whole-body scan or included only the H &N region. This was determined by calculating the length of the scan in the axial direction from the image volume. Used together with the provided nine clinical features, this pipeline produced 2430 features.

Binary-weighted Model: From the literature and our experience, we hypothesize that it is difficult to accurately estimate biomarker importance in outcome prediction. For instance, Adams et al. found the national comprehensive cancer network international prognostic index to be more predictive of progression free survival than whole-body total metabolic tumor volume in diffuse large B-cell lymphoma, while Cottereau et al. observed the opposite [212], [251]. Indeed, noise in the data, censoring of the target, e.g. progression free survival, and relatively low number of training samples might increase the risk of biased estimation of the feature weights. To mitigate this effect, we propose to reduce the learned information to the bare minimum and only estimate a sign to be assigned to each feature for estimating the target. This is the core mechanism of the introduced binary-weighted model.

Definition: Our training dataset includes N samples and M features. Many radiomic features are highly correlated. To comply with the basic assumption of our binary-weighted model, only one among a set of correlated features should be kept because if they are all input to the model, this will artificially give a large weight to the information reflected by the feature. We thus perform feature selection by calculating the absolute value of the Pearson correlation coefficient for all pairs of features. A threshold ρ is used to set the value above which two features are deemed too correlated. In such case, one of the two features is randomly selected and dropped.

Let's C_{index} be the Harrell's concordance index [252]. Each feature x_i is evaluated on its ability to correctly predict the target value y with:

$$c_i = C_{index}(x_i, y) \tag{1}$$

To reduce the risk of wrong estimation of the sign, the features with $|c_i| < C_{min}$ are dropped, where $|c_i| = max\{1 - c_i, c_i\}$ and C_{min} is a hyperparameter in $[0.5, 1]$. The remaining features are assigned a sign as follows:

$$s_i = \begin{cases} +1, & if\ c_i \geq 0.5 \\ -1, & otherwise \end{cases} \tag{2}$$

A normalization step is necessary to scale the feature values to the same range. Otherwise, features with large absolute values would have a higher weight in the final prediction. To do so, the model computes the z-score of each feature:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3}$$

Where μ_i and σ_i are the mean and standard deviation of x_i in the train set. The estimate $\hat{y}$ of the target y is computed with:

$$\hat{y} = \frac{1}{M} \sum_i^M s_i \times z_i \tag{4}$$

The computation of $\hat{y}$, μ_i and σ_i are done by ignoring the missing values of the dataset. This allows the model to use features with missing values.

Here, C_{min} and ρ and are the only two hyperparameters of the model.

Curse of Dimensionality: The curse of dimensionality is a phenomenon where we observe a loss in performance of ML models when too many features are given as an input. This especially occurs in medical datasets when the data are high-dimensional and the number of samples is low [253]. We hypothesize that the binary-weighted model is resilient to this phenomenon. We tested this hypothesis on the train set of the HECKTOR dataset by gradually increasing the number of features input to the model.

Ensembling: To produce a more precise and stable estimate $\hat{y}$, a bagging strategy was adopted as described in Sect. 2.2. An ensemble of E binary-weighted models were trained, each model being trained on a random sample of size N of the training data drawn with replacement. Each model also randomly selected F features to work with. The models were trained on their bootstrap sample from the train set and predicted $\hat{y}$ on the test set. The E predictions from the E models were then aggregated with the median. F is a hyperparameter of the ensemble model. Our experiments on the train set suggested that the higher E, the better the performance. We used $E = 10^5$ on the test set, a number large enough to ensure good results while keeping computational cost reasonable.

Cross-validation: To evaluate a model from the train set, we used a two-hundred-fold Monte Carlo cross-validation with a validation set of size $0.5 \times N$ (CV). This large number of folds was used to ensure precise comparison of the numerous tested algorithms, with reproducible results. The model prediction on the validation set was evaluated with Harrell's C-index. The average score and its confidence interval were reported.

Hyperparameters Optimization: The ensemble model has 3 hyperparameters: F, C_{min} and ρ. To determine the best hyperparameter set, random search was used. 1000 hyperparameter sets were randomly drawn and evaluated using CV. The hyperparameter sets were then ranked by their CV scores. To reduce the risk of overfitting the hyperparameter choice on the train set, the B best hyperparameter sets were selected, and for the prediction on the test set, an ensemble model was trained with each binary-weighted model randomly selecting a hyperparameter set from the selected B. The B value was optimized with an additional CV. Three bagged models were evaluated in the train and test sets of the HECKTOR challenge. While similar, each model used more and more hyperparameter sets in its random search, each time increasing the probability of overfitting on the train set. The number of hyperparameter sets tested was increased gradually through the 3 attempts given to the participating teams.

Feature Importance: While the binary-weighted model only gives weights of –1 or +1, after bagging, an approximation of feature importance can be computed by taking the average sign of each feature across all models. Feature importance was determined on the train set of HECKTOR.

3. Results

3.1 Segmentation Evaluation

In this section, except for the visual evaluation where it was assessed patient-wise, the Dice score was always computed on pseudo-volumes of the validation sets during cross-validation (aggregated Dice score).

Cross-validation: The Dice score across all images through the cross-validation was 0.850 for GTVp and 0.789 for GTVn (0.821 on average). For thorough comparison, Table 8 reports the Dice score across the different centers of acquisition.

Center	Patient #	GTVp Dice	GTVn Dice	average Dice
CHUP	72	0.868	0.687	0.778
CHUV	53	0.823	0.781	0.803
MDA	198	0.821	0.813	0.817
HMR	18	0.846	0.811	0.829
CHUS	72	0.865	0.805	0.835
CHUM	56	0.849	0.831	0.840
HGJ	55	0.883	0.829	0.856
All	**524**	**0.850**	**0.789**	**0.821**

Table 8: Dice scores for primary tumor and lymph node segmentation across the different centers evaluated on a five-fold cross-validation on the train set.

Test: Table 9 displays the class-specific Dice scores for our three submitted models for evaluation on the test set. The model trained on all training data for 1500 epochs achieved the highest scores (highlighted in bold).

Method	GTVp Dice	GTVn Dice	average Dice
Ensembled 5 folds	0.778	0.761	0.769
Bagging 9 samples	0.779	0.759	0.769
Whole train set	**0.777**	**0.763**	**0.770**

Table 9: Dice scores from our 3 methods on the test set of HECKTOR.

3.2 Qualitative Assessment

PET/CT images, ground truth and predicted segmentations are shown in Figure 55 for 5 patients. The examples were selected based on the Dice scores. The top two rows display high Dice scoring patients (average Dice 0.922 and 0.910 respectively), the third row a patient with an average score (0.761), while the fourth (0.303) and fifth (0.000) rows display patients with the lowest scores.

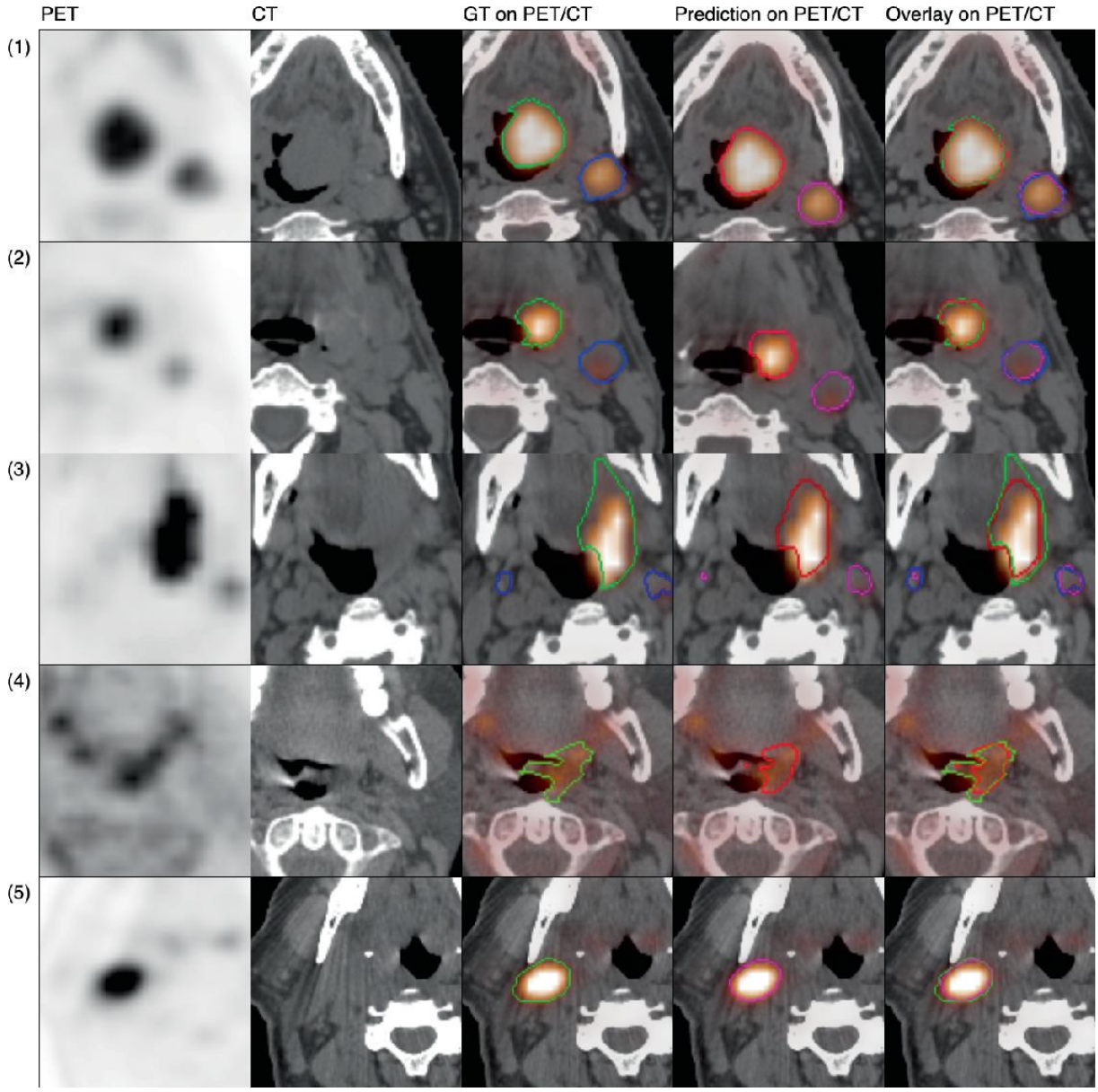


Figure 55: Examples of PET/CT images, ground truth and predicted segmentation for five patients from the validation sets of the five-fold cross-validation. Green and blue ground truth contours correspond to tumor and lymph node respectively. Red and pink contours correspond to the predicted segmentation for tumor and lymph node. (Color figure online)

Results for patients (1) and (2) were very satisfactory. In patient (3), the model accurately identified the two nodes and the tumor but missed some voxels, especially at the sharp edges. In patient (4), false positive node voxels were labeled by the model (not shown in the figure because not in the slice). Last, patient (5) shows an example of accurate detection and segmentation but with complete class mismatch. The green contour representing the tumor is precisely delineated by the model but labelled as a node, as shown by the pink predicted contour, yielding a Dice equal to zero.

3.3 Performance of the Outcome Prediction Model

Table 10 shows the results of the different models tested during the challenge. A binary-weighted model without bagging was evaluated only on the train set and not submitted because its performances were below the bagged models on the train set. The performance of the three submitted bagged models is correlated with the number of hyperparameter sets evaluated on the train set. The best model was the one which had the most extensive search of hyperparameters.

Model	CV C-index train set (CI)	C-index test set	Nb tested sets of hyperparameters
Binary-weighted	0.645 (0.585 - 0.707)		10
Binary-weighted bagged	0.668 (0.605 - 0.730)	0.670	10
Binary-weighted bagged	0.675 (0.613 - 0.731)	0.673	100
Binary-weighted bagged	**0.688 (0.642 - 0.732)**	**0.682**	1000

Table 10: C-index and number of hyperparameters searched for the prediction models evaluated on the train and test set of the HECKTOR challenge. On the train set, the mean C-index over the CV is reported as well as the confidence interval (CI).

3.4 Resilience to the Curse of Dimensionality

Figure 56 shows the result of the experiment using the train set to test our hypothesis stating that binary-weighted models do not suffer from the curse of dimensionality. The performance plateaued when increasing the number of features used by the model up to the maximum number of available features.

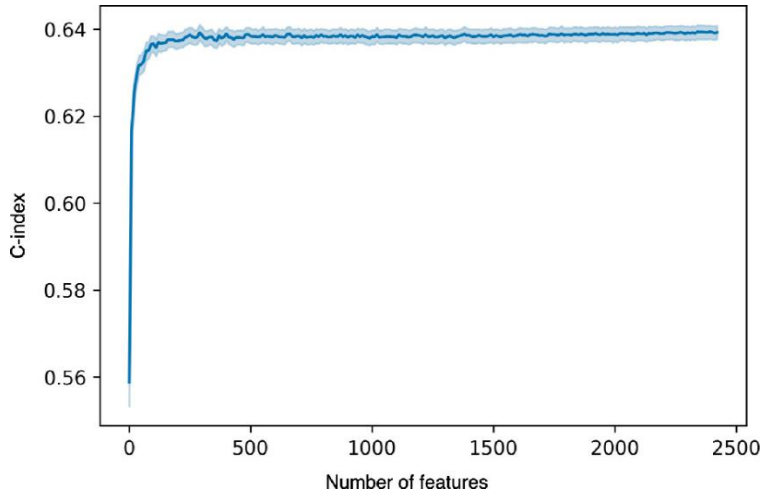


Figure 56: Cross-validated C-index of a binary-weighted model (not bagged) when increasing the number of features. The features and hyperparameters were selected randomly.

3.5 Feature Importance

The importance of the clinical and some representative radiomic features evaluated on the train set is presented in Figure 57. The error bars are not shown because by construction of the model, they are unnecessary (the higher the absolute value, the lower the standard deviation).

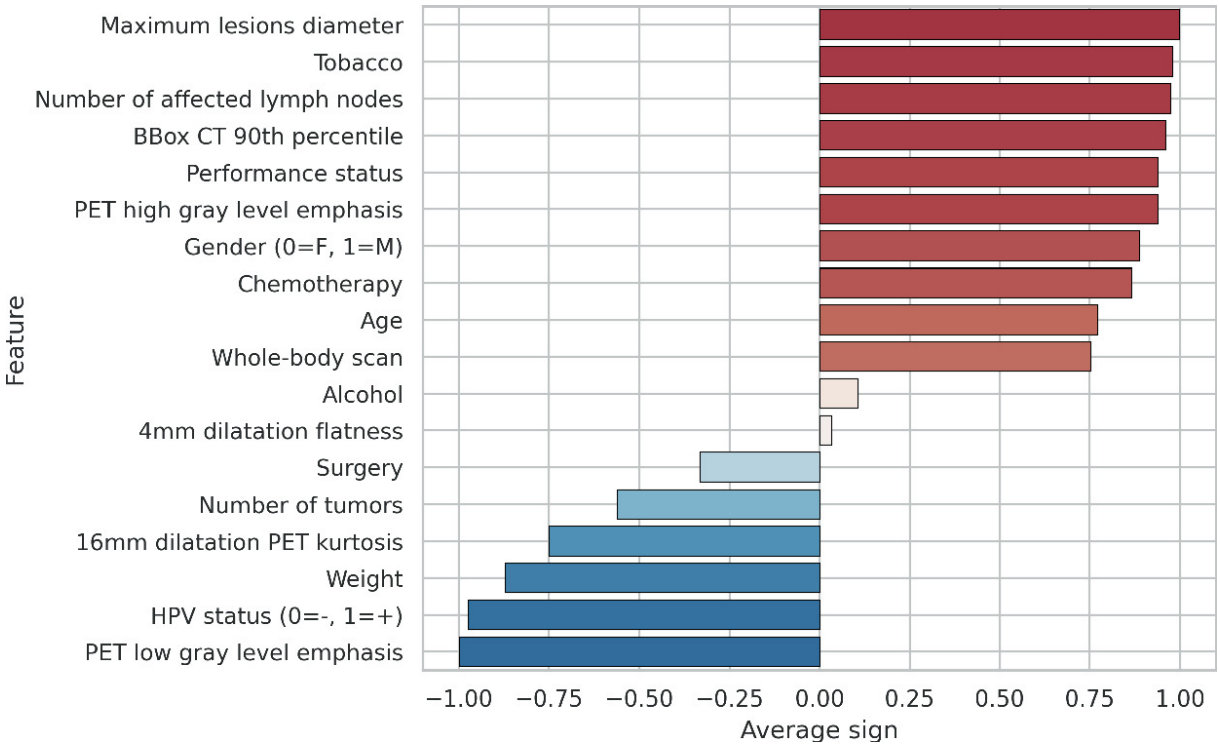


Figure 57: Importance of the clinical and representative radiomic features. A positive value (red) shows a positive correlation with the risk and a negative value (blue) is a negative correlation. The higher the absolute value of the average sign, the more important the feature. "Whole-body scan" is 1 if the scan is whole-body or 0 if only H &N. (Color figure online).

4 Discussion

4.1 Segmentation

Our segmentation method was inspired by Xie and Peng [245] using Isensee et al. [244] framework. Our choice of not using the SE layers and keep PET and CT separated as two channels was based on the intuition that approaching the problem in a straightforward way would increase its robustness. Overall, our segmentation results were satisfactory, ranking fourth in the challenge with 0.770 average Dice, compared to the 0.788 Dice achieved by the winner.

Although the centers had different numbers of patients, Dice scores were consistently lower for lymph nodes than for primary tumor in all centers, demonstrating they are more difficult to segment. Mislabelling of node regions as seen in Figure 55 decreased

Dice value although contours were accurately delineated. One way to address this mislabelling could be to set higher weight to the lymph node class in the loss function.

According to our test results, the deployment strategy did not have a big impact on performance. Indeed, ensembling the cross-validation models, using a bagging strategy while increasing the number of models, or training only one model on the whole dataset, led to very similar performance.

Based on the qualitative visual assessment, our model tends to perform better on smooth connected components. Complex structures and sharp contours are more prone to errors. Processing and training methods adapted to higher resolution input images might have reduced these errors.

4.2 Binary-weighted Model

Our results suggest that the binary-weighted model is a competitive and robust method. This implies that it might indeed be challenging to accurately estimate feature weights. The more degrees of freedom in a model, the higher the risk of overfitting. In problems with weak and noisy targets and low number of training samples, reducing the training to the bare minimum could be of utmost interest. For the HECKTOR challenge, it probably helped mitigate the overfitting.

Figure 56 shows that the binary-weighted model does not suffer from the curse of dimensionality. The vast majority of ML algorithms need some feature selection to avoid a drop in performance due to too many features. We hypothesized that in our binary-weighted model, the features would work together to cancel their noise and biases, analogous to the wisdom of the crowd phenomenon where errors of individuals cancel each other out. Adding more features does not result in loss in performance as in other traditional ML methods.

Features importance shed light on the model interpretation (Figure 57). For instance, a high performance status is associated with worse prognosis. Tobacco is also associated with a higher risk in our model. Large tumor diameter and high SUV values in the lesions are associated with increased risk. Other features, such as chemotherapy, can be interpreted as indirect measure of the patient condition. Interestingly, the number of affected lymph nodes appears to be a strong prognostic factor. In future work, the respective contribution of the different segmentation masks will be investigated. More importantly, separating GTVp and GTVn would make it possible to assess the individual role of these two lesion types.

5 Conclusions

We proposed a new, fully automated framework to predict outcomes in H &N patients from a given PET/CT image and clinical information. It involves deep learning-based GTVp and GTVn segmentation, radiomic feature extraction, and outcome prediction.

Our pipeline including the novel binary-weighted radiomic model outperformed other methods for outcome prediction while providing accurate segmentation, ranking first for prediction and fourth for segmentation in the HECKTOR 2022 challenge. The number of lymph nodes was one of the prognostic features, highlighting the importance of lymph node segmentation for predicting the outcome in H &N cancer.

We created an easy-to-use package for the binary-weighted model, called Individual Coefficient Approximation for Risk Estimation (ICARE). The code is publicly available at: github.com/Lrebaud/ICARE.

8.3 Discussion

In this chapter, we detailed the methodology we used for our participation to the HECKTOR 2022 challenge. A nnUNet with default parameters allowed us to rank 4th for the tumor segmentation task, and the new ICARE model that we developed during the challenge ranked 1st for the outcome prediction task.

Our ranking for the prediction task supports the fact that our intuition to aim for a minimal learning strategy was adapted to the context of outcome prediction given the data that we had available, and confirm our hypothesis: in certain situations, it is better not to learn any weight.

An interesting property of ICARE discovered during the challenge is its ability to handle a large number of features. While many machine learning models, such as a Cox proportional hazard model, might suffer from the curse of dimensionality, ICARE seems to not underperform when the number of features is greater than the number of training samples. We hypothesize that an analogy with the wisdom of the crowd could partly explain this. During his famous demonstration of the wisdom of the crowd effect in 1906, Galton asked a crowd of person to estimate the weight of an ox. While individuals were always wrong, the median of their answers was very close from the truth. One intuitive explanation is that people errors are equally distributed around the true value. If the crowd is large enough, the error will cancel out once the responses are aggregated. A similar phenomenon might happen with ICARE, where the error of each feature is cancelled out by the error of other features, leading to a good estimate of the outcome. This hypothesis requires perfect independence of the features, which is rare in practice. For this reason, a preselection removing correlation between features is often beneficial with the ICARE model.

During the challenge, this property allowed us to develop a new strategy. By constructing different variants of the segmentation masks of the tumors, and measuring the same features in each mask, many versions of each radiomic features were constructed. Because the biases and errors of each segmentation were different, some of their errors probably cancelled out via the large number of features, leading to better predictions.

However, not learning any weight might be too strong of a limitation on the model. In scenarii in which enough data are present or with limited amount of noise, a model assigning a higher weight to the most predictive features would be better than ICARE. We therefore tried to determine in which situations it is preferable to use the ICARE model.

Chapter 9

Comparison of the ICARE model to other machine learning models

9.1 Introduction

To understand under which conditions, it is better to use the weightless ICARE model than a traditional machine learning model, I collected 71 real medical datasets coming from two collections: SurvSet and TCGA. These large, diverse, and realistic datasets allow for a proper comparison of ICARE to other machine learning models. The datasets were composed of multiple features and one censored target to predict (e.g., survival prediction). Nine models were evaluated on these datasets. For a fair comparison between models and to make the comparison closer to real life applications, models and feature preprocessing were optimized on each dataset.

9.2 Article in review

Similar performance of 8 machine learning models on 71 censored medical datasets: a case for simplicity

IN REVIEW

Louis Rebaud[1,2], Nicolò Capobianco[3], [2]Nicolas Captier, [2]Thibault Escobar, Bruce Spottiswoode[4], Irène Buvat[2]

[1]Siemens Healthcare SAS, Saint Denis, France; [2]LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France; [3]Siemens Healthcare GmbH, Germany; [4]Siemens Medical Solutions USA, Inc., Knoxville, Tennessee, United States.

Abstract

In the analysis of medical data with censored outcomes, identifying the optimal machine learning pipeline is a challenging task, often requiring extensive preprocessing, feature selection, model testing, and tuning. To investigate the impact of the choice of pipeline on prediction performance, we evaluated 9 machine learning models on 71 medical datasets with censored targets. Only the decision tree model was consistently underperforming, while the other 8 models performed similarly across datasets, with little to no improvement from preprocessing optimization and hyperparameter tuning. Interestingly, more complex models did not outperform simpler ones, and reciprocally. ICARE, a straightforward model univariately learning only the sign of each feature instead of a weight, demonstrated similar performance to other models across most datasets while exhibiting lower overfitting, particularly in high-dimensional datasets. These findings suggest that using the ICARE model to build signatures between centers could improve reproducibility. Our findings also challenge the traditional approach of extensive model testing and tuning to improve performance.

Introduction

When new biomarkers are identified as related to a time-dependent outcome (e.g., response to treatment, progression-free survival, overall survival), it is crucial to determine how to use them for widespread acceptance and application. A common strategy consists in identifying cut-off values that can categorize patients in different risk categories. Yet, this method faces multiple issues. First, since it assumes a monotonic relationship between the biomarker and the outcome, it might be inappropriate for non-monotonic associations. For instance, total cholesterol level typically rise until middle age, after which it tends to decrease in older individuals [254]. Additionally, using a cut-off creates abrupt changes in risk categories for patients with a biomarker value close to the cut-off value. Agreement on cut-off values between centers is often challenging, requiring corrections for center effects using approaches such as ComBat [255], [256]. Furthermore, categorization might reduce the prognostic power by deleting valuable continuous information [257], [258], [259]. Last, since the human brain can effectively handle only up to four features simultaneously, this method becomes even less effective when many biomarkers are available [260].

For these reasons, machine learning models are a more effective way to leverage and combine biomarkers information into a so-called score or signature [261]. These models offer the possibility to aggregate multiple features and learn the best way to combine them to predict the target (e.g., survival, risk of relapse, response to treatment). However, not any model can be used since the target value is often censored (e.g., we know that the patient was alive until a certain date, but then we do not know if he died, and when). Machine learning models specifically designed or

adapted for censored data must thus be used. The Cox proportional hazard model is frequently used since it effectively handles censored outcome, controls for confounders, and is interpretable. The weights of the model (hazard ratios) can easily be shared (e.g., as a nomogram), which makes the Cox model an easy-to-use and versatile tool to build and share signatures. Yet, Cox models have their own limitations, as they assume a linear relationship between the log-hazard and the continuous explanatory variables, and they are sensitive to collinearity [262]. Many traditional machine learning models have also been adapted to handle censored data (e.g., tree-based models, SVMs) [263].

A major challenge of machine learning based signatures is to make them robust enough with respect to slight technical changes in the data: a model trained on data from one center might not work well on data from another center, even if the patient population is similar. This is a well-known problem in machine learning, referred to as overfitting. Training models on medical datasets is prone to overfitting [264], because of the often-limited number of training examples [114], of the many features and feature combinations that are tested and of the inherent complexity and noise in the target to predict, such as overall survival, which is frequently censored.

To mitigate this effect, the Individual Coefficient Approximation for Risk Estimation (ICARE) model was developed [232]. This model reduces the risk of overfitting by reducing the amount of information learned from the training set to a strict minimum, based on the following reasoning: the less is learned from the training set, the less likely it is to learn something that will not generalize to other cohorts. ICARE does this by univariately learning only a sign (-1 or +1) for each feature instead of a positive or negative weight, the assumption being that we often do not have enough training data to reliably determine if a feature is more predictive than another. During the training step, it also computes two normalization factors (mean and standard-deviation) from the training data to normalize each feature with a z-score, so features with larger values will not have an arbitrary stronger weight in the prediction. This model won the HECKTOR 2022 challenge for predicting the recurrence free survival of head and neck cancer patients based on ^{18}F-FDG PET/CT images [239].

This diversity of machine learning method raises the question of which model should be used when building a new signature based on censored data. In this study, we investigated the impact of the choice of model on the quality of the signature by conducting an extensive benchmark of methods for predicting time-dependent clinical outcome and developing associated signatures. We trained and tested 9 machine learning models on 71 medical datasets retrieved from the SurvSet and TCGA collections. Each dataset was composed of multiple features and a censored target. The data types varied in nature (e.g., radiology, transcriptomic). We thus covered a large number of realistic scenarios using publicly available data presenting a wide variety in terms of nature, number of samples and features, and extent of censoring.

Results

A comprehensive benchmark composed of a large variety of medical datasets

We collected 71 different datasets, 51 originated from the SurvSet collection [265] and 20 were extracted from the TCGA database (https://www.cancer.gov/tcga). Table 11 shows some statistics of the different characteristics of each group of datasets, including the number of features, the number of samples, the proportion of censored samples, and the maximum time-dependent area under the receiver operating characteristics curve (tAUC) obtained by any of the 9 tested models.

Statistic	Dataset group	Number of samples	Number of features	Number of features / Number of samples	Proportion of missing values	Proportion of censored samples	Best tAUC by any model
minimum	SurvSet	92	4	0.0006	0.00	0.00	0.58
	TCGA	86	19324	17.8609	0.00	0.15	0.55
median	SurvSet	461.0	23.0	0.0391	0.00	0.57	0.73
	TCGA	405.5	19497.0	48.1430	0.00	0.62	0.67
mean	SurvSet	1664.75	645.47	3.4351	0.01	0.54	0.74
	TCGA	392.65	19480.3	69.6843	0.00	0.60	0.69
std	SurvSet	2773.82	1935.58	11.5695	0.04	0.25	0.09
	TCGA	221.65	55.45	49.3687	0.00	0.20	0.09
maximum	SurvSet	14294	8664	54.8354	0.26	0.94	0.97
	TCGA	1093	19547	224.6977	0.00	0.86	0.88

Table 11: Statistics of the distribution of the characteristics in the two datasets groups. The performance of the 9 models was estimated in a nested cross-validation with time-dependent area under the receiver operating characteristics curve (tAUC).

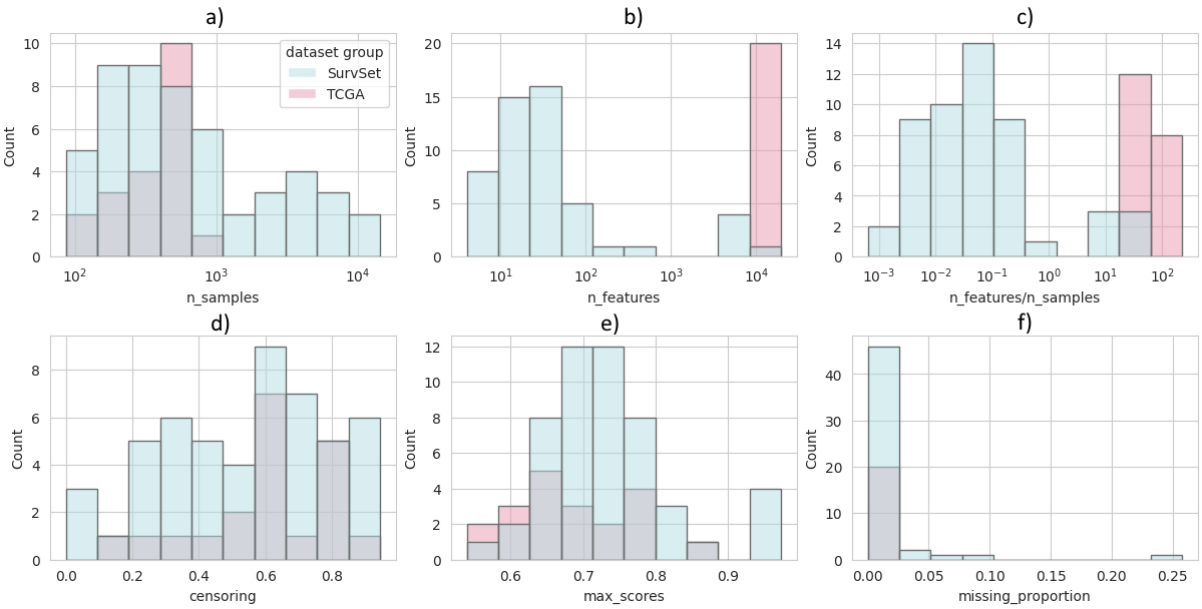


Figure 58: Histograms of the characteristics of the datasets in the SurvSet and TCGA groups.

The histograms of the different characteristics are displayed in Figure 58. For all characteristics, the SurvSet collections had the largest range of values, while the TCGA datasets focused on scenarios with a high number of features compared to the number of samples (Figure 58c). Both groups of datasets included a wide variety of censoring values and of maximum tAUC achieved by any of the 9 models tested, the latter reflecting how easy it is to predict the target. Almost no missing values were present in the datasets (Figure 58f). TCGA datasets had no missing values, and only 15 SurvSet datasets had missing values. For 9 of these datasets, less than 1% of the values were missing, and for 11 of them, less than 5% of values were missing. The largest proportion of missing data was 26% in one SurvSet dataset.

Most machine learning models exhibit comparable performance for a given dataset

On these 71 datasets, we trained, optimized, and evaluated 9 different models to predict the censored target. The machine learning methods included Cox models, decision trees, tree bagging, boosting of linear models, linear and non-linear support vector machines (SVM) and the ICARE model. All models were evaluated with a 10×10 nested cross-validation (see Methods).

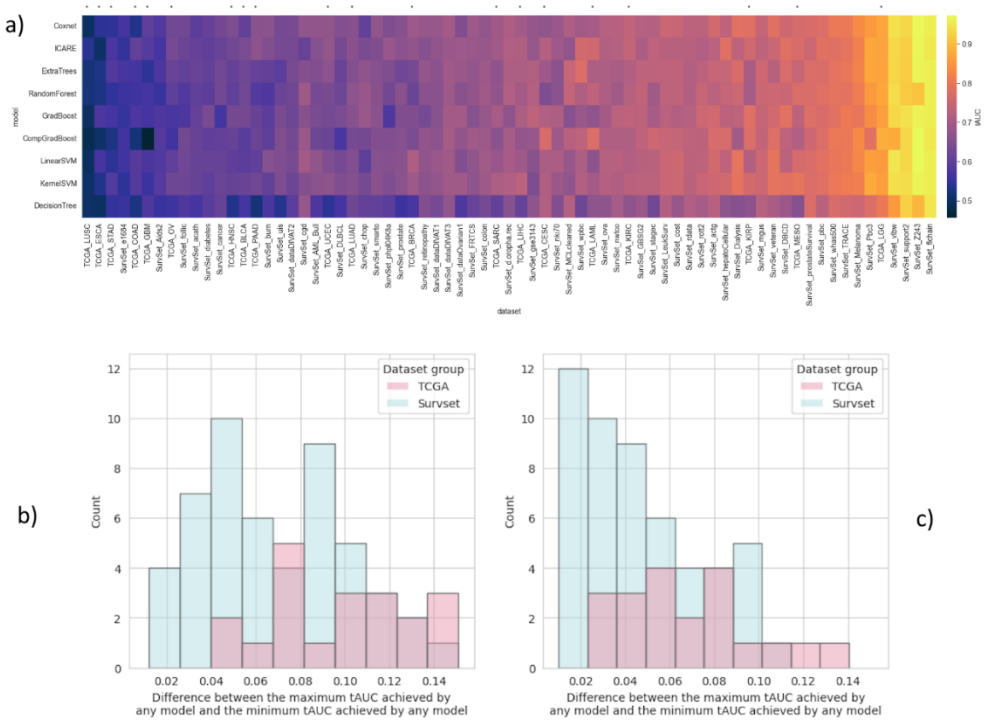


Figure 59: ***a)*** *Average time-dependent AUC (tAUC) for each model and for each dataset. The datasets are sorted by average tAUC across all models. TCGA datasets are indicated with a star at the top of the heatmap.* ***b)*** *Histograms of the difference between the maximum tAUC achieved by any model minus the minimum tAUC achieved by any model, for both TCGA and SurvSet datasets, for all models.* ***c)*** *is the same as b) but the decision tree model was removed.*

Figure 59a shows all the tAUC achieved by each model on each dataset on a heatmap. Datasets are ordered from the lowest (left) to the highest (right) average tAUC across models. The heatmap shows little difference in performance between models for a given dataset (i.e. along a same column), except for the Decision Tree (last row) that was consistently underperforming. If we exclude the DecisionTree model, the difference between the maximum and minimum tAUC achieved by any model on each dataset was less than 0.04 on SurvSet and less than 0.07 on TCGA for half of the datasets. The maximum difference observed were 0.10 on SurvSet et 0.14 on TCGA. Figure 59b and Figure 59c shows the distribution of this difference, with and without DecisionTree included. This difference tended to be smaller in SurvSet datasets than in TCGA datasets.

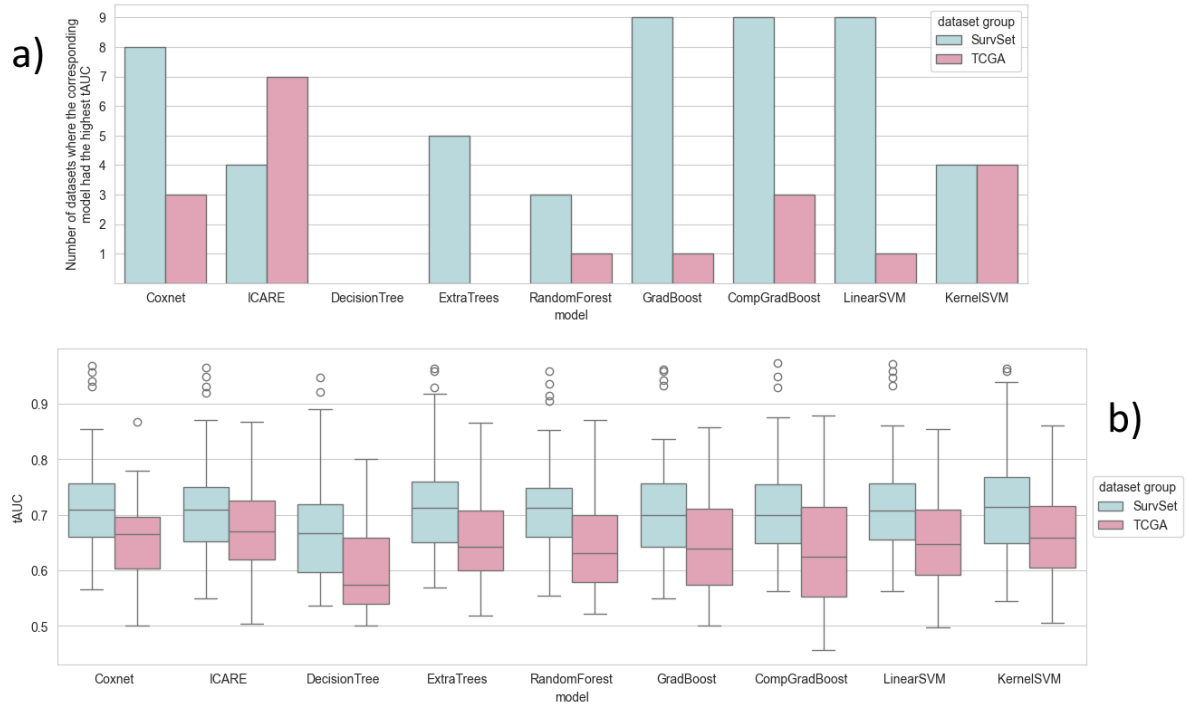


*Figure 60: **a)** Number of datasets for which the model indicated on the x-axis had the highest average time-dependent AUC (tAUC), for the nine models, for each group of datasets. **b)** Boxplots of the average time-dependent AUC (tAUC) achieved by the models shown on the x-axis on the test sets of the nested cross-validation for all datasets, for the nine models and for each dataset group.*

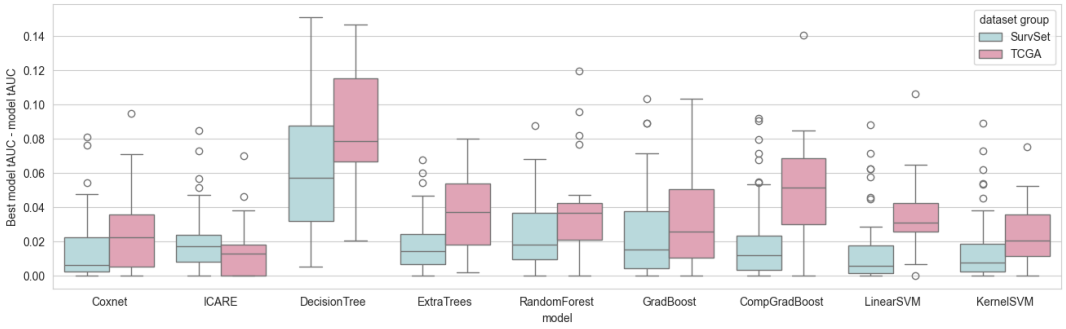
No model was systematically better than any other model. Figure 60a shows the number of datasets for which each model was the one with the highest time-dependent AUC (tAUC). For neither SurvSet nor TCGA did a single model clearly stand out as the best performer. For SurvSet, the most consistent models (i.e. GradBoost, CompGradBoost, LinearSVM) performed only the best on 9 out of 51 datasets while for TCGA, ICARE was the most frequently best performing model, but this occurred only in 7 out of 20 datasets. Decision Tree was never the best performer.

Figure 60b shows the distribution of the average tAUC achieved by all models on all datasets. TCGA datasets tended to have lower tAUC whatever the model compared to SurvSet datasets. This is also observed in Figure 59a where TCGA datasets are more concentrated on the left part of the heatmap. Apart from the decision tree model that had overall lower performance, all models had similar distribution of tAUC across all dataset groups.

The ICARE model compares favorably with more complex methods

Small differences can also be observed if we measure the difference in tAUC between the best performing model on each dataset and the tAUC of all other models on the same dataset. Figure 61a shows the distribution of these differences for all models. For half of the datasets, all models except the decision tree were less than 0.02 points of tAUC below the best model on SurvSet, and less than 0.05 points on TCGA. For 95% of datasets, most models were less than 0.10 points of tAUC below the best score on both SurvSet and TCGA. Tree-based models had larger differences with the best model.

a)



b)

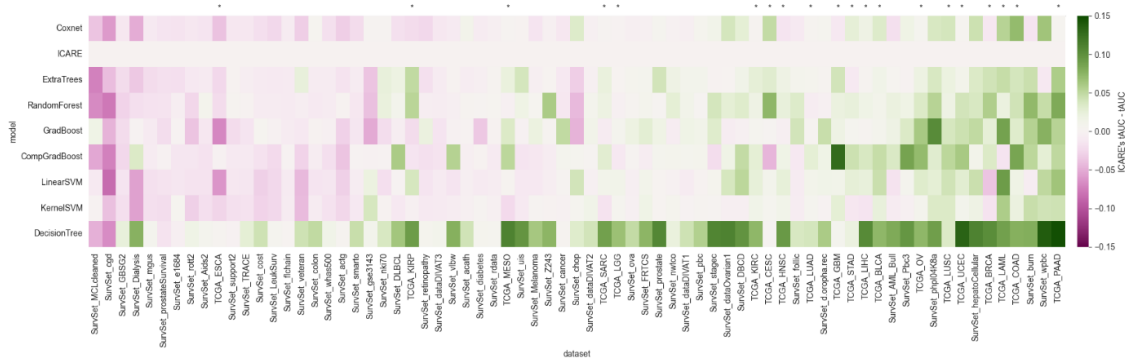


Figure 61: ***a)*** *Boxplots of the differences between the best time-dependent AUC (tAUC) achieved by any model and the tAUC of the model shown on the x-axis on all datasets.* ***b)*** *Differences between the average time-dependent AUC (tAUC) of the ICARE model and the tAUC of other models for each model and for each dataset. The datasets are sorted by average difference of tAUC across all models. TCGA datasets are indicated with a star at the top of the heatmap.*

When focusing on the difference in tAUC between ICARE and the tAUC of all other models on all datasets, we found that the difference was higher than -0.03 of tAUC for 95% of combinations of model and datasets. The difference was positive for 50% of

combinations and above -0.01 for 76% of combinations. The lowest value was -0.09 and the highest was 0.15. Figure 61b shows all these differences on a heatmap. A higher concentration of TCGA datasets can be seen on the right part of the heatmap, where ICARE tended to outperform other models more frequently.

Preprocessing and hyperparameter tuning have no clear impact on the performance

The same models were evaluated on the same datasets but without any feature preprocessing nor hyperparameter tuning. When measuring the difference between the model with default settings and its tuned counterpart, no strong trend suggested a clear benefit of feature preprocessing and hyperparameter tuning. Figure 62a displays the distribution of the differences between a model trained with feature preprocessing and hyperparameter tuning and the same model without any selection nor preprocessing of features and with default hyperparameters. Positive values mean that the preprocessing and tuning improved the performance.

The details can be seen in Figure 62b, which contains a heatmap of all differences in tAUC between a model with and without feature preprocessing and hyperparameter tuning, for all models and all datasets. While some models had substantial gains in tAUC on some datasets, for most combinations of models and datasets, the tAUC was not substantially increased with tuning and feature selection. No dataset had tAUC substantially and systematically increased by preprocessing and tuning.

Only 14% of all combinations of models and datasets benefited from feature preprocessing and hyperparameter tuning by more than 0.05 of tAUC. The full histogram of the gain in tAUC through feature preprocessing and hyperparameter tuning for all combinations of models and datasets is provided in Figure 62c. The median was 0.004, meaning that only half of the combinations of models and datasets benefited from feature preprocessing and hyperparameter tuning, and the other half underwent almost no change or a reduction in tAUC in the process.

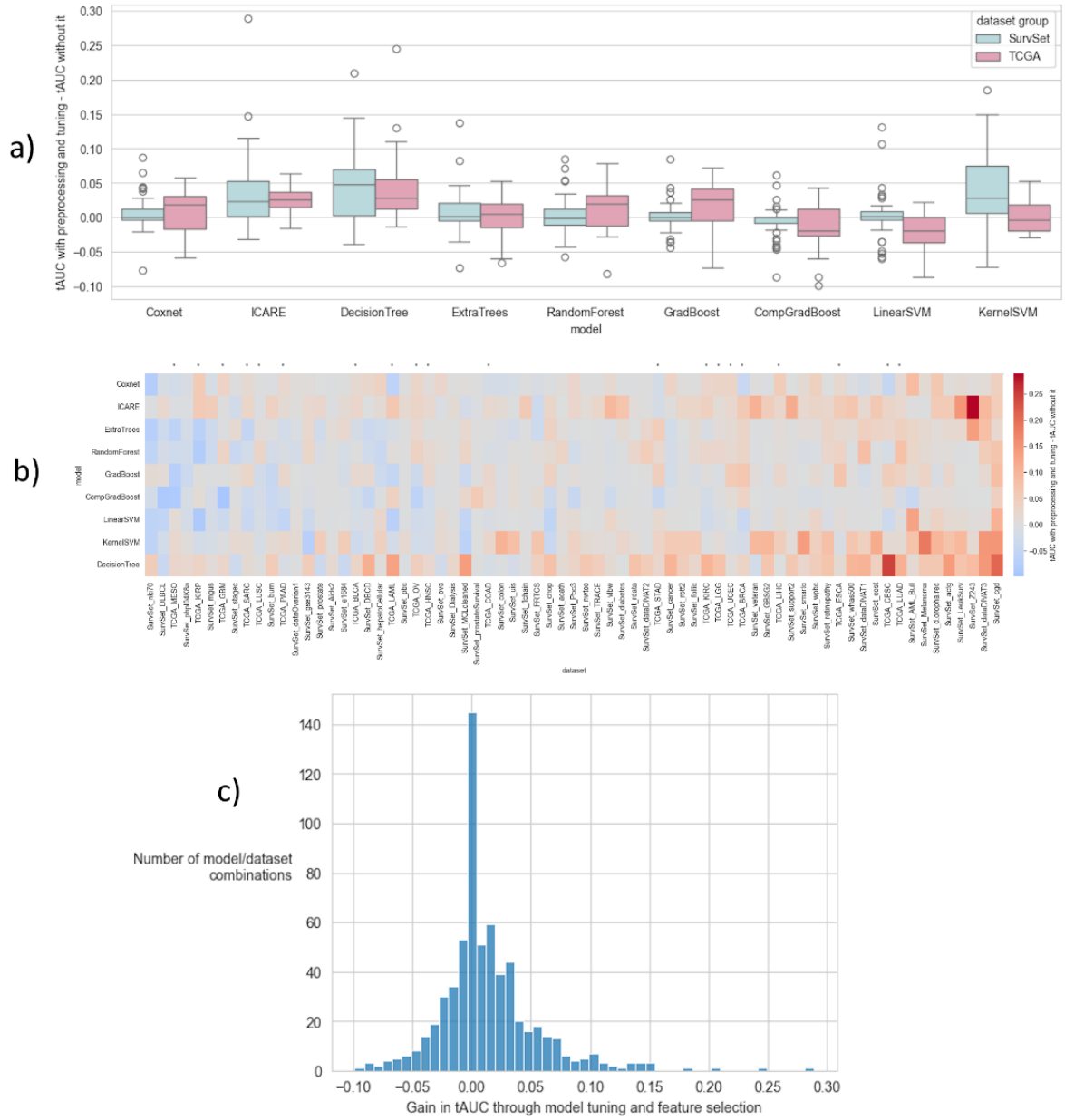
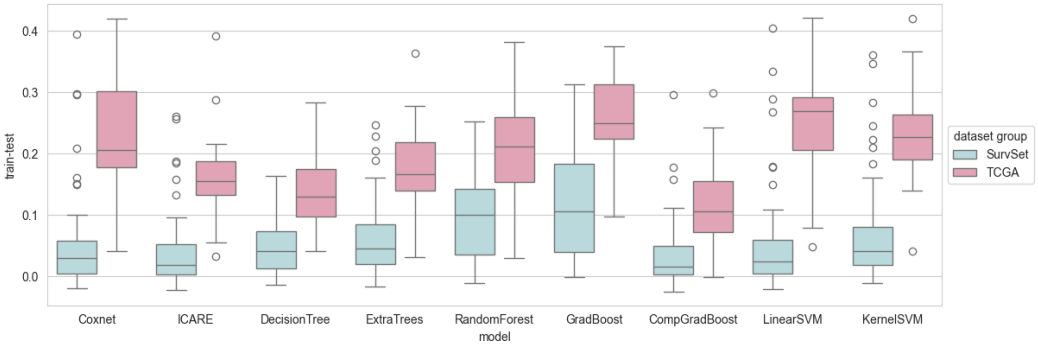


Figure 62: ***a)*** *Boxplots of the differences between the tAUC achieved by the model with feature selection and hyperparameters tuning and the model without it, on all datasets, for each model and for the SurvSet and TCGA dataset groups.* ***b)*** *Differences between the tAUC achieved by the model with feature selection and hyperparameters tuning and the model without it, for all models on each dataset. Datasets are sorted by averaged increase in tAUC when using feature selection and hyperparameters tuning across all models. TCGA datasets are indicated with a star at the top of the heatmap.* ***c)*** *Histogram of gain in time-dependent AUC (tAUC) through feature selection and hyperparameter tuning for all model and dataset combinations. The gain is defined as the difference between the tAUC of the model trained and evaluated with feature selection and hyperparameter tuning and the tAUC of the same model on the same dataset with default settings.*

<u>CompGradBoost, ICARE, ExtraTrees and DecisionTree are more robust to overfitting than other models</u>

More substantial variations between models were observed when comparing differences between the average tAUC achieved on the train set and the average tAUC achieved on the test set. Figure 63a displays the distributions of these differences for all models. The higher the difference, the higher the overfitting. On TCGA datasets, overfitting was greater than on SurvSet datasets. On SurvSet, Cox, ICARE and CompGradBoost overfitted the least, while on TCGA, it was ICARE, CompGradBoost and DecisionTree who overfitted the least. The biggest differences were observed for the Cox, GradBoost and SVMs models on the TCGA datasets, with tAUC on the train set often superior by 0.20 to the tAUC on the test set.

a)



b)

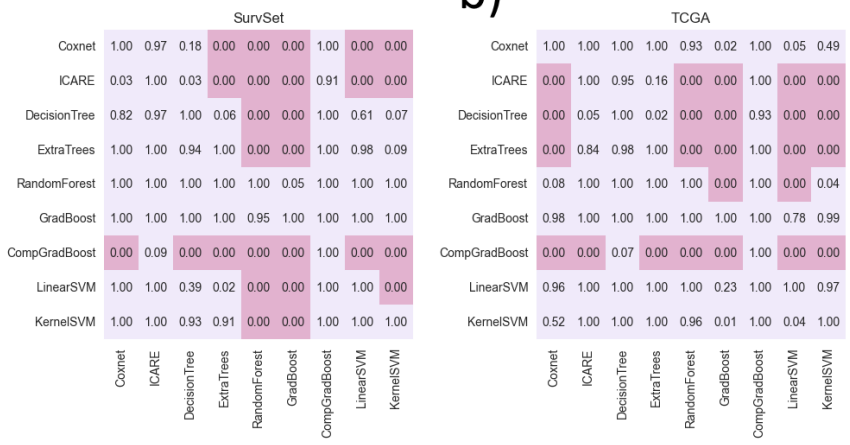


Figure 63: ***a)*** *Boxplots of the differences between the tAUC achieved by the model shown on the x-axis on the train set and on the test set, for the nine models (x-axis) and the SurvSet (blue) and TCGA (pink) dataset groups.* ***b)*** *p-values of a Wilcoxon signed-rank test assessing if the model of the corresponding row had a significantly smaller difference in performance between its test and train tAUC than the model of the corresponding column. Significant p-values are shown in pink cells. The significance was assessed while controlling for multiple testing with two-stage linear step-up procedure (TST) to have less than one false positive.*

Some models overfitted significantly more than others. Figure 63b presents the result of a Wilcoxon test to assess if a model overfitted significantly less than another. Based on these tests, CompGradBoost overfitted significantly less than almost any other model on both SurvSet and TCGA. ICARE overfitted significantly less than SVMs, RandomForest and GradBoost on all datasets. On TCGA, Cox overfitted significantly more than ICARE, DecisionTree, ExtraTrees and CompGradBoost, while on SurvSet, only CompGradBoost was significantly better than Cox. Overall, the models that were the most robust to overfitting in all scenarios were CompGradBoost, ICARE, ExtraTrees and DecisionTree.

Discussion

In this study, we evaluated 9 machine learning models on 71 datasets from the SurvSet and TCGA databases, representing a wide variety of scenarios of prediction tasks involving censored targets. Preprocessing was optimized and models were tuned to maximize model performance in each dataset, allowing for a fair comparison of the models.

An important observation is that the choice of model, of the preprocessing and of model tuning does not have a major impact on the performance most of the time. Extensive testing and tuning of models appear to bring little improvement in the results, as only 14% of combinations of model and datasets had their tAUC increased by more than 0.05 following feature preprocessing and hyperparameter tuning. In addition, almost all models had similar performance and were often close to the best score achieved by any model except the decision tree model, which was consistently underperforming. If we exclude this model, on 50% of datasets, the best model was less than 0.05 of tAUC higher than the lowest tAUC achieved by any model. We also observed major differences between models in terms of overfitting, with some models generalizing significantly better to new unseen data than other models.

Models with non-linearity or features weighting did not substantially outperform the ICARE model. In most cases, ICARE had a score close to the optimum achieved by any model. It was the best performing model in 11 out of the 71 datasets and was less than 0.02 point of tAUC below the best model in half of datasets. This confirms that the strategy behind the model is valid and that reducing the amount of information learned from the training set does not substantially impair the performance of the model. This suggests that in most situations, weighting the features is not necessary and only the sign of the correlation needs to be estimated (as well as their normalization factors, used in ICARE). Based on our results, this does not reduce performance compared to a Cox model and might even increase the chances of replicating the findings in other centers, as the ICARE model overfitted less than a Cox model.

Another implication of this work is that preprocessing, building, testing, and tuning many models might not be an effective time and energy investment. These computationally intensive steps frequently used in machine learning do not appear to

substantially improve the performance. This suggests that the most effective way to improve model performance for outcome prediction tasks involving censored data might not be to look for the best combination of the available features, but rather to search for new biological information that could bring additional knowledge about the patient.

Based on our results, it appears that the ICARE model has the potential to be a solid choice for signature building. Not only its performances were often the best or close to the best performance achieved by any model, but it was also one of the models with the least overfitting on both SurvSet and TCGA datasets, meaning that it would provide realistic estimation of its performance on new cohorts of patients, in other centers. Of all tested models, it is also one of the simplest, as only the sign of the correlation between the target value to predict and the features, as well as the normalization factors of the features (mean and standard deviation) are needed to fully describe the model. Another benefit of the ICARE model not leveraged in this study is its ability to handle missing data, contrary to all the other tested models who requires feature removal or imputation. In this study, the number of missing data was too low to influence the results.

CompGradBoost was also a strong choice for signature building as it was the model with the least overfitting of all evaluated models. However, it is an ensemble of multiple models via boosting and is therefore more complex to interpret and share than a Cox or ICARE model. It was also further away from the maximum score than ICARE and Cox on many high dimensionality (TCGA) datasets.

This study has some limitations. First, the automated preprocessing and model optimization used in this study cannot replicate human ingenuity and experience and our conclusions might not always apply. An expert manually tuning each model on each dataset could achieve better performance than our automated approach, and substantial gains in tAUC might be observed compared to default models. Secondly, our conclusions depend heavily on the 71 selected datasets, that cannot reflect all real-world datasets. Some specific scenarios missing from our collection might have yielded different conclusions. Moreover, models specifically designed for a precise task will probably perform better than the general models evaluated in this study on this specific task.

For these reasons, our experiments should be repeated on other large collections of datasets with more models and other model tuning. To support this effort, we made all our code publicly available on GitHub at:

https://github.com/Lrebaud/survival_benchmark.

Methods

Datasets

A total of 51 datasets of the SurvSet collection (https://github.com/ErikinBC/SurvSet) and 20 datasets of the TCGA database (https://portal.gdc.cancer.gov) were used. In SurvSet, only the datasets composed of medical or biological data were selected. Those with time-dependent features were removed since they were a minority and would have introduced extra complexity to the study, while not being representative of many datasets. On both collections, only the datasets with more than 40 non censored samples were kept. This value was chosen to have enough samples in each dataset for a robust evaluation of model performance, while retaining as many datasets as possible. The complete list of the datasets used in our study is given in Table 12.

SurvSet	TCGA
diabetes, dataDIVAT1, gse3143, GBSG2, LeukSurv, smarto, vlbw, d.oropha.rec, DBCD, prostate, whas500, uis, Dialysis, retinopathy, veteran, DLBCL, dataDIVAT2, nwtco, AML_Bull, dataDIVAT3, burn, MCLcleaned, phpl04K8a, stagec, Pbc3, nki70, cancer, hepatoCellular, mgus, TRACE, colon, support2, acath, cgd, wpbc, prostateSurvival, Aids2, rott2, ova, cost, e1684, chop, Melanoma, FRTCS, Z243, pbc, dataOvarian1, follic, actg, flchain, rdata	UCEC, BLCA, GBM, OV, LIHC, COAD, CESC, LAML, ESCA, LGG, KIRC, MESO, PAAD, LUSC, STAD, LUAD, SARC, HNSC, BRCA, KIRP

Table 12: List of datasets used in each collection.

On TCGA datasets, the features were the RNAseq data and the outcome to predict was the overall survival (OS), cleaned and prepared by Liu and colleagues [266].

Models

Most models implemented in the scikit-survival Python package [225] were included in the study. We used the 0.21.0 version. This collection adapts a wide variety of traditional machine learning models to handle censored targets. Table 13 shows the correspondence between the name of the models in our study with the corresponding model in scikit-survival.

Elastic net penalty was used in the baseline Cox model of this study since a Cox model without any regularization would not have produced satisfactory results in most datasets due to collinearity in the features. In addition, through hyperparameter tuning, both L_1 and L_2 penalties could be explored at the same time.

An implementation of the ICARE model is provided with the code used to process the data and evaluate the models, publicly available on GitHub at https://github.com/Lrebaud/survival_benchmark. In the original ICARE paper, multiple feature selection steps were present inside the model, such as correlation removal and dropping of features with a low C-index. Here, we removed all these steps from ICARE

to have the same feature selection steps for all models as described below.

Name in the study	Name in scikit-survival	scikit-survival's description
Coxnet	CoxnetSurvivalAnalysis	Cox's proportional hazard's model with elastic net penalty.
CompGradBoost	ComponentwiseGradientBoostingSurvivalAnalysis	Gradient boosting with component-wise least squares as base learner.
GradBoost	GradientBoostingSurvivalAnalysis	Gradient-boosted Cox proportional hazard loss with regression trees as base learner.
RandomForest	RandomSurvivalForest	A random survival forest.
ExtraTrees	ExtraSurvivalTrees	An extremely random survival forest.
DecisionTree	SurvivalTree	A survival tree.
LinearSVM	FastSurvivalSVM	Efficient Training of linear Survival Support Vector Machine
KernelSVM	FastKernelSurvivalSVM	Efficient Training of kernel Survival Support Vector Machine

Table 13: List of models from scikit-survival used in this study.

Feature selection and preprocessing

For a fair and realistic evaluation of the models, all models had the same preprocessing steps, and all parameters of these steps and model hyperparameters were tuned with a random search in a nested cross-validation. The preprocessing steps were the following:

- Dropping the features for which the proportion of missing values was above an adjustable cut-off.
- Dropping the features for which the C-index with the target was below an adjustable cut-off.
- Dropping features for which the Spearman's correlation with other features was above an adjustable cut-off.
- Imputation of the missing values with one of the following techniques: mean, median, mode, constant, KNN (5 neighbors), using the implementation of the scikit-learn library [224]. No imputation was performed for ICARE since it can handle missing values.
- All features were normalized with a z-score calculated on the train set.

The normalization of the features being an integral part of the preprocessing, it was present for all models. When models were evaluated without preprocessing, only ICARE normalized the features as this step is an integral part of this model.

Evaluation

All models were evaluated with a 10×10 nested cross-validation with consistent splitting of the data. Model hyperparameters and feature preprocessing parameters

were optimized in the inner loop with a random search of 100 iterations. The model with optimized hyperparameters was then retrained on all the data of the inner loop. Performance of the model was assessed on samples from the inner loop (train set) and outer loop (test data) with time dependent AUC (tAUC). This metric was chosen because it is not sensitive to the proportion of censored samples, contrary to the concordance index [100]. The average value across the 10 outer folds was used to assess the model performance on the train and test sets.

Statistical information

To test if one model was overfitting significantly more than another model, the difference between the average tAUC obtained on the test set and the average tAUC obtained on the train set was computed for all models on all datasets. For each pair of models, a Wilcoxon signed-rank test was used to assess if one model had significantly greater differences than the other, across all datasets.

9.3 Discussion

In this chapter, the evaluation of 9 machine learning models for censored target prediction tasks on 71 real medical datasets revealed several key insights with implications for signature building. A nested-cross validation with extensive hyperparameters tuning allowed for a thorough and realistic estimation of performance for each model. A key finding was the relatively uniform performance of models across most datasets. Specifically, in half of the datasets analyzed, the difference in tAUC between the best and worst-performing models was less than 0.05, excluding the consistently underperforming decision tree model. Similarly, extensive hyperparameter tuning and feature selection and preprocessing did not bring significant improvements, as only 14% of combinations of model and datasets increased in tAUC by more than 0.05, and half of them did not benefit from it. This highlights a limited impact of model choice and extensive tuning on overall predictive performance.

Among the models that we evaluated, the ICARE model proved remarkably effective. It was the top-performing model in 11 out of the 71 datasets, and in 50% of the cases, it was within 0.02 tAUC points of the highest-scoring model. It was also one the models with the least overfitting and the simplest one. These results underscore the ICARE's potential to create signatures more generalizable to other centers, and aligns well with the ranking of ICARE in the HECKTOR 2022 challenge.

The results of the study prompt a reconsideration of the emphasis generally placed on extensive model testing, suggesting a potential shift towards simpler models such as ICARE and the search for new biological insights. However, the automated nature of the study processes and the specific limitations of the datasets that were used highlight the need for broader exploration in future research.

Chapter 10

General discussion, conclusion, and perspectives

The goal of my PhD was to look for new image-based prognostic features in the PET/CT scans and determine how to use them for enhanced patient stratification. After a manual search with limited results, I developed a semi-automated approach to comprehensively search the images for new information. By automatically building thousands of candidate biomarkers from the image data and testing them on two cohorts of FL and DLBCL patients, I was able to identify dozens of new radiomic features significantly associated with the outcome. However, these features, while handcrafted, well defined mathematically and prognostic of the outcome, were not all easy to interpret and the biological information that they encoded was not clear for most of them. Deciphering them through visual examples and comparison with simpler features allowed us to build tens of new surrogate biomarkers much easier to interpret and that were still prognostic of the outcome. Among them, 10 were prognostic on the two cohorts of DLBCL and FL patients, increasing their likelihood to be truly prognostic, and their usefulness as they could be used to stage more patients. These results have thus contributed to the core question of the PhD by developing tools to effectively search for new prognostic information and identifying new potential prognostic biomarkers for lymphoma patients.

While discovering new biological information relevant in the image is a first step, their efficient use in clinic is as important. I made two contributions addressing this point. The first one is the development of the ICARE model. This is a new machine learning model that favors simplicity. The core idea is that in survival prediction, we often do not have enough data to give accurate weights to each feature. The ICARE model tries to reduce the risk of overfitting by learning only a sign for each feature in a univariate way. This minimalistic approach was validated during the HECKTOR 2022 challenge where we ranked 1st for the outcome prediction task. I evaluated ICARE in a comprehensive comparison with 8 other machine learning models on 71 medical datasets. I demonstrated that ICARE was overfitting significantly less than many other machine learning models, and thanks to its simplicity and robustness, it appears to be a good model to build new clinical signatures encompassing multiple biomarkers.

The second contribution to efficiently use the new biomarkers in clinic is the

implementation of my results in a software. I developed an extension for the PARS software offered by Siemens Healthineers which allows for the automated segmentation of organs and lesions on PET/CT images. My extension builds upon these segmentations to automatically calculate feature values from the image. Then the features are fed into prediction models and visualization tools. The results are shown to the user. A screenshot of the extension showing results is displayed in Figure 64.

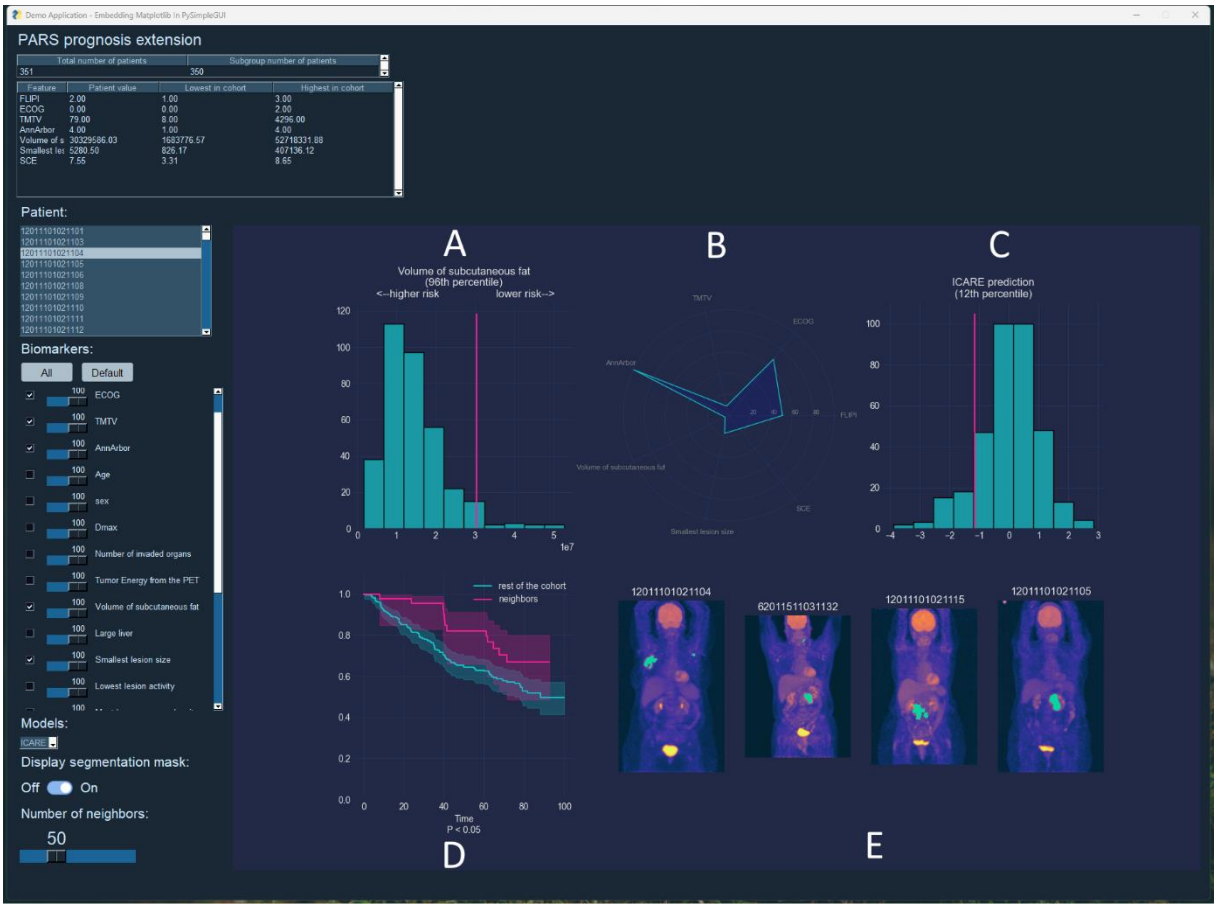


Figure 64: Screenshot of the extension of the PARS software I developed. The A, B, C, D and E letters are not present in the software and are only used in this chapter to reference a specific plot.

On this screen, many features can be displayed at once and used for risk prediction. The various plots show how the patient compares in terms of risk to a reference cohort according to the selected features. This reference cohort, which can be from the center where the tool is used or from outside to leverage more patient data, gives intuition about new biomarkers and new model. For instance, in Figure 64A, we can see that the patient studied (in red) is at lower risk than the rest of the cohort (in blue), according to the volume of subcutaneous fat. More globally, Figure 64B is a radar plot showing the percentile of the patient values for different biomarkers, relative to the reference cohort, and oriented by risk. Hence, the closer to the center, the lower the risk according to the corresponding biomarker, and vice-versa. Therefore, the higher the total surface of the radar plot, the higher the risk. Discrepancies among the biomarker can easily be observed that way. Similarly to Figure 64A, Figure 64C shows how the

patient relates to the rest of the cohort in terms of risk according to a model encapsulating multiple biomarkers. In this example, an ICARE model trained on the reference cohort predicted a favorable outcome for the patient studied (in red) compared to the rest of the cohort (in blue). Various prediction models can also be tested. Figure 64D shows Kaplan-Meier curves of the PFS of patients from the reference cohort split in two groups: those who are similar to the patient studied according to the selected biomarkers (in red), and the rest of the cohort (in blue). Here we can see that patients similar to the selected patients tended to have longer PFS than other patients. Figure 64E shows the MIPs of the PET images and the lesion segmentation of the selected patient and of the 3 patients the most similar according to the selected biomarkers. The reference cohort can also be refined to select reference patients for whom several features have similar values compared to the analyzed patient. For instance, the user can pick only the patients in the reference cohort with similar TMTV and FLIPI score. This is a useful feature in the extension as I believe that the new biomarker interest lies in the stratification of population of patients that seem homogeneous in risk according to currently used biomarkers.

In parallel of this handcrafted biomarker search, I also explored deep learning methods to find new image-based biomarkers. Deep features are a new paradigm in the field of radiomics and hold the promise to discover new and more subtle biomarkers, by leveraging deep learning models to analyze the images. The first project I conducted on this topic is the use of a variational autoencoder to encode the PET images in an expressive latent representation. I specifically focused on β-VAE for their ability to create an intuitive latent space. With this approach, I was able to rediscover two already identified biomarkers: TMTV and Dmax, but I did not find any new prognostic information. We tried to take this idea further with the help of a master student during his internship. Even though we were able to successfully encode the PET images and reconstruct them with minimal loss, the latent space was not found prognostic of the outcome nor linked to any existing clinical features.

To force the model to encode prognostic information, I also tested a supervised approach. Here the model had to predict the outcome of the patient from the MIP of the PET image. I adapted the concordance index, our metric of interest for outcome prediction, into a loss function that could be used to train a deep learning model. I validated this approach during the HECKTOR 2021 challenge where I ranked 5th among 30 teams for outcome prediction from PET images, and I presented the approach during the SNMMI 2022 conference. While this deep learning model did not improve on more traditional models (e.g., Cox model) using radiomic features, it reached similar performance for outcome prediction, while being a deep learning model, meaning that it was creating its own features rather than using hand-crafted ones, like a Cox model would. This was the main interest of this approach: the model learning prognostic information from the images and encoding it into its deep features. Hence, I then tried to understand the learned prognostic information. I tried many techniques to do that, but the most promising one was to use SHAP values.

This technique comes from game theory and quantifies the impact of each pixel on the target. This allows to understand how each region of the PET image impacts the outcome of the patient, according to the model. However, I found that the results were extremely noisy and hard to interpret. These poor results probably come from the fact that the prediction of the model is far from being perfect. To mitigate this problem, I tried an ensemble of SHAP values. By merging the prediction and the interpretation of multiple models, I reduced the variability and the overfitting and noise of each model. I found that this ensembling of SHAP values produced much better interpretation maps, as shown in Figure 65.

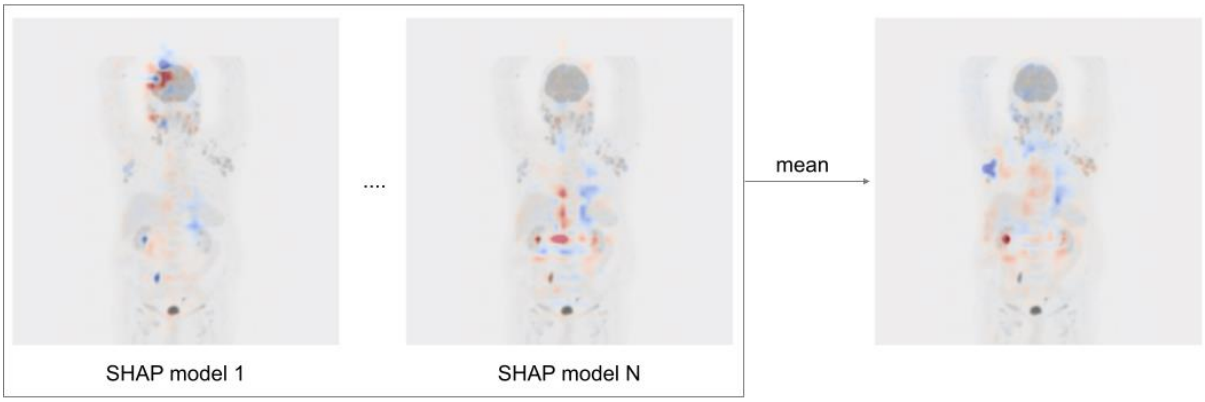


Figure 65: SHAP values of a model predicting the PFS of DLBLC patients superimposed on the MIPs of their PET images. Pixels that increased the risk are displayed in blue, while pixel reducing it are displayed in red. The SUV are displayed with a grey scale, with darker pixels representing higher activity. The two images on the left display individual SHAP map of the same patient, but with a different model, trained with different resampling of the training data. The image on the right is a mean aggregation of 100 SHAP maps of the same patient from 100 different models trained with a 100 different resampling.

However, despite this progress in interpretation, I could still not understand and find any biological information meaningful for prognostic outcome. The inconclusive results of the deep learning approach might be due to the lack of data. These techniques are developed to use thousands of samples, while in medical imaging, we are often working with hundreds of samples. This limited amount of training examples is even more exacerbated by the complexity of the task and the high dimensionality of the medical imaging data. However, I have not explored the pre-training and fine-tuning of models and these techniques could significantly improve results. Moreover, I am extremely curious to see how the current multimodal large language models will contribute to the search of new radiomic features. By being pre-trained on many data, these models could develop intuition useful for the search for new image-based prognostic features. They could give intuitive textual information to explain the content of an image, and what is relevant for prognosis.

The limited amount of data and the noisy nature of the targets seem to also be the reasons why the simplistic approach of the ICARE model works well. By reducing the complexity of the model and therefore the number of degrees of freedom, we

constrained the search space to make it less likely to overfit. An even more interesting property of ICARE is its ability to handle a high number of features. While most models drop in performance when more and more features are added to them, ICARE does not seem to suffer from the curse of dimensionality. We observed this phenomenon during the HECKTOR challenge where we were able to feed thousands of radiomic features to the model and observed either a plateau or an improvement in performance, as illustrated in Figure 66.

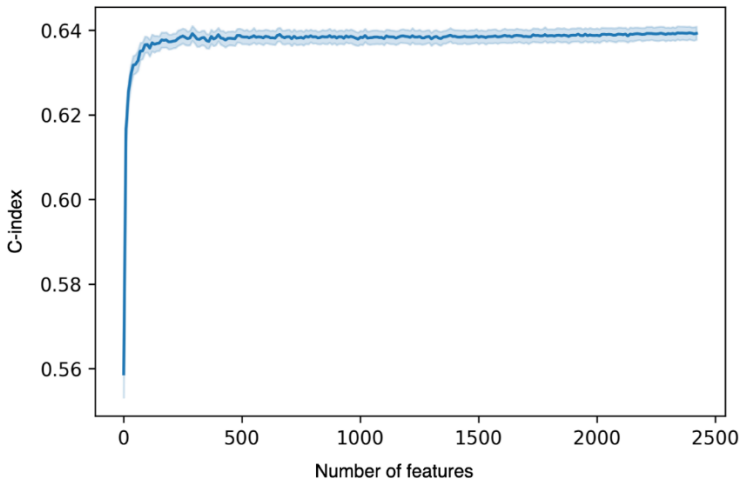


Figure 66: Concordance index of the ICARE model to predict to risk of relapse of the HECKTOR 2022 challenge as a function of the number of image-based features input into the model.

While I could not prove it, I made a hypothesis that could explain this phenomenon. In ICARE, the features signs are evaluated in univariate, independently from the other features. This creates a situation similar to the wisdom of the crowd effect. In his famous experiment in 1906, Galton showed that a crowd of people can accurately guess the weight of an ox. Individual answers were almost all wrong but the median of all the answers was close from the truth. This phenomenon was observed countless times in many situations and is now well documented and understood. One key component to make this effect operate is that each individual should answer independently, without the influence of others. The exact same phenomenon could happen in ICARE, where each feature contributes equally to the final answer without being influenced by other features. It is the opposite of a Cox model for instance, where the weight of each feature is determined not only by the target but also by the other features during training. One reason to explain the effectiveness of the wisdom of the crowd effect is that the answer of each individual is equally distributed around the truth, and by aggregating all the answers, all the individual errors cancel out. The same might happen in ICARE, where the errors in prediction of each feature is cancelled out by the errors of all other features. When more and more features are added, errors do better cancel out. This is what we observed during the HECKTOR challenge. Unfortunately, I could not prove that this is what happens in ICARE, and I hope that future works will answer this question.

If we take a step back from ICARE, simplicity was overall a key ally during my PhD. I noticed in multiple aspects of my work that the simpler the approach, the more effective it was to answer the questions. Simpler features were more effective at predicting the outcome and were more interpretable than sophisticated ones, simpler models were more effective at predicting a target than complicated models, simpler methods were more effective at searching the image space and simpler and more intuitive interpretation techniques were more effective than advanced ones. This is not a call for simplistic approaches, but rather for parsimonious ones in the complex and noisy domain of medical imaging.

This idea also applies to the interpretability of the features. The deep feature approach explored was promising but extremely hard to interpret. I tried advanced technique like GradCam and SHAP values, but despite my efforts, I was not able to find any convincing biological intuition from these features. Handcrafted radiomic features were much easier to interpret. By being defined by simpler rules on specifically defined portion of the image, their core information was easier to understand by their definition only. But an even greater ally was visual examples of the features. By plotting examples of lesions at high or low risk according to the studied feature, strong intuition about the quantified information could be developed. For instance, the shell CT GLDM Dependence Entropy that was identified as being prognostic in both the DLBCL and FL cohorts was difficult to interpret until visual examples were produced. Figure 67 shows an example of the images used to understand this feature. With these examples, it becomes obvious that this feature is quantifying the homogeneity in density around the lesions.

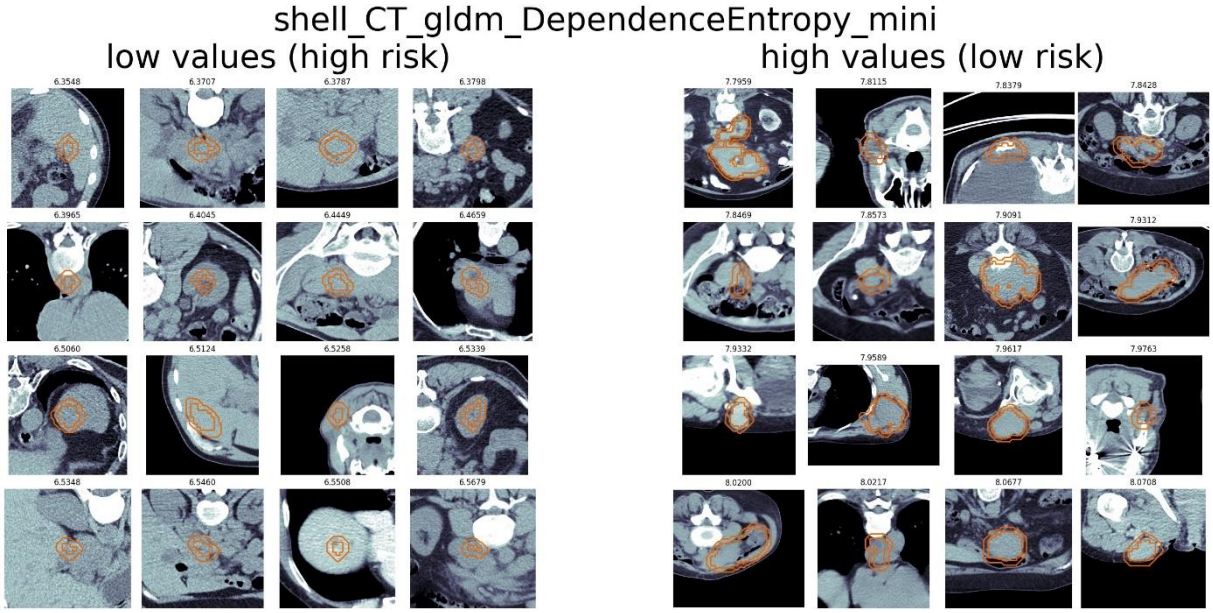


Figure 67: CT slices of lesions of FL patients with low and high values of GLDM Dependence Entropy computed in the 8mm thick shell of tissues surrounding the lesions. Lower values were associated with more homogeneous density and higher risk. Lesion and shell segmentation are depicted in orange.

These contributions open new perspectives. First, the identified biomarkers need to be thoroughly tested on new cohorts of DLBCL and FL patients. If confirmed, these findings could help better understand the diseases but also better predict the outcome of patients and stratify them by risk. It would also be interesting to see if some of them are prognostic in other types of cancer.

Second, the semi-automated methodology developed to search for new biomarkers can easily be applied to other cohorts with different diseases. By making the radiomic feature search easier, faster, and more exhaustive, these tools have the potential to help increase the number of discoveries in the field. As we see, they cannot fully replace the human search. Careful interpretation and manual re-encoding of the features are needed. Yet, they could be a great help to understand where to look for new information.

I believe that my PhD contributions show that there is still more information to leverage from PET and CT images, and we are currently only considering the tip of the iceberg, as there is probably so much more we can learn and measure from these images. I am convinced that in the future, PET/CT scans will play an increasingly crucial role in diagnosing and staging diseases in patients. I hope my contributions prove valuable and I extend my best wishes for success to those who will continue this research.

References

[1] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, "The ever-increasing importance of cancer as a leading cause of premature death worldwide," *Cancer*, vol. 127, no. 16, pp. 3029–3030, 2021, doi: 10.1002/cncr.33587.

[2] J. Ferlay *et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International Journal of Cancer*, vol. 144, no. 8, pp. 1941–1953, 2019, doi: 10.1002/ijc.31937.

[3] Y. F. Gu, F. P. Lin, and R. J. Epstein, "How aging of the global population is changing oncology," *Ecancermedicalscience*, vol. 15, p. ed119, Dec. 2021, doi: 10.3332/ecancer.2021.ed119.

[4] L. Chen, H. M. Linden, B. O. Anderson, and C. I. Li, "Trends in 5-year survival rates among breast cancer patients by hormone receptor status and stage," *Breast Cancer Res Treat*, vol. 147, no. 3, pp. 609–616, Oct. 2014, doi: 10.1007/s10549-014-3112-6.

[5] W. S. Halsted, "I. The Results of Operations for the Cure of Cancer of the Breast Performed at the Johns Hopkins Hospital from June, 1889, to January, 1894," *Ann Surg*, vol. 20, no. 5, pp. 497–555, Nov. 1894, doi: 10.1097/00000658-189407000-00075.

[6] M. O'Leary, M. Krailo, J. R. Anderson, and G. H. Reaman, "Progress in Childhood Cancer: 50 Years of Research Collaboration, a Report From the Children's Oncology Group," *Seminars in Oncology*, vol. 35, no. 5, pp. 484–493, Oct. 2008, doi: 10.1053/j.seminoncol.2008.07.008.

[7] S. K. Zhou *et al.*, "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, May 2021, doi: 10.1109/JPROC.2021.3054390.

[8] X. Ren *et al.*, "Weakly supervised label propagation algorithm classifies lung cancer imaging subtypes," *Sci Rep*, vol. 13, no. 1, Art. no. 1, Mar. 2023, doi: 10.1038/s41598-023-32301-4.

[9] "What is a CT Scan? | Neighborhood Radiologist." Accessed: Feb. 09, 2024. [Online]. Available: https://www.neighborhoodradiologist.com/what-is-a-ct-scan/

[10] S. Maharjan, K. Parajuli, S. Sah, and U. Poudel, "Knowledge of radiation protection among radiology professionals and students: A medical college-based study," *European Journal of Radiology Open*, vol. 7, p. 100287, Jan. 2020, doi: 10.1016/j.ejro.2020.100287.

[11] "X-rays," National Institute of Biomedical Imaging and Bioengineering. Accessed: Feb. 09, 2024. [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/x-rays

[12] R. Conrad Wilhelm, *Deutsch: Röntgen, Wilhelm Conrad: Röntgenphotographie der »Hand von Frau Röntgen«*. 1895. Accessed: Feb. 09, 2024. [Online]. Available:

https://commons.wikimedia.org/w/index.php?title=File:R%C3%B6ntgen,_Wilhelm_Conrad_-_R%C3%B6ntgenphotographie_der_%C2%BBHand_von_Frau_R%C3%B6ntgen%C2%AB_(Zeno_Fotografie).jpg&oldid=777728792

[13] R. A. Schulz, J. A. Stein, and N. J. Pelc, "How CT happened: the early development of medical computed tomography," *JMI*, vol. 8, no. 5, p. 052110, Oct. 2021, doi: 10.1117/1.JMI.8.5.052110.

[14] H. Hu, "Multi-slice helical CT: Scan and reconstruction," *Medical Physics*, vol. 26, no. 1, pp. 5–18, 1999, doi: 10.1118/1.598470.

[15] J. P. Heiken, J. A. Brink, and M. W. Vannier, "Spiral (helical) CT.," *Radiology*, vol. 189, no. 3, pp. 647–656, Dec. 1993, doi: 10.1148/radiology.189.3.8234684.

[16] D. S. Raksha, "The Difference Between CT Scan And X-Ray," Kiran Lab. Accessed: Feb. 09, 2024. [Online]. Available: https://kiranpetct.com/the-difference-between-ct-scan-and-x-ray/

[17] "schoolphysics::Welcome::" Accessed: Feb. 09, 2024. [Online]. Available: https://www.schoolphysics.co.uk/age16-19/Medical%20physics/text/CT_scanning/index.html

[18] L. L. Geyer *et al.*, "State of the Art: Iterative CT Reconstruction Techniques," *Radiology*, vol. 276, no. 2, pp. 339–357, Aug. 2015, doi: 10.1148/radiol.2015132766.

[19] W. De Vos, J. Casselman, and G. R. J. Swennen, "Cone-beam computerized tomography (CBCT) imaging of the oral and maxillofacial region: A systematic review of the literature," *International Journal of Oral and Maxillofacial Surgery*, vol. 38, no. 6, pp. 609–625, Jun. 2009, doi: 10.1016/j.ijom.2009.02.028.

[20] W. R. Webb, W. E. Brant, and N. M. Major, *Fundamentals of Body CT E-Book*. Elsevier Health Sciences, 2019.

[21] K. Huynh, A. H. Baghdanian, A. A. Baghdanian, D. S. Sun, K. P. Kolli, and R. J. Zagoria, "Updated guidelines for intravenous contrast use for CT and MRI," *Emerg Radiol*, vol. 27, no. 2, pp. 115–126, Apr. 2020, doi: 10.1007/s10140-020-01751-y.

[22] D. Picone, "Contrast medium injection in CT performed for bowel obstruction: is it really useful?," p. 844 words, 2015, doi: 10.1594/ECR2015/C-1059.

[23] M. M. Lell and M. Kachelrieß, "Recent and Upcoming Technological Developments in Computed Tomography: High Speed, Low Dose, Deep Learning, Multienergy," *Investigative Radiology*, vol. 55, no. 1, p. 8, Jan. 2020, doi: 10.1097/RLI.0000000000000601.

[24] J. J. Vaquero and P. Kinahan, "Positron Emission Tomography: Current Challenges and Opportunities for Technological Advances in Clinical and Preclinical Imaging Systems," *Annual Review of Biomedical Engineering*, vol. 17, no. 1, pp. 385–414, 2015, doi: 10.1146/annurev-bioeng-071114-040723.

[25] R. Huang *et al.*, "FAPI-PET/CT in Cancer Imaging: A Potential Novel Molecule of the Century," *Frontiers in Oncology*, vol. 12, 2022, Accessed: Feb. 09, 2024. [Online]. Available: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.854658

[26] J. Rong, A. Haider, T. E. Jeppesen, L. Josephson, and S. H. Liang, "Radiochemistry for positron emission tomography," *Nat Commun*, vol. 14, no. 1, Art. no. 1, Jun. 2023, doi: 10.1038/s41467-023-36377-4.

[27] P. Lindholm, H. Minn, S. Leskinen-Kallio, J. Bergman, U. Ruotsalainen, and H. Joensuu, "Influence of the Blood Glucose Concentration on FDG Uptake in Cancer—A PET Study," *Journal of Nuclear Medicine*, vol. 34, no. 1, pp. 1–6, Jan. 1993, Accessed: Feb. 09, 2024. [Online]. Available: https://jnm.snmjournals.org/content/34/1/1

[28] S. Basu, T. C. Kwee, S. Surti, E. A. Akin, D. Yoo, and A. Alavi, "Fundamentals of PET and PET/CT imaging," *Annals of the New York Academy of Sciences*, vol. 1228, no. 1, pp. 1–18, 2011, doi: 10.1111/j.1749-6632.2011.06077.x.

[29] H. Cho *et al.*, "Tau PET in Alzheimer disease and mild cognitive impairment," *Neurology*, vol. 87, no. 4, pp. 375–383, Jul. 2016, doi: 10.1212/WNL.0000000000002892.

[30] T. V. Baxa Jan, *PET/CT image*. 2011. Accessed: Feb. 09, 2024. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:Petct1.jpg&oldid=495898320

[31] C. Catana, A. Drzezga, W.-D. Heiss, and B. R. Rosen, "PET/MRI for Neurologic Applications," *Journal of Nuclear Medicine*, vol. 53, no. 12, pp. 1916–1925, Dec. 2012, doi: 10.2967/jnumed.112.105346.

[32] S. F. Barrington and M. J. O'Doherty, "Limitations of PET for imaging lymphoma," *Eur J Nucl Med Mol Imaging*, vol. 30, no. 1, pp. S117–S127, Jun. 2003, doi: 10.1007/s00259-003-1169-2.

[33] B. Huang, M. W.-M. Law, and P.-L. Khong, "Whole-Body PET/CT Scanning: Estimation of Radiation Dose and Cancer Risk," *Radiology*, vol. 251, no. 1, pp. 166–174, Apr. 2009, doi: 10.1148/radiol.2511081300.

[34] X. Chen *et al.*, "A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy," *Radiotherapy and Oncology*, vol. 160, pp. 175–184, Jul. 2021, doi: 10.1016/j.radonc.2021.04.019.

[35] K. K. D. Ramesh, G. K. Kumar, K. Swapna, D. Datta, and S. S. Rajest, "A Review of Medical Image Segmentation Algorithms," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 27, pp. e6–e6, Apr. 2021, doi: 10.4108/eai.12-4-2021.169184.

[36] X. Liu, L. Song, S. Liu, and Y. Zhang, "A Review of Deep-Learning-Based Medical Image Segmentation Methods," *Sustainability*, vol. 13, no. 3, Art. no. 3, Jan. 2021, doi: 10.3390/su13031224.

[37] X. Luo *et al.*, "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image," *Medical Image Analysis*, vol. 82, p. 102642, Nov. 2022, doi: 10.1016/j.media.2022.102642.

[38] "FDG-PET-CT-LESIONS," The Cancer Imaging Archive (TCIA). Accessed: Feb. 09, 2024. [Online]. Available: https://www.cancerimagingarchive.net/collection/fdg-pet-ct-lesions/

[39] A. Tuladhar, S. Schimert, D. Rajashekar, H. C. Kniep, J. Fiehler, and N. D. Forkert, "Automatic Segmentation of Stroke Lesions in Non-Contrast Computed Tomography Datasets With Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 94871–94879, 2020, doi: 10.1109/ACCESS.2020.2995632.

[40] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, May 2021, doi: 10.1016/j.ejmp.2021.05.003.

[41] M. Havaei *et al.*, "Brain tumor segmentation with Deep Neural Networks," *Medical Image Analysis*, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[42] J. S. Suri, S. K. Setarehdan, and S. Singh, *Advanced algorithmic approaches to medical image segmentation: state-of-the-art applications in cardiology, neurology, mammography and pathology*. Springer Science & Business Media, 2012.

[43] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-UNet: A Dimension-Fusion U Shape Network for Chronic Stroke Lesion Segmentation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 940–950, May 2021, doi: 10.1109/TCBB.2019.2939522.

[44] H. Ilyas *et al.*, "Defining the optimal method for measuring baseline metabolic tumour volume in diffuse large B cell lymphoma," *Eur J Nucl Med Mol Imaging*, vol. 45, no. 7, pp. 1142–1154, Jul. 2018, doi: 10.1007/s00259-018-3953-z.

[45] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated Design of Deep Learning Methods for Biomedical Image Segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[46] J. Wasserthal *et al.*, "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images," *Radiol Artif Intell*, vol. 5, no. 5, p. e230024, Jul. 2023, doi: 10.1148/ryai.230024.

[47] L. K. S. Sundar *et al.*, "Fully Automated, Semantic Segmentation of Whole-Body 18F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence," *Journal of Nuclear Medicine*, vol. 63, no. 12, pp. 1941–1948, Dec. 2022, doi: 10.2967/jnumed.122.264063.

[48] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat Mach Intell*, vol. 1, no. 9, Art. no. 9, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[49] N. Farzaneh, S. Ansari, E. Lee, K. R. Ward, and M. W. Sjoding, "Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome," *npj Digit. Med.*, vol. 6, no. 1, Art. no. 1, Apr. 2023, doi: 10.1038/s41746-023-00797-9.

[50] W. N. Price II, S. Gerke, and I. G. Cohen, "Potential Liability for Physicians Using Artificial Intelligence," *JAMA*, vol. 322, no. 18, pp. 1765–1766, Nov. 2019, doi: 10.1001/jama.2019.15064.

[51] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images Are More than Pictures, They Are Data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[52] J. Song, Y. Yin, H. Wang, Z. Chang, Z. Liu, and L. Cui, "A review of original articles published in the emerging field of radiomics," *European Journal of Radiology*, vol. 127, p. 108991, Jun. 2020, doi: 10.1016/j.ejrad.2020.108991.

[53] B. Wichtmann *et al.*, "Influence of image processing on the robustness of radiomic features derived from magnetic resonance imaging—a phantom study," in *ISMRM 2018*, 2018, p. 5.

[54] B. A. Altazi *et al.*, "Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms," *Journal of applied clinical medical physics*, vol. 18, no. 6, pp. 32–48, 2017.

[55] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.

[56] J. J. M. van Griethuysen *et al.*, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, Oct. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[57] F. Orlhac, C. Nioche, M. Soussan, and I. Buvat, "Understanding Changes in Tumor Texture Indices in PET: A Comparison Between Visual Assessment and Index Values in Simulated and Patient Data," *Journal of Nuclear Medicine*, vol. 58, no. 3, pp. 387–392, Mar. 2017, doi: 10.2967/jnumed.116.181859.

[58] R. El Ayachy *et al.*, "The Role of Radiomics in Lung Cancer: From Screening to Treatment and Follow-Up," *Frontiers in Oncology*, vol. 11, 2021, Accessed: Feb. 10, 2024. [Online]. Available: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.603595

[59] R. Jing *et al.*, "A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules," *Sci Rep*, vol. 11, no. 1, Art. no. 1, Nov. 2021, doi: 10.1038/s41598-021-01470-5.

[60] V. Romeo *et al.*, "Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions," *Eur Radiol*, vol. 31, no. 12, pp. 9511–9519, Dec. 2021, doi: 10.1007/s00330-021-08009-2.

[61] P. Aonpong, Y. Iwamoto, X.-H. Han, L. Lin, and Y.-W. Chen, "Genotype-Guided Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer," *IEEE Access*, vol. 9, pp. 90244–90254, 2021, doi: 10.1109/ACCESS.2021.3088234.

[62] E. A. Hoivik *et al.*, "A radiogenomics application for prognostic profiling of endometrial cancer," *Commun Biol*, vol. 4, no. 1, Art. no. 1, Dec. 2021, doi: 10.1038/s42003-021-02894-5.

[63] S. Sellami *et al.*, "Predicting response to radiotherapy of head and neck squamous cell carcinoma using radiomics from cone-beam CT images," *Acta Oncologica*, vol. 61, no. 1, pp. 73–80, Jan. 2022, doi: 10.1080/0284186X.2021.1983207.

[64] S. Volpe, F. Mastroleo, M. Krengli, and B. A. Jereczek-Fossa, "Quo vadis Radiomics? Bibliometric analysis of 10-year Radiomics journey," *Eur Radiol*, vol. 33, no. 10, pp. 6736–6745, Oct. 2023, doi: 10.1007/s00330-023-09645-6.

[65] D. Pinto dos Santos, M. Dietzel, and B. Baessler, "A decade of radiomics research: are images really data or just patterns in the noise?," *Eur Radiol*, vol. 31, no. 1, pp. 1–4, Jan. 2021, doi: 10.1007/s00330-020-07108-w.

[66] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/radiol.2020191145.

[67] P. Whybra *et al.*, "The Image Biomarker Standardization Initiative: Standardized Convolutional Filters for Reproducible Radiomics and Enhanced Clinical Insights," *Radiology*, vol. 310, no. 2, p. e231319, Feb. 2024, doi: 10.1148/radiol.231319.

[68] C. Nioche *et al.*, "LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity," *Cancer Res*, vol. 78, no. 16, pp. 4786–4789, Aug. 2018, doi: 10.1158/0008-5472.CAN-18-0125.

[69] F. Orlhac *et al.*, "How can we combat multicenter variability in MR radiomics? Validation of a correction procedure," *Eur Radiol*, vol. 31, no. 4, pp. 2272–2280, Apr. 2021, doi: 10.1007/s00330-020-07284-9.

[70] B. Koçak, E. Ş. Durmaz, E. Ateş, and Ö. Kılıçkesmez, "Radiomics with artificial intelligence: a practical guide for beginners," *Diagn Interv Radiol*, vol. 25, no. 6, pp. 485–495, Nov. 2019, doi: 10.5152/dir.2019.19321.

[71] M. Hatt *et al.*, "Joint EANM/SNMMI guideline on radiomics in nuclear medicine," *Eur J Nucl Med Mol Imaging*, vol. 50, no. 2, pp. 352–375, Jan. 2023, doi: 10.1007/s00259-022-06001-6.

[72] E. P. Huang *et al.*, "Criteria for the translation of radiomics into clinically useful tests," *Nat Rev Clin Oncol*, vol. 20, no. 2, Art. no. 2, Feb. 2023, doi: 10.1038/s41571-022-00707-0.

[73] B. Kocak *et al.*, "METhodological RadiomICs Score (METRICS): A quality scoring tool for radiomics research," Nov. 2023. Accessed: Feb. 10, 2024. [Online]. Available: https://hal.science/hal-04305543

[74] M. R. Tomaszewski and R. J. Gillies, "The Biological Meaning of Radiomic Features," *Radiology*, vol. 298, no. 3, pp. 505–516, Mar. 2021, doi: 10.1148/radiol.2021202553.

[75] F. Orlhac, B. Thézé, M. Soussan, R. Boisgard, and I. Buvat, "Multiscale Texture Analysis: From 18F-FDG PET Images to Histologic Images," *J Nucl Med*, vol. 57, no. 11, pp. 1823–1828, Nov. 2016, doi: 10.2967/jnumed.116.173708.

[76] T. Escobar *et al.*, "Voxel-wise supervised analysis of tumors with multimodal engineered features to highlight interpretable biological patterns," *Medical Physics*, vol. 49, no. 6, pp. 3816–3829, 2022, doi: 10.1002/mp.15603.

[77] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[78] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," *Kidney Res Clin Pract*, vol. 36, no. 1, pp. 3–11, Mar. 2017, doi: 10.23876/j.krcp.2017.36.1.3.

[79] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, Art. no. 7639, Feb. 2017, doi: 10.1038/nature21056.
[80] M. L. Giger, "Machine Learning in Medical Imaging," *Journal of the American College of Radiology*, vol. 15, no. 3, Part B, pp. 512–520, Mar. 2018, doi: 10.1016/j.jacr.2017.12.028.
[81] L. Sibille *et al.*, "18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks," *Radiology*, vol. 294, no. 2, pp. 445–452, Feb. 2020, doi: 10.1148/radiol.2019191114.
[82] D. Cielen and A. Meysman, *Introducing Data Science: Big data, machine learning, and more, using Python tools*. Simon and Schuster, 2016.
[83] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *International journal of computer science*, vol. 1, no. 2, pp. 111–117, 2006.
[84] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *WIREs Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012, doi: 10.1002/wics.1198.
[85] T. G. Nick and K. M. Campbell, "Logistic Regression," in *Topics in Biostatistics*, W. T. Ambrosius, Ed., in Methods in Molecular Biology™. , Totowa, NJ: Humana Press, 2007, pp. 273–301. doi: 10.1007/978-1-59745-530-5_14.
[86] "Python Decision Tree Classification Tutorial: Scikit-Learn DecisionTreeClassifier." Accessed: Feb. 10, 2024. [Online]. Available: https://www.datacamp.com/tutorial/decision-tree-classification-python
[87] P. Protopapas, K. Rader, R. Dave, and M. Levine, "Lecture 15: Regression Trees and Random Forests," *Harvard CS109A: Introduction to Data Science*.
[88] R. Odegua, "An empirical study of ensemble techniques (bagging, boosting and stacking)," in *Proc. Conf.: Deep Learn. IndabaXAt*, 2019.
[89] S. Ismail, Z. El Mrabet, and H. Reza, "An Ensemble-Based Machine Learning Approach for Cyber-Attacks Detection in Wireless Sensor Networks," *Applied Sciences*, vol. 13, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/app13010030.
[90] W. S. Noble, "What is a support vector machine?," *Nat Biotechnol*, vol. 24, no. 12, Art. no. 12, Dec. 2006, doi: 10.1038/nbt1206-1565.
[91] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," Medium. Accessed: Feb. 10, 2024. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[92] F. M. Wittmann, "Visualization of SVM Kernels Linear, RBF, Poly and Sigmoid on Python (Adapted from: http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html)," Gist. Accessed: Feb. 10, 2024. [Online]. Available: https://gist.github.com/WittmannF/60680723ed8dd0cb993051a7448f7805
[93] K. E. Koech, "The Basics of Neural Networks (Neural Network Series) — Part 1," Medium. Accessed: Feb. 10, 2024. [Online]. Available: https://towardsdatascience.com/the-basics-of-neural-networks-neural-network-series-part-1-4419e343b2b

[94] M. Faqe Hussein, A. Saeed, and S. Mohamad, "Comparison Markov Chain and Neural Network Models for forecasting Population growth data in Iraq," vol. 13, pp. 1–14, Dec. 2023.
[95] S. P. Jenkins, "Survival analysis," *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, vol. 42, pp. 54–56, 2005.
[96] "GraphPad Prism 10 Statistics Guide - Censored Data." Accessed: Feb. 10, 2024. [Online]. Available: https://www.graphpad.com/guides/prism/latest/statistics/stat_censored_data.htm
[97] A. Nardi and M. Schemper, "Comparing Cox and parametric models in clinical studies," *Statistics in Medicine*, vol. 22, no. 23, pp. 3597–3610, 2003, doi: 10.1002/sim.1592.
[98] L. Rebaud *et al.*, "Multitask learning-to-rank neural network for predicting survival of diffuse large B-cell lymphoma patients from their unsegmented baseline [18F]FDG-PET/CT scans.," *Journal of Nuclear Medicine*, vol. 63, no. supplement 2, pp. 3250–3250, Jun. 2022, Accessed: Feb. 10, 2024. [Online]. Available: https://jnm.snmjournals.org/content/63/supplement_2/3250
[99] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011, doi: 10.1002/sim.4154.
[100] J. Lambert and S. Chevret, "Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves," *Stat Methods Med Res*, vol. 25, no. 5, pp. 2088–2102, Oct. 2016, doi: 10.1177/0962280213515571.
[101] "Evaluating Survival Models - scikit-survival." Dec. 2023. [Online]. Available: https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html
[102] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding Kaplan-Meier curves," *Otolaryngol Head Neck Surg*, vol. 143, no. 3, pp. 331–336, Sep. 2010, doi: 10.1016/j.otohns.2010.05.007.
[103] "Kaplan Meier curves: an introduction." Accessed: Feb. 10, 2024. [Online]. Available: https://rmvpaeme.github.io/KaplanMeier_intro/
[104] D. G. Kleinbaum and M. Klein, *Survival analysis a self-learning text*. Springer, 1996.
[105] J. Li *et al.*, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, p. 94:1-94:45, Dec. 2017, doi: 10.1145/3136625.
[106] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, Dec. 1996, doi: 10.1016/S0895-4356(96)00236-3.
[107] T. J. VanderWeele and I. Shpitser, "On the definition of a confounder," *Ann Stat*, vol. 41, no. 1, pp. 196–220, Feb. 2013, Accessed: Feb. 10, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276366/

[108] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.
[109] K. E. S. Pilario, Y. Cao, and M. Shafiee, "A kernel design approach to improve kernel subspace identification," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 7, pp. 6171–6180, 2020.
[110] R. Turner *et al.*, "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020," in *NeurIPS 2020 Competition and Demonstration Track*, PMLR, 2021, pp. 3–26.
[111] R. K. Samala, H.-P. Chan, L. Hadjiiski, and S. Koneru, "Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks," in *Medical Imaging 2020: Computer-Aided Diagnosis*, SPIE, Mar. 2020, pp. 279–284. doi: 10.1117/12.2549313.
[112] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning." arXiv, Nov. 10, 2020. doi: 10.48550/arXiv.1811.12808.
[113] Y. Zhong, J. He, P. Chalise, Y. Zhong, J. He, and P. Chalise, "Nested and Repeated Cross Validation for Classification Model With High-dimensional Data," *Revista Colombiana de Estadística*, vol. 43, no. 1, pp. 103–125, Jun. 2020, doi: 10.15446/rce.v43n1.80000.
[114] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41746-022-00592-y.
[115] P. Good, *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.
[116] "Permutation tests and independent sorting of data," The DO Loop. Accessed: Feb. 10, 2024. [Online]. Available: https://blogs.sas.com/content/iml/2021/06/07/permutation-tests-sorting-data.html
[117] J. P. Shaffer, "Multiple Hypothesis Testing," *Annual Review of Psychology*, vol. 46, no. 1, pp. 561–584, 1995, doi: 10.1146/annurev.ps.46.020195.003021.
[118] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[119] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, pp. 491–507, Sep. 2006, doi: 10.1093/biomet/93.3.491.
[120] K. C. Thandra, A. Barsouk, K. Saginala, S. A. Padala, A. Barsouk, and P. Rawla, "Epidemiology of Non-Hodgkin's Lymphoma," *Medical Sciences*, vol. 9, no. 1, Art. no. 1, Mar. 2021, doi: 10.3390/medsci9010005.
[121] P. Nenclares and K. J. Harrington, "The biology of cancer," *Medicine*, vol. 48, no. 2, pp. 67–72, Feb. 2020, doi: 10.1016/j.mpmed.2019.11.001.
[122] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[123] E. N. Kontomanolis *et al.*, "Role of Oncogenes and Tumor-suppressor Genes in Carcinogenesis: A Review," *Anticancer Research*, vol. 40, no. 11, pp. 6009–6015, Nov. 2020, doi: 10.21873/anticanres.14622.
[124] N. Parsa, "Environmental Factors Inducing Human Cancers," *Iran J Public Health*, vol. 41, no. 11, pp. 1–9, Nov. 2012, Accessed: Feb. 10, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3521879/
[125] A. Patel, "Benign vs Malignant Tumors," *JAMA Oncology*, vol. 6, no. 9, p. 1488, Sep. 2020, doi: 10.1001/jamaoncol.2020.2592.
[126] D. M. Hausman, "What is cancer?," *Perspectives in biology and medicine*, vol. 62, no. 4, pp. 778–784, 2019.
[127] S. S. Devesa, W. J. Blot, B. J. Stone, B. A. Miller, R. E. Tarone, and J. F. Fraumeni Jr., "Recent Cancer Trends in the United States," *JNCI: Journal of the National Cancer Institute*, vol. 87, no. 3, pp. 175–182, Feb. 1995, doi: 10.1093/jnci/87.3.175.
[128] B. M. Rothschild, B. J. Witzke, and I. Hershkovitz, "Metastatic cancer in the Jurassic," *The Lancet*, vol. 354, no. 9176, p. 398, 1999.
[129] E. J. Odes *et al.*, "Earliest hominin cancer: 1.7-million-year-old osteosarcoma from Swartkrans Cave, South Africa," *South African Journal of Science*, vol. 112, no. 7–8, pp. 1–5, 2016.
[130] C. Prates, S. Sousa, C. Oliveira, and S. Ikram, "Prostate metastatic bone cancer in an Egyptian Ptolemaic mummy, a proposed radiological diagnosis," *International Journal of Paleopathology*, vol. 1, no. 2, pp. 98–103, 2011.
[131] A. R. David and M. R. Zimmerman, "Cancer: an old disease, a new disease or something in between?," *Nature Reviews Cancer*, vol. 10, no. 10, pp. 728–733, 2010.
[132] C. M. Galmarini, "Lessons from Hippocrates: Time to change the cancer paradigm," *International Journal of Chronic Diseases*, vol. 2020, 2020.
[133] J. Lipsick, "A history of cancer research: tumor suppressor genes," *Cold Spring Harbor Perspectives in Biology*, vol. 12, no. 2, p. a035907, 2020.
[134] D. E. Sabath, "Philadelphia Chromosome," in *Brenner's Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Eds., San Diego: Academic Press, 2013, p. 308. doi: 10.1016/B978-0-12-374984-0.01155-4.
[135] D. Hanahan, "Hallmarks of cancer: new dimensions," *Cancer discovery*, vol. 12, no. 1, pp. 31–46, 2022.
[136] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *cell*, vol. 100, no. 1, pp. 57–70, 2000.
[137] J. P. Williamson, M. Phillips, D. Hillman, and P. Eastwood, "Managing obstruction of the central airways," *Internal medicine journal*, vol. 40, no. 6, pp. 399–410, 2010.
[138] D. D. Correa, "Neurocognitive function in brain tumors," *Current neurology and neuroscience reports*, vol. 10, pp. 232–239, 2010.
[139] "End-of-Life Care - NCI." Accessed: Feb. 10, 2024. [Online]. Available: https://www.cancer.gov/about-cancer/advanced-cancer/care-choices/care-fact-sheet

[140] C. Y. Fangand and R. A. Schnoll, "Impact of psychological distress on outcomes in cancer patients," *Expert review of pharmacoeconomics & outcomes research*, vol. 2, no. 5, pp. 495–506, 2002.
[141] K. D. Stein, K. L. Syrjala, and M. A. Andrykowski, "Physical and psychological long-term and late effects of cancer," *Cancer*, vol. 112, no. S11, pp. 2577–2592, 2008.
[142] C. Pitceathly and P. Maguire, "The psychological impact of cancer on patients' partners and other key relatives: a review," *European Journal of cancer*, vol. 39, no. 11, pp. 1517–1524, 2003.
[143] L. Fass, "Imaging and cancer: a review," *Molecular oncology*, vol. 2, no. 2, pp. 115–152, 2008.
[144] R. Vaidyanathan, R. H. Soon, P. Zhang, K. Jiang, and C. T. Lim, "Cancer diagnosis: from tumor to liquid biopsy and beyond," *Lab on a Chip*, vol. 19, no. 1, pp. 11–34, 2019.
[145] D. Chakravarty and D. B. Solit, "Clinical cancer genomic profiling," *Nature Reviews Genetics*, vol. 22, no. 8, pp. 483–501, 2021.
[146] J. Zugazagoitia, C. Guedes, S. Ponce, I. Ferrer, S. Molina-Pinelo, and L. Paz-Ares, "Current challenges in cancer treatment," *Clinical therapeutics*, vol. 38, no. 7, pp. 1551–1566, 2016.
[147] P. Krzyszczyk *et al.*, "The growing role of precision and personalized medicine for cancer treatment," *Technology*, vol. 6, no. 03n04, pp. 79–100, 2018.
[148] L. Wyld, R. A. Audisio, and G. J. Poston, "The evolution of cancer surgery and future perspectives," *Nature reviews Clinical oncology*, vol. 12, no. 2, pp. 115–124, 2015.
[149] R. Baskar, K. A. Lee, R. Yeo, and K.-W. Yeoh, "Cancer and radiation therapy: current advances and future directions," *International journal of medical sciences*, vol. 9, no. 3, p. 193, 2012.
[150] P. Nygren, "What is cancer chemotherapy?," *Acta Oncologica*, vol. 40, no. 2–3, pp. 166–174, 2001.
[151] K. Esfahani, L. Roudaia, N. Buhlaiga, S. Del Rincon, N. Papneja, and W. Miller, "A review of cancer immunotherapy: from the past, to the present, to the future," *Current Oncology*, vol. 27, no. s2, pp. 87–97, 2020.
[152] A.-M. Tsimberidou, "Targeted therapy in cancer," *Cancer chemotherapy and pharmacology*, vol. 76, pp. 1113–1132, 2015.
[153] G. P. Canellos, "Lymphoma: present and future challenges," presented at the Seminars in hematology, Elsevier, 2004, pp. 26–31.
[154] B. W. using this image in external sources it can be cited as:Blausen com staff, *Lymphatic System*. 2013. Accessed: Feb. 10, 2024. [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=28086436
[155] U. photographer/artist, *Electron microscopic image of a single human lymphocyte. Type: Black & White Print*. 1976. Accessed: Feb. 10, 2024. [Online]. Available: https://commons.wikimedia.org/w/index.php?title=File:SEM_Lymphocyte.jpg&oldid=597788760
[156] M. Null and M. Agarwal, "Anatomy, lymphatic system," 2018.

[157] R. L. E. Cano and H. D. E. Lopera, "Introduction to T and B lymphocytes," in *Autoimmunity: From Bench to Bedside [Internet]*, El Rosario University Press, 2013.
[158] A. Carbone *et al.*, "Follicular lymphoma," *Nat Rev Dis Primers*, vol. 5, no. 1, p. 83, Dec. 2019, doi: 10.1038/s41572-019-0132-x.
[159] M. Al-Hamadani, T. M. Habermann, J. R. Cerhan, W. R. Macon, M. J. Maurer, and R. S. Go, "Non-Hodgkin lymphoma subtype distribution, geodemographic patterns, and survival in the US: A longitudinal analysis of the National Cancer Data Base from 1998 to 2011," *American journal of hematology*, vol. 90, no. 9, pp. 790–795, 2015.
[160] E. N. Mugnaini and N. Ghosh, "Lymphoma," *Primary Care: Clinics in Office Practice*, vol. 43, no. 4, pp. 661–675, 2016.
[161] Y. Zhang, Y. Dai, T. Zheng, and S. Ma, "Risk factors of non-Hodgkin's lymphoma," *Expert opinion on medical diagnostics*, vol. 5, no. 6, pp. 539–550, 2011.
[162] P. de MM Boccolini *et al.*, "Pesticide use and non-Hodgkin's lymphoma mortality in Brazil," *International journal of hygiene and environmental health*, vol. 216, no. 4, pp. 461–466, 2013.
[163] R. F. Ambinder, "Epstein-barr virus and hodgkin lymphoma," *ASH Education Program Book*, vol. 2007, no. 1, pp. 204–209, 2007.
[164] R. W. Harbron and E. Pasqual, "Ionising radiation as a risk factor for lymphoma: A review," *Journal of Radiological Protection*, vol. 40, no. 4, p. R151, 2020.
[165] W. D. Lewis, S. Lilly, and K. L. Jones, "Lymphoma: diagnosis and treatment," *American family physician*, vol. 101, no. 1, pp. 34–41, 2020.
[166] L. Colomo *et al.*, "Clinical impact of the differentiation profile assessed by immunophenotyping in patients with diffuse large B-cell lymphoma," *Blood, The Journal of the American Society of Hematology*, vol. 101, no. 1, pp. 78–84, 2003.
[167] J. Iqbal, Z. Liu, K. Deffenbacher, and W. C. Chan, "Gene expression profiling in lymphoma diagnosis and management," *Best Practice & Research Clinical Haematology*, vol. 22, no. 2, pp. 191–210, 2009.
[168] J. W. Friedberg and V. Chengazi, "PET scans in the staging of lymphoma: current status," *The Oncologist*, vol. 8, no. 5, pp. 438–447, 2003.
[169] S. F. Barrington and J. Trotman, "The role of PET in the first-line treatment of the most common subtypes of non-Hodgkin lymphoma," *The Lancet Haematology*, vol. 8, no. 1, pp. e80–e93, 2021.
[170] L. Wang *et al.*, "Advances in targeted therapy for malignant lymphoma," *Signal transduction and targeted therapy*, vol. 5, no. 1, p. 15, 2020.
[171] S. M. Ansell and Y. Lin, "Immunotherapy of lymphomas," *The Journal of Clinical Investigation*, vol. 130, no. 4, pp. 1576–1585, 2020.
[172] J. N. Brudno and J. N. Kochenderfer, "Chimeric antigen receptor T-cell therapies for lymphoma," *Nature reviews Clinical oncology*, vol. 15, no. 1, pp. 31–46, 2018.
[173] S. J. Schuster, "Bispecific antibodies for the treatment of lymphomas: Promises and challenges," *Hematological Oncology*, vol. 39, pp. 113–116, 2021.
[174] H. Hatic, D. Sampat, and G. Goyal, "Immune checkpoint inhibitors in lymphoma: challenges and opportunities," *Annals of translational medicine*, vol. 9, no. 12, 2021.

[175] G. Kanas, W. Ge, R. G. Quek, K. Keeven, K. Nersesyan, and J. E. Arnason, "Epidemiology of diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) in the United States and Western Europe: population-level projections for 2020–2025," *Leukemia & Lymphoma*, vol. 63, no. 1, pp. 54–63, 2022.
[176] I. S. Lossos and R. D. Gascoyne, "Transformation of follicular lymphoma," *Best practice & research Clinical haematology*, vol. 24, no. 2, pp. 147–163, 2011.
[177] N. Masir *et al.*, "BCL2 protein expression in follicular lymphomas with t (14; 18) chromosomal translocations," *British journal of haematology*, vol. 144, no. 5, pp. 716–725, 2009.
[178] L. Pasqualucci *et al.*, "Genetics of follicular lymphoma transformation," *Cell reports*, vol. 6, no. 1, pp. 130–140, 2014.
[179] T. J. Bakhshi and P. T. Georgel, "Genetic and epigenetic determinants of diffuse large B-cell lymphoma," *Blood cancer journal*, vol. 10, no. 12, p. 123, 2020.
[180] H. Tilly *et al.*, "Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Annals of oncology*, vol. 26, pp. v116–v125, 2015.
[181] M. Dreyling *et al.*, "Newly diagnosed and relapsed follicular lymphoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†☆," *Annals of Oncology*, vol. 32, no. 3, pp. 298–308, 2021.
[182] J. O. Armitage, "Staging non-Hodgkin lymphoma," *CA: a cancer journal for clinicians*, vol. 55, no. 6, pp. 368–376, 2005.
[183] C. M. Kelly and A. Shahrokni, "Moving beyond Karnofsky and ECOG performance status assessments with new technologies," *Journal of oncology*, vol. 2016, 2016.
[184] S. Ahn, Y.-S. Lee, Y. H. Chun, K. S. Lim, W. Kim, and J.-L. Lee, "Predictive factors of bacteraemia in low-risk patients with febrile neutropenia," *Emergency Medicine Journal*, 2011.
[185] Leukemia & Lymphoma Society, "Hodgkin Lymphoma Staging." Accessed: Feb. 10, 2024. [Online]. Available: https://www.lls.org/lymphoma/hodgkin-lymphoma/diagnosis/hodgkin-lymphoma-staging
[186] B. D. Cheson *et al.*, "Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification," *Journal of clinical oncology*, vol. 32, no. 27, p. 3059, 2014.
[187] S. F. Barrington *et al.*, "Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group," *Journal of clinical oncology*, vol. 32, no. 27, p. 3048, 2014.
[188] International Non-Hodgkin's Lymphoma Prognostic Factors Project, "A predictive model for aggressive non-Hodgkin's lymphoma," *New England Journal of Medicine*, vol. 329, no. 14, pp. 987–994, 1993.
[189] M. B. Møller, B. E. Christensen, and N. T. Pedersen, "Prognosis of localized diffuse large B-cell lymphoma in younger patients," *Cancer*, vol. 98, no. 3, pp. 516–521, 2003, doi: 10.1002/cncr.11497.

[190] Z. Zhou *et al.*, "An enhanced International Prognostic Index (NCCN-IPI) for patients with diffuse large B-cell lymphoma treated in the rituximab era," *Blood, The Journal of the American Society of Hematology*, vol. 123, no. 6, pp. 837–842, 2014.
[191] P. Solal-Céligny *et al.*, "Follicular lymphoma international prognostic index," *Blood*, vol. 104, no. 5, pp. 1258–1265, 2004.
[192] N. G. Mikhaeel *et al.*, "Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL," *Eur J Nucl Med Mol Imaging*, vol. 43, no. 7, pp. 1209–1219, Jul. 2016, doi: 10.1007/s00259-016-3315-7.
[193] L. Kostakoglu *et al.*, "Baseline PET-Derived Metabolic Tumor Volume Metrics Predict Progression-Free and Overall Survival in DLBCL after First-Line Treatment: Results from the Phase 3 GOYA Study," *Blood*, vol. 130, p. 824, Dec. 2017, doi: 10.1182/blood.V130.Suppl_1.824.824.
[194] C. Schmitz *et al.*, "Dynamic risk assessment based on positron emission tomography scanning in diffuse large B-cell lymphoma: Post-hoc analysis from the PETAL trial," *European Journal of Cancer*, vol. 124, pp. 25–36, Jan. 2020, doi: 10.1016/j.ejca.2019.09.027.
[195] A. S. Cottereau *et al.*, "Prognostic model for high-tumor-burden follicular lymphoma integrating baseline and end-induction PET: a LYSA/FIL study," *Blood*, vol. 131, no. 22, pp. 2449–2453, May 2018, doi: 10.1182/blood-2017-11-816298.
[196] J.-H. Liang *et al.*, "Prognostic Value of Baseline and Interim Total Metabolic Tumor Volume and Total Lesion Glycolysis Measured on 18F-FDG PET-CT in Patients with Follicular Lymphoma," *Cancer Res Treat*, vol. 51, no. 4, pp. 1479–1487, Mar. 2019, doi: 10.4143/crt.2018.649.
[197] A. Cottereau *et al.*, "Metabolic tumor volume predicts outcome in patients with advanced stage follicular lymphoma from the RELEVANCE trial," *Annals of Oncology*, vol. 35, no. 1, pp. 130–137, 2024.
[198] A.-S. Cottereau *et al.*, "18F-FDG PET Dissemination Features in Diffuse Large B-Cell Lymphoma Are Predictive of Outcome," *J Nucl Med*, vol. 61, no. 1, pp. 40–45, 2020, doi: 10.2967/jnumed.119.229450.
[199] D. Albano *et al.*, "18F-FDG PET/CT Maximum Tumor Dissemination (Dmax) in Lymphoma: A New Prognostic Factor?," *Cancers*, vol. 15, no. 9, p. 2494, 2023.
[200] L. Rebaud *et al.*, "Evaluation of the prognostic value of tumor fragmentation on [18F]-FDG PET/CT on an independent cohort of diffuse large B-cell lymphoma patients," 2022.
[201] L. Zanoni *et al.*, "PET/CT in Non-Hodgkin Lymphoma: An Update," *Semin Nucl Med*, vol. 53, no. 3, pp. 320–351, May 2023, doi: 10.1053/j.semnuclmed.2022.11.001.
[202] H. Wang, Y. Zhou, L. Li, W. Hou, X. Ma, and R. Tian, "Current status and quality of radiomics studies in lymphoma: a systematic review," *European Radiology*, vol. 30, pp. 6228–6240, 2020.
[203] M. Piñeiro-Fiel, A. Moscoso, V. Pubul, Á. Ruibal, J. Silva-Rodríguez, and P. Aguiar, "A systematic review of PET textural analysis and radiomics in cancer," *Diagnostics*, vol. 11, no. 2, p. 380, 2021.

[204] P. Decazes *et al.*, "Tumor fragmentation estimated by volume surface ratio of tumors measured on 18F-FDG PET/CT is an independent prognostic factor of diffuse large B-cell lymphoma," *European journal of nuclear medicine and molecular imaging*, vol. 45, pp. 1672–1679, 2018.

[205] "Diffuse Large B-Cell Lymphoma - Cancer Stat Facts," SEER. Accessed: Jan. 18, 2024. [Online]. Available: https://seer.cancer.gov/statfacts/html/dlbcl.html

[206] N. G. Mikhaeel *et al.*, "Proposed New Dynamic Prognostic Index for Diffuse Large B-Cell Lymphoma: International Metabolic Prognostic Index," *J Clin Oncol*, vol. 40, no. 21, pp. 2352–2360, Jul. 2022, doi: 10.1200/JCO.21.02063.

[207] L. Vercellino *et al.*, "Predictive factors of early progression after CAR T-cell therapy in relapsed/refractory diffuse large B-cell lymphoma," *Blood Advances*, vol. 4, no. 22, pp. 5607–5615, Nov. 2020, doi: 10.1182/bloodadvances.2020003001.

[208] S. Wang *et al.*, "The prognostic value of splenic abnormalities in pretreatment 18F-FDG PET/CT in patients with complete response diffuse large B-cell lymphoma," *Clinical Radiology*, vol. 78, no. 5, pp. 375–380, May 2023, doi: 10.1016/j.crad.2023.01.010.

[209] R. Shen *et al.*, "Influence of oncogenic mutations and tumor microenvironment alterations on extranodal invasion in diffuse large B-cell lymphoma," *Clinical and Translational Medicine*, vol. 10, no. 7, p. e221, 2020, doi: 10.1002/ctm2.221.

[210] C. Thieblemont *et al.*, "Lenalidomide Maintenance Compared With Placebo in Responding Elderly Patients With Diffuse Large B-Cell Lymphoma Treated With First-Line Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 35, p. JCO2017726984, Apr. 2017, doi: 10.1200/JCO.2017.72.6984.

[211] R.-O. Casasnovas *et al.*, "FDG-PET–driven consolidation strategy in diffuse large B-cell lymphoma: final results of a randomized phase 2 study," *Blood*, vol. 130, no. 11, pp. 1315–1326, Sep. 2017, doi: 10.1182/blood-2017-02-766691.

[212] A.-S. Cottereau *et al.*, "Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†," *Annals of Oncology*, vol. 32, no. 3, pp. 404–411, Mar. 2021, doi: 10.1016/j.annonc.2020.11.019.

[213] R. Beare, B. Lowekamp, and Z. Yaniv, "Image Segmentation, Registration and Characterization in R with SimpleITK," *J Stat Softw*, vol. 86, p. 8, Aug. 2018, doi: 10.18637/jss.v086.i08.

[214] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research," *J Digit Imaging*, vol. 31, no. 3, pp. 290–303, Jun. 2018, doi: 10.1007/s10278-017-0037-8.

[215] B. Lowekamp, D. Chen, L. Ibanez, and D. Blezek, "The Design of SimpleITK," *Frontiers in Neuroinformatics*, vol. 7, 2013, Accessed: Oct. 16, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2013.00045

[216] C. Thieblemont *et al.*, "A tumor volume and performance status model to predict outcome before treatment in diffuse large B-cell lymphoma," *Blood Advances*, vol. 6, no. 23, pp. 5995–6004, Nov. 2022, doi: 10.1182/bloodadvances.2021006923.
[217] L. Vercellino *et al.*, "High total metabolic tumor volume at baseline allows discrimination of survival even in patients aged 60 to 80 years responding to R-CHOP.," Jan. 2020, doi: 10.1182/blood.2019003526.
[218] S. Hatzl *et al.*, "Prognostic Value of Baseline and Interim Positron Emission Tomography Markers in Diffuse Large B-cell Lymphoma Patients: A Real-world Perspective," *Hemasphere*, vol. 5, no. 8, p. e621, Jul. 2021, doi: 10.1097/HS9.0000000000000621.
[219] L. Guerra *et al.*, "The impact of spleen metabolic tumor volume on total metabolic tumor volume and prognosis in patients with follicular lymphoma enrolled in FOLL 12 trial," *Hematological Oncology*, vol. 41, no. S2, pp. 339–340, 2023, doi: 10.1002/hon.3164_242.
[220] S. Yamanaka *et al.*, "The prognostic significance of whole-body and spleen MTV (metabolic tumor volume) scanning for patients with diffuse large B cell lymphoma," *Int J Clin Oncol*, vol. 26, no. 1, pp. 225–232, Jan. 2021, doi: 10.1007/s10147-020-01807-6.
[221] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat, "Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis," *J Nucl Med*, vol. 55, no. 3, pp. 414–422, Mar. 2014, doi: 10.2967/jnumed.113.129858.
[222] M. A. Hernán, S. Hernández-Díaz, M. M. Werler, and A. A. Mitchell, "Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology," *American Journal of Epidemiology*, vol. 155, no. 2, pp. 176–184, Jan. 2002, doi: 10.1093/aje/155.2.176.
[223] S. Horvath, *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.
[224] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Nov. 06, 2023. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html
[225] S. Pölsterl, "scikit-survival: a library for time-to-event analysis built on top of scikit-learn," *J. Mach. Learn. Res.*, vol. 21, no. 1, p. 212:8747-212:8752, Jan. 2020.
[226] M. Sasanelli *et al.*, "Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma," *Eur J Nucl Med Mol Imaging*, vol. 41, no. 11, pp. 2017–2022, Nov. 2014, doi: 10.1007/s00259-014-2822-7.
[227] C. Casulo *et al.*, "Validation of POD24 as a robust early clinical end point of poor survival in FL from 5225 patients on 13 clinical trials," *Blood, The Journal of the American Society of Hematology*, vol. 139, no. 11, pp. 1684–1693, 2022.
[228] H. Li *et al.*, "Prediction of prognosis and pathologic grade in follicular lymphoma using 18F-FDG PET/CT," *Frontiers in Oncology*, vol. 12, p. 943151, 2022.

[229] R. Zhu *et al.*, "Assessment of correlation between early and late efficacy endpoints to identify potential surrogacy relationships in non-Hodgkin lymphoma: a literature-based meta-analysis of 108 phase II and phase III studies," *The AAPS journal*, vol. 19, pp. 669–681, 2017.

[230] L. Rebaud *et al.*, "ROBI: a Robust and Optimized Biomarker Identifier to increase the likelihood of discovering relevant radiomic features.," *LINK PREPRINT*.

[231] B. D. Cheson *et al.*, "Report of an international workshop to standardize response criteria for non-Hodgkin's lymphomas," *Journal of clinical oncology*, vol. 17, no. 4, pp. 1244–1244, 1999.

[232] L. Rebaud, T. Escobar, F. Khalid, K. Girum, and I. Buvat, "Simplicity Is All You Need: Out-of-the-Box nnUNet Followed by Binary-Weighted Radiomic Model for Segmentation and Outcome Prediction in Head and Neck PET/CT," in *Head and Neck Tumor Segmentation and Outcome Prediction*, V. Andrearczyk, V. Oreiller, M. Hatt, and A. Depeursinge, Eds., in Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 121–134. doi: 10.1007/978-3-031-27420-6_13.

[233] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Advances in neural information processing systems*, vol. 12, 1999.

[234] K. Pak *et al.*, "Prognostic value of metabolic tumor volume and total lesion glycolysis in head and neck cancer: a systematic review and meta-analysis," *Journal of Nuclear Medicine*, vol. 55, no. 6, pp. 884–890, 2014.

[235] S. Pellegrino *et al.*, "Total metabolic tumor volume by 18F-FDG PET/CT for the prediction of outcome in patients with non-small cell lung cancer," *Annals of Nuclear Medicine*, vol. 33, pp. 937–944, 2019.

[236] A. Cottereau *et al.*, "Prognostic value of baseline total metabolic tumor volume (TMTV0) measured on FDG-PET/CT in patients with peripheral T-cell lymphoma (PTCL)," *Annals of Oncology*, vol. 27, no. 4, pp. 719–724, 2016.

[237] L. Kostakoglu *et al.*, "Total metabolic tumor volume as a survival predictor for patients with diffuse large B-cell lymphoma in the GOYA study," *Haematologica*, vol. 107, no. 7, p. 1633, 2022.

[238] V. Oreiller *et al.*, "Head and neck tumor segmentation in PET/CT: the HECKTOR challenge," *Medical image analysis*, vol. 77, p. 102336, 2022.

[239] V. Andrearczyk *et al.*, "Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images," in *Head and Neck Tumor Segmentation and Outcome Prediction*, V. Andrearczyk, V. Oreiller, M. Hatt, and A. Depeursinge, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 1–37. doi: 10.1007/978-3-030-98253-9_1.

[240] Z. Guo, N. Guo, K. Gong, and Q. Li, "Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network," *Physics in Medicine & Biology*, vol. 64, no. 20, p. 205015, 2019.

[241] J. Ren, J. G. Eriksen, J. Nijkamp, and S. S. Korreman, "Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation," *Acta Oncologica*, vol. 60, no. 11, pp. 1399–1406, 2021.

[242] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.
[243] M. Antonelli *et al.*, "The medical segmentation decathlon. Nat Commun 13: 4128," 2022.
[244] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
[245] J. Xie and Y. Peng, "The head and neck tumor segmentation based on 3D U-Net," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, Springer, 2021, pp. 92–98.
[246] M. Vallières *et al.*, "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep 7: 10117," 2017.
[247] A. Diamant, A. Chatterjee, and M. Vallières, "Deep learning in head & neck cancer outcome prediction. Sci Rep 9," 2019.
[248] N. Saeed, R. Al Majzoub, I. Sobirov, and M. Yaqub, "An ensemble approach for patient prognosis of head and neck tumor using multimodal data," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, Springer, 2021, pp. 278–286.
[249] M. A. Naser *et al.*, "Progression free survival prediction for head and neck cancer using deep learning based on clinical and PET/CT imaging data," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, Springer, 2021, pp. 287–299.
[250] M. R. Salmanpour, G. Hajianfar, S. M. Rezaeijo, M. Ghaemi, and A. Rahmim, "Advanced automatic segmentation of tumors and survival prediction in head and neck cancer," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, Springer, 2021, pp. 202–210.
[251] H. J. Adams *et al.*, "Prognostic superiority of the N ational C omprehensive C ancer N etwork I nternational P rognostic I ndex over pretreatment whole-body volumetric–metabolic FDG-PET/CT metrics in diffuse large B-cell lymphoma," *European journal of haematology*, vol. 94, no. 6, pp. 532–539, 2015.
[252] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.
[253] V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, vol. 4, no. 1, p. 153, 2021.
[254] D. A. Glei, N. Goldman, Y.-H. Lin, and M. Weinstein, "Age-related Changes in Biomarkers: Longitudinal Data From a Population-based Sample," *Res Aging*, vol. 33, no. 3, pp. 312–326, May 2011, doi: 10.1177/0164027511399105.
[255] M. Meignan, A.-S. Cottereau, L. Specht, and N. G. Mikhaeel, "Total tumor burden in lymphoma – an evolving strong prognostic parameter," *BJR*, vol. 94, no. 1127, p. 20210448, Nov. 2021, doi: 10.1259/bjr.20210448.
[256] F. Orlhac *et al.*, "A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies," *Journal of Nuclear Medicine*, vol. 63, no. 2, pp. 172–179, Feb. 2022, doi: 10.2967/jnumed.121.262464.

[257] D. G. Altman, "Categorising continuous variables.," *Br J Cancer*, vol. 64, no. 5, p. 975, Nov. 1991, Accessed: Feb. 05, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1977443/

[258] F.-S. Ou, S. Michiels, Y. Shyr, A. A. Adjei, and A. L. Oberg, "Biomarker Discovery and Validation: Statistical Considerations," *Journal of Thoracic Oncology*, vol. 16, no. 4, pp. 537–545, Apr. 2021, doi: 10.1016/j.jtho.2021.01.1616.

[259] M.-Y. C. Polley and J. J. Dignam, "Statistical Considerations in the Evaluation of Continuous Biomarkers," *Journal of Nuclear Medicine*, vol. 62, no. 5, pp. 605–611, May 2021, doi: 10.2967/jnumed.120.251520.

[260] G. S. Halford, R. Baker, J. E. McCredden, and J. D. Bain, "How Many Variables Can Humans Process?," *Psychol Sci*, vol. 16, no. 1, pp. 70–76, Jan. 2005, doi: 10.1111/j.0956-7976.2005.00782.x.

[261] A. Antoranz, T. Sakellaropoulos, J. Saez-Rodriguez, and L. G. Alexopoulos, "Mechanism-based biomarker discovery," *Drug Discovery Today*, vol. 22, no. 8, pp. 1209–1215, Aug. 2017, doi: 10.1016/j.drudis.2017.04.013.

[262] B. T. Babalola and B. W. Yahya, "Effects of Collinearity on Cox Proportional Hazard Model with Time Dependent Coefficients: A Simulation Study," *Journal of Biostatistics and Epidemiology*, vol. 5, no. 2, pp. 172–182, 2019, doi: 10.18502/jbe.v5i2.2348.

[263] P. Wang, Y. Li, and C. K. Reddy, "Machine Learning for Survival Analysis: A Survey," *ACM Comput. Surv.*, vol. 51, no. 6, p. 110:1-110:36, Feb. 2019, doi: 10.1145/3214306.

[264] A. C. C. Coolen, J. E. Barrett, P. Paga, and C. J. Perez-Vicente, "Replica analysis of overfitting in regression models for time-to-event data," *J. Phys. A: Math. Theor.*, vol. 50, no. 37, p. 375001, Aug. 2017, doi: 10.1088/1751-8121/aa812f.

[265] E. Drysdale, "SurvSet: An open-source time-to-event dataset repository." arXiv, Mar. 06, 2022. doi: 10.48550/arXiv.2203.03094.

[266] J. Liu *et al.*, "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics," *Cell*, vol. 173, no. 2, pp. 400-416.e11, Apr. 2018, doi: 10.1016/j.cell.2018.02.052.

Supplemental

Supplemental of Chapter 6:

To facilitate the use of ROBI, we provide here some intuition and range of values for the parameters of the pipeline. The parameters can be grouped in three sets:

- Parameters to control for biomarker candidates reproducing known information: (S, M, W)
  These three parameters can help to adjust how we want the selected candidates to be different from already known biomarkers. The lower the values of these parameters, the lower the chance of having a selected candidate reproducing already quantified information. It is thus important to identify these known biomarkers and measure them in the cohort that is studied. On the other hand, this selection step can be completely ignored if we want to try to discover features that could replace known ones or if no biomarker relevant to the task of interest have already been identified.

    - **S**: This is the maximum absolute Spearman correlation, within [0,1], a candidate can have with a known feature. Default value is 0.5.
    - **M**: This is the maximum VIF score within [0, +∞] a candidate can have with known features. It helps to prevent candidates to encode multiple known biomarkers at the same time. Default value is 5. Is it commonly accepted that M ≤ 1 indicates no collinearity, M ≥ 2.5 shows high collinearity and M = +∞ means that the feature can be perfectly defined by the other covariates. Existing multicollinearity between known features must be taken into account.
    - **W**: it encodes how sensitive to other covariates the weight of the candidate in a multivariate model can be. If this weight changes a lot when other covariates are introduced, it means that the candidate is confounded by some of these covariates. A common value is 10%, so this weight should not change by more than 10% when known features are introduced in the model. Lower values can be used to be more restrictive, but values should not be higher than 10%. The range of W is therefore [0, 10].

- Parameters to increase the number of selected biomarkers: (C, Q)
  These two parameters can help to increase the number of selected biomarkers.
  - **C:** this parameter controls the CCO optimization. Candidates sharing similar information will be grouped in clusters based on their absolute Spearman correlation. C defines the maximum correlation between two candidates of different clusters. The higher the value, the lower the number of clusters. This parameter is helpful when many candidates reflect the same information. Instead of evaluating the same information 10 times through 10 different candidates, thus losing statistical power, only the candidate of the cluster with the highest prognostic value is used. Therefore, less candidates are evaluated in the subsequent TST step, making the TST less demanding on the p-values, thus increasing the number of selected candidates. It also helps capturing a higher number of distinct biological information: instead of selecting 10 candidates that all reflect the same feature, 10 candidates encoding 10 different features could be selected. Default values is 0.5 and C should be within [0, 1], 1 meaning that all candidates will be in a single cluster, and 0 that each cluster will only include one candidate (so no clustering is performed).
  - **Q:** Possible values of this parameter are within [0, 1]. The parameter controls how permissive selection is. Lower values will be more restrictive, lowering the number of false positives at the cost of less selected biomarkers, while higher values will accept more candidates, increasing the number of selected biomarkers and the number of false positives. This can be adjusted to different needs. By default, ROBI will perform the selection with 50 values between 0.01 and 0.5 and will report the number of selected candidates and estimated number of false positives for each value. We recommend choosing a Q value for which the number of false positives is acceptable for the study and using the corresponding selected candidates.

- Parameters to improve the quality of the evaluation: (P, T)
  These parameters are the easiest to set: the higher the better. The higher the values of P and T, the more precise the evaluation of the candidates, at the cost of increased computing time.

  - **P:** It defines the number of permutations performed to determine the significance of the prognostic values of the candidates. The p-value of a

candidate is calculated by measuring the proportion of permuted features with a C-index higher than the C-index of the candidate. To have a precise estimate, it is good to have at least one hundred permuted features with a C-index higher than the C-index of the best candidates. Default value if 10^3 but it can be greatly increased without increasing the computation time too much, especially with a GPU. As mentioned in the discussion, P=10^7 can be handled by a PC. ROBI package automatically recommends a value depending on the number of candidates.

- **T:** It defines how many permutations of the dataset will be performed to estimate the number of false positives. This parameter is the one with the highest impact on the computational cost. Default value is 10^3. A value of 10^4 can still be calculated in a few hours. It is not necessary to go beyond this value as the gain in precision will not be worth the longer computation time.

Supplemental of Chapter 7:

Selected biomarker	C-index	Sign of the correlation with the risk
duodenum_shape_Sphericity	0.58 (p < 0.01)	+
lesion_CT_firstorder_Kurtosis_range	0.57 (p < 0.01)	+
lesion_CT_glcm_Idn_range	0.59 (p < 0.01)	+
lesion_CT_glcm_Imc1_range	0.59 (p < 0.01)	+
lesion_CT_glrlm_ShortRunHighGrayLevelEmphasis_maxi	0.59 (p < 0.01)	+
lesion_PT_firstorder_10Percentile_mini	0.56 (p < 0.02)	-
lesion_PT_firstorder_Skewness_mini	0.56 (p < 0.01)	-
lesion_PT_glcm_InverseVariance_range	0.57 (p < 0.01)	+
lesion_PT_glrlm_RunEntropy_range	0.57 (p < 0.01)	+
liver_PT_firstorder_Skewness	0.58 (p < 0.01)	+
liver_shape_Maximum3DDiameter	0.57 (p < 0.01)	+
oneroi_shape_Flatness	0.58 (p < 0.01)	+
pancreas_PT_firstorder_Energy	0.57 (p < 0.01)	+
pancreas_PT_glcm_SumSquares	0.57 (p < 0.01)	+
pancreas_shape_Elongation	0.58 (p < 0.01)	+
shell_CT_firstorder_Mean_range	0.57 (p < 0.01)	+
shell_CT_firstorder_Median_mini	0.59 (p < 0.01)	-
shell_CT_firstorder_Minimum_range	0.56 (p < 0.01)	+
shell_CT_firstorder_Minimum_std	0.58 (p < 0.01)	+
shell_CT_gldm_DependenceEntropy_mini	0.60 (p < 0.01)	-
shell_PT_firstorder_10Percentile_mini	0.59 (p < 0.01)	-
shell_PT_firstorder_Energy_mini	0.58 (p < 0.01)	-
shell_PT_glcm_Correlation_mini	0.59 (p < 0.01)	-
shell_PT_gldm_DependenceNonUniformityNormalized_range	0.57 (p < 0.01)	+
shell_shape_Elongation_maxi	0.59 (p < 0.01)	+
small_bowel_shape_SurfaceVolumeRatio	0.58 (p < 0.01)	-
urinary_bladder_PT_glszm_GrayLevelNonUniformity	0.57 (p < 0.01)	+
volume_fat	0.58 (p < 0.01)	-

Table S14: The 28 candidates selected on the FL cohort. In the name of the biomarker, multiple terms are separated by an underscore. The first term describes in which region the biomarker was computed (lesion, organ, shell, …). The second describes the modality (PET, CT values or shape). The third one is the PyRadiomics name of the feature. The fourth if any, explains the aggregation method used to aggregate the lesion level biomarker to the patient level (e.g., minimum value across all lesions, maximum, standard-deviation of the values). The C-index for PFS prediction is reported with its p-value, as well as the sign of the correlation with the risk (PFS).

Selected biomarker	C-index	Sign of the correlation with the risk
colon_PT_glrlm_GrayLevelVariance	0.58 (p < 0.01)	+
colon_shortestDistanceToTumor	0.59 (p < 0.01)	-
esophagus_CT_gldm_SmallDependenceLowGrayLevelEmphasis	0.59 (p < 0.01)	-
esophagus_CT_glrlm_RunEntropy	0.58 (p < 0.01)	-
esophagus_PT_glcm_Imc2	0.60 (p < 0.01)	+
insidemuscle_PT_glrlm_GrayLevelVariance	0.58 (p < 0.01)	+
kidney_left_CT_glcm_ClusterShade	0.59 (p < 0.01)	+
lesion_CT_gldm_SmallDependenceHighGrayLevelEmphasis_maxi	0.59 (p < 0.01)	+
lesion_CT_glszm_SizeZoneNonUniformity_range	0.63 (p < 0.01)	+
lesion_CT_glszm_SmallAreaHighGrayLevelEmphasis_range	0.58 (p < 0.01)	+
lesion_PT_firstorder_RobustMeanAbsoluteDeviation_mini	0.60 (p < 0.01)	-
lesion_shape_Sphericity_maxi	0.58 (p < 0.01)	+
liver_shape_MajorAxisLength	0.58 (p < 0.01)	+
lung_left_shape_MinorAxisLength	0.57 (p < 0.02)	+
lung_right_CT_glcm_ClusterProminence	0.59 (p < 0.01)	+
lung_right_volTumorInside/vol_organ	0.58 (p < 0.02)	+
shell_CT_firstorder_10Percentile_maxi	0.58 (p < 0.01)	+
shell_CT_firstorder_Kurtosis_maxi	0.60 (p < 0.01)	+
shell_CT_firstorder_MeanAbsoluteDeviation_mini	0.60 (p < 0.01)	-
shell_CT_glcm_ClusterProminence_maxi	0.61 (p < 0.01)	+
shell_CT_glcm_ClusterShade_maxi	0.58 (p < 0.01)	+
shell_CT_gldm_LowGrayLevelEmphasis_mini	0.60 (p < 0.01)	-
shell_PT_firstorder_10Percentile_range	0.58 (p < 0.01)	+
shell_oneroi_PT_firstorder_Energy	0.58 (p < 0.01)	+
shell_oneroi_PT_glcm_Idmn	0.58 (p < 0.01)	+
stomach_CT_firstorder_RobustMeanAbsoluteDeviation	0.62 (p < 0.01)	-
trachea_volTumorInside	0.60 (p < 0.01)	+
urinary_bladder_gotTumor	0.58 (p < 0.01)	+

Table S15: The 28 candidates selected on the DLBCL cohort. In the name of the biomarker, multiple terms are separated by an underscore. The first term describes in which region the biomarker was computed (lesion, organ, shell, ...). The second describes the modality (PET, CT values or shape). The third one is the PyRadiomics name of the feature. The fourth, if any, explains the aggregation method used to aggregate the lesion level biomarker to the patient level (e.g., minimum value across all lesions, maximum, standard-deviation of the values). The C-index for PFS prediction is reported with its p-value, as well as the sign of the correlation with the risk (PFS).

Biomarker	Balanced accuracy (p-value)
lesion_CT_glcm_Idn_range	0.61 (p < 0.01)
lesion_CT_glcm_Imc1_range	0.59 (p < 0.01)
lesion_CT_glrlm_ShortRunHighGrayLevelEmphasis_maxi	0.58 (p < 0.03)
lesion_PT_glcm_InverseVariance_range	0.59 (p < 0.02)
pancreas_PT_firstorder_Energy	0.58 (p < 0.02)
pancreas_shape_Elongation	0.60 (p < 0.01)
shell_CT_firstorder_Median_mini	0.58 (p < 0.02)
shell_CT_gldm_DependenceEntropy_mini	0.62 (p < 0.01)
shell_PT_glcm_Correlation_mini	0.60 (p < 0.01)
shell_PT_gldm_DependenceNonUniformityNormalized_range	0.59 (p < 0.01)
shell_shape_Elongation_maxi	0.62 (p < 0.01)
small_bowel_shape_SurfaceVolumeRatio	0.58 (p < 0.02)
volume_fat	0.60 (p < 0.01)

Table S16: Biomarkers that significantly discriminated FL patients responding to treatment vs FL patients with progressive disease.

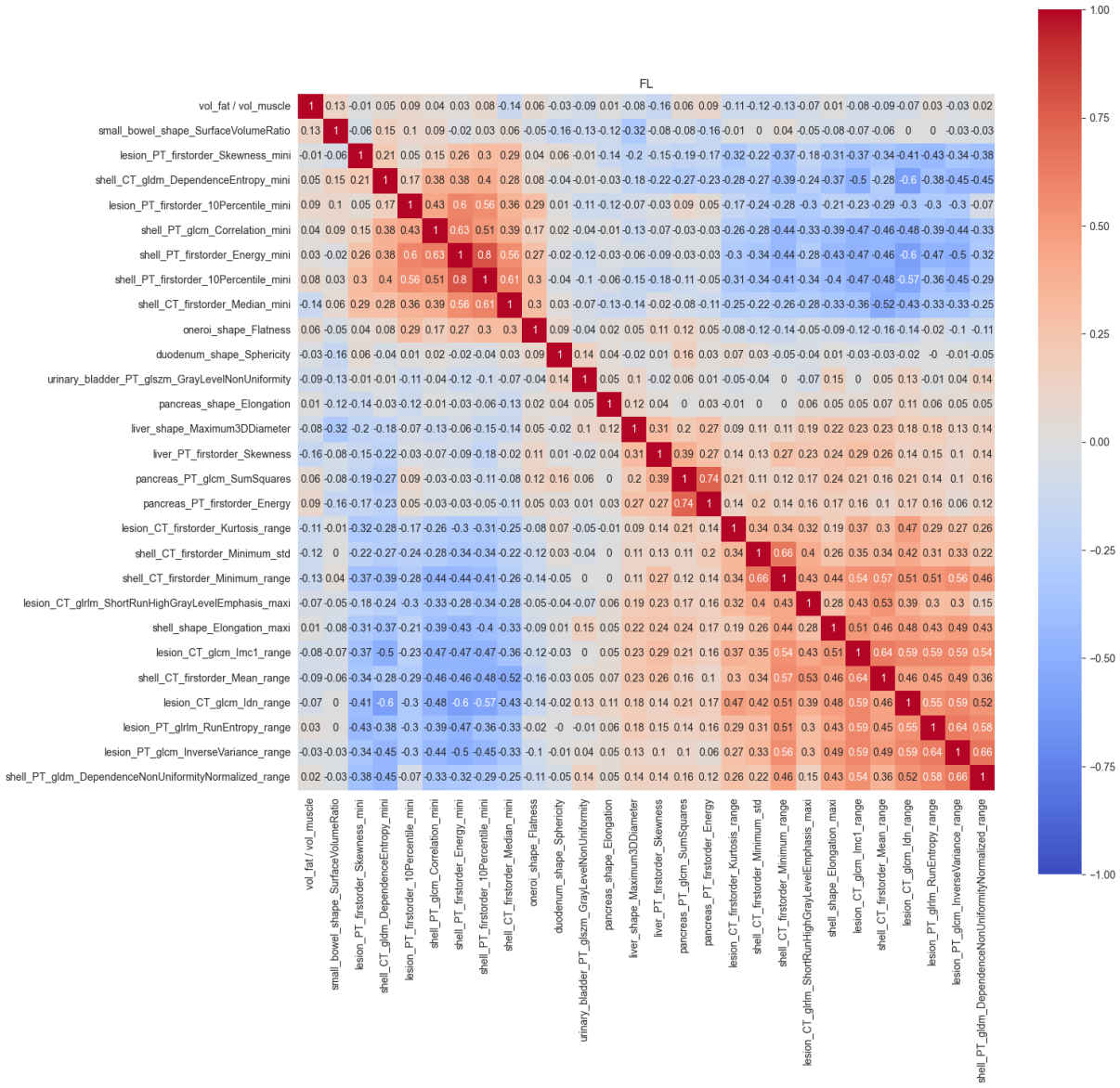


Figure S68: Correlogram of biomarkers selected on the FL cohort, based on their Spearman correlation on the FL cohort.

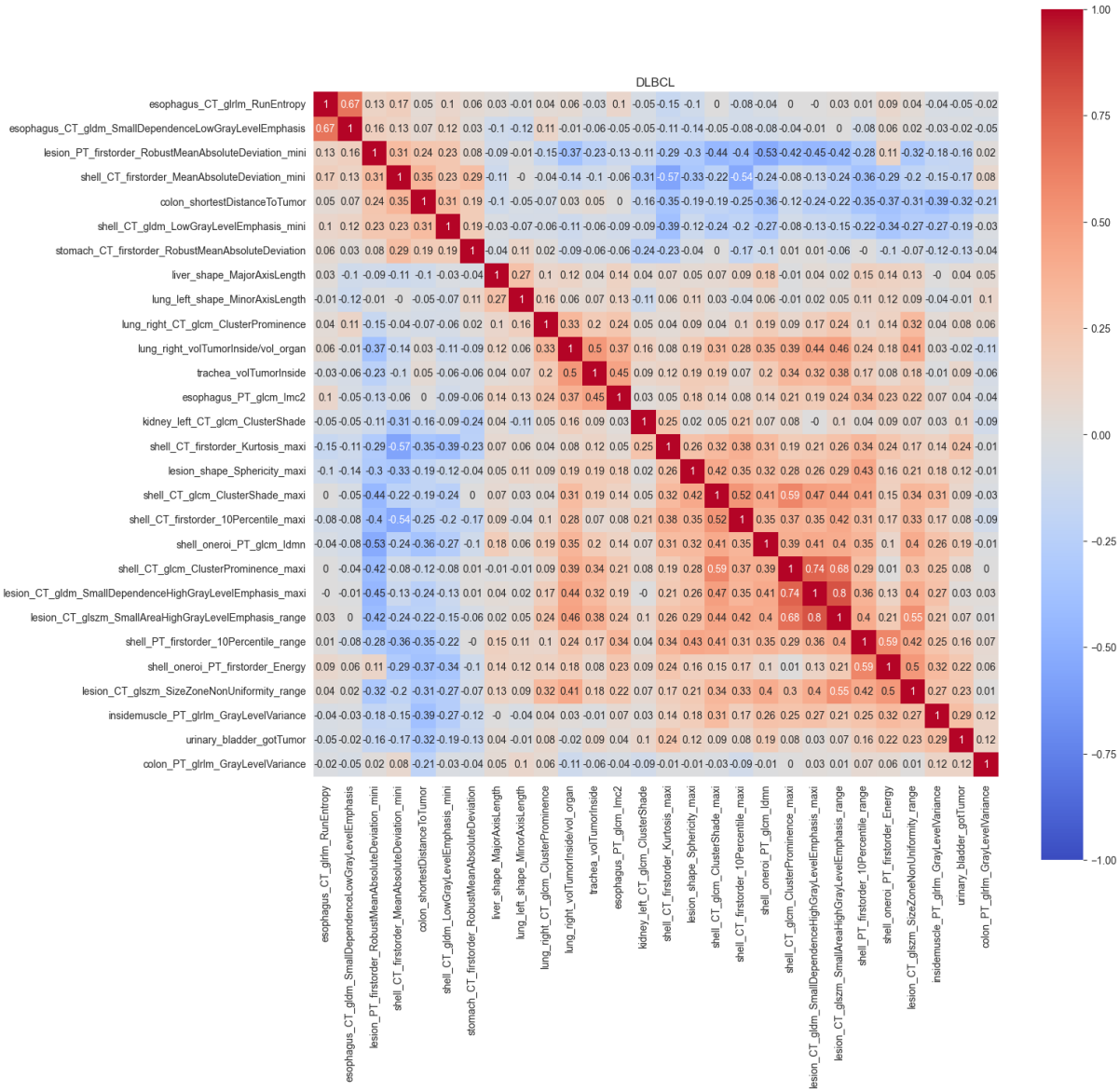


Figure S69: Correlogram of biomarkers selected on the DLBCL cohort, based on their Spearman correlation on the DLBCL cohort.

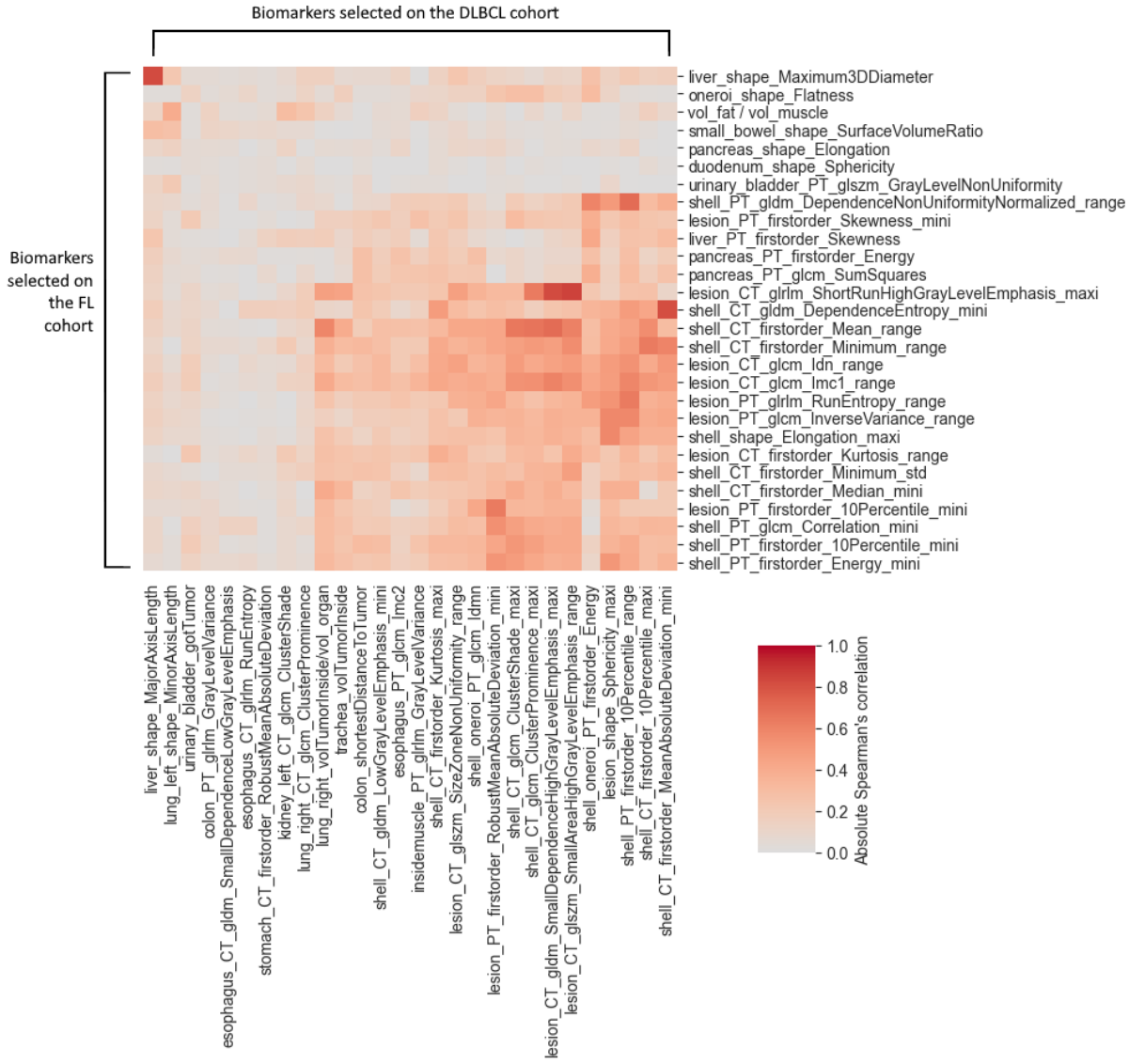


Figure S70: Correlogram of biomarkers selected on the DLBCL and FL cohorts, with their absolute Spearman correlation calculated with the two cohorts merged into one. A large fraction of the prognostic information identified is shared between the two cohorts.

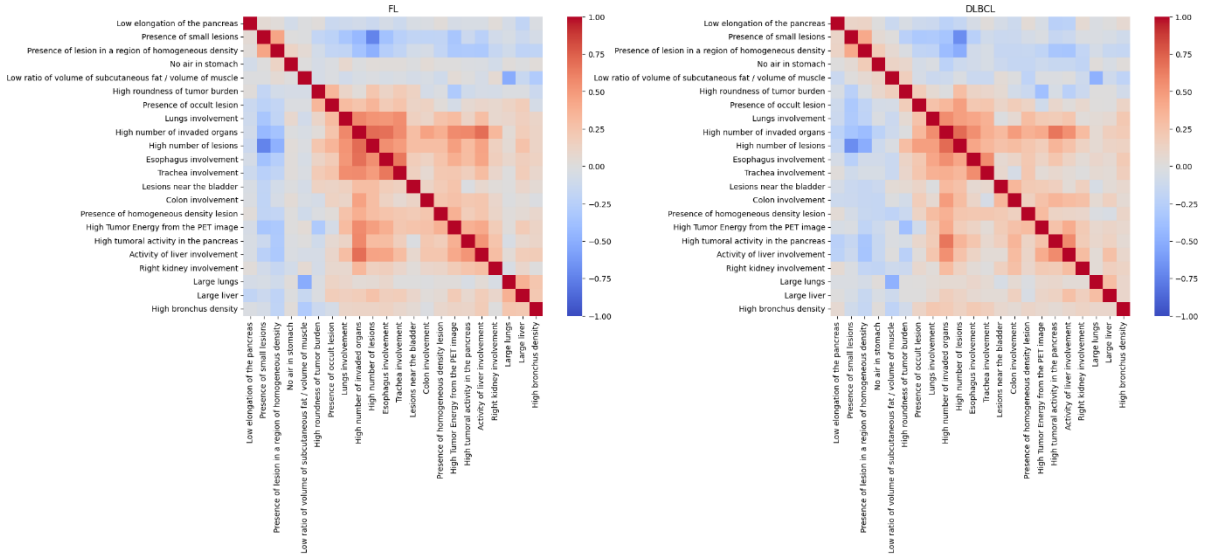


Figure S71: Correlogram of the surrogate biomarkers, based on their Spearman correlation on both the FL and DLBCL cohorts.

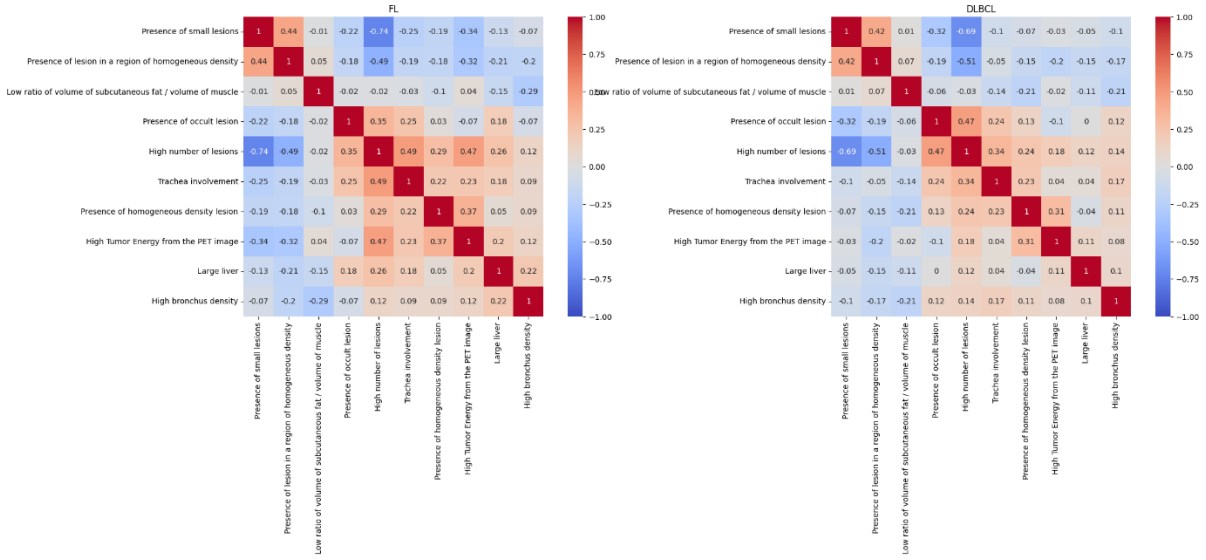


Figure S72: Correlogram of the 10 surrogate biomarkers prognostic on both FL and DLBCL cohorts, based on their Spearman correlation on both the FL and DLBCL cohorts.