



UNIVERSITAT DE
BARCELONA

Machine Learning for predicting Immunotherapy response from routine Medical Imaging

Marta Ligeró Hernández

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tdx.cat) i a través del Dipòsit Digital de la UB (diposit.ub.edu) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tdx.cat) y a través del Repositorio Digital de la UB (diposit.ub.edu) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tdx.cat) service and by the UB Digital Repository (diposit.ub.edu) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Machine Learning for predicting Immunotherapy response from routine Medical Imaging

Doctoral thesis dissertation presented by

Marta Ligeró Hernández

to apply for the degree of Doctor in the University of Barcelona

Directed by Raquel Perez-Lopez and Roser Sala-Llonch
Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO)
Department of Biomedicine, Faculty of Medicine and Health Sciences, University of
Barcelona
Doctoral Program in Biomedicine
Facultat de Medicina i Ciències de la Salut. Universitat de Barcelona
January 2024

Marta Ligeró Hernández
mligerhe@gmail.com
ORCID iD: 0000-0001-9824-7316
© Marta Ligeró Hernández 2024

FUNDING

The research presented in this thesis dissertation was financially supported by the Pre-doctoral fellowship PERIS 2020: SLT017/20/000080 — “*Precision Medicine and Artificial Intelligence for characterizing cancer response to immunotherapy*”.

The international secondment at Clinical AI - Kather Lab, EKFZ, Technical University of Dresden (Dresden, Germany) was founded by the BBVA Cancer Immunotherapy and Immunology (CAIMI) Secondment Internships Program by Fundació BBVA.

TABLE OF CONTENTS

FUNDING	ii
LIST OF FIGURES	vi
LIST OF APPENDICES	vii
LIST OF ABBREVIATIONS	viii
STUDIES	x

CHAPTER

1 Introduction	1
1.1 Motivation	1
1.2 Precision medicine and advanced cancer treatments	5
1.2.1 Cancer Immunotherapy	5
1.2.2 Biomarkers of response to Immune Checkpoints Inhibitors	5
1.2.3 Angiogenesis and its interaction with the immune system	8
1.3 The role of oncological imaging	9
1.3.1 Computed Tomography	9
1.3.2 Positron Emission Tomography	10
1.3.3 Magnetic Resonance Imaging	11
1.3.4 Advancing imaging through Radiomics: A quantitative approach	11
1.4 Development of radiomic signatures for outcome prediction	14
1.4.1 Standardization and harmonization initiatives	14
1.4.2 Dimensionality reduction methods	15
1.4.3 Machine learning models and Survival analysis for Radiomics	17
1.4.4 Feature aggregation methods and mixed effects models	19
1.4.5 The role of Deep Learning in overcoming Radiomics limitations	20
1.5 Integrating Radiology and Pathology: Advancing imaging biomarkers in oncology	23
1.5.1 Digital pathology imaging	23
1.5.2 Deep Learning methods for Computational Pathology	23
1.6 State-of-the-art imaging biomarkers for immunotherapy response	26

1.6.1	CT-based radiomic biomarkers for predicting response to immunotherapy	26
1.6.2	Computational pathology for predicting immunotherapy response	27
1.6.3	Multimodal models for predicting response to immunotherapy	29
2	Hypothesis and Objectives	31
2.1	Hypothesis	31
2.2	Objectives	32
3	General Materials and Methods	33
3.1	Data description	33
3.2	Imaging data acquisition protocol	37
3.3	Clinical data and outcome definition	37
3.4	Image pre-processing and feature extraction for CT and PET	38
3.5	Image pre-processing and feature extraction for IHC	38
3.6	Statistical analysis	39
4	Results	40
4.1	Study 1 - Minimizing acquisition-related radiomics variability by image re-sampling and batch effect correction to allow for large-scale data analysis	41
4.2	Study 2 - A CT-based radiomics signature is associated with response to Immune Checkpoint Inhibitors in advanced solid tumors	53
4.3	Study 3 - PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma	65
4.4	Study 4 - Radiomics and outcome prediction to anti-angiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumours: findings from the phase II TALENT trial	74
4.5	Study 5 - Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression	83
4.6	Supervisors statement	105
5	Summary of Results	106
6	Discussion	112
6.1	Challenges in the development of radiomics predictive signatures	113
6.1.1	Radiomics variability	113
6.1.2	Radiomic feature aggregation methods	113
6.1.3	Definition of dimensionality reduction and classifier selection of radiomics studies	115
6.2	The translational impact of radiomics and AI for cancer treatment	115
6.2.1	Development of predictive models across diverse tumor types	115
6.2.2	Endpoint definition for immunotherapy response predictive model	116
6.2.3	Multi-center studies to improve model generalizability	117

6.2.4	Integration of radiomics and clinical data	118
6.3	Biological validation of radiomic features	118
6.4	Limitations and future perspectives in the development of AI-based clinical biomarkers	119
7	Conclusion	121
	Bibliography	122
	 APPENDICES	 139

LIST OF FIGURES

FIGURE

1	Evolution of cancer therapy from one-size-fits-all to precision cancer therapy . .	2
2	Kaplan-Meier Overall Survival Analysis Comparing First-Line Immunotherapy to Chemotherapy	3
3	Mechanism of Immune Checkpoint Inhibitors	6
4	The Role of Medical Imaging in Oncology	9
5	Radiomic Feature Computation Workflow	13
6	End-to-End Radiomics Pipeline for Direct Prediction of Clinical Endpoints from CT Images	14
7	Conventional and Deep Radiomics Pipelines for Clinical Endpoint Prediction .	20
8	Convolutional Neural Network (CNN) Structure Applied to Radiology Images .	21
9	Computational Workflow from Whole-Slide Image (WSI) to Patient-Level Prediction	25
10	Comparative Workflow of Supervised Learning Studies for Predicting Immunotherapy Response from Clinical Image Data: Pathology and Radiology . .	29
11	Radiology and Pathology Visual Features in Image Data	30

LIST OF APPENDICES

A A whirl of radiomics-based biomarkers in cancer immunotherapy, why is large scale validation still lacking?	139
B Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer	164
C Enhancing immunotherapy response prediction via multimodal integration of radiology and pathology deep learning models	173

LIST OF ABBREVIATIONS

AI Artificial Intelligence	IHC Immunohistochemistry
ANG2 angiopoietin 2	ITH Intratumor Heterogeneity
CAR Chimeric Antigen Receptor	KNN K-Nearest Neighbors
CPS Combined Positive Score	LASSO Least Absolute Shrinkage and Selection Operator
CNN Convolutional Neural Network	ML machine learning
CR Complete Response	MMRd mismatch repair deficiency
CT Computed Tomography	MSI Microsatellite instability
DL Deep Learning	mRMR minimum Redundancy Maximum Relevance
DICOM Digital Imaging and Communications in Medicine	MRI Magnetic Resonance Imaging
FDA Food and Drug Administration	NGTDM Neighboring Gray Level Dependence Matrix
FDG 18F-fluorodeoxyglucose	NSCLC Non-Small Cell Lung Cancer
GLCM Gray Level Co-occurrence Matrix	OS Overall Survival
GLDM Gray Level Dependence Matrix	PCA Principal Component Analysis
GLRLM Gray Level Run Length Matrix	PD-1/PD-L1 Programed Death Ligand 1
GLSZM Gray Level Size Zone Matrix	PET Positron Emission Tomography
HE Hematoxylin and Eosin	PFS Progression Free Survival
HR hazard ratio	PR Partial Response
HU Hounsfield Units	RECIST Response Evaluation Criteria in Solid Tumours
IBSI Image Biomarker Standardization Initiative	ROIs regions of interest
ICC Interclass Correlation Coefficient	RQS Radiomics Quality Score
ICI Immune checkpoint inhibitors	SD Stable Disease

SUV_{max} the maximum standardized uptake value	TPS Tumor Proportion Score
SVM Support Vector Machine	TMTV total metabolic tumor volume
TCGA The Cancer Genome Atlas	US Ultrasounds
TIL Tumour-infiltrating lymphocytes	VEGF vascular endothelial growth factor
TMB Tumor Mutational Burden	WSI Whole Slide Image

STUDIES

Thesis in compendium of publications format. The thesis consists of four objectives and five articles.

Marta Ligeró, Olivia Jordi-Ollero, Kinga Bernatowicz, Alonso Garcia-Ruiz, Eric Delgado-Muñoz, David Leiva, Richard Mast, Cristina Suarez, Roser Sala-Llonch, Nahum Calvo, Manuel Escobar, Arturo Navarro-Martin, Guillermo Villacampa, Rodrigo Dienstmann, Raquel Perez-Lopez. *Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis.* European Radiology (2021)

Impact Factor (quartile): 7.13 (Q1)

Marta Ligeró, Alonso Garcia-Ruiz, Cristina Viaplana, Guillermo Villacampa, Maria V. Raciti, Jaid Landa, Ignacio Matos, Juan Martin-Liberal, Maria Ochoa-de-Olza, Cinta Hierro, Joaquin Mateo, Macarena Gonzalez, Rafael Morales-Barrera, Cristina Suarez, Jordi Rodon, Elena Elez, Irene Braña, Eva Muñoz-Couselo, Ana Oaknin, Roberta Fasani, Paolo Nuciforo, Debora Gil, Carlota Rubio-Perez, Joan Seoane, Enriqueta Felip, Manuel Escobar, Josep Tabernero, Joan Carles, Rodrigo Dienstmann, Elena Garraalda, Raquel Perez-Lopez. *A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors.* Radiology (2021)

Impact Factor (quartile): 29.146 (Q1)

Marta Ligeró, Marc Simó, Cecilia Carpio, Gloria Iacoboni, Maria Balaguer, Victor Navarro, Mario Sanchez, Sabela Bobillo, Ana Marín-Niebla, Josu Iraola, Pau Abrisqueta, Roser Sala, Francesc Bosch, Raquel Perez-Lopez, Pere Barba. *PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma.* Electronic Journal of Haematology (eJHaem) (2023).

Marta Ligeró, Jorge Hernando, Eric Delgado, Alonso Garcia-Ruiz, Xavier Merino-Casabiel, Toni Ibrahim, Nicola Fazio, Carlos Lopez, Alexandre Teulé, Juan W. Valle, Salvatore Tafuto, Ana Custodio, Nicholas Reed, Markus Raderer, Enrique Grande, Rocio Garcia-Carbonero, Paula Jimenez-Fonseca, Alejandro Garcia-Alvarez, Manuel Escobar, Oriol Casanovas, Jaume Capdevila and Raquel Perez-Lopez. *Radiomics and outcome prediction to antiangiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumours: findings from the phase II TALENT trial*. BJC Reports (2023)

Marta Ligeró, Garazi Serna, Omar S.M. El Nahhas, Irene Sansano, Siarhei Mauchanski, Cristina Viaplana, Julien Calderaro, Rodrigo Toledo, Rodrigo Dienstmann, Rami S. Vanguri, Jennifer L. Sauter, Francisco Sanchez-Vega, Sohrab P. Shah, Santiago Ramón y Cajal, Elena Garralda, Paolo Nuciforo, Raquel Perez-Lopez, Jakob Nikolas Kather. *Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression*. Cancer Research Communications (2023)

CHAPTER 1

Introduction

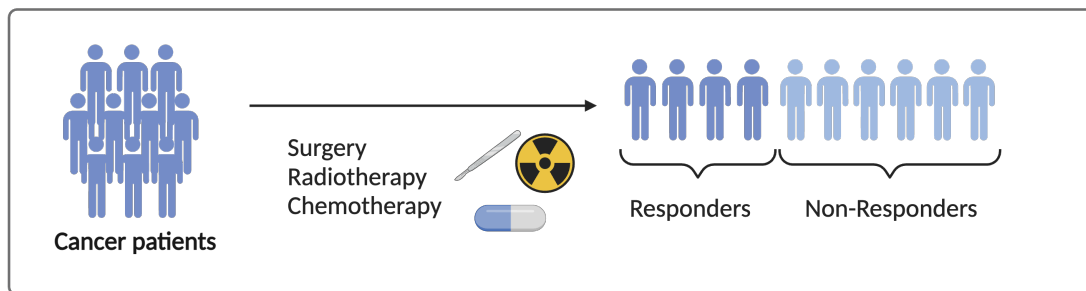
1.1 Motivation

The field of oncology has witnessed a significant transformation over the years, shifting from traditional one-size-fits-all treatments such as surgery, chemotherapy, and radiotherapy to more personalized and targeted therapies, including immunotherapy (**Figure 1**) [1]. Targeted therapies have brought about promising advancements in oncology, significantly improving long-term survival rates across various tumor types, even in the most aggressive cases. Illustrative examples include advanced stage melanoma (**Figure 2A**) [2], Non-Small Cell Lung Cancer (NSCLC) (**Figure 2B**) [3], and high microsatellite instability (MSI) colorectal cancer tumors (**Figure 2C**) [4]. Before the advent of targeted therapies, these tumor types were associated with aggressive behavior and had limited treatment options, leading to poor prognoses. However, the development of targeted therapies, such as immune checkpoint inhibitors (ICIs), have significantly improved the response rates, extended survival [5], and even led to remarkable responses in some cases, exemplifying the transformative potential of personalized medicine [6–8].

Despite this success, personalized medicine still faces challenges in accurately associating tumor characterization with treatment response patterns. The complexity of cancer biology, tumor heterogeneity and the impact of tumor microenvironment contribute to significant variability in patient responses to treatment [9–13]. As a result, there is a need to find and develop better predictive biomarkers for treatment response.

Although different medical fields, such as genomics, pathology and radiology, have contributed to improve tumor characterization, the ability to accurately predict treatment outcomes remains limited [14–17]. Addressing these challenges requires comprehensive efforts to gather and analyze existing evidence from clinical trials and real-world data. By doing so, we could gain valuable insights into treatment outcomes across diverse patient populations, ultimately improving patient stratification, maximizing treatment efficacy, and minimizing

Conventional therapy



Precision cancer therapy

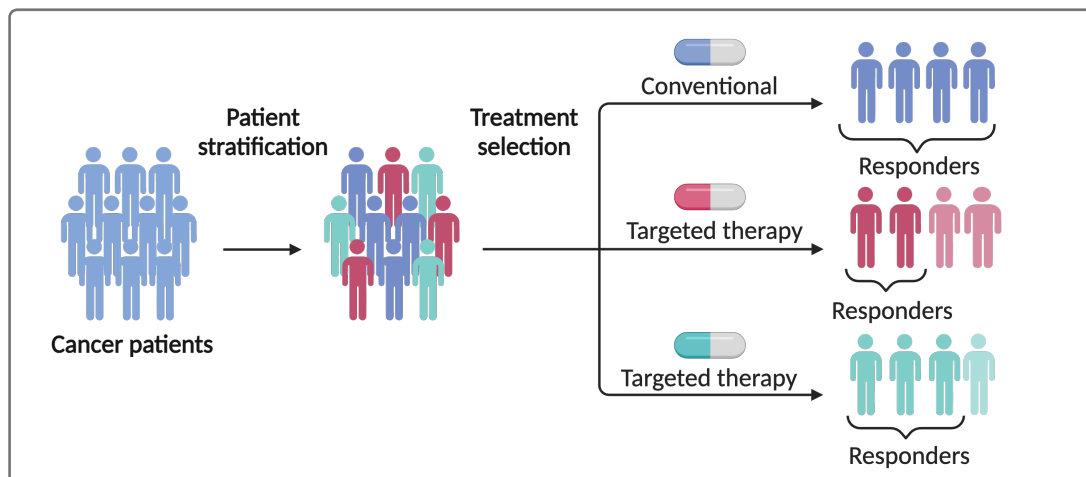


Figure 1: Evolution of cancer therapy from one-size-fits-all to precision cancer therapy.

This schematic representation highlights the transition from conventional, one-size-fits-all cancer therapies (such as surgery, radiotherapy, and chemotherapy) to the era of precision cancer therapy. Advances in precision medicine facilitated the identification of individualized biomarkers associated with the tumor phenotype, enabling more personalized and effective treatments. Ultimately enhancing treatment responses and improving the overall survival of cancer patients. The figure has been created with using Biorender.

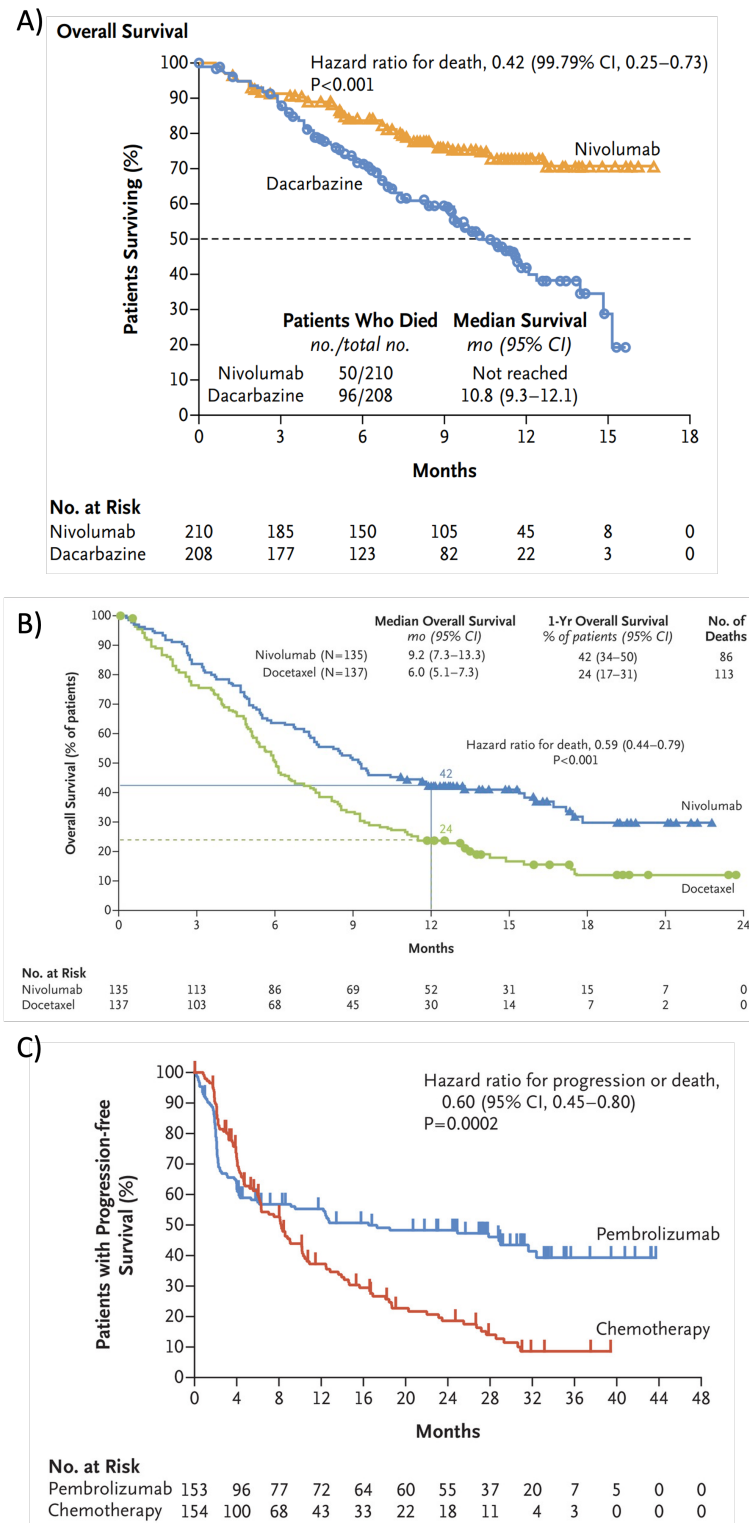


Figure 2: Kaplan-Meier Overall Survival Analysis Comparing First-Line Immunotherapy to Chemotherapy. Kaplan-Meier curves derived from clinical trial results depict the efficacy of first-line immunotherapy, specifically Nivolumab and Pembrolizumab, in comparison to chemotherapy. The data highlights significantly improved survival outcomes in melanoma (A)[2], non-small cell lung cancer (B) [3] and renal cancer (C)[4]

potential harm for those who may not benefit from certain therapies [18].

In this context, image processing and Artificial Intelligence (AI) emerge as promising tools for biomarker development [18, 19]. Medical imaging offers non-invasive insights into tumor biology and disease stage. While visual evaluation of images has been a common practice, by implementing image quantification analysis and machine learning methods on routine medical images, we could uncover patterns from large amounts of data that are not captured by the human eye. This approach has the potential to enhance our understanding of tumor biology and provide biomarker candidates for cancer treatment.

This thesis aims to advance the field of predictive biomarker development, with a primary focus on immunotherapy, combining the use of AI and quantitative representation of medical images to comprehensively characterize tumors. While the main emphasis lies in Computed Tomography (CT) imaging, the thesis also explores the feasibility of using other imaging modalities like Positron Emission Tomography (PET) and histopathology images for biomarker development. Furthermore, the thesis research also explores the potential applicability and generalizability of the proposed methods for creating biomarkers for different targeted therapies, such as Chimeric Antigen Receptor (CAR) T-cells therapy and anti-angiogenic treatments.

1.2 Precision medicine and advanced cancer treatments

1.2.1 Cancer Immunotherapy

Cancer immunotherapy, particularly immune checkpoint inhibitors (ICI), has emerged as the gold standard for treating various cancers, specifically for lung, renal, and melanoma [3, 8, 20, 21]. These therapies aim to unleash an immune response against tumor cells by leveraging the power of the body’s immune system. Concretely, ICIs block specific immune checkpoints and remove the brakes on the immune system, allowing it to recognize and target cancer cells more effectively (**Figure 3**). The notable success achieved with ICI has raised optimism for its potential application in treating numerous types of cancer. Currently, six ICIs have already been approved by the Food and Drug Administration (FDA) and became standard of care in clinical practice for some tumor types such as melanoma and lung cancer [3, 21, 22]. However, only a low percentage of patients (15-25%) respond adequately to treatment [23, 24]. The variability in patient responses has brought about a need to identify biomarkers capable of predicting individual responses to ICI [25].

1.2.2 Biomarkers of response to Immune Checkpoints Inhibitors

In response to the identified need, in recent years, several molecular and cellular biomarkers have been investigated to identify tumor properties that might influence the efficacy of immunotherapy. Highlighted below are some of the most studied biomarkers as reviewed by Havel J et al. [26].

Tumor Mutation Burden (TMB), which quantifies the total number of mutations in cancer genomes, has been described as a potential indicator of ICIs response [27]. The relevance of TMB as a biomarker could be linked to the fact that immunotherapy aims to reinvigorate antitumor immune responses, which are mainly targeted by tumor-specific neo-antigen peptides that arise from somatic mutations in cancer genomes. High TMB leads to increased production of neoantigens, which can attract a stronger immune response when targeted by ICIs. High TMB has proved significant associations with ICI response in different tumor types [28–30] and has been approved by FDA as treatment selection biomarker in specific cases [31].

Notably, some specific cancers such as melanoma and NSCLC have exhibited more favorable responses to ICIs compared to other tumor types. This difference in response rates could be attributed to the underlying nature of these cancers, often driven by carcinogens like ultraviolet light and cigarette smoke, resulting in a high mutation burden within the

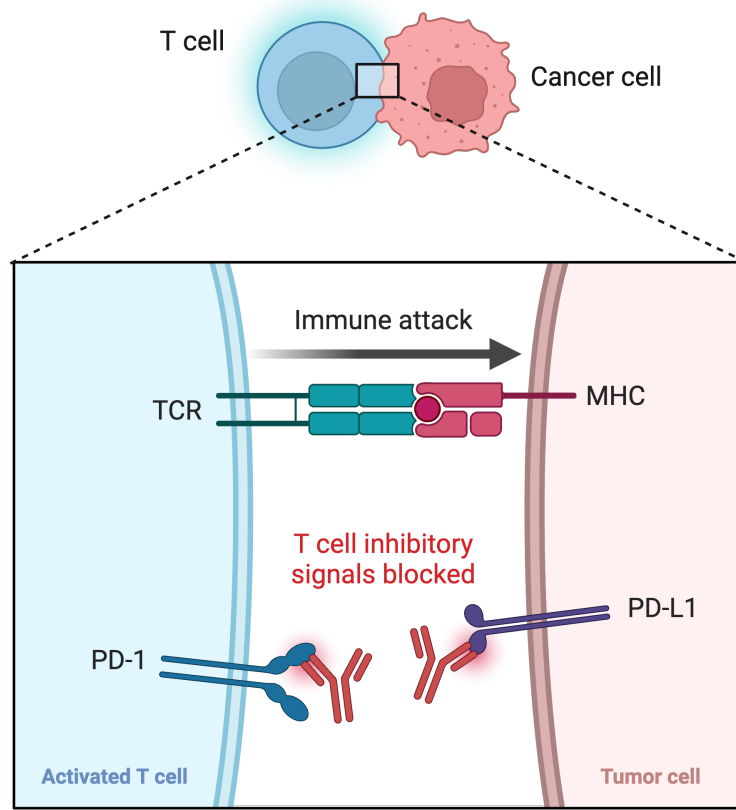


Figure 3: Mechanism of Immune Checkpoint Inhibitors. This schematic representation illustrates the interaction between a tumor cell and a T-cell lymphocyte, highlighting the immune checkpoint mechanism. When the T-cell receptor (TCR) binds to the antigen on cancer cells, T-cell activation occurs, leading to an immune attack and induction of cell death. However, inhibition can occur if the PD-1 receptor binds to the PD-L1 receptor on the cancer cell, preventing T-cell-induced cell death. Immune checkpoint inhibitors can counteract this inhibition by binding to either PD-L1 or PD-1, thereby blocking the inhibitory signals and restoring T-cell activity. *The figure was created using Biorender.*

tumor cells [32, 33]. However, other studies have revealed that TMB alone might not clearly discriminate between responders and non-responders [34]. Therefore, other factors likely play a role in determining the response patterns to immunotherapy.

Due to TMB limitation for predicting ICIs response, more specific quantification of the number of mutations targeted specifically by T cells, known as true neoantigen burden, was explored. For neoantigens to be effective and trigger immune response, they must be present in a majority of tumor cells [35]. Consequently, high neoantigen **Intratumoral Heterogeneity (ITH)** may be associated with poor responses to immunotherapy [36, 37]. It has been shown that homogeneous tumors —tumors with consistent neoantigens from shared mutations (clonal mutations)— may respond better to ICIs than heterogeneous tu-

mors—with diverse neoantigens from different mutations (subclonal mutations) [35]. The ability of heterogeneous tumors to host multiple immune evasion mechanisms could explain their associations with poor ICI outcomes.

Additionally, tumors with **mismatch repair deficiency (MMRd)** exhibiting **microsatellite instability (MSI)** have shown high sensitivity to ICIs therapy regardless of the tumor origin [38, 39]. These findings support the association between TMB and ICIs response, as high MSI tumors have a hypermutated phenotype, resulting in the generation of numerous neopeptides that will act as a target for the immune system [40].

Other factors beyond mutations and tumor heterogeneity might play a key role in determining response patterns to ICIs. Some of the most studied biomarkers are **tumor immune microenvironment** related biomarkers.

Program Death Ligand 1 (PD-L1) expression is one of the most established biomarkers for predicting response to ICIs [29, 41, 42], and has received FDA approval as a companion diagnostic tool for anti-PD1 therapy for patients with NSCLC [43, 44]. However, the use of PD-L1 as a predictive biomarker remains suboptimal [45], and there are still challenges in defining specific thresholds for response prediction [46]. These challenges may arise due to the lack of standardization and variability in the measurement techniques used to assess PD-L1 expression [47, 48]. Additionally, it remains unclear whether the relevance of PD-L1 expression lies in the tumor cells themselves or in the tumor-infiltrating immune cells [49]. Taken together, these factors further complicate the use of PD-L1 expression as a standalone predictor for ICIs response.

On the other hand, considering the immune-enhancing properties of immunotherapy, **tumour-infiltrating lymphocytes (TILs)** such as CD8+ and CD4+ memory T cells have been strongly defined as biomarkers of ICIs response, specially in colorectal and melanoma tumor types [50–54]. The presence and location of immune cells within the tumor immune microenvironment can categorize tumors as immune-inflamed, immune-excluded and immune-desert [55, 56], which is directly related to response to anti-PD1/PDL1 treatment.

In conclusion, the complexity of tumor biology and immune-related processes makes it challenging to identify comprehensive standalone biomarkers for predicting response to immunotherapy. Combining multiple biomarkers has shown more promising results in response prediction [14, 57], but there is still potential for improvement. Integrating medical imaging and AI with the established biomarkers, offers an opportunity to enhance treatment outcomes.

1.2.3 Angiogenesis and its interaction with the immune system

Another hallmark of cancer is the potential of tumors to induce the formation of new blood vessels from pre-existing endothelial cells, known as angiogenesis [58]. The progression of tumors is supported by the recruitment of blood vessels. These vessels provide the tumor with oxygen, nutrients, and growth factors that can stimulate growth of adjacent tumor cells and sustains metastatic dissemination of the tumor cells. Furthermore, tumor-associated neovasculature is abnormal in both morphology and functionality, compared to normal vasculature. These abnormalities restrict the entry of cytotoxic drugs and immune cells and can promote an abnormal tumor microenvironment that might modulate the effectiveness of other drugs such as immunotherapy [59, 60]. That explains that hypoxia is one of the mechanisms of resistance of the tumor towards immunotherapy treatment.

The main molecular drivers of tumor neovasculature are proangiogenic factors such as vascular endothelial growth factor (VEGF) and angiopoietin 2 (ANG2). Therefore, treatment strategies targeting these molecules have been developed in the last decade, also known as anti-angiogenic treatment [61]. These treatments have shown encouraging results in well-vascularized tumors types, including but not exclusive to clear cell renal carcinoma [62, 63], ovarian [64, 65] and cervical cancer [Tewari2014-mp], hepatocellular carcinoma [66, 67] and neuroendocrine tumors [68, 69]. These tumors are by biology highly vascularized (renal and ovarian cancer) or are characterized by vascular enhancement on imaging studies (liver and endocrine tumors).

Additionally, the adoption of combined therapies, anti-angiogenics and immunotherapy, has been increasing in the last years, particularly in the management of hepatocarcinoma [70]. Evidence indicates that this combined approach enhances effective immune infiltration in various cancer types, including colorectal cancer, renal cell carcinoma, and non-small cell lung cancer [71–73]. This is due to the effect of tumor vasculature normalization derived from anti-angiogenic treatment, which can modify the protumoral/antitumoral immune balance, and reactivate and favor the intratumoral trafficking of cytotoxic T cells.

1.3 The role of oncological imaging

Medical imaging plays a key role in the oncology field regarding tumor detection, characterization, and monitoring (Figure 4). Radiology imaging captures **whole-body 3D information** in a single **non-invasive measurement**, enabling **repeated examinations** along the course of the disease [74]. Imaging can also provide **biological tissue properties** like cell density, metabolism, vascularity. Commonly used imaging modalities in clinical practice include Ultrasounds (US), Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET).

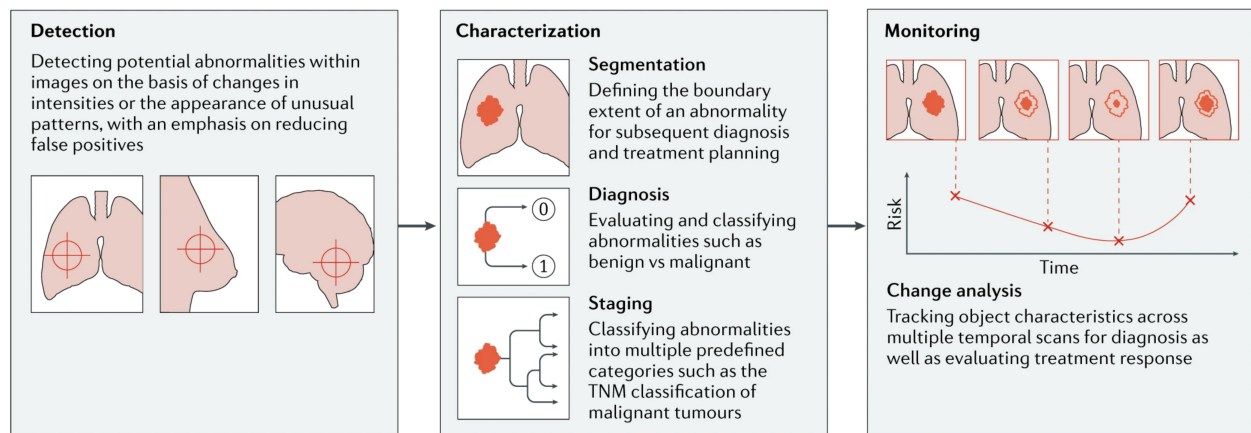


Figure 4: The Role of Medical Imaging in Oncology. This figure provides a comprehensive overview of the primary applications of radiology in cancer treatment. Initially, radiology images serve the crucial function of detecting abnormalities in the body, which may be indicative of malignancies. Subsequently, the detected abnormality undergoes characterization through segmentation, allowing for the diagnosis of its nature as benign or malignant and facilitating the staging of the disease. Throughout the course of patient treatment, medical imaging plays a pivotal role in evaluating disease progression, identifying responsive patients, and aiding in clinical decision-making for determining the next steps in patient care. Figure adapted from Hosny et al., "Artificial Intelligence in Radiology" *Nat Rev Cancer* (2018) [74].

1.3.1 Computed Tomography

Computed Tomography (CT) is the preferred imaging modality in oncology, which provides an anatomical representation of the whole disease based on density and contrast. CT imaging represents the density of the body tissue through the absorption of X-ray radiation and it is measured using the Hounsfield Units (HU). The CT scanner emits narrow beams of X-rays while rotating around the patient. After crossing the different body tissues, the attenuated X-rays are captured by detectors, also known as scintillation detectors. As in any tomographic imaging technique, the resulting data is stored in the form of a sinogram,

a graphical representation of the raw information in the angular domain. To reconstruct the image into a three-dimensional space, advanced mathematical algorithms are applied to the sinogram data. This involves sophisticated techniques like filtered back projection or iterative reconstruction [75].

As CT imaging is widely available and cost-effective in clinical practice, it is the main tool for evaluating disease evolution. The globally recognized standard for assessing treatment response is the **Response Evaluation Criteria in Solid Tumours (RECIST) 1.1** guidelines [76, 77], primarily based on CT scans. RECIST aims to standardize the tumor response assessment in solid tumors. In RECIST disease evaluation, the radiologist visually analyzes the lesions, reports their presence or absence, and identifies up to 5 target lesions that will be followed up during treatment. The target lesions need to be measurable lesions ($> 1\text{cm}$ in the largest diameter) and a maximum of 2 lesions per organ can be defined as target lesions. The rest of the lesions will be considered non-target lesions and also followed up. In every evaluation, the largest diameter of the target lesions will be computed and compared with the previous CT scan. The change in the sum of the largest diameter measured is then used to classify patients in the following four categories: Complete Response, CR (disappearance of all lesions) or Partial Response, PR ($\leq 30\%$ decrease), Stable Disease, SD (not significant growth or shrinkage of the target lesions), or Progressive Disease (PD) ($\geq 20\%$ increase).

However, an arguable weakness of CT imaging is its limitation to anatomical information and narrows tumor characterization to tumor size, density, and shape. For that reason, other imaging modalities are being introduced as baseline images due to their ability to provide more biological properties of the tumor, such as PET and MR images.

1.3.2 Positron Emission Tomography

Positron Emission Tomography (PET) is an imaging technique that relies on the simultaneous detection of photons produced through the annihilation of positrons. In order to capture signals from the body, PET imaging utilizes radioactive substances, such as ^{18}F , which emit positrons. These radioactive components can be combined with other molecules, like FDG (Fluorodeoxyglucose), allowing them to be absorbed by the body. This process enables the visualization of metabolic activity and distribution of the radiolabeled substance within the body, providing valuable insights for diagnostic and research purposes [78].

As a result, **^{18}F -fluorodeoxyglucose (FDG) PET** imaging also plays a pivotal role in cancer diagnosis, staging and treatment monitoring, especially for tumor types like Lymphoma where conventional imaging may fail to effectively capture the whole disease. By

using FDG as a radiotracer, this technique enables the assessment of glucose metabolism in tissues, aiming to identify higher metabolic rates in cancerous cells. Notably, FDG-PET provides quantitative imaging parameters, such as total metabolic tumor volume (TMTV) and the maximum standardized uptake value (SUVmax), which have shown associations with treatment outcomes and prognosis [79–81]. However, the focus of PET characteristics on tumor burden and the metabolic consumption of the tumor, along with the technique’s heterogeneity [82] and radiation exposure, may limit its potential as a predictive biomarker.

1.3.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging is another medical imaging modality widely used in clinical practice. MRI has enormous potential to surpass CT imaging due to its multi-sequence properties and non-ionizing nature. Furthermore, MRI provides not only detailed anatomical representation but also valuable biological insights. Some examples are perfusion MRI for vascularity information and diffusion MRI for cellularity assessment, among others. However, unlike PET imaging, MRI only furnishes indirect measures of tumor metabolism. The main drawback of MRI, particularly for AI and clinical applications, lies in the acquisition time and cost. Acquiring MRI images is more time-consuming and expensive than CT, making their use less common and limiting the development of MRI-based AI predictive biomarkers from routine clinical large databases.

1.3.4 Advancing imaging through Radiomics: A quantitative approach

In recent years, medical imaging in oncology has undergone a transformative shift. With the advent of big data and AI, its potential has evolved beyond patient diagnosis and tumor evaluation towards the quantification of CT and PET for the development of imaging biomarker. By analyzing medical imaging data, more comprehensive information of tumor properties can be obtained. Additionally, the implementation of machine learning to identify data patterns can facilitate the association between imaging phenotypes and treatment response.

Radiomics was a groundbreaking approach in medical imaging that went beyond the traditional qualitative characterization of tumor phenotypes in radiology, also known as semantic features. Radiomics introduced a quantitative approach for evaluating radiological images, employing handcrafted features to describe tumor properties related to tumor shape, intensity, and texture **Figure 5A**. First introduced for oncological imaging by Aerts et al.

in 2014 [83] Radiomics has since rapidly expanded its applications over the past decades, transforming medical image analysis for precision oncology [84].

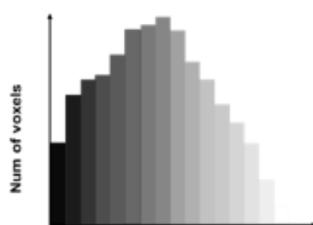
Radiomics converts imaging data into a high-dimensional, mineable feature space through automated data-characterization algorithms. It derives quantitative information, known as radiomic features, pertaining to the tumoral regions of interest (ROIs). Radiomic features are computed using handcrafted algorithms from texture matrices [85, 86]. When applying texture quantification to medical imaging data, images are treated as 3D matrices of voxel values, the HU in the case of CT. From those matrices, the statistical and spatial distribution of the HU values inside the ROIs are computed [87].

One widely used texture matrix is the **Gray Level Co-occurrence Matrix (GLCM)**, introduced by Haralick et al. in 1973 [88]. GLCM computes the frequency at which a voxel with a particular value occurs adjacent to another voxel with a different value in 13 different angles. From the GLCM, various textural features like homogeneity, variance, contrast, energy and entropy, can be derived, allowing for the quantification of the image heterogeneity levels. **Figure 5B** shows a visual representation of the computation of GLCM.

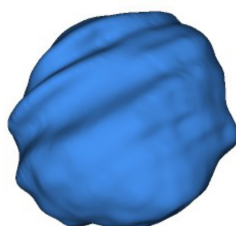
Beyond GLCM, several other texture matrices have been defined to explore different ways of computing the distribution of pixel pairs in an image. For instance, the **Gray Level Run Length Matrix (GLRLM)** and **Gray Level Size Zone Matrix (GLSZM)** focus on quantifying the amount of connected pixels with similar gray levels, providing information about the image coarseness and smoothness, respectively. On the other hand, the **Gray Level Dependence Matrix (GLDM)** measures the dependence between neighboring pixels with similar gray levels, offering insights into image complexity. Similarly, the **Neighboring Gray Level Dependence Matrix (NGTDM)** extends this analysis by considering the dependence between neighboring pixels with different gray levels, allowing for the characterization of dissimilarity between adjacent pixels. Overall, these texture matrices facilitate the computation of handcrafted features, providing valuable insights about image texture and heterogeneity [90].

A) Radiomics features

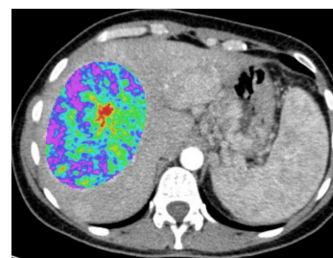
tumor density, cellularity, metabolism



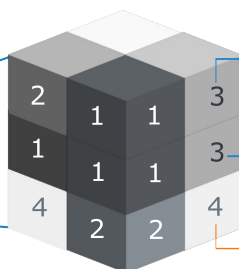
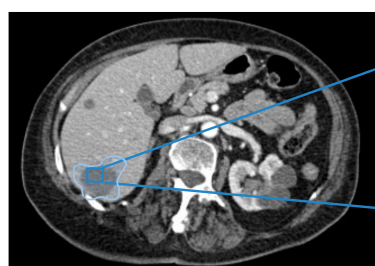
tumor shape



tumor heterogeneity: texture



B) Gray Level Co-occurrence Matrix - GLCM



Neighbor pixel value (j)

	1	2	3	4
1	1	0	2	0
2	1	0	0	1
3	0	0	0	0
4	¹ 0	1	0	0

Figure 5: Radiomic Feature Computation Workflow. This schematic illustrates the process of radiomic feature quantification. Initially, radiomic quantification involves the segmentation of the tumoral area from the medical image. Within this segmented region, features are extracted, encompassing statistical first-order distribution of density values and shape properties of the tumor. Furthermore, second-order features, exploring the relationship between voxel distributions, enable the quantification of more intricate representations, like tumor heterogeneity from the GLCM matrix. Figure adapted from Ghaffari and Ligerio et al., "Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer," Clinical Cancer Research [89].

1.4 Development of radiomic signatures for outcome prediction

The methods implemented for radiomics usually follow a two-step approach. First, a quantifiable set of **handcrafted features** is computed from imaging data. Second, **machine learning (ML) methods** are trained to predict clinically relevant categories based on these features [91]. **Figure 6** shows the end-to-end radiomic pipeline to predict clinical endpoints from CT images.

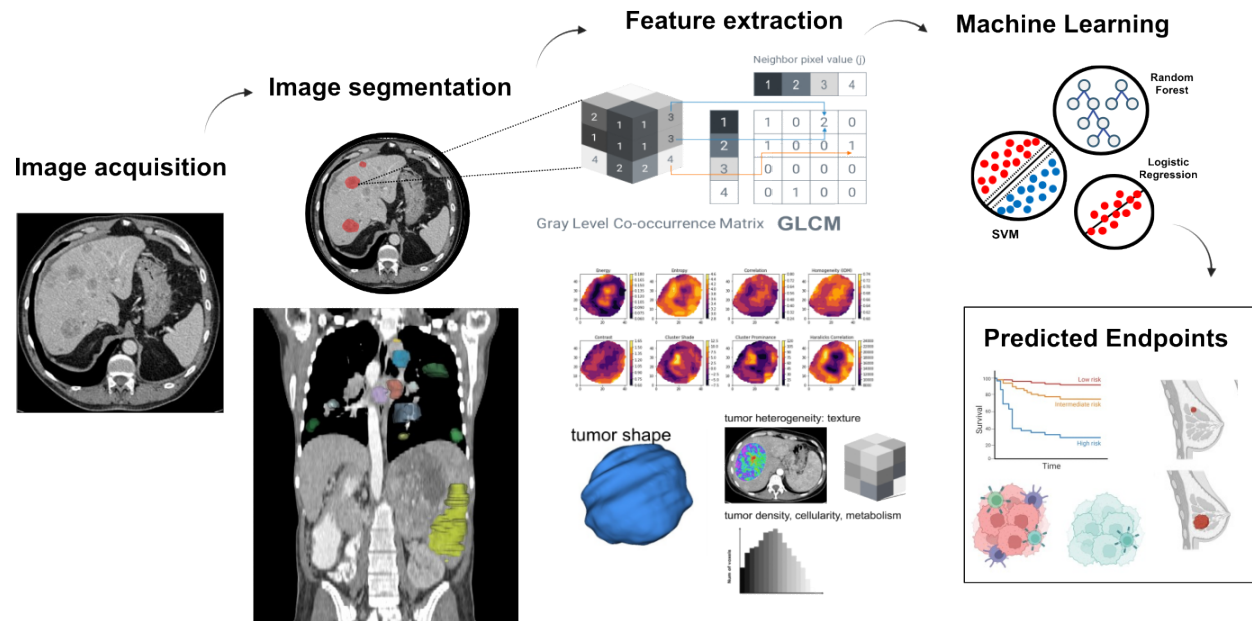


Figure 6: End-to-End Radiomics Pipeline for Direct Prediction of Clinical Endpoints from CT Images. This diagram illustrates the step-by-step process involved in creating a radiomics-based predictive model. The process begins with the acquisition, preprocessing, and segmentation of images. Subsequently, radiomic features are extracted from these segmentations. Finally, a machine learning method is employed, incorporating feature selection and modeling techniques, to train and predict the target clinical endpoints.

1.4.1 Standardization and harmonization initiatives

With the rapid growth of the radiomics field and the increase of multi-centric data sharing initiatives, encompassing images from different scanner manufacturers and protocols, there was an increasing demand to standardize radiomics metrics. This standardization is crucial to ensure the robustness and reproducibility of the technique.

The **Image Biomarker Standardization Initiative (IBSI)** guidelines [92] were introduced with the purpose to standardize radiomic features computation. These guidelines

encompass a set of 169 radiomic features that have been standardized across several computational software packages. However, despite these standardization efforts, radiomic features can still be influenced by the image-acquisition techniques and the reconstruction parameters, especially in multi-centric studies involving machines from different vendors. Consequently, to address this challenge, the field began studying radiomics sources of variability.

Several studies raised the need for harmonization efforts in radiomic features due to the high inter-observer and inter-scan variability in radiomic features [93, 94]. Most of these studies only focused on assessing radiomics variability by reporting the % of robust features by the Interclass Correlation Coefficient (ICC). However, none of the previous studies tried to find solutions for correcting this variability. Consequently, inspired by the genomics field [95, 96], first proof of concept studies showed that the implementation of **ComBat harmonization** methods could compensate for scanner effects in radiomic features [97].

Furthermore, due to the increasing development of imaging biomarkers studies, a **Radiomics Quality Score (RQS)** was defined to assess the reliability of predictive studies [84]. The RQS sets criteria for more reliable studies, including standardized imaging protocols, robust segmentation, clear methodology reporting, and external validation, among others. However, it is noteworthy that not all published studies meet all the RQS requirements. Radiomics datasets often remain limited to small sample sizes and lack external validation, potentially hindering their integration into clinical practice. For that reason, new guidelines to develop more robust radiomics studies have been defined [98], including the CLEAR checklist [99].

1.4.2 Dimensionality reduction methods

Accounting for the high-dimensional representation of radiomic features, dimensionality reduction techniques are commonly employed. These methods can be either **supervised**, using the target outcome to select data, or **unsupervised**, independent of any specific outcome.

Principal Component Analysis (PCA) is a common unsupervised technique used for dimensionality reduction. It transforms the original features into orthogonal components (principal components), which are linear combinations of the original features that capture the maximum variance directions in the data [100]. Therefore, it allows for the transformation of large feature sets into smaller numbers of vectors that can explain most of the variance in the data. Nonetheless, it is important to note that while PCA reduces complexity, the resultant principal components might not always directly correspond to interpretable biological or clinical features. This highlights the trade-off between reducing dimensionality and maintaining interpretability.

A more explainable method commonly used in radiomics studies is the use of supervised feature selection methods like **minimum Redundancy Maximum Relevance (mRMR)** or **Least Absolute Shrinkage and Selection Operator (LASSO)** models [101]. The mRMR algorithm is an entropy based feature selection method that employs mutual information (MI) to rank features based on their relevance to the outcome variable while reducing redundancy [102]. Yet, the main drawback of using mRMR is the need for human input to select the number of features to include. Considering Equation 1 where y is the target variable, $\beta_0, \beta_1, \dots, \beta_n$ the coefficients, and $x_0, x_1, \dots, x_n$ the features, LASSO adds the penalty term on the absolute values of the coefficients (Equation 2), where λ is the regularization parameter. LASSO aims to find the coefficients that minimize the objective function (Equation 3), adding the L1 penalty $\lambda \cdot \sum_j |\beta_j|$ to the sum of squares error between the predicted value and the real value $\sum_i (y_i - (\beta_0 + \sum_j \beta_j X_{ji}))^2$.

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n, \quad (1)$$

$$L_1 = \lambda \cdot (|\beta_1| + |\beta_2| + \dots + |\beta_n|), \quad (2)$$

$$\min \left\{ \frac{1}{2n} \cdot \sum_i (y_i - (\beta_0 + \sum_j \beta_j X_{ji}))^2 + \lambda \cdot \sum_j |\beta_j| \right\}, \quad (3)$$

The penalty term drives less important feature coefficients towards zero, effectively eliminating them from the model. By adjusting the weight parameters λ , LASSO can shrink the coefficients of certain features, leading to improved model interpretability and better generalization to new data [103]. Nonetheless, LASSO presents some limitations. For example, when the number of samples in the data is smaller than the number of features ($n < p$), LASSO saturates at n number of variables. Moreover, when there are many correlated features, LASSO selects one, potentially missing relevant features. To overcome these limitations, ElasticNet was proposed by adding shrinking properties and considering grouping data. This is achieved by introducing a new regularization term, λ_2 , to the objective function (Equation 4) [104].

$$\min \left\{ \frac{1}{2n} \cdot \sum_i (y_i - (\beta_0 + \sum_j \beta_j X_{ji}))^2 + \lambda_1 \cdot \sum_j |\beta_j| + \lambda_2 \cdot \sum_j |\beta_j|^2 \right\}, \quad (4)$$

Nevertheless, it is noteworthy the numerous amount of dimensionality reduction techniques and the lack of a widely accepted standard method for radiomics, despite several studies attempting to benchmark these approaches [91, 105].

1.4.3 Machine learning models and Survival analysis for Radiomics

After dimensionality reduction, radiomics studies often implement ML methods. These methods can also be classified as supervised or unsupervised. Despite some studies used K-Nearest Neighbors (KNN) or hierarchical clustering to find similar patterns in data without outcome input, most of the radiomics-based predictive models in the literature are based on supervised methods. Therefore, the model selection will be more focused on the clinical target to predict, which can be either continuous (e.g., time to response or biomarker expression) or binary (e.g., response yes or no).

In the field of medical research, the most relevant target variable to assess is response to treatment through survival metrics like **Progression Free Survival, PFS** (time between treatment start day and progression date) or **Overall Survival, OS** (time between treatment start day and exitus date). These metrics stand out because they account for situations in which the event under study has not yet occurred, such as instances where patients have not yet experienced disease progression or remain alive. These instances are referred to as “censored” data.

The **Cox proportional-hazards model** [106] is a widely used statistical method in medical research to find how various factors relate to patient survival, while also accommodating instances of censored data. The model computes how a specific factor, known as covariate, influences the probability of an event, like disease progression or death, happening at a specific time, termed as **hazard rate**.

As depicted in Equation 5, the hazard function, $h(t)$, describes the risk of an event occurring at a specific time (t , representing the survival time). This risk is determined by a set of covariates $(x_1, x_2, \dots, x_p)$ and their corresponding coefficients $(b_1, b_2, \dots, b_p)$, which measure the impact of each covariate. The h_0 represents the baseline hazard.

$$h(t) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + \dots + b_nx_n), \quad (5)$$

Each covariate contributes to the computation of a hazard ratio (HR), calculated as $\exp(b_i)$. The HR indicates how much a change in the covariate affects the event rate. An HR greater than 1 ($HR > 1$) indicates that an increase in that covariate will negatively impact patient survival, while an HR below 1 ($HR < 1$) indicates a positive effect with an increase of the covariate. An HR of 1 indicates no effect of the covariate within the model.

While continuous models offer promising insights, there is a notable trend towards using discretized endpoints in radiomics studies. This preference might be attributed to the complexity of the radiomic data, leading researchers to opt for simpler and more interpretable

models. Furthermore, most radiomics studies aim to enhance patient stratification into treatment response-related subgroups, making classification problems a suitable choice, even when the response is evaluated continuously.

Some of the most commonly used ML methods in radiomics analysis are listed below:

Logistic Regression is a statistical method used for binary classification tasks adapted from linear regression (Equation 1 and 6). It estimates the likelihood of an event occurring based on a set of covariates. Through the process of maximizing the log likelihood function (Equation 7), the model determines coefficients (b) for each covariate, thereby aligning predicted probabilities with actual outcomes.

$$f(x) = \frac{1}{1 + e^{-\beta x}}, \quad (6)$$

$$l(\beta) = \sum_{i=1}^n y_i \beta x_i - \log(1 + e^{\beta x_i}), \quad (7)$$

The **Support Vector Machine (SVM)** is a machine learning technique designed to identify a hyperplane within an N-dimensional space that efficiently separates data points belonging to distinct classes. SVM is particularly effective when addressing complex datasets that might lack linear separability. SVM models are optimized by maximizing the margin between the hyperplane and the nearest data points of each class, often referred to as support vectors. The hinge loss function plays a key role in this optimization, emphasizing the importance of maximizing not only the separation distance between classes but also the margins of the hyperplane, enhancing model generalization ability [107].

Random Forest (RF) is an ensemble learning technique based on classification and regression trees. In this method, each tree is trained on a bootstrap sample, and optimal variables for each split are determined from a random subset of all variables. In the classification task, the predicted class is selected by majority voting from all trees. Furthermore, beyond its predictive capabilities, RF can also assess variable importance measures, allowing the ranking of variables by their predictive significance and increasing interpretability [108].

Despite the wide range of feature selectors and ML models available, there is a predominant trend in these studies to apply LASSO for feature selection, followed by a logistic regression for classification. Multiple studies have highlighted the benefits of utilizing LASSO as

the feature selection method due to its efficacy in high-dimensional data regression, thereby mitigating the risk of overfitting. Similarly, several studies have explored the performance of different classification algorithms for predicting response, including SVM, RF and KNN. Interestingly, most of these studies have indicated that logistic regression demonstrated either similar or slightly better performance compared to other more complex classifiers [109–112].

1.4.4 Feature aggregation methods and mixed effects models

Radiomic feature aggregation methods are commonly needed to make patient-level predictions accounting for the whole burden of disease from single lesion-level predictions. Radiomic features are extracted for each lesion and consecutively aggregated to a single patient feature. There are different ways of aggregating features, the most relatively easy options are using statistical metrics such as average (Equation 8) or volume weighted average (Equation 8), where f is the feature vector, N the total number of lesions per patient, the volume weights and V the volume of a lesion.

$$f_{aggregated} = \sum_i^N \frac{f_i}{N}, \quad (8)$$

$$f_{aggregated} = \sum_i^N \omega_i f_i ; \quad \omega_i = \frac{V_i}{\sum_i^N V_i}, \quad (9)$$

Other alternatives are to treat the lesions as an independent instance and aggregate the predictions in the end by averaging the prediction scores. However, it is important to consider that features from the same patient are not independent between them, and this non-independence of features should be included in the models as such. **Generalized linear mixed-effects models** are statistical models that account for both fixed and random effects. Fixed effects are the variables that are expected to have an impact on the target variable to predict (i.e. the radiomic features). Random effects correspond to the grouping variables that need to be controlled in the model (i.e. patient). In this way, the model accounts for the variance among groups (patients) when doing the predictions as defined in Equation 10, where $X\beta$ correspond to the fixed effects matrix and coefficients, respectively, Zu correspond to the random effect matrix and coefficients, respective and ϵ to the error.

$$y = X\beta + Zu + \epsilon \quad (10)$$

1.4.5 The role of Deep Learning in overcoming Radiomics limitations

Despite their promising results, handcrafting sets of features come with certain drawbacks. First, they require the definition of ROIs or tumor segmentation, a process that can be time-consuming, especially when dealing with large datasets. Additionally, handcrafted features are limited to the predefined equations, limiting the ability to learn new features from images associated with the outcome. Consequently, Deep Learning (DL) methods have emerged as a solution for various image processing algorithms that typically require segmentations or manual interventions (**Figure 7**). In the field of cancer imaging, these approaches are often referred to as **Deep Radiomics**, when DL methods are employed to predict the target label directly from image data [113].

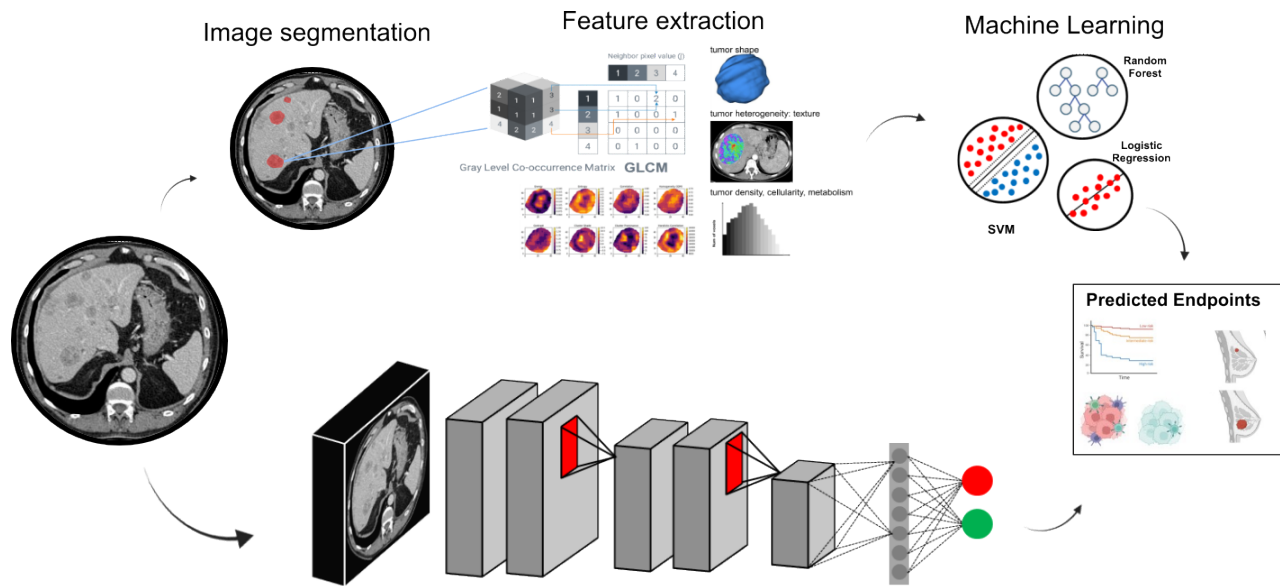


Figure 7: Conventional and Deep Radiomics Pipelines for Clinical Endpoint Prediction. This figure highlights the distinctions between traditional radiomics pipelines and deep learning methods in predicting clinical endpoints. While conventional radiomics involves image segmentation, handcrafted feature extraction and selection before machine learning implementation, deep learning directly learns representations from images associated with clinical endpoints. This approach enables the model to derive learned features directly from images, avoiding the use of predefined quantification and eliminating the need for manual annotations.

Convolutional Neural Networks (CNNs) are a type of DL models that can learn a larger number of hierarchical patterns or features at different levels of abstraction from raw image data. These models assign a label to each input image, encompassing different classes, such as responder/non-responder or disease/no-disease. They can also handle regression metrics, such as specific risk scores, or perform semantic segmentations, where each voxel receives a label differentiating for example tumor from healthy tissue.

The CNNs consist of different layers including convolutional layers, pooling layers, and fully connected layers (**Figure 8**). In the **convolutional layers**, filters (kernels) are applied to the image to extract features by sliding across the image and performing element-wise multiplication and summation. This process captures local patterns like edges and textures. Subsequently, **max pooling layers** reduce the spatial dimension of the feature maps, making the computation less complex. Finally, the **fully connected layer** is trained to classify the high dimensional features to the target label. Then, through **backpropagation**, the CNNs optimize their layer weights by minimizing the error between the predicted label and the ground truth. Thereby, in contrast to radiomic models, the extracted features are selected and weighted on the basis of the task at hand, which provides more flexibility and can improve performance [114].

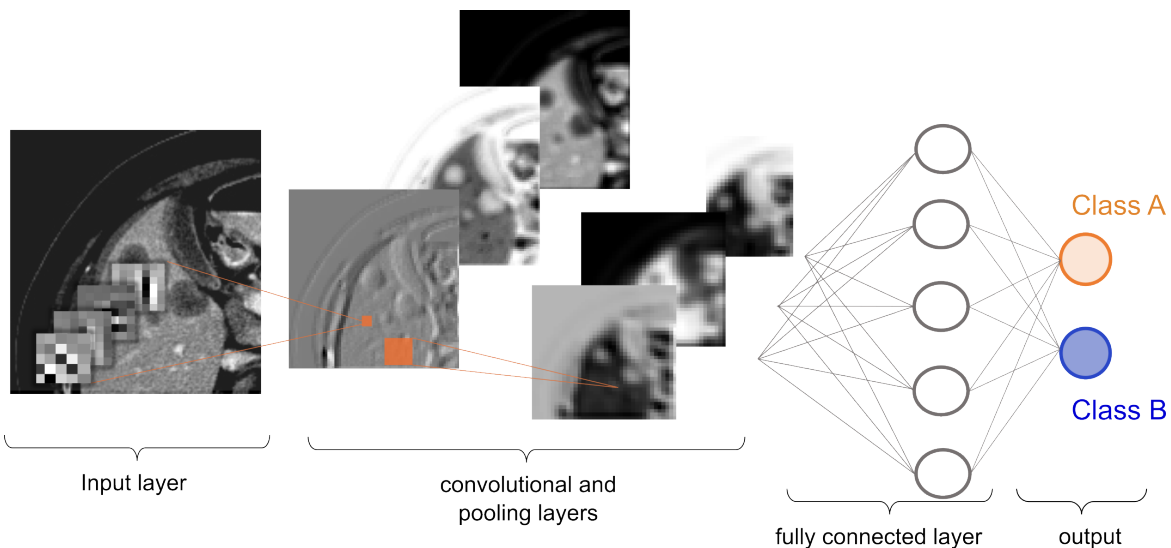


Figure 8: Convolutional Neural Network (CNN) Structure Applied to Radiology Images. The CNN comprises three key stages. Initially, image kernels or filters are employed in the convolutional layer to extract texture representations from the input image. These convolutional layers are coupled with pooling layers to reduce the complexity of the feature representations. Finally, the fully connected layer is trained to predict outcomes based on the earlier extracted representations. Figure adapted from Ghaffari and Ligerio et al., "Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer," Clinical Cancer Research [89].

A key limitation of deep learning models, particularly CNNs, is that they are powerful but data-hungry algorithms that require substantial amounts of labeled data to train effectively. In the field of medical imaging, the development of imaging biomarkers based on CNNs is still not well-established due to the scarcity of structured and labeled imaging data, presenting a significant challenge.

Therefore, most deep learning applications in radiology have focused on lesion detection and diagnosis. Tasks like detecting breast lesions from mammography [115] or lung nodules from CT scans [116] have seen wide implementation, with some tools gaining FDA approval and becoming integrated into clinical practice.

1.5 Integrating Radiology and Pathology: Advancing imaging biomarkers in oncology

Histopathology images offer valuable insights into tumor biology, providing a detailed view of cellular and subcellular phenotypes that cannot be achieved with radiology. However, one of the significant drawbacks of histopathology is that it requires invasive procedures to obtain tissue samples. Additionally, unlike radiomic biomarkers, which can be re-computed in serial imaging during the course of the disease, histopathology image biomarkers are usually only measured on the initial stages of the tumor disease or just before starting treatment. Therefore, combining both modalities could provide a more comprehensive view of the tumor and patient status by integrating information of the disease at different levels.

1.5.1 Digital pathology imaging

In clinical practice, the most commonly used histological images are hematoxylin and eosin (H&E) stained images. H&E stained images provide essential information about tissue morphology at a molecular level and are mainly used for tumor diagnosis, staging, and prognosis. Additionally, there are other staining techniques, such as Immunohistochemistry (IHC), that can provide more specific insights into cellular and molecular structures. IHC is commonly employed to detect specific proteins or antigens in the tissue, aiding in the identification of molecular biomarkers like PD-L1.

1.5.2 Deep Learning methods for Computational Pathology

In lines with the radiology field, as early as 2006, several years before DL was used as a tool in biomedicine, handcrafted image analysis pipelines with classical ML methods were implemented in digitalized histopathology images to predict biomarkers from slides [51, 117, 118]. However, digital pathology, due to its high resolution and size, rapidly showed a huge potential for DL methods [119–121].

The main advantages of the pathology images compared to radiology images is that a single whole slide image (WSI) can be divided into multiple images with high resolution, also known as tiles. Then, the model can be fitted with a large image variety from the same sample, increasing the number of cases that the model sees (**Figure 9A**).

Furthermore, The Cancer Genome Atlas (TCGA) Program [122, 123], which provides over 20,000 primary cancer and matched normal samples spanning 33 cancer types with molecular and genetic characterization of the tumors, has enabled the field of digital pathology to advance with the most recent AI developments in computer vision. Most of the first studies

started with the implementation of pre-trained CNNs to extract features from all tiles and then aggregating the results to a patient level [124, 125]. However, with the advances in AI regarding attention mechanism, benchmark studies have been done for weakly supervised learning in histopathology [19].

A notable AI advancement in digital pathology is the utilization of pre-trained feature extractors using self-supervised learning. Self-supervised learning is a deep learning method that derives meaningful feature representations from images by exploiting extensive unlabeled data [126]. One of the most common approaches involves utilizing contrastive learning. This technique entails augmenting images to create task variations and identifying features that bring the original image and its augmentations closer together while distinguishing them from other images. By employing these techniques, the model captures relevant details from the image and generates feature representations by extracting frozen layers from the model. In digital pathology, several pre-trained self-supervised learning models have been implemented using TCGA data. One of the most common feature extractors currently are RetCCL [127] and CTransPath [128], both of them are contrastive learning models that allow for WSI retrieval developed by Wang et al. (**Figure 9B**).

Another significant advancement within the AI domain, widely adopted in computational pathology, is the integration of attention-based mechanisms. Significant improvements in biomarker prediction from histology have been reported upon the implementation of techniques like attention-based multiple instance learning and, more recently, transformers.

Attention-based multiple instance learning is a deep learning method designed to address grouped data information to predict a single target [129]. It identifies patterns within the data and assigns attention scores to each instance in the group (referred to as tiles in the context of histology). The attention scores will determine the relevance of each tile to predict the overall target label. This approach is particularly effective in digital pathology, where accurately labeling individual tiles is challenging. Through this method, tiles containing non-essential data, such as non-tumoral tissue or background elements, are assigned lower attention scores, allowing the model to prioritize relevant areas of the WSI [130] (**Figure 9C**).

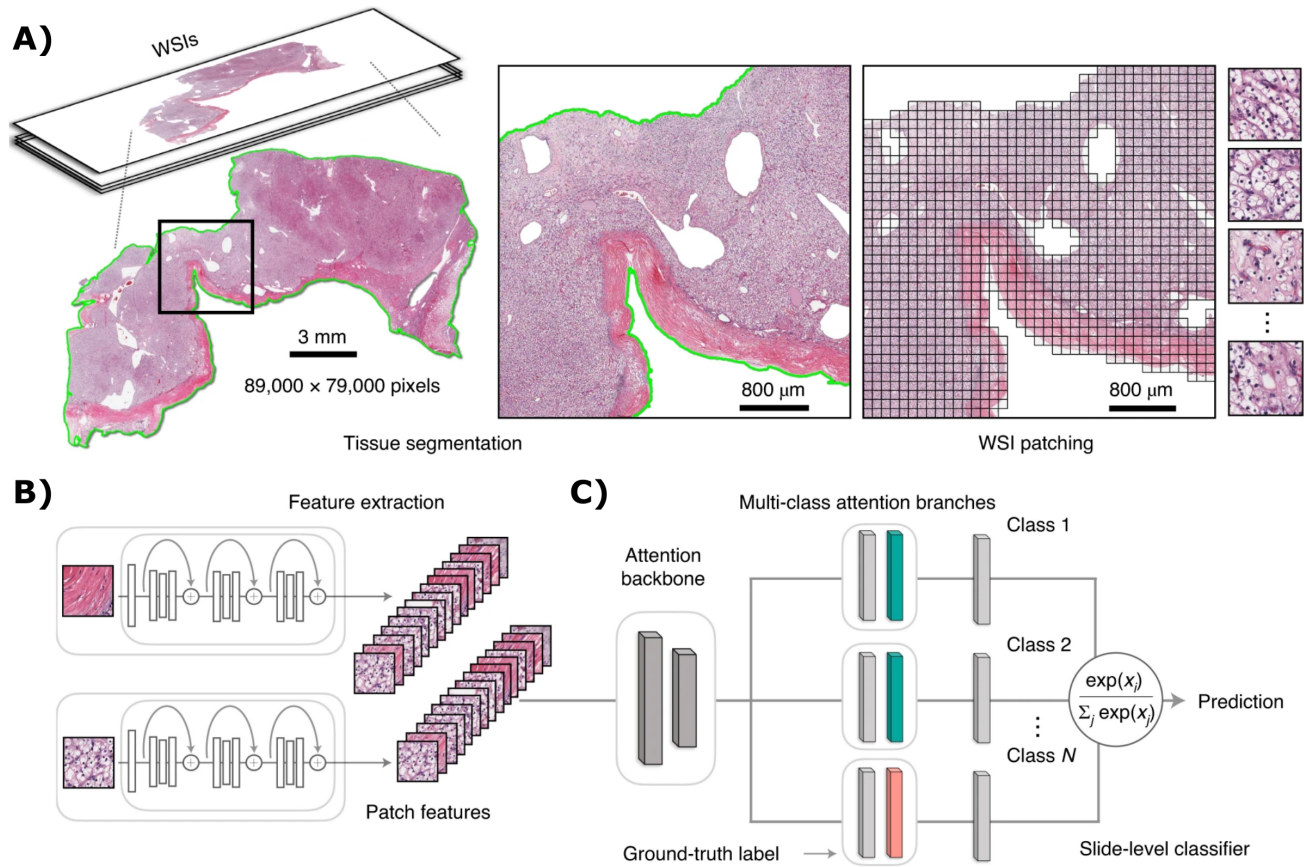


Figure 9: Computational Workflow from Whole-Slide Image (WSI) to Patient-Level Prediction. (A) Processing pathology WSI involves image preprocessing, including WSI tessellation and hematoxylin and eosin (H&E)-staining color normalization. (B) Feature extraction employs self-supervised learning to obtain feature vector representations for each tile in the WSI. (C) The feature vectors, along with the target biomarker, are then input into the weakly-supervised model. This model aggregates tile feature representations using attention-based methods to generate a slide-level prediction. Figure adapted from Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images” Nature Biomedical Engineering [Lu2021-fb].

1.6 State-of-the-art imaging biomarkers for immunotherapy response

1.6.1 CT-based radiomic biomarkers for predicting response to immunotherapy

The growing interest in radiomics as a tool for comprehending tumor biology, along with the lack of reliable immunotherapy response biomarkers, led to the emergence of multiple studies investigating the correlations between radiomic features and the effectiveness of ICIs. These studies aimed either to predict direct treatment responses or to evaluate the levels of surrogate biomarkers such as TILs and PD-L1 [89].

In one of the first studies in the field of radiomics for immunotherapy response, Sun et al. [131] trained a radiomic signature to predict CD8 cell tumor infiltration in a population of 135 patients with advanced solid malignant tumors. This signature was then validated across distinct cohorts: one involving 119 patients from TCGA with accessible CT scans and transcriptomics data, and another in a cohort comprising 100 patients distinguished by their TIL phenotypes (classified as immune-inflamed or immune-desert). Finally, as a biological explainability study, they explored the associations between the radiomic signature and treatment response in an internal cohort of 137 immunotherapy-treated patients. Notably, the signature showed an AUC (Area Under the Receiver Operating Characteristic Curve) of 0.67 (95% CI 0.57–0.77) for TCGA cohort and 0.76 (95%CI 0.66-0.86) in discriminating between immune-inflamed and immune-desert tumors. In the context of predicting immunotherapy outcomes, the authors identified a significant link between individuals who demonstrated a high and prolonged response to treatment and their baseline radiomic scores.

Similarly, a parallel study conducted by Trebeschi et al. [132] also explored radiomics as a predictive signature, but this time with a focus on direct response to immunotherapy from CT-based radiomic features. This signature was developed in 1055 primary and metastatic lesions from 203 patients with advanced melanoma and NSCLC treated with anti-PD1 therapy. This signature aimed to predict tumor growth or shrinkage post-ICI treatment. Furthermore, the lesion based nature of the study allowed for exploring dissociated response – where patients exhibited different patterns of response or progression across different lesions. Finally, they explore how the signature is associated with gene-set enrichment pathways. The outcomes of this investigation were notably promising, with AUC values ranging from 0.64 to 0.83.

Moreover, with the aim of providing the model with all the potential relevant data and knowing the effect of surrounding tumor microenvironment for immunotherapy response,

certain studies have also explored the value of incorporating peritumoral area information into predictive models for predicting response to immunotherapy [109, 133]. However, these studies have proved to still get more relevant information from the intratumoral features. Additionally, studies have shown that intratumoral 3D radiomic features provide more informative insights compared to using only 2D radiomic features [110].

A novel approach was introduced by Himoto et al. [134], studying the inter- and intra-lesion heterogeneity as a new radiomics variable. This study showed that lower inter-tumor heterogeneity was significantly associated with durable clinical responses and longer treatment duration. These findings are in line with the hypothesis that homogeneous tumors are more likely to obtain better response to immunotherapy, as previously stated by molecular biomarkers [35, 37].

Finally, another topic refers to the optimal number of timepoints that should be used to make the predictions. The majority of studies in radiomics research have focused on the development of predictive biomarkers using baseline scans, which refer to the scans obtained just before initiating treatment. This approach facilitates improved patient selection for treatment decision-making. However, some studies have demonstrated enhanced outcomes by analyzing changes in the radiomic tumor phenotype between baseline and early follow-up time points, commonly known as early readouts or delta radiomic signatures [135, 136]. Such approaches enable to capture the response or progression patterns that may go unnoticed by radiologists, thereby potentially preventing patients to stay longer under ineffective treatment. Some of these studies have shown that tracking these changes in CT scans provides better prognostic value compared to the current standard of care, RECIST [136, 137]. It is important to note, however, that these early readouts do not represent true predictive biomarkers per se, but rather serve as indicators of early response. At the time these early readouts are assessed, treatment decisions have already been made, and the patient is already receiving immunotherapy.

After the publication of the first papers, dozens of similar papers have been published in this regard developing models for specific tumor types, including larger cohorts or new advancements. Refer to section A for a systematic review of radiomics studies for immunotherapy response prediction.

1.6.2 Computational pathology for predicting immunotherapy response

In the realm of immunotherapy, one of the applications of DL methods for pathology is to facilitate the quantification of established biomarkers in IHC slides. For example, the

expression level of the PD-L1 protein on tumor cells and immune cells (combined positive score), is assessed by manual observation of IHC. Several studies have used DL to automate the subjective scoring of PD-L1 status [138–141].

Similarly, other studies have explored the implementation of AI methods in digital pathology to quantify TIL count. First, using handcrafted image analysis pipelines with classical ML methods [51] and more recently, numerous studies have used DL for TIL quantification [142, 143]. In addition to quantifying lymphocytes, ML methods were also used to quantify other types of innate and lymphoid immune cells, demonstrating that cell counts are prognostic of survival in multiple tumor types [117, 118].

However, the abilities of DL are not limited to simple quantification tasks such as counting cells. DL can learn morphological patterns from the tissue that allows for cell numbers but also, cellular shape and textures, relative cell positions and phenotypes of connective tissue. For example, by training DL systems on raw histopathology slides, it is possible to infer the MSI status in colorectal, gastric, and endometrial cancer [124, 144, 145].

Similarly, DL has been used to predict TMB directly from H&E in multiple major tumor types, including lung, breast, and colorectal cancer [146, 147]. Also, PD-L1 expression levels have been inferred from H&E directly, without the need for a dedicated IHC [148]. In addition, the expression level of gene signatures which predict ICI response can be inferred by DL from H&E slides, in hepatocellular carcinoma [149], lung cancer [150], and multiple other tumor types [145]. This could help to solve some of the practical problems associated with widespread clinical use of these signatures.

All of these above-mentioned approaches use DL to predict a known biological marker from H&E images. An alternative strategy is to train DL directly on response or outcome data. This has been attempted by at least two studies [151, 152] that used CNNs or graph neural networks (GNN) to predict immunotherapy responses. These studies reported an AUC of 0.778 for prediction of responders in melanoma and an AUC of 0.69 for predicting response in lung cancer. In general, such end-to-end prediction studies are quite difficult to perform for practical reasons: they require the DL system to be trained on clinical outcomes, ideally RECIST data. In practice, it is very difficult to collect a sufficient number of pathology tissue specimens of ICI-treated patients with matched response data. In addition, cancer immunotherapy is often administered in late lines of therapy, after patients have received one or more regimens of chemotherapy and other treatments. However, tissue for histopathology is typically acquired at the initial diagnosis, and re-biopsies at later time points are not commonly performed in most tumor types. This means that DL systems are trained on treatment-naïve tumor tissue while patients might start immunotherapy only months later after multiple previous lines of treatment have failed. This is a strong conceptual limitation

that could only be resolved by a repetitive sampling of tissue, or a move of immunotherapies towards earlier lines of treatment. [89]

1.6.3 Multimodal models for predicting response to immunotherapy

Overall, both radiology and pathology DL models have demonstrated promising potential as biomarkers for predicting immunotherapy response (**Figure 10**, section B). These modalities can spatially depict tumors on different scales, providing different information about the same tumor (**Figure 11**). By integrating both modalities, we could enhance disease characterization while overcoming individual modality drawbacks.

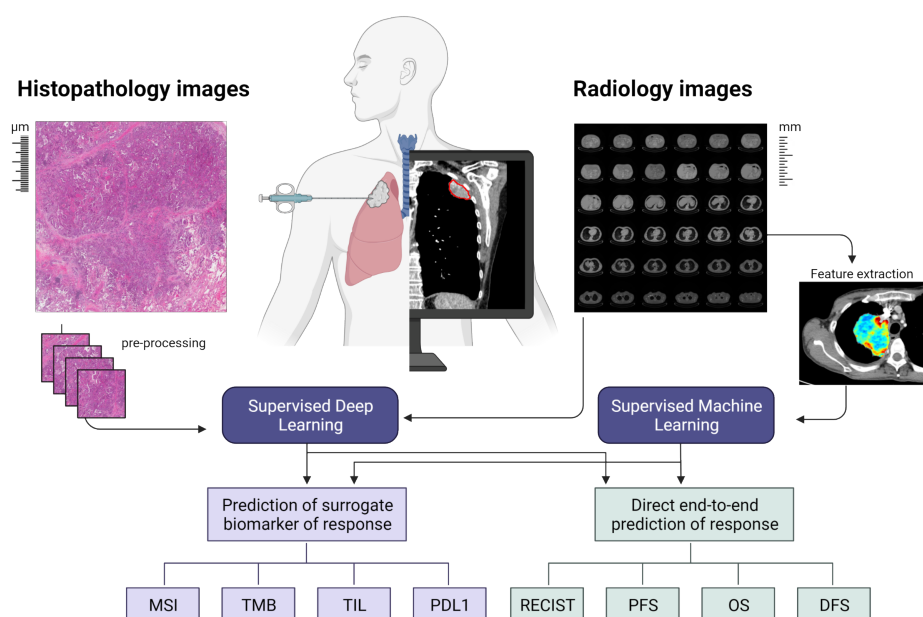


Figure 10: Comparative Workflow of Supervised Learning Studies for Predicting Immunotherapy Response from Clinical Image Data: Pathology and Radiology. This figure delineates the predictive model workflows for immunotherapy response using pathology (left) and radiology (right) image data. Accessible for nearly any patient with a solid tumor in clinical practice, both types of image data contribute to the development of predictive models. These models can anticipate immunotherapy response through surrogate marker prediction (e.g., MSI, TMB, TIL, PDL1) or directly predicting immunotherapy response metrics such as RECIST, PFS, OS, and DFS. Figure adapted from Ghaffari and Ligerio et al., "Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer," Clinical Cancer Research [89].

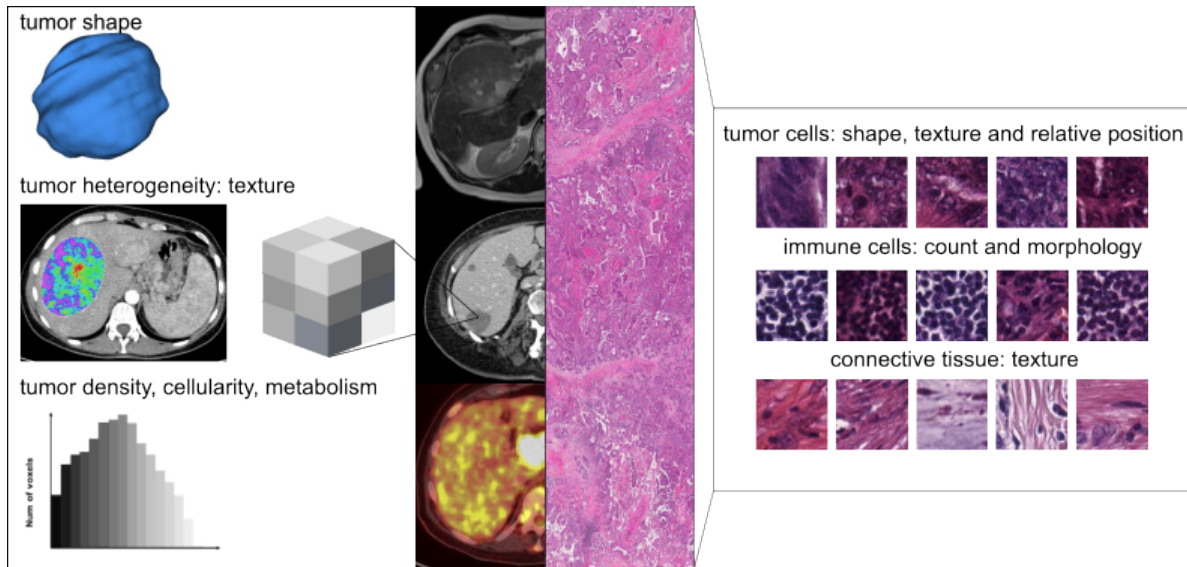


Figure 11: Radiology and Pathology Visual Features in Image Data. Radiology and pathology provide different types of information regarding the tumor phenotype. Radiology can extract 3D spatial information related to tumor heterogeneity, while pathology enables the extraction of cellular-level information concerning both tumoral and immune cells within the tissue. Figure adapted from Ghaffari and Ligerio et al., "Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer," *Clinical Cancer Research* [89].

Additionally, multimodal models extend beyond image-based biomarkers. Radiomic signatures have exhibited enhanced performance when combined with clinical data. Similarly, pathology-based DL models have demonstrated improved performance in predicting overall survival when combined with genomic data [153].

Several proof-of-concept studies have shown promising outcomes in integrating radiology and pathology to predict response and prognosis for NSCLC and ovarian cancer [154, 155]. In both studies, the integration of modalities outperformed unimodal predictions. However, the limited dataset size in these studies (around 100-200 patients) restricts model complexity.

On the other hand, complex multimodal architectures, particularly those centered on early fusion techniques, hold significant potential to enhance model performance by enabling the extraction of combined features from both modalities [156]. However, multimodal models are much more complex compared to unimodal models, due to a larger number of trainable parameters, which raises the risk of overfitting. Consequently, larger and more representative cohorts are essential to mitigate overfitting risks in multimodal approaches. The current lack of availability of extensive datasets encompassing multiple modalities hampers the development of multimodal models in a clinical setting [157].

CHAPTER 2

Hypothesis and Objectives

2.1 Hypothesis

Precision oncology has revolutionized the landscape of cancer treatment. Emerging therapeutic approaches, such as immunotherapy or anti-angiogenic agents, have exhibited remarkable outcomes in solid tumors. Despite the comprehensive understanding of the underlying biological mechanism driving these techniques, the translation of targeted therapies into clinical practice have resulted in benefits for only a subset of patients.

Efforts have been dedicated to develop biomarkers for predicting treatment outcomes and improving patient selection. While significant progress has been made in tailoring therapeutic strategies to specific molecular and biological subtypes, no robust biomarker obtained to date can accurately anticipate tumor response.

The general hypothesis within this thesis is founded on the premise that extracting robust quantitative data from standard medical images, we can effectively capture tumor phenotypes. These phenotypes could serve as indicators to identify patients who are more likely to respond to immunotherapy and other targeted therapies, thereby advancing the field of precision oncology.

Specifically, we hypothesize that:

- Robust radiomic features, accounting for multi-centric acquisition and standardized preprocessing, will enhance the performance of predictive models.
- It is possible to capture patterns of response to immunotherapy from medical imaging representations of tumor phenotypes.
- Application of radiomic predictive models can generalize to other imaging modalities and specific treatment responses.
- Response to immunotherapy can be predicted from different sources of medical imaging, such as radiological and pathological imaging.

2.2 Objectives

The general objective of this thesis is to evaluate the potential of quantitative medical image and machine learning modeling in deriving accurate tumor phenotype representations for patient stratification in cancer treatment. More specifically, the objectives of this thesis are:

- **Objective 1:** To identify potential sources of variability in CT imaging acquisition that could affect Radiomics reproducibility and explore methods for variability correction.
- **Objective 2:** To develop and validate radiomic signatures derived from CT scans for predicting patient response in a population with different tumor types treated with immunotherapy.
- **Objective 3:** To explore the use of radiomics in anti-angiogenics targeted therapies and PET imaging.
- **Objective 4:** To investigate the performance of DL-methods applied to histological images to predict response to immunotherapy

CHAPTER 3

General Materials and Methods

All the studies included in this thesis were retrospectively collected from the corresponding electronic patient records from different studies. The Vall d’Hebron Institute of Oncology (VHIO) institutional review board approved all studies for retrospective data [PR(AG)70/2018, PR(AG)29/2020, PR(AG)404/2020]. All patients included in the clinical trials provided written informed consent. Need for informed consent for the computational analysis of the images in the retrospective study was waived.

3.1 Data description

The main sources of medical image included in this thesis correspond to CT scans (Radiology), PET scans (Nuclear Medicine), H&E stained WSI and PD-L1 stained IHC WSI (Pathology). During this thesis we collected routine clinical data from several cohorts corresponding to clinical trials and standard of care treated patients carried at the Vall d’Hebron University Hospital or in multi-centric studies. External open-sources datasets from other institutions were also included. In this section, datasets are classified based on the type of treatment, clinical trial study and treated hospital. The datasets described in this section were totally or partially included in the studies of this thesis, according to the availability of the data at the time of the study. **Table 1** provides an overview of all the included cohorts and information regarding tumor type, available imaging modality and study.

The datasets related to immunotherapy response prediction used in Study 1, 2 and 5 can be classified as follows:

VHIO Retrospective Immunotherapy cohort: Advanced metastatic patients with solid tumors from different tumor types (pan-cancer), including lung, melanoma, bladder among others, treated with Immune-checkpoints inhibitors (ICIs), anti-PD-1/PD-L1, alone or in combination with other ICI target therapies at Vall d’Hebron

Universitary Hospital in Phase I, II and III clinical trials or standard of care treatment from August 2012 to May 2018.

VHIO PREDICT trial cohort: Prospective dataset with advanced metastatic patients with solid tumors from different tumor types (pan-cancer) treated with anti-PD-1/PD-L1 ICIs in combination with other ICI target therapies at Vall d’Hebron Universitary Hospital from August 2020 to May 2023.

VHIO Cytotoxic-RNA signature validation cohort: Patients with different tumor types (lung cancer majority) treated at Vall d’Hebron University Hospital with available gene expression profiling from solid tumor biopsy samples where contrast-enhanced CT was performed within 4 months before biopsy. None of these patients had been included in the previous cohorts. This cohort was used for biological explainability study of the signature developed in Study 2.

MSKCC Lung cohort: Open source available dataset of advanced metastatic patients with lung cancer treated with anti-PD-1/PD-L1 ICIs at Memorial Sloan Kettering Cancer Center [155] Data regarding treatment response evaluation, CT scans, and PD-L1 IHC samples were available for this cohort. This cohort was used in Study 5 for training a DL predictive model for PD-L1 and explore immunotherapy response associations.

For these cohorts, data regarding treatment response evaluation, CT scans, H&E and PD-L1 stained IHC WSI samples were available and collected retrospectively. Imaging data (CT, H&E and PD-L1 stained IHC WSI) prior to immunotherapy treatment were collected. All metastatic lesions were delineated and radiomics analysis was performed. CT images from this cohort were used for training models in Study 2 and for tumor type classification purposes in Study 1. For available cases, PD-L1 stained IHC samples were collected and included in Study 5 for model validation purposes. Cohorts containing both radiology and pathology data partially overlap in Study 2 and Study 5, enabling the possibility of conducting multimodal analyses in subsequent studies.

The dataset included for predicting response to other targeted therapies are classified as follows:

VHIO CAR T-cell Lymphoma cohort: Patients with relapsed or refractory (R/R) large B-cell lymphoma (LBCL), de novo or transformed from an indolent lymphoma, treated with CD19-targeted second-generation CAR T cells from July 2018 to November 2021 at Vall d’Hebron University Hospital. Baseline PET-CT scans prior to CAR

T-cell treatment were collected. The largest lesion was delineated and radiomics analysis was performed. Data regarding treatment response evaluation and PET-CT scans were available. This cohort was used in Study 3 for training and validating a PET-based signature for predicting response to CAR T-cells.

Multi-centric TALENT trial Neuroendocrine cohort: Patients with gastroenteropancreatic neuroendocrine tumours (GEP-NETs) treated with the multikinase inhibitor lenvatinib in a multicentre, international, phase II clinical trial conducted from October 2015 to August 2020, identified as NCT02678780. Data regarding treatment response evaluation and CT scans were available for this cohort. Baseline CT scans prior to anti-angiogenic treatment were collected. All liver lesions were delineated and radiomics analysis was performed. CT images from this cohort were used for training and validating the model in Study 4.

VHIO Sunitinib Neuroendocrine cohort: Patients with pancreatic NETs treated with the multikinase inhibitor sunitinib as standard of care at the Vall d’Hebron Institute of Oncology (Barcelona, Spain) from October 2011 to September 2020. Data regarding treatment response evaluation and CT scans were available for this cohort. Baseline CT scans prior to anti-angiogenic treatment were collected. All liver lesions were delineated and radiomics analysis was performed. CT images from this cohort were used for externally testing the model in Study 4.

Study included	Cohort Name	Tumor type	Treatment type	N	Imaging Modality (number of cases)
Study 1 Study 2 Study 5 Annex III	VHIO Immunotherapy cohort (retrospective and PREDICT)	Pan-cancer Lung Bladder Melanoma	Immunotherapy (anti-PD-1/PD-L1) Combined Immunotherapy (anti-PD-1/PD-L1 + other targets)	424	CT (369) IHC (106) H&E (234)
Study 2	VHIO-Cytotoxic RNA signature	Pan-cancer	NR	20	CT
Study 5	MSKCC-Lung	Lung	Immunotherapy (anti-PD-1/PD-L1)	233	CT (187) IHC (233)
Study 3	VHIO-Lymphoma	R\R DLBCL	Immunotherapy (CAR T-cells)	97	PET-CT
Study 4	TALENT-Neuroendocrine	NETs	Anti-angiogenics	65	CT
Study 4	Sunitinib-Neuroendocrine	NETs	Anti-angiogenics	24	CT

Table 1: Description of cohorts collected, preprocess and included in the thesis studies or in academic related work. The table describes for each cohort the studies where the data was included, the primary tumor type and the corresponding treatment. Finally, the total number of patients and the modalities included. For datasets with different modalities, the number of available cases for each modality is specified in brackets.

N = number of cases, PD-1 = programmed cell death protein-1, PD-L1 = programmed cell death protein ligand-1, CAR = Chimeric antigen receptor, CT = Computed Tomography, IHC= immuno histochemistry, H&E = Hematoxylin and Eosin staining, PET = Positron = Positron Emission Tomography, R\R LBCL = relapsed or refractory large B-cell lymphoma, NETs =Neuroendocrine Tumors, NR = not relevant.

3.2 Imaging data acquisition protocol

For the development of CT-imaging biomarkers (studies 2 and 4), CT scans were obtained within 28 days before the treatment start date. Contrast-enhanced CT scans were acquired by using 16- or 64-channel CT scanners (Siemens, Philips, and GE Healthcare) in the axial plane with tube voltage of 100–130 kV, section thickness of 2–5 mm, and section interval of 1.25–5 mm.

In study 3, the PET-CT exam required a minimum of 6-hour fast prior to the intravenous administration of 3.7 MBq/kg (222–370 MBq) of ^{18}F -FDG. Glucose values below 140 mg/dL were required in all cases prior to administration of the radiopharmaceutical. Before scanning, the patients were at rest for a minimum of 60 min. Images were obtained using a Siemens Biograph mCT, which combines a spiral CT of 64 slices (210 keV, 120 mAs, care dose) with a dedicated PET, from the skull to the upper third of both femurs.

PD-L1 stained IHC image acquisition is specific for every center and can be found in Study 5 methods section.

3.3 Clinical data and outcome definition

Clinical and demographic data were collected from electronic patient records. These included age, sex, tumor type, Eastern Cooperative Oncology Group performance status, number of organs with metastasis. Study specific clinical data was also collected such as lab data: lactate dehydrogenase level, leukocyte count, albumin level, lymphocyte count, platelet count, and neutrophil count at baseline, RNA sequencing data, pathologist PD-L1 Tumor Proportion Score (TPS) and Combined Positive Score (CPS) quantification, tumor grade and Ki-67. Outcome data included PFS and OS. The PFS was defined as the time between trial start date and tumor progression date. A patient is considered to have benefited from the treatment (clinical benefit) if achieving either a CR, PR or SD by RECIST 1.1 for a duration exceeding the median PFS. The median PFS was obtained according to the average response rates of the treatment (5 months for immunotherapy trials ([158]) or the specific median PFS corresponding to phase II and III clinical trials (15.7 months for TALENT Phase II trial [69], 11.4 months for Sunitinib Phase III trial [68], and 3 months for CAR T-Cells [79]).

3.4 Image pre-processing and feature extraction for CT and PET

All acquired CT and PET images were stored in the standardized Digital Imaging and Communications in Medicine (DICOM) format. A DICOM file contains reconstructed imaging data and metadata regarding acquisition and patient information. DICOM images were retrieved from the imaging collecting system used in our institution (Picture archiving and communication system, PACs). Images were pseudo-anonymized with an identifiable patient ID and transferred to a password protected VHIO storage server. Anonymized DICOM images were converted to NIfTI (Neuroimaging Informatics Technology Initiative) files and the metadata regarding image acquisition was stored in json files using dcm2niix conversion software [159]. The NIfTI file is a standard format for storing 3D and 4D volumetric data, commonly used in the medical field beyond neuroimaging.

Regarding CT scans tumor segmentation, all well-defined target lesions per patient were delineated and supervised by an expert radiologist by using the 3DSlicer version 4.8.1 [153] semiautomatic contouring tool. Small metastases (i.e., largest diameter < 1 cm) or lesions with artifacts were excluded.

In PET images, Standard Uptake Value (SUV) was computed voxel-wise from 18F-FDG PET/CT images normalized to patient’s weight and injected dose [160] prior to image segmentation. Then, lesions were delineated by an experienced nuclear medicine physician using a custom workflow from MIM Encore™ software (MIM Software Inc., Cleveland, OH, USA), which selected all metabolic activity above the 41% maximum SUV threshold.

CT images were resampled to isometric voxels of $1 \times 1 \times 1 \text{ mm}^3$ by using spline interpolation. PET images were resample to $2 \times 2 \times 2 \text{ mm}^3$ due to its lower resolution [161]. CT Hounsfield units and PET SUV values were binarized to discrete values of 10 or 25 HU, depending on the study purpose [162]. Radiomic features (first order, shape, and texture) were derived by using the Pyradiomics package (version 2.2.0) for Python (version 3.7.1), compliant with the IBSI guidelines [92]. After feature extraction, vectors of 105 features were obtained from every lesion.

3.5 Image pre-processing and feature extraction for IHC

The PD-L1 stained IHC WSI were stored in ndpi format and identified by the biopsy number. Imaging samples were annotated in polygonal regions of interest using the open source

QuPath software v0.3.2. These annotations were used to crop the image excluding any other tissue from the WSI that was not from the patient, i.e. reference samples. PD-L1 stained WSI were tessellated into non-overlapping tiles of 224x224 pixels corresponding to an edge length of 256 μm from the ROIs. Blurry tiles or tiles without tissue samples, defined as tiles containing less than 2% of edges in the image, were discarded using Canny edge detection [163].

Feature extraction was performed using the Retrieval with Clustering-guided Contrastive Learning (RetCCL) model [127], an ImageNet-weighted ResNet50 pre-trained in a self-supervised manner on a pathology dataset with 32,000 H&E-stained histology slides across various cancer types. After feature extraction, vectors of 2048 features were obtained from every 224x224 pixels tile image.

3.6 Statistical analysis

Reproducible radiomic features were identified as ICC values greater than 0.7-0.8 [164]. Feature selection in radiomics analysis was implemented by LASSO regression and mRMR. Model fitting in radiomics analysis was performed with Logistic regression. Feature aggregation methods like median of feature values, linear mixed model and attention MIL were implemented in the studies.

The metric to evaluate the developed models was the Area Under the curve (AUC) and 95% CI (DeLong method) from the Receiver Operating Characteristic (ROC) curve. P values were assessed by the Mann-Whitney U test to test statistical significance from random prediction (AUC=0.5). Internal validation of radiomic signatures was performed using cross-validation analysis.

Cox proportional HR model was implemented to associate the prediction scores with PFS and OS. Univariate and multivariable analyses were performed to explore the additional value of clinical variables to the radiomic score.

Moreover, survival analysis was performed by using the Kaplan–Meier method after predictive score dichotomization. Patients who did not reach the endpoint were censored to the last follow-up date or end of treatment for OS and PFS, respectively.

All the statistical analysis and model implementation was developed using R v3.5.0 or Python 3.8. All codes can be found at <https://github.com/mligerhe/>.

CHAPTER 4

Results

4.1 Study 1

Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis

This study corresponds to **objective 1** of the thesis. This research aimed to find potential sources of variability in radiomic features and explore methods for image pre- and post-processing to harmonize data towards more reliable radiomics model development.

Reference:

Ligero, M., Jordi-Ollero, O., Bernatowicz, K. et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 31, 1460–1470 (2021).



Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis

Marta Liger¹ · Olivia Jordi-Ollero² · Kinga Bernatowicz¹ · Alonso Garcia-Ruiz¹ · Eric Delgado-Muñoz¹ · David Leiva³ · Richard Mast⁴ · Cristina Suarez⁵ · Roser Sala-Llonch⁶ · Nahum Calvo³ · Manuel Escobar⁴ · Arturo Navarro-Martin⁷ · Guillermo Villacampa⁸ · Rodrigo Dienstmann⁸ · Raquel Perez-Lopez^{1,4} 

Received: 14 February 2020 / Revised: 23 June 2020 / Accepted: 10 August 2020 / Published online: 9 September 2020
© The Author(s) 2020

Abstract

Objective To identify CT-acquisition parameters accounting for radiomics variability and to develop a post-acquisition CT-image correction method to reduce variability and improve radiomics classification in both phantom and clinical applications.

Methods CT-acquisition protocols were prospectively tested in a phantom. The multi-centric retrospective clinical study included CT scans of patients with colorectal/renal cancer liver metastases. Ninety-three radiomics features of first order and texture were extracted. Intraclass correlation coefficients (ICCs) between CT-acquisition protocols were evaluated to define sources of variability. Voxel size, ComBat, and singular value decomposition (SVD) compensation methods were explored for reducing the radiomics variability. The number of robust features was compared before and after correction using two-proportion *z* test. The radiomics classification accuracy (*K*-means purity) was assessed before and after ComBat- and SVD-based correction.

Results Fifty-three acquisition protocols in 13 tissue densities were analyzed. Ninety-seven liver metastases from 43 patients with CT from two vendors were included. Pixel size, reconstruction slice spacing, convolution kernel, and acquisition slice thickness are relevant sources of radiomics variability with a percentage of robust features lower than 80%. Resampling to isometric voxels increased the number of robust features when images were acquired with different pixel sizes ($p < 0.05$). SVD-based for thickness correction and ComBat correction for thickness and combined thickness–kernel increased the number of reproducible features ($p < 0.05$). ComBat showed the highest improvement of radiomics-based classification in both the phantom and clinical applications (*K*-means purity 65.98 vs 73.20).

Conclusion CT-image post-acquisition processing and radiomics normalization by means of batch effect correction allow for standardization of large-scale data analysis and improve the classification accuracy.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00330-020-07174-0>) contains supplementary material, which is available to authorized users.

✉ Raquel Perez-Lopez
rperez@vhio.net

¹ Radiomics Group, Vall d’Hebron Institute of Oncology (VHIO), Hospital Universitari Vall d’Hebron, Vall d’Hebron Barcelona Hospital Campus (Spain), Barcelona, Spain

² Medical Physics and Radiation Protection Department, Catalan Institute of Oncology (ICO), Duran i Reynals Hospital, Barcelona, Spain

³ Radiology Department, Bellvitge University Hospital, Barcelona, Spain

⁴ Radiology Department, Vall d’Hebron University Hospital, Barcelona, Spain

⁵ Medical Oncology, Vall d’Hebron Institute of Oncology (VHIO), Hospital Universitari Vall d’Hebron, Vall d’Hebron Barcelona Hospital Campus (Spain), Barcelona, Spain

⁶ Department of Biomedicine, Faculty of Medicine, University of Barcelona, Barcelona, Spain

⁷ Radiation Oncology Department, Catalan Institute of Oncology (ICO), Duran i Reynals Hospital, Barcelona, Spain

⁸ Oncology Data Science (ODysSey) Group, Vall d’Hebron Institute of Oncology (VHIO), Hospital Universitari Vall d’Hebron, Vall d’Hebron Barcelona Hospital Campus (Spain), Barcelona, Spain

Key Points

- *The voxel size (accounting for the pixel size and slice spacing), slice thickness, and convolution kernel are relevant sources of CT-radiomics variability.*
- *Voxel size resampling increased the mean percentage of robust CT-radiomics features from 59.50 to 89.25% when comparing CT scans acquired with different pixel sizes and from 71.62 to 82.58% when the scans were acquired with different slice spacings.*
- *ComBat batch effect correction reduced the CT-radiomics variability secondary to the slice thickness and convolution kernel, improving the capacity of CT-radiomics to differentiate tissues (in the phantom application) and the primary tumor type from liver metastases (in the clinical application).*

Keywords Radiologic phantom · X-ray computed tomography · Image processing · Metastasis

Abbreviations

GLCM	Gray-level co-occurrence matrix
GLDM	Gray-level dependence matrix
GLRLM	Gray-level run length matrix
GLSZM	Gray-level size zone matrix
ICC	Intraclass correlation coefficient
NGTDM	Neighboring gray-tone different matrix
PCA	Principal component analysis
SVD	Singular value decomposition
TCGA-KIRC	The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma
VOI	Volume of interest

Introduction

Radiomics is revolutionizing medical image assessment and interpretation, moving from a subjective evaluation to a quantifiable -omics image assessment method [1, 2]. Multiple studies have shown that radiomics provides meaningful information about cancer and correlates with histological and molecular tumor phenotypes, creating opportunities to develop novel predictive and prognostic biomarkers for cancer [3, 4]. The maximum benefit for cancer patients has been shown when tailoring treatments to specific cancer characteristics [5]. Thus, radiomics can play a key role in improving personalized medicine. However, radiomics features are influenced by the image-acquisition technique and the reconstruction parameters [6–9]. Studies performed at a single institution usually do not account for this source of variability, and then, the results entail low scalability of the signatures for multicentric applications.

To achieve meaningful generalizable radiomics-based tools, large-scale studies are necessary [10]. These require multicenter data collection, which implies scans acquired with different protocols, particularly when including retrospective data. Different strategies have been followed to minimize the effects of radiomics variability. Aerts et al considered radiomics variability as a feature selection tool by using test–retest analysis, eliminating radiomics features with high

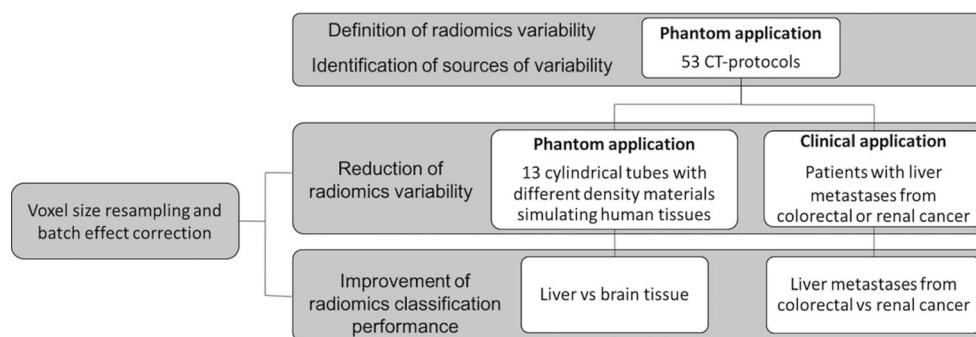
variability based on their cohort results [5, 11]. Sun et al introduced the image-acquisition parameters as a confounding variable into the model [3], and Choe et al explored convolutional neural networks-based kernel conversion for reducing radiomics variability [12].

There is an unmet need to establish robust pre- or post-image-acquisition methods for radiomics data harmonization. In this study, we explore the main image-acquisition factors that generate radiomics variability. These variability-causing factors are called “batch effects” [13]. Different batch effect correction techniques have been developed, allowing for genomics and proteomics data harmonization [14]. These batch effect correction techniques aim to remove the variance of the signal caused by the variability between batches to improve the biological signal. Alter et al defined the singular value decomposition (SVD)-based batch effect removal, where the principal components associated with the batch variability are filtered from the data and the matrix is reconstructed without these factors [15]. Johnson et al implemented the ComBat algorithm for batch correction based on an empirical Bayes approach to standardize the means and variances across batches to reduce the batch effect error [16, 17]. However, there is little evidence of the application of these methods towards reducing radiomics variability [18].

In this study, we aim to describe the image-acquisition-based sources of CT-radiomics variability. We also explore the role of image resampling and batch effect as post-image-acquisition correction methods for reducing radiomics variability, thereby improving the classification accuracy of radiomics in phantom and clinical applications.

Materials and methods

Multiple images of a phantom were acquired with different CT-acquisition protocols to explore the radiomics variability and identify the sources of variability according to the CT-acquisition parameters. Then, image post-processing and batch correction methods were implemented to reduce the radiomics variability in phantom and clinical applications.

Fig. 1 Methodology flowchart

Finally, we explored the improvement of radiomics classification performance by reducing the radiomics variability (Fig. 1).

The retrospective clinical study was approved by the institutional review board. Informed consent for the computational analysis of the CT images was waived.

Phantom image acquisition

The Gammex Model 467 Tissue Characterization Phantom (Gammex RMI) was used to describe intra-scanner variability for different acquisition parameters. This phantom includes a matrix with 13 rods of 33 cm diameter with different density materials simulating human tissues (Supplementary Material 1).

Phantom CT scans were acquired in a 16-channel Philips CT scanner by fixing all the acquisition parameters except the one tested. The tested acquisition parameters included voltage, current, slice thickness, and voxel size (accounting for the slice spacing and the pixel size) with a total of 25 different acquisition protocols (Supplementary Material 2). The minimum and maximum values of the acquisition parameters (voltage, slice thickness, slice spacing, and pixel size) were reconstructed with all the available Philips-specific reconstruction kernels (A, B, C, D, and E) to study the kernel variability with different acquisition parameter sets. The rest was reconstructed with kernel “A,” leading to 53 different protocols (Supplementary Material 3).

Clinical image acquisition

The clinical study included 43 patients (mean [range] age 66.41 [41–77] years; 46.51% [20/43] female, 53.49% [23/43] male) with 97 liver metastases (mean [range] lesions per patient 2.26 [1–7]) from colorectal adenocarcinoma (53.61% [52/97]) and clear cell renal carcinoma (46.39% [45/97]) [19].

Contrast-enhanced CT scans were collected retrospectively and acquired at Vall d’Hebron University Hospital and Bellvitge University Hospital between November 2013 and September 2019 with two specific acquisition protocols from General Electric and Siemens CT scanners. Additionally, all

CT scans from the open-access database The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) [20] were acquired with a General Electric scanner. Only patients with liver metastases from this open-access database were included in the analysis. Detailed information of CT-scan acquisition and reconstruction protocols for each hospital and the TCGA-KIRC database are defined in Supplementary Material 4.

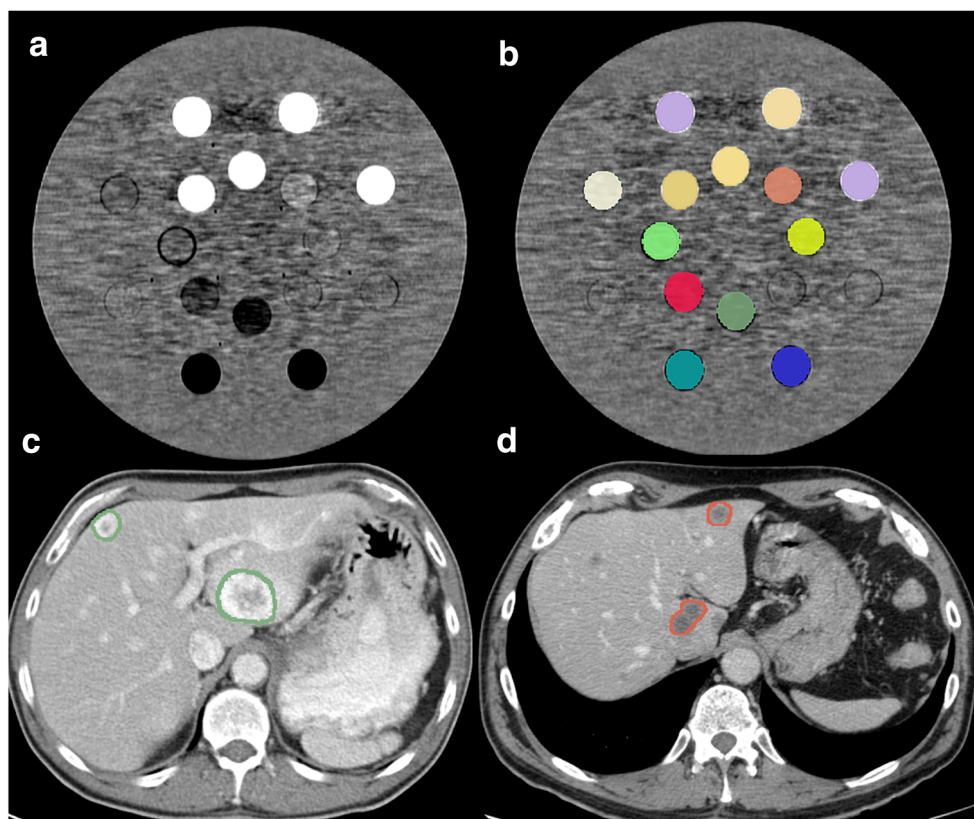
Image processing and radiomics features extraction

In the phantom application, the 13 rods were delineated (including the entire rod) with a semi-automatic contouring function from 3DSlicer v4.8.1 [21], obtaining one volume of interest (VOI) per rod. The same VOI per rod was used to extract the radiomics features from the phantom in all the studied CT-acquisition protocols (Fig. 2a, b). Image registration was not needed, given that the scans were acquired with the same starting and ending position, and the phantom’s position did not change between scans.

In the clinical application, all well-defined liver metastases were included in the analyses (Fig. 2c, d). Small metastases (i.e., largest diameter < 1 cm) or with artifacts were excluded. Lesions were delineated with the 3DSlicer v4.8.1 semi-automatic contouring function [21] supervised by a radiologist physician with 10-year experience in oncological imaging.

In the phantom application, radiomics data from images without voxel resampling were extracted to study the impact of image resampling on radiomics data. For batch correction analysis, the images and masks were resampled to isometric voxels of $1 \times 1 \times 1 \text{ mm}^3$ using spline interpolation and nearest-neighbor interpolation, respectively. Image values were discretized to a bin size of 50 HU; afterwards, the CT-radiomics features from the VOIs were extracted. The radiomics features, including first-order and texture analyses, were derived using an in-house program based on the Pyradiomics package for Python [22]. For texture feature extraction, five gray-level matrices (gray-level co-occurrence matrix [GLCM], gray-level dependence matrix [GLDM], gray-level run length matrix [GLRLM], gray-level size zone matrix [GLSZM], and neighboring gray-tone different matrix

Fig. 2 Axial CT of the Gammex 467 Tissue Characterization Phantom showing the thirteen tissue and water materials (a) with the segmented volumes of interest (VOI) for the different rod materials (b). Axial enhanced CT of the abdomen showing the target liver metastases (green and red masks) of a patient with clear cell renal carcinoma (c) and a patient with colorectal adenocarcinoma (d)



[NGTDM]) were calculated in three dimensions. Ninety-three radiomics variables were obtained for each VOI, including variables from the first-order histogram and the five gray-level matrices (Supplementary Material 5).

Sources of variability identification

Relevant sources of variability were identified by the intraclass correlation coefficient (ICC) of all the radiomics features between different acquisition parameters accounting for pixel size, reconstruction slice spacing (interpolated from the raw CT-image data without voxel resampling), acquisition slice thickness, convolution kernel, current, and voltage. The CT-acquisition variables that presented less than 80% of robust radiomics features (i.e., less than 80% of the features with $ICC > 0.8$ [23]) were defined as relevant sources of variability (batches) for further correction.

Techniques for radiomics variability correction

Image resampling

To correct variability from parameters related to voxel size, radiomics data were extracted from images resampled to isotropic voxels of $1 \times 1 \times 1 \text{ mm}^3$. Acquisition voxel size variability was analyzed separately by pixel size and slice spacing. Radiomics data of all the phantom materials with different

acquisition pixel sizes (0.35×0.35 , 0.78×0.78 , and $1 \times 1 \text{ mm}^2$) and slice spacings (1, 1.25, 2, 2.5, and 5 mm) were included while the rest of parameters remained fixed.

To assess the effect of resampling data on variability correction, the ICCs between groups of different acquisition pixel sizes and slice spacings were computed before and after resampling. Principal component analysis (PCA) was implemented to qualitatively show the reduction of variability on radiomics data variance caused by resampling the acquisition voxel size.

Batch effect removal

To correct variability sources related to image acquisition and reconstruction, two methods of batch correction were applied: singular value decomposition-based (SVD-based) correction [15] and ComBat correction [16]. ComBat correction was applied using the SVA package from R version 3.6.1. [17]. For SVD-based correction, principal components (PC) with higher correlation with batches (i.e., convolution kernel and slice thickness defined as per the ICC analysis) were removed from the PCA space, and the matrix was reconstructed back to the feature space (Supplementary Material 6). ComBat correction with parametric adjustments was applied three times considering the sources of variability as batches (i.e., convolution kernel and slice thickness and the slice thickness–convolution kernel combination).

To evaluate data correction, the ICCs between groups of acquisition parameters were assessed before and after the application of two different batch corrections. Principal component analysis (PCA) was implemented to qualitatively show the reduction of variability on radiomics data variance caused by batch correction techniques.

Two-proportion z test was applied to compare the percentage of robust features before and after correction. The p value threshold for significance was established at 0.05. Adjustment for multiple testing was performed by controlling the false discovery rate at 0.05 according to the Benjamini and Hochberg method.

Validation analysis

Unsupervised K -means clustering purity was used to evaluate the improvement in data classification after batch variability correction with SVD and ComBat. To measure clustering purity, each cluster is assigned to the most frequent class in the cluster. Then, the accuracy is measured by counting the number of correctly classified data in the assigned class. The performance of the unsupervised clustering before and after the implementation of batch correction was analyzed in a phantom and a clinical application. The clustering performance was also analyzed in non-resampled data.

Phantom application

Two similar phantom materials (liver and brain) were included. Batch effect correction by ComBat was applied three times considering different sources of variability as batches: convolution kernel (five batches: A, B, C, D, E), slice thickness (three batches: 2, 3, 5 mm), and the combination of convolution kernel with slice thickness (15 batches; all possible combinations of convolutional kernel and slice thickness).

Clinical application

The clinical application aimed to analyze the performance of clustering different primary tumor types (colorectal versus renal) based on liver metastasis radiomics data. Batch effect correction by ComBat was applied three times considering different sources of variance as batches: manufacturer-dependent convolution kernel (two batches: General Electric and Siemens), slice thickness (four batches: 1.25, 2, 2.5, 5 mm), and the combination of convolution kernel with slice thickness (eight batches; all possible combinations of manufacturers and slice thickness).

The K -means clustering was computed 1000 times, and the highest purity of the clustering appearing on more than 20% of the iterations was chosen for the comparison between the initial data and the data after different batch correction techniques (SVD-based, ComBat) [24].

Results

Population (phantom and clinical applications)

In the phantom application, a total of 53 different CT scans of the 13 phantom materials were acquired in a Big Bore 16 CT scanner (Philips) with different acquisition parameters and reconstruction kernels.

The clinical population included 97 liver metastases from 43 patients. CT scans were retrospectively collected from Vall d'Hebron University Hospital (26/43) and Bellvitge University Hospital (12/43). In addition, five cases from the open-access database TCGA-KIRC were also included. CT images were acquired in CT scanners from two manufacturers: 60.46% (26/43) from Sensation 64 CT scanner (Siemens) and 39.53% (17/43) from Light Speed Pro 16 CT scanner (General Electric) (Table 1).

Defined sources of variability

For the phantom data, including all materials, ICCs between the different batches were assessed. Pixel size, slice spacing, slice thickness, convolution kernel, and voltage presented a low percentage of robust radiomics features (i.e., less than 80% of the radiomics features with ICC > 0.8) in at least one of the combinations from the ranging CT-acquisition parameters (Fig. 3).

Voxel size was defined as a relevant source of variability with a percentage of robust features ranging from 48.40 to 78.49% for pixel size and from 43.01 to 86.02% for slice spacing.

According to the slice thickness, the percentage of robust features ranged from 75.25% (when 2 mm and 5 mm were compared) to 88.17% (when 2 mm and 3 mm were compared). The percentage of robust radiomics features according to the convolution kernel ranged between 55.92% (when A and D were compared) and 97.85% (when A and B were compared).

The acquisition voltage of 90 kV showed the highest variability on radiomics data (65.59% of reproducible features). The standard voltage in clinical protocol range (i.e., 120–140 kV) showed higher radiomics robustness (81.72% of reproducible features). The percentages of robust features are described in Supplementary Material 7.

Therefore, voxel size, slice thickness, and convolution kernel were defined as the sources of variability with the highest impact on radiomics data reproducibility; voxel size was corrected by resampling, whereas slice thickness and convolution kernel were considered for batch correction.

Evaluation of processing effects on radiomics data correction

Image resampling

Voxel size resampling increased the mean percentage of robust features from 59.50 to 89.25% for pixel size and from

Table 1 Population description by tumor type (clinical application)

	Colorectal adenocarcinoma	Clear cell renal cancer	Total
<i>N</i> patients	24/43 (55.81%)	19/43 (44.19%)	43/43 (100%)
<i>N</i> patients Vall d’Hebron University Hospital	18/26 (69.23%)	8/26 (30.77%)	26/43 (60.47%)
<i>N</i> patients Bellvitge University Hospital	6/12 (50.00%)	6/12 (50.00%)	12/43 (27.90%)
<i>N</i> patients TCGA-KIRC	0/5 (0%)	5/5 (100%)	5/43 (11.63%)
<i>N</i> lesions	52/97 (53.61%)	45/97 (46.39%)	97/97 (100%)
<i>N</i> lesions Siemens	28/51 (54.90%)	23/51 (45.10%)	51/97 (52.58%)
<i>N</i> lesions GE	24/46 (52.17 %)	22/46 (47.83%)	46/97 (47.42%)

TCGA-KIRC The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma, GE General Electric

71.62 to 82.58% for slice spacing (Fig. 4). The percentage of robust radiomics features (ICC > 0.8) before and after resampling data to isometric voxels of $1 \times 1 \times 1 \text{ mm}^3$ are defined in Table 2.

Batch effect removal

Batch effect removal was implemented in all phantom materials based on the previously defined sources of variability (slice thickness and convolution kernel). For SVD-based batch correction, from the principal component analysis (PCA), PC2 significantly associated with convolution kernel ($p < 0.001$), and PC3 and PC4 significantly associated with slice thickness ($p < 0.001$); these principal components were removed at the transformed space, and the data matrix was back reconstructed to the feature space. The SVD-based correction technique increased the mean number of robust features (Table 3). Importantly, when analyzing images with different slice thicknesses, the mean number of robust radiomics features increased: from 82.79% without correction to 92.83% with SVD-based batch correction. When analyzing images restructured with different convolutional kernels, the mean

percentage of robust radiomics features increased: from 78.45% without correction to 85.25% with SVD-based correction (Table 3, Supplementary Material 8).

The ComBat correction technique increased the mean percentage of robust features to 95.34% for slice thickness and to 89.55% for convolution kernel when considering as batches the convolution kernel–slice thickness combination (Table 3, Supplementary Material 8).

Improvement in K-means clustering performance

To test the classification performance based on radiomics features and the potential improvement by reducing radiomics variability by batch correction, a classification of similar density tissues was performed.

Phantom application

The phantom application included liver and brain phantom tissues. SVD-based correction was applied, removing the PC2 (24.84%) and PC3 (13.53%), which associated significantly with slice thickness ($p < 0.001$), and PC5 (3.21%) that

Fig. 3 Intraclass correlation coefficients (ICCs) of the radiomics features of first order and texture matrices (gray-level co-occurrence matrix [GLCM], gray-level dependence matrix [GLDM], gray-level run length matrix [GLRLM], gray-level size zone matrix [GLSZM], neighboring gray-tone different matrix [NGTDM]) between extreme CT-acquisition parameters in the phantom application

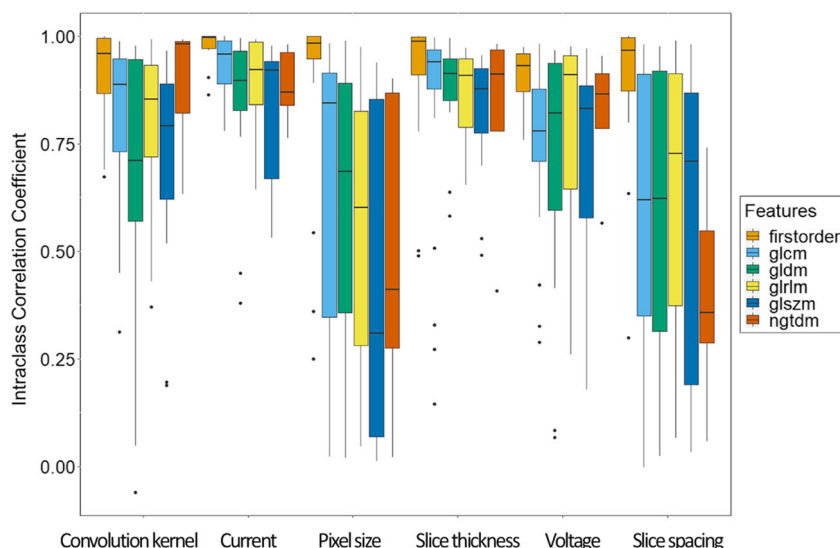


Table 2 Effect of voxel resampling to $1 \times 1 \times 1 \text{ mm}^3$ on the percentage of robust features when comparing radiomics data from phantom CT scans with different voxel sizes. The percentage of robust radiomics features was compared before and after correction using two-proportion z test (p value < 0.05 in italics)

% reproducible features (ICC > 0.8) comparing CT scans of the phantom with different acquisition parameters			
Pixel size (mm^2)	Non-resampled data	Pixel size resampled data $1 \times 1 \text{ mm}^2$	p value*
0.39×0.39 – 0.78×0.78	51.61 (48/93)	87.10 (81/93)	< 0.01
0.78×0.78 – 1×1	78.49 (73/93)	94.62 (88/93)	< 0.01
0.39×0.39 – 1×1	48.40 (45/93)	86.02 (80/93)	< 0.01
Slice spacing (mm)	Non-resampled data	Slice spacing resampled data 1 mm	p value
1–1.25	86.02 (80/93)	88.17 (82/93)	0.83
1.25–2	79.60 (74/93)	98.92 (92/93)	< 0.01
2–2.5	83.87 (78/93)	87.09 (81/93)	0.83
2.5–5	65.59 (61/93)	82.80 (77/93)	0.03
1–5	43.01 (40/93)	55.91 (52/93)	0.18

ICC intraclass correlation coefficient

*Adjustment for multiple testing was performed in each variability factor by controlling the false discovery rate according to the Benjamini and Hochberg method

associated significantly with convolution kernel ($p < 0.001$). ComBat was applied three times for batch correction, defined as convolution kernel, slice thickness, and convolution kernel–slice thickness combination. The K -means purity for tissue classification after SVD and ComBat batch correction is described in Table 4. Importantly, ComBat correction considering convolution kernel–slice thickness combination as batch effects showed the highest clustering purity (85.85%) (Fig. 5).

In the phantom study, the clustering performance from resampled images did not show improvement from non-resampled image data clustering (Supplementary Material 9).

Clinical application

In the clinical application of tumor type classification based on liver metastasis, the SVD was applied removing the PC1 (34.21%) that associated significantly with slice thickness

Fig. 4 Principal component analysis (PCA) before and after resampling to $1 \times 1 \times 1 \text{ mm}^3$ voxels of CT images acquired with different pixel sizes (a) and slice spacings (b). PC4 (explaining 7.14% of the radiomics data variance) is associated with the different acquisition pixel sizes before resampling. PC2 (18.42%) is associated with the distribution of the different acquisition pixel heights. After resampling, the acquisition voxel size (accounting for pixel size and slice spacing) is not associated with the variance explained by the PCA

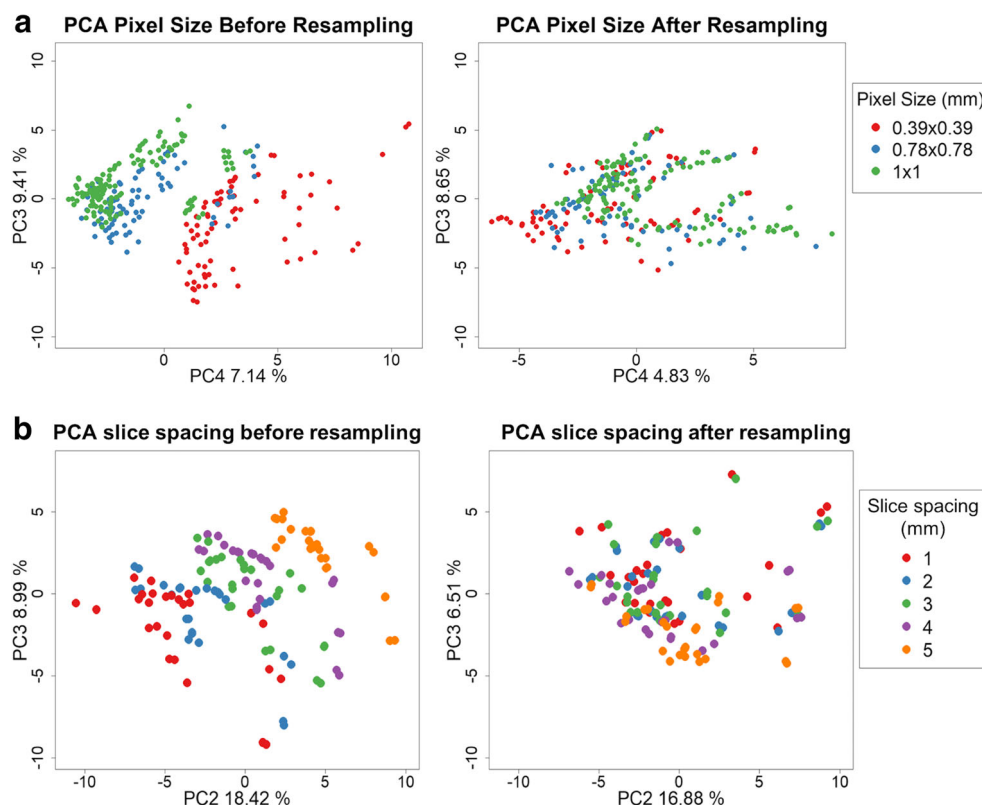


Table 3 Percentage of robust radiomics features (ICC > 0.8) without batch correction and after batch correction with singular value decomposition (SVD) and ComBat methods. The percentage of robustradiomics features was compared before and after correction using two-proportion z test (p value < 0.05 in italics)

% Reproducible features comparing CT scans of the phantom with different acquisition parameters							
Initial data		Batch effect correction methods					
		SVD Thickness–kernel	SVD Kernel	SVD Thickness	ComBat Thickness–kernel	ComBat Kernel	ComBat Thickness
Slice thickness (mm)							
2–3	88.17	97.85	96.77	98.92*	97.85	94.62	97.85
	<i>p</i> value**	0.06	0.16	0.01	0.026	0.28	0.02
3–5	84.95	93.55	88.17	96.77*	95.70	93.55	96.77*
	<i>p</i> value	0.10	1	0.01	0.026	0.28	0.02
2–5	75.27	87.10	75.27	90.32	92.47*	82.80	89.25
	<i>p</i> value	0.09	1	0.01	0.008	0.28	0.02
Convolution kernel							
A–B`	97.85	98.92*	97.85	98.92*	97.85	97.85	96.77
	<i>p</i> value	1	1	1	1	1	1
B–C	93.55	95.70	94.62	95.70	98.92*	96.77	92.47
	<i>p</i> value	0.87	1	1	0.14	0.58	1
C–D	72.04	81.72	78.49	78.49	86.02*	86.02*	73.12
	<i>p</i> value	0.38	0.62	1	0.07	0.05	1
D–E	79.57	86.02	84.95	82.80	90.32	91.40*	78.49
	<i>p</i> value	0.58	0.62	1	0.11	0.05	1
A–E	83.87	88.17	89.25	83.87	92.47	95.70*	82.79
	<i>p</i> value	0.74	0.3-90	1	0.14	0.04	1
B–D	62.37	75.27	73.11	64.51	81.72	83.87*	61.29
	<i>p</i> value	0.21	0.62	1	0.01	< 0.01	1
D–A	55.91	70.97	65.59	62.37	79.57	83.87*	59.13
	<i>p</i> value	0.21	0.62	1	0.01	< 0.01	0.767

ICC intraclass correlation coefficient, SVD singular value decomposition

*The highest increase of robust radiomics features

**Adjustment for multiple testing was performed in each variability factor and correction method by controlling the false discovery rate according to the Benjamini and Hochberg method

and convolution kernel ($p < 0.001$). Batch correction by ComBat was applied three times, considering as batches the convolution kernel, slice thickness, and convolution kernel–slice thickness combination.

The K -means improvement for tissue classification after SVD and ComBat batches correction are described in Table 4. Importantly, ComBat correction for the convolution kernel showed the best performance for primary tumor type

Table 4 K -means purity for phantom tissue (brain vs liver) and tumor type (colorectal carcinoma vs clear cell renal carcinoma) classification before and after batch correction

K -means purity							
	Initial data	SVD Thickness–kernel	SVD Kernel	SVD Thickness	Combat Thickness–kernel	Combat Kernel	Combat Thickness
Phantom	83.02	78.30	78.30	85.85*	85.85*	84.90	82.07
Tumor type	65.98	62.89	62.89	62.89	67.01	73.20*	67.01

SVD singular value decomposition

*The highest improvement of K -means purity classification

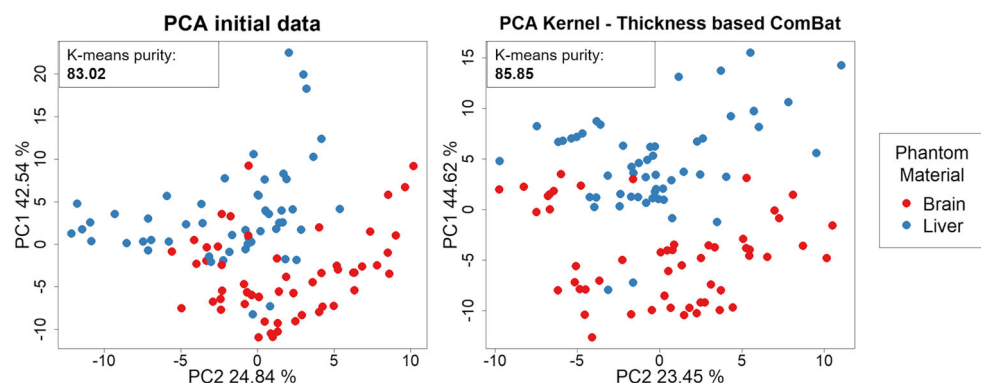


Fig. 5 Principal component analysis (PCA) of the brain and liver material radiomics distribution before and after convolution kernel–slice thickness ComBat correction. The distance between the radiomics data of the brain and liver materials from CT scans with different acquisitions protocols

increases after applying batch correction (i.e., the radiomics distribution better reflects differences between materials and not due to the CT-acquisition parameters)

classification based on radiomics data from liver metastasis (purity = 73.20%) (Fig. 6).

In the clinical application, the clustering performance from resampled images did not show improvement from non-resampled image data clustering (Supplementary Material 9).

Discussion

The capacity to extract a large amount of valuable quantitative data from medical images, such as CT, is revolutionizing the way medical scans can be evaluated. However, the development of reliable imaging biomarkers requires robust CT-based radiomics data. In this study, we defined the main sources of CT-radiomics variability based on a comprehensive phantom study with multiple CT-acquisition protocols. We also evaluated the influence of image resampling and the effect of radiomics data normalization by means of batch effect correction to reduce the variability and improve the tissue-classification capacity of radiomics in a phantom and clinical application.

We have shown that voxel size, convolution kernel, and slice thickness are relevant sources of variability. The voxel size has the highest impact on radiomics variability, particularly on texture features. Convolution kernel also affects first-order features, which are overall the most robust features regardless of the CT-acquisition protocol. Similarly to Berenguer et al [8], we also found that the radiomics features presented more variability when evaluated in CT scans acquired with low voltage values (90 kV). Importantly, the radiomics features were more robust when the voltage was within the range applied in standard clinical practice (i.e., 120–140 kV).

The pixel size varies in each scan and for each patient due to the changing field of view, limiting the possibility to pre-define this parameter. This study shows that image processing techniques regarding voxel resampling reduce the variability caused by the acquisition voxel size. A possible explanation for this variability decrease could be the resolution homogenization and the smoothness in gray-level transitions in the resampling direction. Therefore, in the *z*-direction, despite that we cannot restore the missing information in a large voxel size

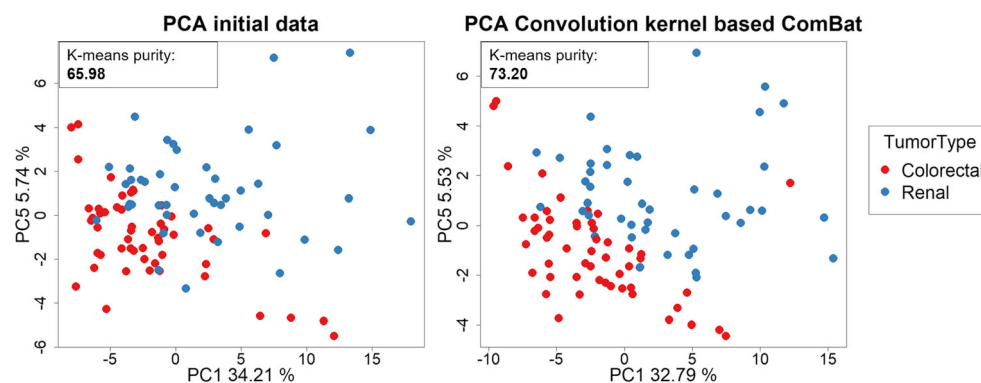


Fig. 6 Principal component analysis (PCA) of the liver metastasis radiomics distribution from CT scans of patients with colorectal adenocarcinoma and clear cell renal carcinoma. PCA before and after convolution kernel ComBat correction. The distribution of the groups

of patients with different tumor types differs more after batch correction. The first component (PC1 [%]) of data variance can differentiate better between groups (colorectal versus renal) after correction

(e.g., 5 mm), when resampling, the texture analysis considers the changes in gray levels inside the original voxel size and avoids abrupt gray-level changes to make them more comparable to the *x*- and *y*-directions.

In order to correct the variability caused by the reconstruction kernel and acquisition slice thickness, SVD-based and ComBat batch correction techniques were applied to radiomics data considering both image parameters as batches. In line with Orlhac et al [18], we demonstrate that after applying ComBat, the distribution of the data (using PCA) was modified to differentiate the phantom materials based on radiomics features. In our study, we also show that the reproducibility improves by means of ICCs. However, we aimed to study not only the variability correction by ComBat but also how the tissue-classification performance of radiomics improves after this variability correction. We have shown that ComBat correction for kernel and for kernel–slice thickness combinations in both the phantom and clinical applications outperforms the classification accuracy of radiomics data. The SVD-based correction improved the reproducibility of the radiomics features, although this could have suffered from overcorrection, leading to a loss of biological meaning and decreasing the tissue-clustering accuracy of radiomics.

Deep learning techniques have been developed to reduce radiomics variability by reconstructing images to the same convolution kernels [12]. However, the need for large data sets and the wide variety of intra- and inter-manufacturer reconstruction kernels limits the application of these techniques. ComBat correction can be applied in smaller datasets due to the non-parametric adjusting methods used to correct data variance associated to a particular factor.

The results of our study are promising, but we acknowledge some limitations. First, we implemented a variability correction method in both phantom and clinical applications. The phantom was used to assess the intra-scanner variability from one CT vendor while the clinical application analyzed the inter-scanner variability for ComBat and SVD-based correction. Further studies with intra- and inter-manufacturer CT scans could be performed to extend the application of batch correction methods. Moreover, there are several reconstruction kernels along manufacturers that could be considered comparable, as proposed by Mackin et al [25]. This would reduce the inter-manufacturer variability and would facilitate the definition of batches based on inter-manufacturer similar kernels for large-scale multicenter studies. Second, the radiomics variability correction was clinically tested as a method to improve the tumor type classification. Although this highlights the impact of post-acquisition CT-radiomics normalization by means of batch correction, further applications need to be tested and validated in larger populations. Finally, the described normalization methods have been tested in CT images; it is of interest to test these in multi-image modalities including MRI.

In conclusion, the main sources of CT-radiomics variability are slice thickness and reconstruction kernels. The application of image post-processing and the ComBat correction method minimizes radiomics data variability regardless of the differences in the CT-image-acquisition protocols. These methods are easy to apply to expand the potential of radiomics implementation in new retrospective and prospective multicenter large-scale studies where the variability of the acquisition protocols and scanners is the major limitation.

Funding This study was supported by the Banco Bilbao Vizcaya Argentaria and Fundacio La Caixa. Raquel Perez-Lopez is supported by Prostate Cancer Foundation Young Investigator Award, CRIS Foundation Talent Award (TALENT-05), Fero Foundation, and the Instituto de Salud Carlos III-Investigación en Salud (PI18/01395).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Raquel Perez-Lopez, principal investigator of the radiomics group at Vall d'Hebron Institut of Oncology, Barcelona, Spain

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry Guillermo Villacampa kindly provided statistical advice for this manuscript.

Informed consent Written informed consent was waived by the institutional review board.

Ethical approval Institutional review board approval was obtained.

Methodology

- prospective/retrospective
- experimental
- multicenter study

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 14:749–762

2. Rizzo S, Botta F, Raimondi S et al (2018) Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2:36
3. Sun R, Limkin EJ, Vakalopoulou M et al (2018) A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* 19:1180–1191
4. Trebeschi S, Drago SG, Birkbak NJ et al (2019) Predicting response to cancer immunotherapy using non-invasive radiomic biomarkers. *Ann Oncol*. <https://doi.org/10.1093/annonc/mdz108>
5. Aerts HJWL, Velazquez ER, Leijenaar RT et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 5:4006
6. Mackin D, Fave X, Zhang L et al (2015) Measuring computed tomography scanner variability of radiomics features. *Invest Radiol* 50:757–765
7. Pavic M, Bogowicz M, Wurms X et al (2018) Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol* 57:1070–1074
8. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J et al (2018) Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology* 288:407–415
9. Kim H, Park CM, Lee M et al (2016) Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One* 11:e0164924
10. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts H (2018) Data analysis strategies in medical imaging. *Clin Cancer Res* 24:3492–3499
11. Lee SH, Cho HH, Lee HY, Park H (2019) Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. *Cancer Imaging* 19:54
12. Choe J, Lee SM, Do KH et al (2019) Deep learning-based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology* 292:365–373
13. Lazar C, Meganck S, Taminiau J et al (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 14:469–490
14. Luo J, Schumacher M, Scherer A et al (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J* 10:278–291
15. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97:10101–10106
16. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–127
17. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28:882–883
18. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology* 291:53–59
19. Ortiz-Ramon R, Larroza A, Ruiz-Espana S, Arana E, Moratal D (2018) Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study. *Eur Radiol* 28:4514–4523
20. Akin O, Elnajjar P, Heller M et al (2016) Radiology data from The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma [TCGA-KIRC] collection. The Cancer Imaging Archive
21. Fedorov A, Beichel R, Kalpathy-Cramer J et al (2012) 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 30:1323–1341
22. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 77:e104–e107
23. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15:155–163
24. Rathore S, Akbari H, Rozycki M et al (2018) Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci Rep* 8:5087
25. Mackin D, Ger R, Gay S et al (2019) Matching and homogenizing convolution kernels for quantitative studies in computed tomography. *Invest Radiol* 54:288–295

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4.2 Study 2

A CT-based radiomics signature is associated with response to Immune Checkpoint Inhibitors in advanced solid tumors

This study corresponds to **objective 2**. In this research, we explore the implementation of CT-based radiomic features to develop predictive signatures of response to immunotherapy in a population of advanced cancer patients with different tumor types and validate it in two separate cohorts of bladder and lung cancer.

Reference:

Ligero, M., Garcia-Ruiz, A., Viaplana, C., et al. A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors. Radiology. Vol 299, No 1 (2021).

A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors

Marta Ligeró, MSc • Alonso García-Ruiz, MSc • Cristina Viaplana, MSc • Guillermo Villacampa, MSc • Maria V. Raciti, MD • Jaid Landa, MD • Ignacio Matos, MD • Juan Martín-Liberal, MD, PhD • Maria Ochoa-de-Olza, MD • Cinta Hierro, MD • Joaquín Mateo, MD, PhD • Macarena González, MD • Rafael Morales-Barreira, MD • Cristina Suárez, MD • Jordi Rodon, MD, PhD • Elena Elez, MD, PhD • Irene Braña, MD, PhD • Eva Muñoz-Couselo, MD • Ana Oaknin, MD • Roberta Fasani, MD • Paolo Nuciforo, MD, PhD • Debora Gil, PhD • Carlota Rubio-Perez, PhD • Joan Seoane, PhD • Enriqueta Felip, MD, PhD • Manuel Escobar, MD • Josep Tabernero, MD, PhD • Joan Carles, MD • Rodrigo Dienstmann, MD • Elena Garralda, MD • Raquel Pérez-López, MD, PhD

From the Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Cellex Center, Natzarret 115-117, Barcelona 08035, Spain (M.L., A.G.R., R.P.L.); Oncology Data Science (ODysSeY) Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain (C.V., G.V., R.D.); Institute of Radiology, Foundation IRCCS Policlinic San Matteo, Pavia, Italy (M.V.R.); Department of Radiology, Bellvitge University Hospital, L'Hospitalet de Llobregat, Spain (J.L.); Department of Medical Oncology, Vall d'Hebron University Hospital and Institute of Oncology (VHIO), Barcelona, Spain (I.M., J.M.L., M.O.d.O., C.H., J.M., M.G., R.M.B., C.S., J.R., E.E., I.B., E.M.C., A.O., E.F., J.T., J.C., E.G.); Department of Molecular Oncology, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain (R.F., P.N.); Computer Vision Center, Department of Computer Science, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain (D.G.); Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron University Hospital (HUVH), Autonomous University of Barcelona, Institució Catalana de Recerca i Estudis Avançats (ICREA) and CIBERONC, Barcelona, Spain (C.R.P., J.S.); Department of Radiology, Vall d'Hebron University Hospital, Barcelona, Spain (M.E., R.P.L.); and Department of Medicine, Autonomous University of Barcelona (UAB), Barcelona, Spain (J.T.). Received March 9, 2020; revision requested April 6; revision received November 2; accepted November 11. **Address correspondence to** R.P.L. (e-mail: rperez@vhio.net).

Study supported by the Banco Bilbao Vizcaya Argentaria, Fundació La Caixa (RTI2018-095209-B-C21, FIS-G64384969), Generalitat de Catalunya (2017-SGR-1624), and CERCA-Programme. J.M. and R.P.L. supported by Prostate Cancer Foundation Young Investigator Awards. R.P.L. supported by a CRIS Foundation Talent Award (TALENT19-05) and the Instituto de Salud Carlos III-Investigación en Salud (PI18/01395). D.G. is supported by Serra Hunter Fellowship.

Conflicts of interest are listed at the end of this article.

See also the editorial by Summers in this issue.

Radiology 2021; 299:109–119 • <https://doi.org/10.1148/radiol.2021200928> • Content codes:  

Background: Reliable predictive imaging markers of response to immune checkpoint inhibitors are needed.

Purpose: To develop and validate a pretreatment CT-based radiomics signature to predict response to immune checkpoint inhibitors in advanced solid tumors.

Materials and Methods: In this retrospective study, a radiomics signature was developed in patients with advanced solid tumors (including breast, cervix, gastrointestinal) treated with anti-programmed cell death-1 or programmed cell death ligand-1 monotherapy from August 2012 to May 2018 (cohort 1). This was tested in patients with bladder and lung cancer (cohorts 2 and 3). Radiomics variables were extracted from all metastases delineated at pretreatment CT and selected by using an elastic-net model. A regression model combined radiomics and clinical variables with response as the end point. Biologic validation of the radiomics score with RNA profiling of cytotoxic cells (cohort 4) was assessed with Mann-Whitney analysis.

Results: The radiomics signature was developed in 85 patients (cohort 1: mean age, 58 years $\pm$ 13 [standard deviation]; 43 men) and tested on 46 patients (cohort 2: mean age, 70 years $\pm$ 12; 37 men) and 47 patients (cohort 3: mean age, 64 years $\pm$ 11; 40 men). Biologic validation was performed in a further cohort of 20 patients (cohort 4: mean age, 60 years $\pm$ 13; 14 men). The radiomics signature was associated with clinical response to immune checkpoint inhibitors (area under the curve [AUC], 0.70; 95% CI: 0.64, 0.77; $P < .001$). In cohorts 2 and 3, the AUC was 0.67 (95% CI: 0.58, 0.76) and 0.67 (95% CI: 0.56, 0.77; $P < .001$), respectively. A radiomics-clinical signature (including baseline albumin level and lymphocyte count) improved on radiomics-only performance (AUC, 0.74 [95% CI: 0.63, 0.84; $P < .001$]; Akaike information criterion, 107.00 and 109.90, respectively).

Conclusion: A pretreatment CT-based radiomics signature is associated with response to immune checkpoint inhibitors, likely reflecting the tumor immunophenotype.

© RSNA, 2021

Online supplemental material is available for this article.

Cancer immunotherapy is one of the most promising anticancer treatment strategies. The landmark example of immunotherapy success was the development of immune checkpoint inhibitors (programmed cell death-1 [PD-1] and programmed cell death ligand-1 [PD-L1]). Several monoclonal anti-PD-1 and anti-PD-L1

antibodies, such as pembrolizumab, nivolumab, and atezolizumab, are now approved for the treatment of different tumor types, including urinary tract and lung cancer (1–10). Despite the excellent results of immune checkpoint inhibitor trials, some caveats remain about optimal patient stratification (11,12). Although some

Abbreviations

AUC = area under the curve, PD-1 = programmed cell death protein-1, PD-L1 = programmed cell death protein ligand-1, RECIST = Response Evaluation Criteria in Solid Tumors

Summary

A CT-based radiomics signature, based on histogram, shape, and local-regional texture features, showed an association with immune response in advanced solid tumors.

Key Results

- In this retrospective study of 85 patients, a CT-based 14-feature radiomics signature (histogram, shape, and local-regional texture) was associated with response to immune checkpoint inhibitors in advanced solid tumors (area under curve [AUC], 0.70; $P < .001$; sensitivity, 75%; specificity, 53%).
- Radiomics combined with clinical data (baseline albumin and lymphocytes) improved response classification performance compared with radiomics or clinical data alone (AUC, 0.74 vs 0.70 and 0.64, respectively).

patients achieve dramatic responses, others do not respond at all. In addition, immune checkpoint inhibitors can cause serious side effects in some patients (13).

Further immunotherapy agents are being tested in combination with PD-1/PD-L1 inhibitors to improve the response of immunotherapy-sensitive tumors and to convert immunotherapy-insensitive tumors into responsive tumors (14). Given this, it is critical to optimally identify patients more likely to respond to immune checkpoint inhibitors to maximize patient benefit, minimize risks, and accelerate drug development. Currently, tumor PD-L1 expression, microsatellite instability status, and tumor mutational burden are imperfect predictive biomarkers of response to anti-PD-1/PD-L1 antibodies, with varying performance across different tumor types (15,16).

Radiomics derives quantitative variables from imaging data of tumor tissues by the application of data mining algorithms to medical images (17–21). Much of the discussion about personalized medicine has focused on molecular characterization by using genomic and proteomic technologies. However, these often require biopsies or invasive operations to extract and analyze small portions of

tumor tissue, which do not allow for complete tumor characterization. Tumors are spatially and temporally heterogeneous (22,23), giving imaging an advantage in guiding personalized treatment because it yields a comprehensive view of the entire tumor. Radiomics can provide noninvasive tools to study tumor biology, capture tumor heterogeneity, and monitor tumor evolution and response to therapy.

In this study, we aimed to develop and test a radiomics signature from baseline (pretreatment) CT scans to predict response to anti-PD-1/PD-L1 in patients with advanced solid tumors. For a biologic validation of this radiomics signature, associations with tumor-infiltrating lymphocytes and RNA profiling of cytotoxic cells were performed. We also explored the integration of CT-based radiomics and clinical data into a multidimensional signature for response.

Materials and Methods

The institutional review board approved this retrospective study. All patients included in the clinical trials provided written informed consent. Need for informed consent for the computational analysis of the images was waived. Clinical trials included in the study are listed in Table E1 (online).

Study Sample

Cohort 1 (training set) included patients treated with anti-PD-1/PD-L1 monotherapy in phase I clinical trials from August 2012 to May 2018. Cohort 2 (test set 1) was composed of patients with bladder cancer treated with anti-PD-1/PD-L1 from June 2013 to May 2018. Cohort 3 (test set 2) was composed of patients with lung cancer treated with anti-PD-1/PD-L1 from December 2013 to April 2016. Cohort 4 (biologic validation set) was

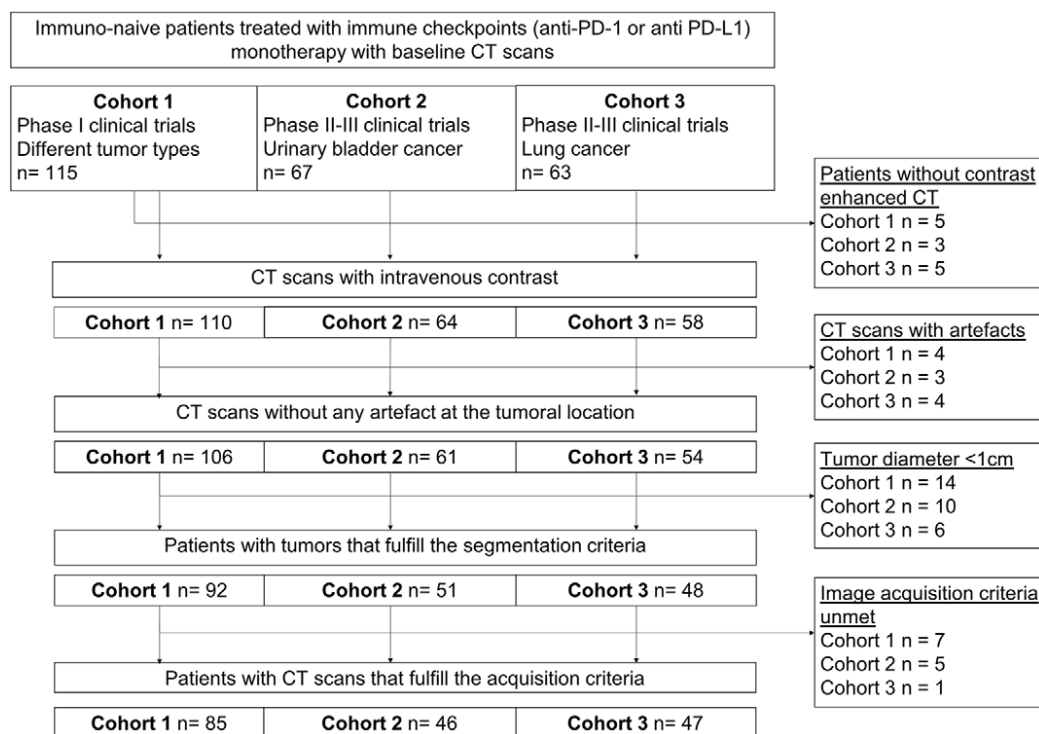


Figure 1: Flowchart shows inclusion/exclusion for all three patient cohorts. PD-1 = programmed cell death protein-1, PD-L1 = programmed cell death protein ligand-1.

composed of patients with different tumor types and gene expression profiling from solid tumor biopsy samples where contrast-enhanced CT was performed within 4 months before biopsy. None of these patients had been included in cohorts 1, 2, or 3. All patients were treated at the Vall d'Hebron Institute of Oncology (Barcelona, Spain).

Exclusion criteria for this study comprised patients who had received previous immunotherapies, nonenhanced CT, CT with artifacts, non-measurable lesions per Response Evaluation Criteria in Solid Tumors (RECIST) 1.1, and lesions with cystic changes or cavitation. We also excluded patients in whom the biopsy had been performed before the CT to avoid biopsy-related changes. The study flowchart is shown in Figure 1.

Image Data Acquisition and Analysis

CT scans obtained within 28 days before the treatment start date were analyzed. Contrast-enhanced CT scans were acquired by using 16- or 64-channel CT scanners (Siemens, Philips, and GE Healthcare) in the axial plane with tube voltage of 100–130 kV, section thickness of 2–5 mm, and section interval of 1.25–5 mm (Table E2 [online]). All well-defined lesions per patient were delineated by a radiologist (R.P.L., with 10 years of experience in oncologic imaging), blinded to the clinical outcome, by using the 3DSlicer version 4.8.1 (24) semiautomatic contouring tool.

Images were resampled to isometric voxels of $1 \times 1 \times 1 \text{ mm}^3$ by using spline interpolation. Hounsfield units were binarized to discrete values of 10 HU (25). CT-based radiomics features (first order, shape, and texture) were derived by using the Pyradiomics package (version 2.2.0) for Python (version 3.7.1), compliant with the Image Biomarker Standardization Initiative guidelines (26,27) (Table

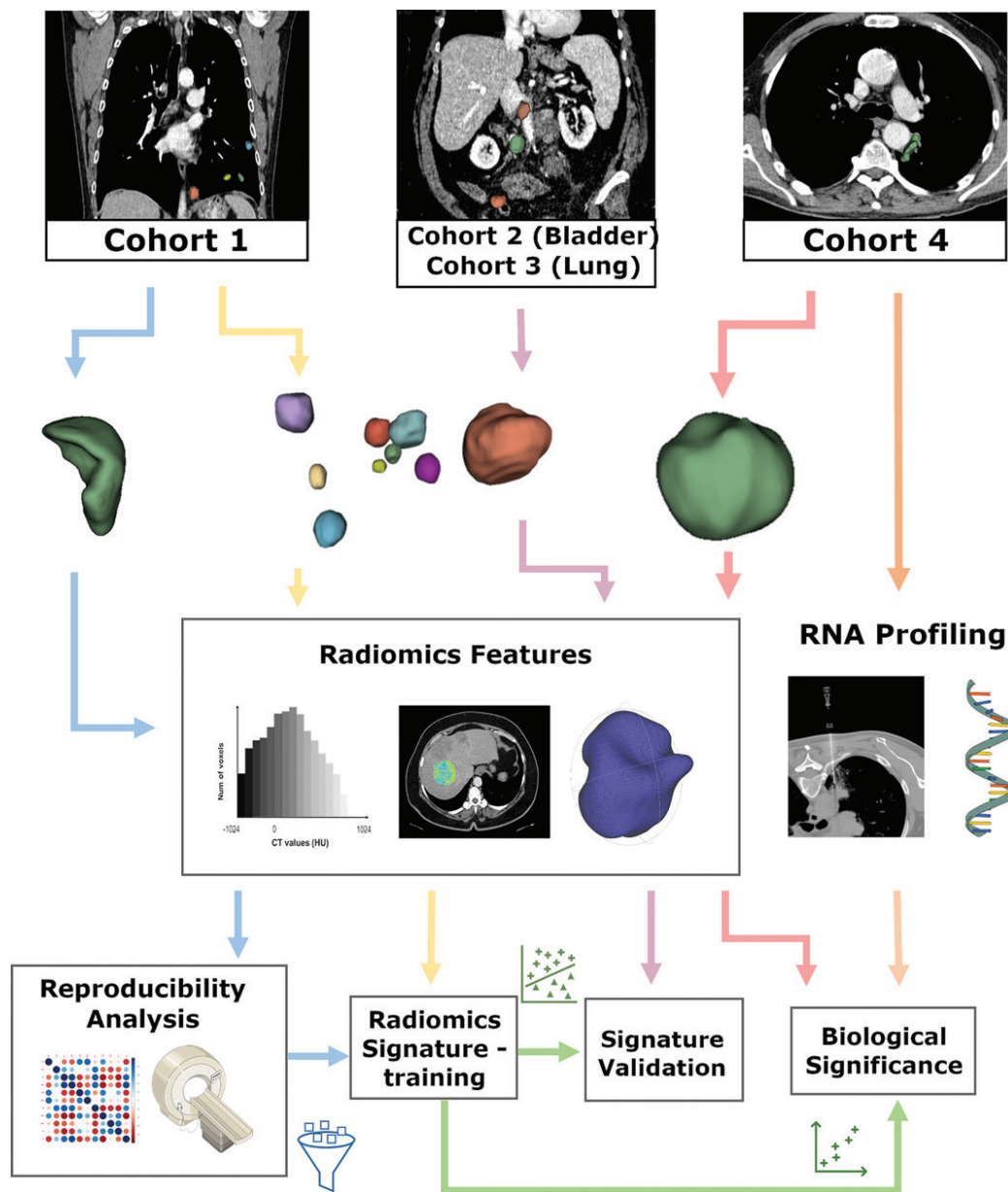


Figure 2: Schema shows radiomics workflow. The model was trained in cohort 1. Features were selected according to the reproducibility analysis. Then, the signature was tested in cohorts 2 and 3. Finally, a biologic validation was implemented by using RNA profiling in cohort 4.

E3 [online]). Radiomics data were standardized to zero-centered distributions with logarithmic transformations.

Clinical Data

Clinical and laboratory data were collected from electronic patient records. These included age, sex, tumor type, Eastern Cooperative Oncology Group performance status, number of organs with metastasis, lactate dehydrogenase level, leukocyte count, albumin level, lymphocyte count, platelet count, and neutrophil count at baseline. Outcome data included clinical benefit and overall survival. Clinical benefit was defined as stable disease after 5 months of treatment or partial or complete response within the first 5 months of treatment according to RECIST 1.1 (28,29).

Table 1: Study Sample Characteristics

Parameter	Cohort 1 (n = 85)	Cohort 2 (n = 46)	Cohort 3 (n = 47)	Cohort 4 (n = 20)
Mean age (y)*†	58 ± 13 (29–85)	70 ± 12 (44–91)	64 ± 11 (39–88)	60 ± 14 (25–85)
Sex				
Female	42 (49)	9 (20)	7 (15)	6 (30)
Male	43 (51)	37 (80)	40 (85)	14 (70)
Primary tumor type				
Breast	14 (16)	...	...	...
Cervix	10 (12)	...	...	...
Other	10 (12)	...	...	1 (5)
Gastrointestinal	9 (11)	...	...	...
Melanoma	8 (9)	...	...	1 (5)
Thyroid	6 (7)	...	...	...
Lung	6 (7)	...	47 (100)	15 (75)
Bladder	6 (7)	46 (100)	...	...
Head and neck	5 (6)	...	...	...
Ovary	3 (4)	...	...	3 (15)
Pleura	2 (2)	...	...	...
Thymus	2 (2)	...	...	...
Kidney	2 (2)	...	...	...
Adrenal	2 (2)	...	...	...
Treatment‡				
PD-1 immune checkpoint inhibitor	38 (45)	21 (46)	35 (75)	...
PD-L1 immune checkpoint inhibitor	47 (55)	25 (54)	12 (25)	...
Best response				
Complete response	5 (6)	7 (15)	2 (4)	...
Partial response	7 (8)	15 (33)	11 (24)	...
Stable disease	41 (48)	6 (13)	17 (36)	...
Progressive disease	32 (38)	18 (39)	17 (36)	...
Lesion location				
Node	86 (36)	66 (47)	46 (44)	4 (20)
Liver	53 (22)	30 (21)	22 (21)	1 (5)
Lung	60 (25)	28 (20)	26 (25)	11 (55)
Other	37 (16)	16 (12)	10 (10)	4 (20)
No. of lesions	236	140	104	20
Mean lesion size (mm)*	35 (10–139)	35 (10–259)	39 (13–96)	40 (14–99)
Mean no. of lesions per patient*	2.77 (1–17)	3.04 (1–11)	2.21 (1–10)	1

Note.—Data in parentheses are percentages unless otherwise indicated. Cohort 1 is training, cohort 2 is bladder test set, cohort 3 is lung test set, and cohort 4 is biologic validation. PD-1 = programmed cell death protein–1, PD-L1 = programmed cell death protein ligand–1.

*Data in parentheses are range.

†Data are ± standard deviation.

‡Atezolizumab, avelumab, durvalumab, retifanlimab, pembrolizumab, nivolumab, cemiplimab, PDR001, FAZ053.

Patient follow-up was considered as the time between treatment baseline (treatment start date) and discontinuation.

Exploring Biologic Significance of CT-based Radiomics Signature

Assessment and immunophenotyping of tumor-infiltrating lymphocytes by NanoString (NanoString Technologies) gene expression profiling of tumor samples were performed at the Vall d'Hebron Institute of Oncology (30). The CT-based radiomics score and the enrichment score of cytotoxic cells were calculated from the same biopsied lesion. The enrichment score was computed as the mean of the log-2 z-score of the transformed expression values of the gene set, after background

thresholding and housekeeping adjustment. The cytotoxic cell gene set included *CD8A*, *GZMA*, *GZMB*, and *PRF1* genes. Methods for assessment of tumor-infiltrating lymphocytes and radiomics–tumor-infiltrating lymphocyte associations are described in Appendix E1 (online).

Statistical Analysis

Test-retest reproducibility based on the acquisition protocols and interobserver reproducibility were assessed by using intra-class correlation coefficients (31) (Appendix E1 [online]). An elastic-net model was implemented for feature selection in both radiomics and clinical-radiomics signature. The regularization parameters λ and α were determined by using cross validation.

Table 2: CT-based Radiomics Signature and Radiomics-Clinical Signature, Variables, and Coefficients

Variable	β Model Coefficient
CT-based radiomics signature	
First-order entropy	11.73 (−4.14, 27.6)
First-order total energy	4.18 (−0.78, 9.13)
First-order uniformity	5.56 (−4.11, 15.22)
GLCM inverse difference moment	−4.66 (−18.37, 9.06)
GLCM joint entropy	−4.67 (−14.18, 4.84)
GLDM dependence nonuniformity normalized	−9.24 (−21.23, 2.75)
GLDM dependence variance	−6.9 (−15.62, 1.81)
GLDM large dependence emphasis	3.82 (−6.1, 13.73)
Elongation	0.54 (−9.04, 10.13)
Least axis length*	0.62 (−7.55, 8.78)
Major axis length*	3.16 (−23.96, 30.29)
Maximum 2D diameter column	0.4 (−5.76, 6.56)
Minor axis length*	−7.53 (−35.08, 20.01)
Surface area	−1.66 (−26.27, 22.95)
Liver	−1.77 (−7.21, 3.67)
Other	5.33 (−1.38, 12.05)
CT-based radiomics and clinical signature	
Mean radiomics score	2.06 (0.64, 3.48)
Baseline lymphocytes ($\times 10^9/L$)	0.49 (−0.16, 1.14)
Baseline albumin level (g/dL)	0.39 (−0.1, 0.88)

Note.—Data in parentheses are 95% CIs. GLCM = gray level co-occurrence matrix, GLDM = gray level dependence matrix, 2D = two-dimensional.

*Major axis length defined as the largest axis length of the region-of-interest–enclosing ellipsoid; minor axis length defined as the second-largest axis length of the region-of-interest–enclosing ellipsoid; least axis length defined as the smallest axis length of the region-of-interest–enclosing ellipsoid.

The radiomics signature was implemented by using generalized linear mixed models, including robust radiomics features (defined as intraclass correlation coefficient ≥ 0.7) selected by elastic net, lesion location (node, liver, lung, and others) and with consideration of patient identification as a random effect to adjust for the inpatient variability for lesions from the same patient. The end point of the model was clinical benefit.

The clinical-radiomics signature was developed by using generalized linear models that included mean radiomics score per patient and clinical data with clinical benefit as the end point. The mean radiomics score per patient was derived from the probability (p) defined as follows:

$$\ln\left(\frac{p}{1-p}\right) = X\beta + Zu \quad (1),$$

$$Rad_{score}_{pat} = \sum_{i=1}^{N_{lesions}} \frac{Rad_{score}_i}{N_{lesions}} \quad (2),$$

where X is a matrix of $N_{pat} \times p$ of p predictor variables, N_{pat} is number of patients, and β is a $p \times 1$ vector of the fixed-effects

coefficients. Z is the $N_{pat} \times q$ matrix for q random effects, and u is a $q \times 1$ vector of the random-effect coefficients. The radiomics score per patient (Rad_{score}_{pat}) is the average of all the lesion radiomics scores in a patient as defined in Equation (2) where $N_{lesions}$ is the number of lesions and Rad_{score}_i is the radiomic score per lesion.

The area under the curve (AUC) and 95% CI (DeLong method) were determined from the receiver operating characteristic curve, and the P value was assessed by using the Mann-Whitney U test. The decision threshold to determine the sensitivity and specificity was defined by maximizing the sensitivity given a minimal value of specificity of 50%.

The model performances obtained by using radiomics predictive score only and by combining radiomics score with clinical data were compared by using Akaike information criterion. Hosmer-Lemeshow goodness-of-fit test was reported as the calibration statistic. The radiomics signature was internally validated with a stratified 10-fold cross-validation analysis (Appendix E1 [online]).

A patient-level analysis was performed by using logistic regression to obtain odds ratios.

A Cox proportional hazard regression model was used to associate overall survival with the radiomics and radiomics-clinical score. The proportionality of hazard assumptions was assessed with Schoenfeld residuals. Univariate and multivariate analyses were performed for each variable from the radiomics-clinical signature. Overall survival analysis was performed by using the Kaplan–Meier method after score dichotomization. Patients who did not reach the end point were censored to the last follow-up date. The Harrell C-index was assessed to compare both scoring systems.

Associations between radiomics score with tumor-infiltrating lymphocytes and cytotoxic immunophenotype enrichment were evaluated by using Mann-Whitney analysis. The P value α level was considered to indicate significance at .05.

Statistical analyses were performed by using R software, version 3.5.0. The R packages used are listed in Table E4 (online). The code can be found at <https://github.com/mligerohe/ImmunoCTrad>. The study workflow is summarized in Figure 2.

Results

Study Patient Characteristics

Of 115 patients from cohort 1, 67 patients from cohort 2, and 63 patients from cohort 3, we excluded 13 because of non-contrast CT scans, 11 because of artifacts at the tumor site, 30 without measurable lesions per RECIST 1.1, and 13 whose scans did not fulfill the image acquisition criteria (Fig 1).

In cohort 1, a total of 236 lesions (mean number of lesions per patient, three; range, one to 17 lesions per patient) from the 85 patients (mean age, 58 years $\pm$ 13 [standard deviation]; 42 women [mean age, 55 years $\pm$ 12] and 43 men [mean age, 62 years $\pm$ 13]) were included in the final analysis. The median follow-up period was 4 months (interquartile range, 2–12 months); 51% (43 of 85) of the patients showed a clinical benefit and 49% (42 of 85) of the patients did not show a complete response, partial response, or stable disease after 5 months of treatment.

In cohort 2, a total of 140 lesions (mean number of lesions per patient, three; range, one to 11 lesions per patient) from the 46 patients (mean age, 70 years $\pm$ 12; nine women [mean

age, 70 years $\pm$ 9] and 37 men [mean age, 70 years $\pm$ 12]) were included in the final analysis. The median follow-up time was 4 months (interquartile range, 2–10 months); 54% (25 of 46)

of the patients showed a clinical benefit, and 46% (21 of 46) did not.

In cohort 3, a total of 104 lesions (mean number of lesions per patient, two; range, one to 10 lesions per patient) from the 47 patients (mean age, 64 years $\pm$ 11; seven women [mean age, 55 years $\pm$ 6] and 40 men [mean age, 66 years $\pm$ 11]) were included in the final analysis. The median follow-up time was 5 months (interquartile range, 2–10 months);

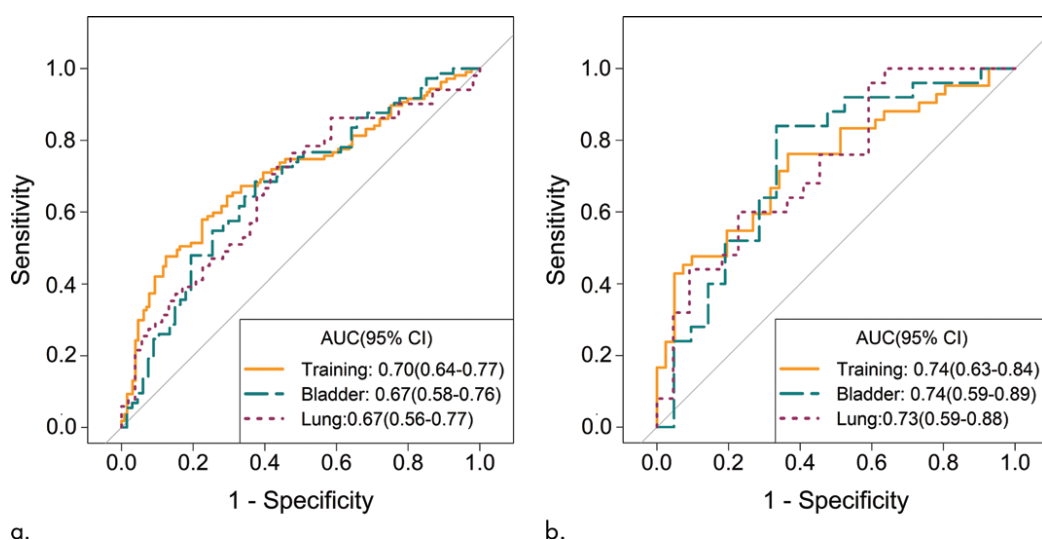


Figure 3: Receiver operating characteristic curves of **(a)** radiomics and **(b)** radiomics-clinical signature for predicting response to immune checkpoint inhibitors in cohort 1 (training set), cohort 2 (bladder test set), and cohort 3 (lung test set). AUC = area under the curve.

Table 3: Radiomics, Radiomics-Clinical, and Clinical-only Signature Performance

Cohort	AUC	P Value	Sensitivity (%)	Specificity (%)	NPV (%)	PPV (%)	Positive LR	Negative LR
Radiomics signature*								
Cohort 1	0.70 [0.64, 0.77]	<.001	75 (80/107) [65, 83]	53 (68/129) [44, 62]	72 (68/95) [61, 80]	57 (80/141) [48, 65]	1.58 [1.28, 1.96]	0.48 [0.33, 0.69]
Cohort 2	0.67 [0.58, 0.76]	<.001	73 (53/73) [61, 82]	55 (37/67) [43, 67]	65 (37/57) [51, 77]	64 (53/83) [53, 74]	1.62 [1.2, 2.19]	0.5 [0.32, 0.76]
Cohort 3	0.67 [0.56, 0.77]	<.001	73 (37/51) [58, 84]	57 (30/53) [42, 70]	68 (30/44) [52, 81]	62 (37/60) [48, 74]	1.67 [1.18, 2.37]	0.48 [0.29, 0.8]
Radiomics-clinical signature†								
Cohort 1	0.74 [0.63, 0.84]	<.001	77 (33/43) [61, 88]	59 (25/42) [42, 74]	71 (25/35) [53, 85]	66 (33/50) [50, 78]	1.87 [1.23, 2.74]	0.38 [0.22, 0.74]
Cohort 2	0.74 [0.59, 0.89]	<.001	84 (21/25) [64, 95]	67 (14/21) [43, 85]	78 (14/18) [52, 94]	75 (21/28) [55, 89]	2.52 [1.34, 4.72]	0.24 [0.09, 0.62]
Cohort 3	0.73 [0.59, 0.88]	<.001	68 (17/25) [46, 85]	55 (12/22) [32, 76]	60 (12/20) [36, 81]	63 (17/27) [42, 81]	1.5 [0.88, 2.54]	0.59 [0.3, 1.17]
Clinical signature‡								
Cohort 1	0.64 [0.52, 0.76]	.01	70 (30/43) [53, 82]	55 (23/42) [37, 69]	66 (23/35) [45, 79]	61 (30/49) [45, 74]	1.50 [1.01, 2.19]	0.55 [0.34, 0.98]
Cohort 2	0.67 [0.5, 0.84]	.02	84 (21/25) [64, 95]	48 (10/21) [26, 70]	71 (10/14) [42, 92]	66 (21/32) [47, 81]	1.6 [1.03, 2.5]	0.34 [0.12, 0.92]
Cohort 3	0.65 [0.49, 0.81]	.04	76 (19/25) [55, 91]	50 (11/22) [28, 72]	65 (11/17) [38, 86]	63 (19/30) [44, 80]	1.52 [0.95, 2.44]	0.48 [0.21, 1.08]

Note.—Data in parentheses are numerator/denominator; data in brackets are 95% CIs. LR = likelihood ratio, NPV = negative predictive value, PPV = positive predictive value.

*Radiomics signature including 14 variables from histogram, shape, and local-regional texture.

†Radiomics-clinical signature including mean radiomics score per patient, albumin level, and lymphocyte count.

‡Clinical signature including albumin level and lymphocyte counts.

53% (25 of 47) of the patients showed a clinical benefit, and 47% (22 of 47) did not.

In cohort 4, 20 patients (mean age, 60 years $\pm$ 13; six women [mean age, 55 years $\pm$ 9] and 14 men [mean age, 63 years $\pm$ 15]) were included in the final analysis for biologic validation. One lesion per patient was delineated, corresponding to the biopsied tumor in which the RNA profiling was performed. The study sample characteristics are summarized in Table 1.

Predictive Model Development and Testing

Twenty-six radiomics features showed an intraclass correlation coefficient of 0.7 or greater in the reproducibility analysis (Appendix E1, Table E5 [online]). Elastic-net parameters were set to λ of 0.001 and α of 1 for the radiomics signature, and 14 from 26 variables were retained (Table 2).

In terms of classification performance, the training set had an AUC of 0.70 (95% CI: 0.64, 0.77; $P < .001$), the bladder test set had an AUC of 0.67 (95% CI: 0.58, 0.76; $P < .001$), and the lung test set had an AUC of 0.67 (95% CI: 0.56, 0.77; $P < .001$) (Fig 3a). Sensitivity, specificity, negative predictive value, and positive predictive value for identifying patients with clinical benefit with an optimal cutoff of 0.29 are summarized in Table 3. The radiomics signature showed no evidence of poor fit (Hosmer-Lemeshow, $P > .05$).

Integrating Radiomics and Clinical Data

The radiomics-clinical elastic-net feature selection retained with nonzero coefficients (elastic-net to $\lambda = 0.08$, $\alpha = .9$) the mean radiomics score per patient and two clinical features (albumin and lymphocytes; Table 2).

The radiomics-clinical model showed an AUC of 0.74 (95% CI: 0.63, 0.84; $P < .001$) in the training set and AUCs of 0.74 (95% CI: 0.59, 0.89; $P < .001$) and 0.73 (95% CI: 0.59, 0.88; $P = .002$) for predicting clinical benefit in cohorts 2 and 3, respectively. Sensitivity, specificity, negative predictive value, and positive predictive value for identifying patients with clinical benefit with an optimal cutoff of 0.45 are summarized in Table 3. The radiomics-clinical model showed no evidence of poor fit (Hosmer-Lemeshow, $P > .05$). The radiomics-clinical signature presented a better performance compared with the radiomics-only signature (Akaike information criterion, 107.00 and 109.90, respectively).

Clinical signature performance is reported in Table 3 and Figure E1 (online).

In the patient-level analysis, an increase in the mean score per patient was associated with clinical benefit in all cohorts

(cohort 1: odds ratio, 1.22 [95% CI: 1.07, 1.41], $P = .004$; cohort 2: odds ratio, 1.34 [95% CI: 1.10, 1.70], $P = .007$; cohort 3: odds ratio, 1.22 [95% CI: 1.03, 1.47], $P = .02$) (Fig 4).

The radiomics-clinical score was associated with overall survival in all cohorts (cohort 1: hazard ratio, 0.14 [95% CI: 0.04, 0.51], $P = .003$; cohort 2: hazard ratio, 0.02 [95% CI: 0.00, 0.20], $P = .001$; cohort 3: hazard ratio, 0.13 [95% CI: 0.03, 0.57], $P = .01$) (Table 4). Univariate and multivariate Cox proportional hazard ratios are in Table E6 (online).

Patients with a high radiomics-clinical score (score > 0.45) showed improved overall survival compared with those with a low radiomics-clinical score (score ≤ 0.45); the 24-month overall survival in cohort 2 was 53% (95% CI: 35, 79) versus 25% (95% CI: 10, 62; $P = .002$), and in cohort 3 it was 36% (95% CI: 21, 60) versus 5% (95% CI: 1, 34; $P = .02$) (Fig 5).

Exploring the Biologic Significance of the CT-based Radiomics Signature

Tumor RNA profiling through NanoString and CT scans was available for 20 patients (cohort 4) (sample characteristics are

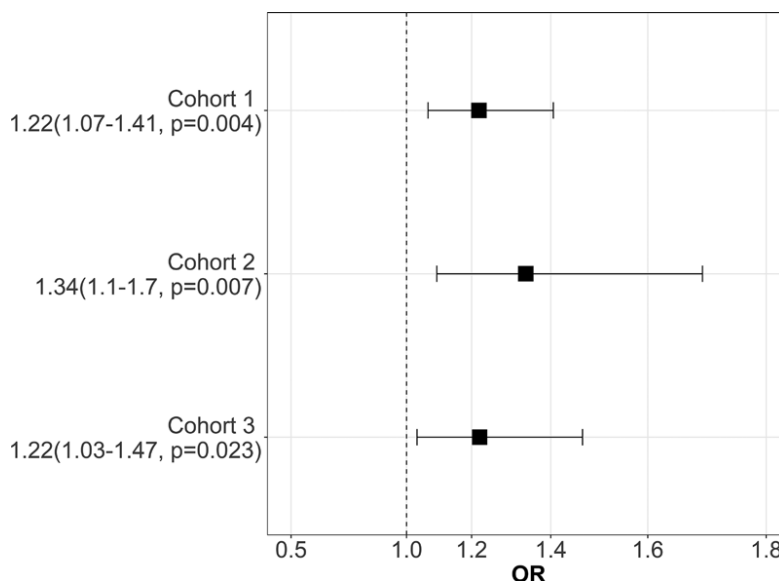


Figure 4: Association of the mean radiomics score per patient with clinical benefit for patient-level analysis in cohort 1 (phase I training set), cohort 2 (bladder test set), and cohort 3 (lung test set). Data in parentheses are 95% CIs. OR = odds ratio.

Table 4: Hazard Ratio and C-index for Association of Mean Radiomics and Radiomics-Clinical Score per Patient with Overall Survival in Cohorts 1, 2, and 3

Variable	Hazard Ratio	P Value	C-index
Radiomics score			
Cohort 1	0.56 (0.28, 1.14)	.11	0.61 (0.52, 0.7)
Cohort 2	0.11 (0.03, 0.37)	<.001	0.74 (0.65, 0.82)
Cohort 3	0.38 (0.16, 0.88)	.03	0.61 (0.51, 0.7)
Radiomics-clinical score			
Cohort 1	0.14 (0.04, 0.51)	.003	0.65 (0.56, 0.73)
Cohort 2	0.02 (0, 0.2)	.001	0.72 (0.62, 0.82)
Cohort 3	0.13 (0.03, 0.57)	.01	0.65 (0.57, 0.73)

Note.—Data in parentheses are 95% CIs.

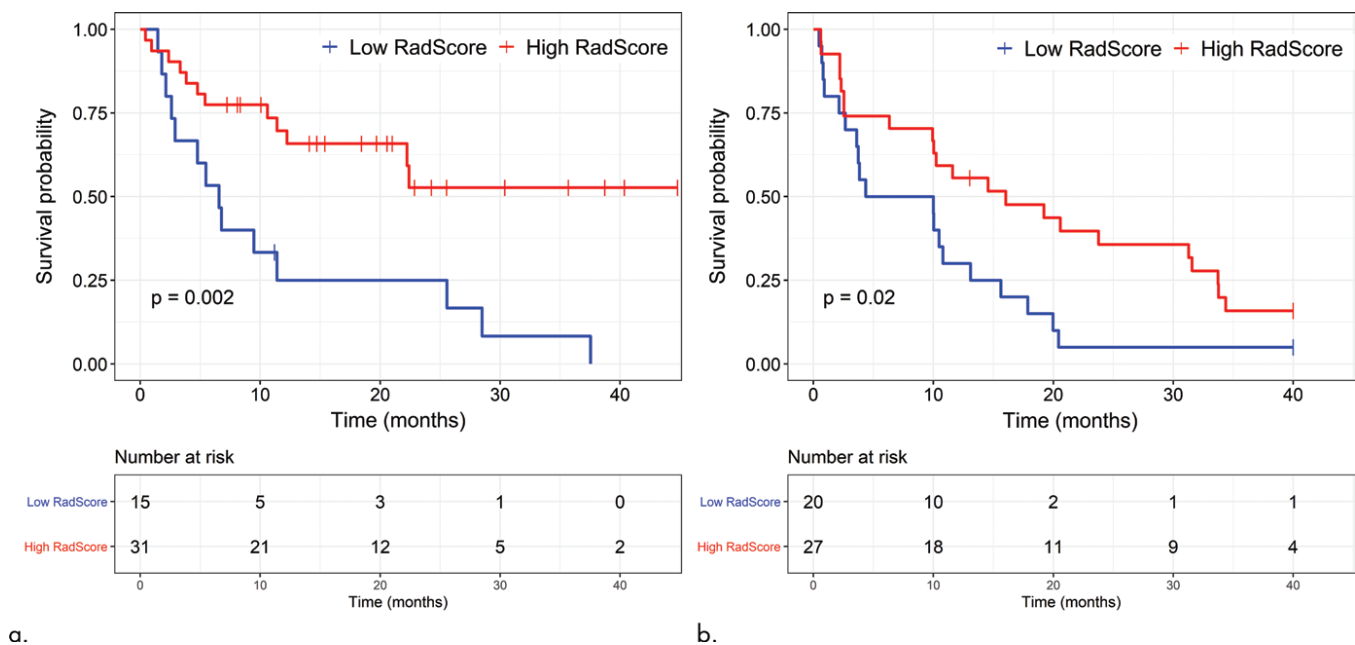


Figure 5: Kaplan-Meier overall survival curve analysis in the cohort 2 (a) and cohort 3 (b) test sets. The CT-based radiomics and clinical score was defined as high (>0.45) or low (≤ 0.45), according to the receiver operating characteristic cutoff maximizing sensitivity. RadScore = radiomics score.

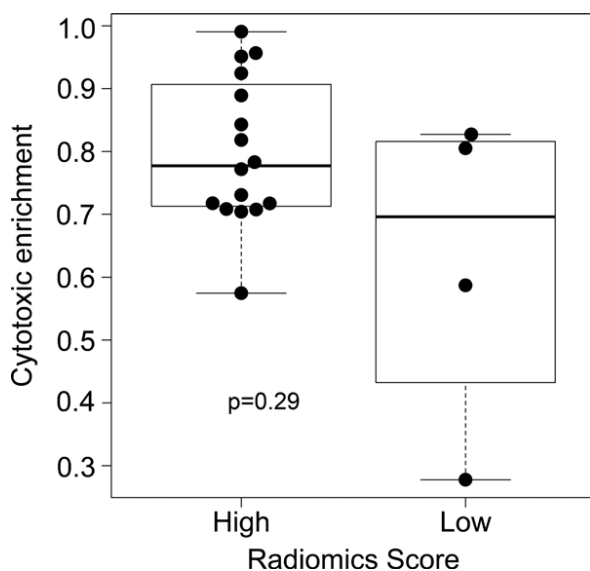


Figure 6: Association between cytotoxic immunophenotype enrichment score and the CT-based radiomics score in cohort 4. The CT-based radiomics score was defined as high (>0.29) or low (≤ 0.29), according to the receiver operating characteristic cutoff maximizing sensitivity. Dots correspond to the Cytotoxic enrichment value of each patient's lesion and whiskers corresponds to the minimum and maximum values of each radiomics score group.

in Table 1). In this cohort, tumors with a high radiomics score (>0.29 according to the cutoff point of the receiver operating characteristic curve) showed higher values of immune cytotoxic cells compared with the tumors with low radiomics score, but this difference was not significant (0.80 ± 0.12 vs 0.62 ± 0.25 ; $P = 0.29$) (Fig 6).

Tumor samples for tumor-infiltrating lymphocyte quantification were available from 75 patients in cohorts 1 and 2 (Table E7 [online]). No association was identified between tumor-infiltrating

lymphocytes and high versus low mean radiomics score (score, >0.29 or ≤ 0.29) (mean, 2.37 ± 1.12 vs 1.89 ± 1.19 ; $P = .36$) (Appendix E1, Fig E2 [online]).

Discussion

Although immune checkpoint inhibitors have transformed the management of many tumor types, response rates are still sub-optimal, and better treatment selection is required. In this study, we developed a radiomics signature that potentially allows stratification of patients who are more likely to benefit from immune checkpoint inhibitors, achieving a 75% sensitivity and 53% specificity. The radiomics signature consisted of 14 features related to first-order histogram, shape, and local-regional texture (gray level-dependent matrix and gray level co-occurrence matrix derived). The negative coefficient of gray level-dependent matrix dependence variance and dependence nonuniformity normalized features, indicating homogeneous tumors, were associated with a higher probability of immune response. Similarly, first-order uniformity that reflects tumor homogeneity was associated with a higher probability of response.

As a first step toward clinical application of our model, we tested it on two independent samples of patients with urinary bladder and lung cancer. In these test cohorts, the radiomics signature identified patients who benefited from treatment, with 73% sensitivity and 55% specificity in the bladder cancer cohort and 73% sensitivity and 57% specificity in the lung cancer cohort. The moderately high sensitivity of this assay indicates a potential to better identify patients who may benefit from immune checkpoint inhibitors and in whom this treatment may be prioritized.

In this study, we showed that integrating clinical data (albumin and lymphocytes at baseline) improved performance compared with clinical-only and radiomics-only signatures (AUC, 0.74 vs 0.64 and 0.70, respectively).

To assess for a biologic basis of the radiomics signature, we explored its association with cytotoxic immunophenotyping by gene expression profiling as a marker of tumor immunogenicity (32,33). There was no statistically significant relationship between high CT-based radiomics score and enriched cytotoxic phenotype.

Few studies have been published on radiomics as a tool to assist in the medical treatment of patients with cancer undergoing immunotherapies. Sun et al (18) developed a CD8+ tumor infiltration radiomics signature. In line with our results, Sun et al demonstrated that tumor homogeneity was associated with high CD8+ tumor infiltration. Similarly, image acquisition parameters were considered for their radiomics signature development. Unfortunately, we could not correlate our signature with that of Sun et al because peritumoral radiomics features were included in their model, whereas we restricted our analysis to tumor features only because the periphery of tumors can contain artifacts (from contrast-enhanced vessels, fat, or air when the tumor is located at the organ edge).

With the advent of new tools for automatic or semiautomatic image segmentation and computational image analysis and quantification of radiomics signature scores, our signature could be incorporated into clinical trials and clinical practice in the near future. Whereas our data are reassuring regarding the biologic validity of our radiomics signature, further correlative studies in larger populations should now be performed in other tumor types including correlations with PD-L1 expression, microsatellite instability status, and tumor mutation burden, all of which are approved as predictive biomarkers for immune checkpoint inhibitor therapy (34).

Our study had some limitations. First, because of the observational and retrospective nature of the work, no standardization of image acquisition parameters was possible. Second, we developed a radiomics signature in a sample of multiple tumor types and validated in only two specific tumor types (bladder and lung cancer), which makes the training cohort more heterogeneous. Third, this predictive radiomics signature may be useful for anti-PD-1/PD-L1 patient selection but will require further testing in other cohorts as newer immunotherapies emerge. Last, our primary end point, clinical benefit, may not be used in pivotal trials, although we also identified an association between our signature and overall survival.

In conclusion, the CT-based radiomics signature developed in this study was associated with response to immune checkpoint inhibitors in patients with advanced solid tumors. The integration of multidimensional data including radiomics, clinical variables and genomic characterization may be key to optimize immune checkpoint inhibitor treatment selection and, ultimately, improve survival and quality of life in patients with cancer.

Acknowledgment: Guillermo Villacampa provided statistical advice for this manuscript.

Author contributions: Guarantors of integrity of entire study, M.L., A.G.R., M.V.R., J.L., R.M.B., R.P.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related

to the work are appropriately resolved, all authors; literature research, M.L., M.V.R., I.M., J.M., J.R., E.M.C., A.O., J.S., M.E., J.T., R.D., E.G., R.P.L.; clinical studies, M.L., C.V., J.L., I.M., J.M.L., M.O.d.O., J.M., M.G., C.S., J.R., E.E., I.B., E.M.C., A.O., J.T., E.G.; experimental studies, M.L., A.G.R., C.V., M.V.R., I.M., R.M.B., A.O., R.F., P.N., J.S., J.T., R.P.L.; statistical analysis, M.L., A.G.R., G.V., I.M., A.O., D.G., C.R.P., J.S., J.T., R.D., R.P.L.; and manuscript editing, M.L., M.V.R., I.M., J.M.L., J.M., R.M.B., C.S., E.E., E.M.C., A.O., J.T., J.C., R.D., E.G., R.P.L.

Disclosures of Conflicts of Interest: M.L. disclosed no relevant relationships. A.G.R. disclosed no relevant relationships. C.V. disclosed no relevant relationships. G.V. Activities related to the present article: disclosed fees for participation in review activities from AstraZeneca and MSD. Activities not related to the present article: disclosed employment from SOLTI group. Other relationships: disclosed no relevant relationships. M.V.R. disclosed no relevant relationships. J.L. disclosed no relevant relationships. I.M. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed ESMO fellowship in translational research. Other relationships: disclosed no relevant relationships. J.M.L. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from Bristol-Myers-Squibb, Novartis, Pierre Fabre, Roche; disclosed payment for lectures from Astellas, Bristol-Myers-Squibb, MSK, Novartis, Pierre Fabre, Pfizer, Roche; disclosed travel expenses from Bristol-Myers-Squibb, MSD, Novartis, Pierre Fabre, Pfizer, Roche, Ipsen. Other relationships: disclosed no relevant relationships. M.O.d.O. disclosed no relevant relationships. C.H. disclosed no relevant relationships. J.M. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from AstraZeneca, Roche, Amgen, Janssen, Clovis Oncology, MSD; grants/grants pending from AstraZeneca, Pfizer; payment for lectures from Astellas, AstraZeneca. Other relationships: disclosed no relevant relationships. M.G. Activities related to the present article: disclosed travel support from Roche, Astellas, Bayer. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. R.M.B. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed employment from VHIO; grants from MSD; travel support from Roche, Sanofi, Aventis, MSD. Other relationships: disclosed no relevant relationships. C.S. disclosed no relevant relationships. J.R. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from Novartis, Eli Lilly, Orion Pharmaceuticals, Servier Pharmaceuticals, Peptomy, Merck Sharp & Dohme, Kelun Pharmaceutical/Klus Pharma, Spectrum Pharmaceuticals Inc, Pfizer, Roche Pharmaceuticals, Ellipses Pharma, Certera, Bayer, Molecular Partners, NovellusDX, Ioncure SA; disclosed grants/grants pending from Bayer, Novartis, Blueprint Pharmaceuticals, Spectrum Pharmaceuticals, Tocagen, Synphogen, BioAlta, Pfizer, GenMab, CytomX, Kelun-Biotech, Takeda-Millennium, GlaxoSmithKline, Ipsen; disclosed travel support from ESMO, Department of Defense, Merck Sharp & Dohme, Louisiana State University, Kelun Pharmaceutical/Klus Pharma, Huntsman Cancer Institute, Cancer Core Europe, Karolinska Cancer Institute, King Abdullah International Medical Research Center, Bayer, WIN Consortium, Jansen, Molecular Partners; disclosed other from European Journal of Cancer, VHIO/Ministerio De Empleo Y Seguridad Social, Chinese University of Hong Kong, Solti, Elsevier, GlaxoSmithKline. Other relationships: disclosed no relevant relationships. E.E. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from Hoffman La Roche, Bristol Myers Squibb, Servier, Amgen, Merck Serono, Array Biopharma, Sanofi; honoraria from Array Biopharma, MSD, Abbvie, Amgen, GlaxoSmithKline, AstraZeneca, Merck Sharp and Dohme, Bristol Myers Squibb, Novartis, Boehringer Ingelheim, Hoffman La-Roche. Other relationships: disclosed no relevant relationships. I.B. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from MSD, Orion Pharma, Rakuten Aspyrian; expert testimony from Roche; grants/grants pending from MSD, BMS, Novartis, AstraZeneca, Regeneron, GlaxoSmithKline, Kura, Pfizer, Orion Pharma, Merck Serono, Celgene, Incyte, Shattuck Lab, Janssen, Rakuten Aspyrian, VCN Biosciences; payment for lectures from BMS, AstraZeneca, Merck Serono; travel support from Merck Serono, AstraZeneca; and other from Instituto Salud Carlos III - Programa Rio Hortega Contract (grant CM15/00255). Other relationships: disclosed no relevant relationships. E.M.C. Activities related to the present article: disclosed consulting fees or honorarium from BMS, MSD, Novartis, Roche, Sanofi, Pierre-Fabre; disclosed travel support from BMS, MSD, Novartis. Activities not related to the present article: disclosed payment for expert testimony from BMS, Sanofi; disclosed payment for lectures from BMS, MSD, Sanofi, Roche, Pierre-Fabre. Other relationships: disclosed no relevant relationships. A.O. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed advisory board service for Roche, AstraZeneca, PharmaMar, Clovis Oncology, Tesaro, Immunogen

and Genmab; travel support from Roche, AstraZeneca, and PharmaMar; institutional financial interest from Abbvie Deutschland, Abilify Pharmaceuticals, Advaxis, Aeterna Zentaris, AMGEN, Aprea Therapeutics, Clovis Oncology, Eisai, F. Hoffmann – La Roche, Regeneron Pharmaceuticals, Immunogen, Merck, Sharp and Dohme de España, Millennium Pharmaceuticals, Pharma Mar, Tesaro, BMS, Bristol Meyers Squibb. **R.F.** disclosed no relevant relationships. **P.N.** disclosed no relevant relationships. **D.G.** disclosed no relevant relationships. **C.R.P.** disclosed no relevant relationships. **J.S.** Activities related to the present article: disclosed ISCIII grant. Activities not related to the present article: disclosed board membership from Northern Biologics; consultancies from Mosaic Biomedicals, SLU; grants/grants pending from Hoffman La Roche, RidgeLine, Isarna Therapeutics, Mosaic Biomedicals, Fundación Científica De La Aecc; disclosed patents issued. Other relationships: disclosed no relevant relationships. **E.F.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed grants/grants pending from Fundación Merck Salud, Merck Grant for Oncology Innovation; payment for lectures from Abbvie, Astra Zeneca, Blue Print Medicines, Boehringer Ingelheim, Bristol-Myers Squibb, Eli Lilly And Company, F. Hoffman -La Roche, Guardant Health, Janssen Biotech, Medscape, Merck Kgaa, Merck Sharp And Dohme, Novartis, Pfizer, Prime Oncology, Samsung, Takeda Oncology, Touchime. Other relationships: disclosed independent member of the board of directors from Grifols. **M.E.** disclosed no relevant relationships. **J.T.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies from Array Biopharma, AstraZeneca, Bayer, BeiGene, Boehringer Ingelheim, Chugai, Genentech, Inc., Genmab A/S, Halozyme, Imugene Limited, Inflection Biosciences Limited, Ipsen, Kura Oncology, Lilly, MSD, Menarini, Merck Serono, Merrimack, Merus, Molecular Partners, Novartis, Peptomyc, Pfizer, Pharmacycics, ProteoDesign SL, Rafael Pharmaceuticals, F. Hoffmann-La Roche, Sanofi, SeaGen, Seattle Genetics, Servier, Symphogen, Taiho, VCN Biosciences, Biocartis, Foundation Medicine, HalioDX SAS, Roche Diagnostics. Other relationships: disclosed no relevant relationships. **J.C.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed board memberships from Bayer, Johnson & Johnson, Bristol-Myers Squibb, Astellas Pharma, Pfizer, Sanofi, MSD Oncology, Roche, AstraZeneca; consultancy from Astellas; grants/grants pending from AB Science, Aragon Pharmaceuticals, Arog Pharmaceuticals, Astellas Pharma, Astrazeneca, Aveo Pharmaceuticals, Bayer, Blueprint Medicines, BN Immunotherapeutics, Boehringer Ingelheim España, Bristol-Myers Squibb International, Clovis Oncology, Cougar Biotechnology, Deciphera Pharmaceuticals, Exelixis, F. Hoffmann-La Roche, Genentech, Glaxosmithkline, Incyte Corporation, Janssen-Cilag International, Karyopharm Therapeutics, Laboratoires Leurquin Mediolanum, Lilly, Medimmune, Millennium Pharmaceuticals, Nanobiotix, Novartis Farmacéutica, Pfizer, SLU, Puma Biotechnology, Sanofi-Aventis, SFJ Pharma, Teva Pharma; payment for lectures from Bayer, Astellas, MSD; travel support from BMS, Roche, AstraZeneca. Other relationships: disclosed no relevant relationships. **R.D.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancy for Roche; grants/grants pending from Roche, Ipsen, Amgen, Servier, Sanofi, Merck Sharp & Dohme, Boehringer Ingelheim. Other relationships: disclosed no relevant relationships. **E.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed consultancies for F. Hoffmann-La Roche, Ellipses Pharma, Neomed Therapeutics Inc, Boehringer Ingelheim, Janssen Global Services, Bristol-Meyers Squibb; travel support from Merck Sharp & Dohme, GlycoTope, Menarini. Other relationships: disclosed no relevant relationships. **R.P.L.** disclosed no relevant relationships.

References

- Long GV, Weber JS, Larkin J, et al. Nivolumab for patients with advanced melanoma treated beyond progression: analysis of 2 phase 3 clinical trials. *JAMA Oncol* 2017;3(11):1511–1519.
- Postow MA, Chesney J, Pavlick AC, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *N Engl J Med* 2015;372(21):2006–2017 [Published correction appears in *N Engl J Med* 2018;379(22):2185.].
- Ribas A, Puzanov I, Dummer R, et al. Pembrolizumab versus investigator-choice chemotherapy for ipilimumab-refractory melanoma (KEYNOTE-002): a randomised, controlled, phase 2 trial. *Lancet Oncol* 2015;16(8):908–918.
- Bellmunt J, de Wit R, Vaughn DJ, et al. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *N Engl J Med* 2017;376(11):1015–1026.
- Powles T, Durán I, van der Heijden MS, et al. Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *Lancet* 2018;391(10122):748–757 [Published correction appears in *Lancet* 2018;392(10156):1402.].
- Motzer RJ, Escudier B, McDermott DF, et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N Engl J Med* 2015;373(19):1803–1813.
- Motzer RJ, Tannir NM, McDermott DF, et al. Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. *N Engl J Med* 2018;378(14):1277–1290.
- Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N Engl J Med* 2015;373(2):123–135.
- Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med* 2015;373(17):1627–1639.
- Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 2016;387(10027):1540–1550.
- Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* 2017;18(3):e143–e152 [Published correction appears in *Lancet Oncol* 2019;20(5):e242.].
- Nishino M, Hatabu H, Hodi FS. Imaging of cancer immunotherapy: current approaches and future directions. *Radiology* 2019;290(1):9–22.
- Jardim DL, de Melo Gagliato D, Giles FJ, Kurzrock R. Analysis of drug development paradigms for immune checkpoint inhibitors. *Clin Cancer Res* 2018;24(8):1785–1794.
- Tang J, Shalabi A, Hubbard-Lucey VM. Comprehensive analysis of the clinical immuno-oncology landscape. *Ann Oncol* 2018;29(1):84–91.
- Buder-Bakhaya K, Hassel JC. Biomarkers for clinical benefit of immune checkpoint inhibitor treatment—a review from the melanoma perspective and beyond. *Front Immunol* 2018;9:1474.
- Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol* 2016;17(12):e542–e551.
- Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14(12):749–762.
- Sun R, Limkin EJ, Vakalopoulou M, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* 2018;19(9):1180–1191.
- Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014;5(1):4006.
- Tang C, Hobbs B, Amer A, et al. Development of an immune-pathology informed radiomics model for non-small cell lung cancer. *Sci Rep* 2018;8(1):1922.
- Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol* 2019;30(6):998–1004.
- Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;366(10):883–892 [Published correction appears in *N Engl J Med* 2012;367(10):976.].
- Caswell DR, Swanton C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Med* 2017;15(1):133.
- Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 2012;30(9):1323–1341.
- Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJWL. Data analysis strategies in medical imaging. *Clin Cancer Res* 2018;24(15):3492–3499.
- van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017;77(21):e104–e107.
- Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295(2):328–338.
- Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45(2):228–247.
- Ohorodnyk P, Eisenhauer EA, Booth CM. Clinical benefit in oncology trials: is this a patient-centred or tumour-centred end-point? *Eur J Cancer* 2009;45(13):2249–2252.
- Pascual-García M, Bonfill-Teixidor E, Planas-Rigol E, et al. LIF regulates CXCL9 in tumor-associated macrophages and prevents CD8⁺ T cell tumor-infiltration impairing anti-PD1 therapy. *Nat Commun* 2019;10(1):2416.

31. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and prognosis of quantitative features extracted from CT images. *Transl Oncol* 2014;7(1):72–87.
32. Charoentong P, Finotello F, Angelova M, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;18(1):248–262.
33. Angelova M, Charoentong P, Hackl H, et al. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol* 2015;16:64.
34. Tray N, Weber JS, Adams S. Predictive biomarkers for checkpoint immunotherapy: current status and challenges for clinical application. *Cancer Immunol Res* 2018;6(10):1122–1128.

4.3 Study 3

PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma

This study corresponds to **objective 3**. In this research, we investigated the generalizability of radiomic features to Positron Emission Tomography (PET) component in a setting of large B-cell lymphoma patients treated with CAR T-cell therapy.

Reference:

Ligero M., Simó M., Carpio C. et al. PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma. Electronic Journal of Haematology (eJHaem) (2023).

RESEARCH ARTICLE

eJHaem



PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma

Marta Liger¹ | Marc Simó² | Cecilia Carpio³ | Gloria Iacoboni³ |
 Maria Balaguer-Montero¹ | Victor Navarro⁴ | Mario Andres Sánchez-Salinas³ |
 Sabela Bobillo³ | Ana Marín-Niebla³ | Josu Iraola-Truchuelo³ | Pau Abrisqueta³ |
 Roser Sala-Llonch^{5,6} | Francesc Bosch³ | Raquel Perez-Lopez¹ | Pere Barba³

¹Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron Barcelona Hospital Campus (VHUB), Barcelona, Spain

²Nuclear Medicine Department, Vall d'Hebron University Hospital, Autonomous University of Barcelona, Barcelona, Spain

³Department of Hematology, Experimental Hematology, Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron University Hospital, Barcelona, Barcelona, Spain

⁴Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain

⁵Faculty of Medicine, Department of Biomedicine, Institute of Neurosciences, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona, Barcelona, Spain

⁶Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain

Correspondence

Dr. Raquel Perez-Lopez, Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Cellex Center, Natzaré 115–117, Barcelona 08035, Spain.

Email: rperez@vhio.net

Abstract

Chimeric antigen receptor (CAR) T-cell therapy is a promising treatment option for relapsed or refractory (R/R) large B-cell lymphoma (LBCL). However, only a subset of patients will present long-term benefit. In this study, we explored the potential of PET-based radiomics to predict treatment outcomes with the aim of improving patient selection for CAR T-cell therapy. We conducted a single-center study including 93 consecutive R/R LBCL patients who received a CAR T-cell infusion from 2018 to 2021, split in training set (73 patients) and test set (20 patients). Radiomics features were extracted from baseline PET scans and clinical benefit was defined based on median progression-free survival (PFS). Cox regression models including the radiomics signature, conventional PET biomarkers and clinical variables were performed for most relevant outcomes. A radiomics signature including 4 PET-based parameters achieved an AUC = 0.73 for predicting clinical benefit in the test set, outperforming the predictive value of conventional PET biomarkers (total metabolic tumor volume [TMTV]: AUC = 0.66 and maximum standardized uptake value [SUV_{max}]: AUC = 0.59). A high radiomics score was also associated with longer PFS and OS in the multivariable analysis. In conclusion, the PET-based radiomics signature predicted efficacy of CAR T-cell therapy and outperformed conventional PET biomarkers in our cohort of LBCL patients.

KEYWORDS

CAR T cells, DLBCL, non-Hodgkin lymphoma, PET, progression-free survival, radiomics

Marta Liger, Marc Simó, Raquel Perez-Lopez, and Pere Barba contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *eJHaem* published by British Society for Haematology and John Wiley & Sons Ltd.

1 | INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma [1]. Despite durable responses to first-line immunochemotherapy in approximately 60% of cases [2, 3], less than half of relapsed/refractory (R/R) patients are salvaged with intensive chemotherapy followed by autologous stem cell transplant consolidation [4]. Chimeric antigen receptor (CAR) T-cell therapy is the established third-line treatment for this disease [5, 6] and has shown significantly improved results in comparison with chemotherapy for early relapsed and refractory patients in the second-line setting [7, 8]. However, not all patients experience the same clinical benefit with this cellular therapy, raising the need to find predictive biomarkers to aid patient selection, leading toward a better efficacy and safety profile.

Quantitative imaging parameters from ^{18}F -fluorodeoxyglucose (FDG) positron emission tomography (PET) such as total metabolic tumor volume (TMTV) and the maximum standardized uptake value (SUV_{max}) have been associated with outcomes in patients with DLBCL [9–12]. However, these standard quantitative PET characteristics are limited to tumor burden and maximum metabolic consumption, but do not provide details about the phenotype or spatial distribution of the malignancy. Also, the reported results of these parameters are heterogeneous, and their routine clinical use is still limited [9, 11, 13–15].

PET-based radiomics provides quantitative imaging data regarding tumor shape, heterogeneity, and metabolic consumption distribution. These parameters can provide insight into the tumor biology, including meaningful information about the cancer cell metabolism and tumor microenvironment for further identification of responsive phenotypes. Some studies have defined PET-based radiomic features associated with a favorable response to standard chemotherapy in patients with DLBCL [16, 17]. However, scarce data is available in the specific setting of CAR T-cell therapy [18].

In this study, we analyzed FDG-PET-based radiomic phenotypes in R/R DLBCL patients treated with CD19-targeted CAR T-cell ther-

apy, aiming to identify those patients who would benefit most from treatment. Furthermore, we looked at the added value of textural features to standard prognostic factors for efficacy and safety after therapy.

2 | METHODS

2.1 | Study population

The study included all consecutive patients with R/R diffuse large B-cell lymphoma (DLBCL), de novo or transformed from an indolent lymphoma, who received a single infusion of CD19-targeted second-generation CAR T cells from July 2018 to November 2021 at Vall d'Hebron University Hospital and had at least 1 disease assessment after infusion. For both cytokine release syndrome (CRS) and immune effector cell-associated neurotoxicity syndrome (ICANS), grading was carried out according to the American Society for Transplantation and Cellular Therapy (ASTCT) criteria and management followed local guidelines [19, 20]. All patients included in the study provided informed consent. The study was approved by the ethical committee of the Vall d'Hebron University Hospital (PR[AG]404/2020) and conducted in accordance with the Declaration of Helsinki.

2.2 | Endpoints and definitions

Clinical data were collected from the electronic medical records. These included age, sex, disease stage before CAR T-cell therapy, previous lines of treatment, previous stem cell transplant (SCT), Eastern Cooperative Oncology Group (ECOG) status, lactate dehydrogenase (LDH), tumor histology (transformed from indolent lymphoma or DLBCL de novo high-grade B-cell lymphoma double hit/triple hit [DH/TH]) and cell of origin (germinal center B-cell [GCB] or non-GCB). The primary endpoint was progression-free survival (PFS), defined as time between

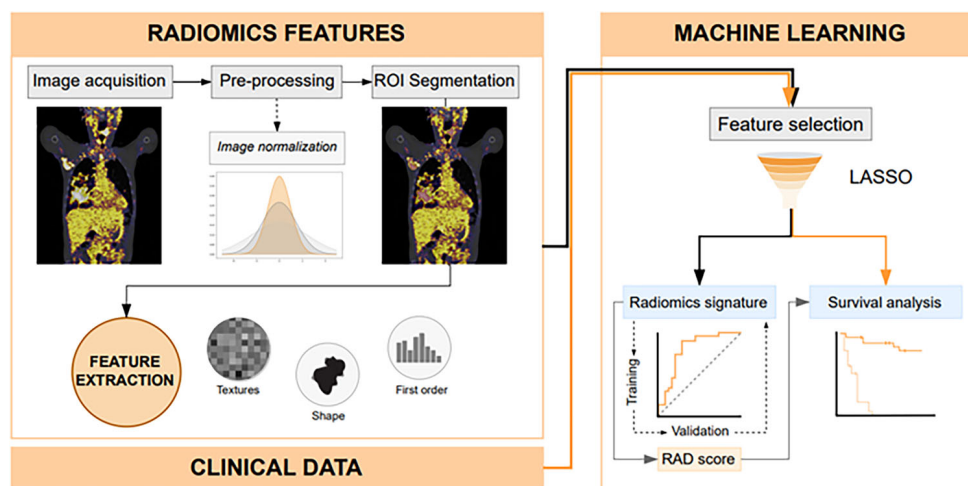


FIGURE 1 Radiomics analysis workflow.

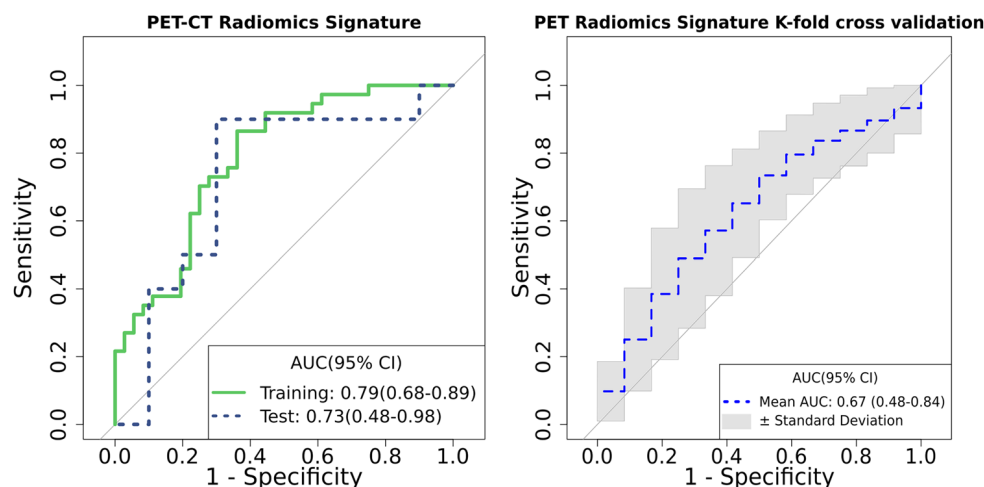


FIGURE 2 ROC curves for 3-month progression-free survival (PFS) prediction of the radiomics signature in the training and test sets (A) and fivefold cross-validation analysis (B).

TABLE 1 Baseline patient characteristics.

Characteristics	N = 93	Training N = 73	Test N = 20	p*
Age, years, median (IQR)	59 (50–68)	57 (47–67)	62 (56–66)	0.115
Age, ≥65 years, n (%)	30 (32)			
Sex, male, n (%)	63 (68)	48 (66)	25 (34)	0.590
Disease stage before CAR T-cell therapy				
Stages I–II, n (%)	25 (27)	21 (29)	4 (20)	
Stages III–IV, n (%)	68 (73)	52 (71)	16 (80)	0.573
Previous lines of treatment				
>2, n (%)	31 (33)	23 (32)	8 (40)	0.593
Previous SCT, n (%)	30 (32)	21 (29)	9 (45)	0.186
ECOG, ≥1, n (%)	61 (66)	46 (63)	15 (75)	0.428
Primary refractory, n (%)	55 (59)	41 (56)	14 (70)	0.312
Median LDH/ULN (IQR), U/L	1.39 (0.93–2.28)	1.39 (0.94–2.24)	1.04 (0.93–2.39)	0.728
LDH > 2 × ULN, n (%)	32 (34)	26 (36)	6 (30)	0.792
IPI prognostic score, n (%)				
0–2	49 (52)	37 (51)	12 (60)	
3–5	44 (47)	36 (49)	8 (40)	0.614
Bulky disease (>7 cm), n (%)	30 (32)	23 (32)	9 (45)	0.294
Histology, n (%)				
Transformed from indolent lymphoma	20 (21)	15 (21)	5 (25)	0.760
DLBCL de novo	61 (66)	48 (65)	(66)	1
High-grade B-cell lymphoma (DH/TH)	12 (13)	10 (14)	2 (10)	1
Cell of origin, n (%)				
GCB	60 (64)	49 (67)	11 (55)	
Non-GCB	25 (27)	19 (26)	6 (30)	
Not available	8 (9)	5 (6)	3 (15)	0.416
Costimulatory domain, n (%)				
41BB	62 (67)	49 (67)	13 (65)	
CD28	31 (33)	24 (33)	7 (35)	1

*p Values were obtained from ANOVA test for continuous variables and Fisher's exact test for categorical variables.

TABLE 2 Radiomics versus conventional PET performance using AUC, accuracy (ACC), sensitivity (SE), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV).

	Training							Test						
	AUC [95% CI]	<i>p</i> Value	ACC (%)	SE (%)	SP (%)	NPV (%)	PPV (%)	AUC [95% CI]	<i>p</i> Value	ACC (%)	SE (%)	SP (%)	NPV (%)	PPV (%)
RAD score	0.79 [0.68–0.89]	<0.001	75	87	64	82	71	0.73 [0.48–0.98]	0.045	65	40	90	80	60
SUV_{max}	0.66 [0.53–0.79]	0.992	67	59	75	82	71	0.59 [0.32–0.86]	0.759	55	70	40	57	54
TMTV	0.66 [0.53–0.78]	0.99	66	81	50	72	63	0.66 [0.38–0.94]	0.891	60	80	40	67	57

Abbreviations: RAD, radiomics score; SUV_{max}, maximum standard uptake value; TMTB, total metabolic tumor burden. In bold *p*-value below 0.05 (i.e., statistically significant).

CAR T-cell infusion and disease progression or death, whichever occurred first. Secondary endpoints included overall survival (OS), defined as the time between CAR T-cell infusion and death, best response achieved after infusion and clinically significant adverse events, defined as grade ≥ 2 CRS and/or ICANS in this study.

2.3 | Disease assessment and evaluation

All patients underwent a PET-CT scan at our institution after the last bridging regimen, within 7 days of starting lymphodepleting chemotherapy. Disease evaluation after CAR T-cell therapy was scheduled at 1, 3, 6, 12, 18, and 24 months after infusion. The imaging reports were based on the Lugano recommendations for response assessment and graded according to the 5-point Deauville score [21]. Patients achieving a complete metabolic response (CMR, Deauville scores 1–3) or partial metabolic response (PMR) were considered responders to CAR T-cell therapy.

2.4 | PET/CT exam and feature acquisition

The PET-CT exam required a minimum of 6-hour fast prior to the intravenous administration of 3.7 MBq/kg (222–370 MBq) of ¹⁸F-fluorodeoxyglucose (¹⁸F-FDG). Glucose values below 140 mg/dL were required in all cases prior to administration of the radiopharmaceutical. Before scanning, the patients were at rest for a minimum of 60 min. Images were obtained using a Siemens Biograph mCT, which combines a spiral CT of 64 slices (210 keV, 120 mAs, care dose) with a dedicated PET, from the skull to the upper third of both femurs. The images generated were evaluated by a nuclear medicine physician in a syngo.via Siemens Healthcare workstation [9].

Standard Uptake Value (SUV) was computed voxel-wise from ¹⁸F-FDG PET/CT images normalized to patient's weight and injected dose [22]. Lesions were delineated in the PET scans by an experienced nuclear medicine physician using a custom workflow from MIM Encore™ software (MIM Software Inc., Cleveland, OH, USA), which selected all metabolic activity above the 41% maximum SUV threshold.

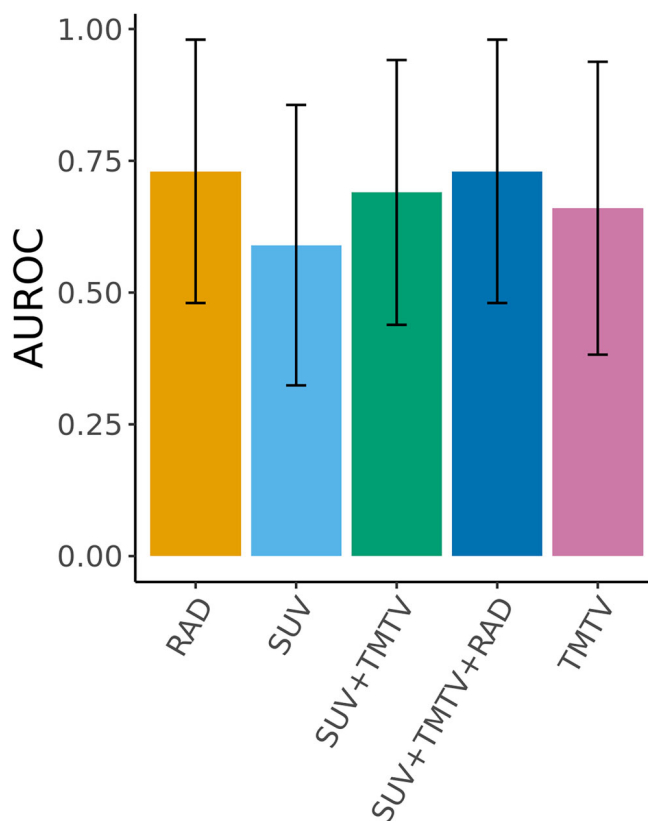


FIGURE 3 Area under the receiver operating characteristic (AUROC) curve for 3-month progression-free survival (PFS) prediction of conventional PET biomarkers (SUV_{max} and TMTV) and radiomics score individually or in combination. RAD, radiomics score; SUV, standard uptake value; TMTB, total metabolic tumor burden.

Images and segmentations were reviewed by a nuclear medicine specialist. Before feature extraction, images were resampled to $2 \times 2 \times 2$ mm³ using spline interpolation and discretized to a fixed bin number of 25 [23]. From the lesion with the largest volume, 105 radiomics features were extracted. These features correspond to quantitative characteristics regarding first order SUV distribution and metabolic consumption heterogeneity in the tumor measured by Gray-Level Co-occurrence Matrix [GLCM], Gray-Level Dependence Matrix [GLDM],

Gray-Level Run Length Matrix [GLRLM], and Gray-Level Size Zone Matrix [GLSZM]). All features were computed as described by van Griethuysen et al. [24]. Conventional PET features (SUV_{max} and TMTV) were also computed from the whole metabolic disease. Image pre-processing and feature extraction were computed using Pyradiomics v3.0.1 [24] package and Python v3.8.8.

2.5 | Statistical analysis

Data was split into training and test sets (80% and 20%, respectively). The training set was used to build the model and the test set to validate its performance. Population characteristics of both sets were balanced according to baseline international prognostic index (IPI) score, costimulatory domain and treatment response. In the training cohort, the least absolute shrinkage and selection operator (LASSO)-regression were used for radiomic feature selection. LASSO regularization parameter (λ) was optimized by maximizing the area under the curve (AUC) from cross-validation. Logistic regression including the selected features was trained and tested for classifying 3-month PFS (median PFS of this cohort). The performance of conventional PET features and the radiomic score were compared using the AUC from the receiver operating characteristic (ROC) curve. The 95% confidence intervals and the p value were assessed with the DeLong method and the Mann-Whitney U test, respectively. Stratified repeated fivefold cross-validation was implemented to explore the model generalizability for different data splits. Finally, univariate and multivariate Cox proportional hazard regression analysis were performed with clinical and PET-based data (including the radiomics signature) to investigate the impact of these factors on continuous PFS and OS. Kaplan-Meier curves were reported for the multivariate Cox prediction after discretization by Youden's threshold optimization (Figure 1). The potential impact of radiomic features on safety outcomes (grade ≥ 2 CRS and/or ICANS) were evaluated using the Mann-Whitney U test.

3 | RESULTS

3.1 | Patient characteristics

Of 101 consecutive patients with R/R DLBCL treated with CD19-targeted CAR T cells during the study period, 7 were excluded because pretreatment PET scans were not available or disease was not measurable, and 1 patient died prior to the first response evaluation (Figure 2). Finally, 93 patients were included in the study.

Patients' characteristics are summarized in Table 1. The median age was 59 years (interquartile range [IQR] 50–68) and 68% were male. Median number of prior lines was 2 (IQR 2–3), 73% had an advanced stage of disease, and 33% had undergone a previous autologous stem cell transplant (ASCT). Regarding the construct, 33% received axicabtagene ciloleucel and 67% received a CD19-targeted construct with 4-1BB costimulatory domain, including tisagenlecleucel and an inves-

tigational product. Median values for SUV_{max} and TMTV were 16 [IQR 21–22] and 177 cm^3 [45–549], respectively.

3.2 | Efficacy and safety data

Best response after CAR T-cell therapy was CMR in 37 (40%) patients, PMR in 33 (35%) (overall response rate 75%) and PMD in 23 (25%). With a median follow-up of 17.6 months (95% CI 12–24.3), median PFS and OS of the whole cohort were 3.9 months (95% CI 3.03–14.1) and 15.4 months (95% CI 10.8–NR), respectively.

Regarding toxicity, any grade of CRS and ICANS occurred in 68 (73%) and 33 (35%) patients, respectively. Grade ≥ 2 CRS and ICANS developed in 29 (31%) and 25 (27%) patients, respectively.

3.3 | Radiomics signature development

Seventy-three and 20 patients were randomly selected to build (training set) and validate (test set) the model. Baseline characteristics were similar between the training and test sets (Table 1). Among the 105 computed radiomics features, four of them were identified as those with the highest association with PFS using LASSO with the lambda of $\lambda = 0.102$ as described in the methods. The 4 selected features included maximum intensity, skewness, major axis length, and large dependence low gray-level emphasis (LDLGLLE) from the texture matrix GLDM. The radiomics score can be obtained by using Equation (1)

$$RadScore = \frac{1}{1 + \exp^{-y}} \quad (1)$$

$$y = -0.042 + -0.580 \times \text{Shape Major Axis Length} + \\ -0.331 \times \text{First order Maximum} + \\ -0.646 \times \text{First order Skewness} \times \\ -0.565 \times \text{GLDM LDLGLE}.$$

The PET-based radiomics signature points out that having larger lesions (measured by major axis length) negatively contribute to have a worse prognosis. Similarly, lesions with high SUV_{max} (maximum intensity) or negatively skewed distribution (skewness), which means more density of voxels in with high SUV values are more likely to not benefit more from CAR T-cell therapy (Table 3). Additionally, the model also includes GLDM large dependence low gray-level emphasis, which indicates that lesions with more homogeneity in lower SUV_{max} values will have less risk of progress. The radiomics score reached an AUC of 0.79 [95% CI 0.68–0.89] and 0.73 [95% CI 0.48–0.98] in the training and test sets, respectively, for prediction of 3-month PFS. Moreover, the radiomics signature showed a stable classification performance with an AUC of 0.67 [95% CI 0.48–0.84] in the cross-validation analysis (Figure 2). Accuracy, sensitivity, and specificity for the optimal cut-off of 0.43 are described in Table 2.

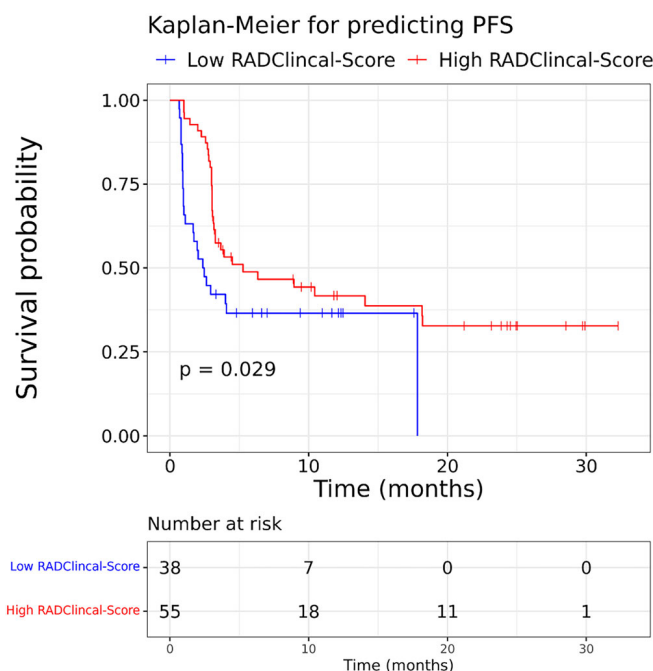


FIGURE 4 Kaplan-Meier curves for the multivariate analysis predicted progression-free survival (PFS) values.

3.4 | Integration of radiomics and conventional PET features to predict CAR T-cell efficacy

Thereafter, we investigated the performance of conventional PET features to predict PFS in our population. PET established biomarkers including baseline SUV_{max} of all lesions (AUC 0.59 [95% CI 0.32–0.86]) and TMTV (AUC 0.66 [95% CI 0.38–0.94]), individually or in combination (AUC 0.69 [95% CI 0.44–0.94]), showed a lower predictive capacity than the radiomics score for 3-month PFS. Furthermore, the addition of preestablished PET parameters to the radiomics score did not increase the predictive capacity of the latter in the training nor test sets (Figure 3).

3.5 | Impact of imaging and clinical factors on efficacy

In the univariate analysis, a low radiomics score (HR 2.97 [95% CI 1.74–5.1], $p < 0.001$), high baseline TMTV (HR 2.01 [95% CI 1.27–3.17], $p = 0.003$) and SUV_{max} (HR 1.03 [95% CI 1.00–1.05], $p = 0.022$), as well as ECOG ≥ 1 at lymphodepletion (HR 1.91 [95% CI 1.06–3.44], $p = 0.031$), a higher number of previous lines of therapies (HR 1.71 [95% CI 1.29–2.27], $p < 0.001$) and 4-1BB as costimulatory domain (HR 4.17 [95% CI 2.04–8.53], $p < 0.001$) were associated with a shorter PFS. The low radiomics score and costimulatory domain 4-1BB retained significance in the multivariate LASSO-Cox regression analysis for PFS (Figure 4, Table 3).

Regarding OS, a low radiomics score (HR 2.83 [95% CI 1.52–5.27], $p < 0.001$), high baseline TMTV (HR 2.33 [95% CI 1.35–3.86],

$p = 0.001$), and SUV_{max} (HR 1.03 [95% CI 1.01–1.06], $p = 0.024$), as well as for ECOG ≥ 1 (HR 2.39 [95% CI 1.14–5.03], $p = 0.021$), higher number of previous lines of treatment (HR 1.89 [95% CI 1.36–2.63], $p < 0.001$) and 4-1BB as costimulatory domain (HR 2.81 [95% CI 1.24–6.35], $p = 0.013$) were associated with a lower OS in the univariate analysis. In the multivariate LASSO-Cox model for OS, the radiomics score, ECOG ≥ 1 , 4-1BB as costimulatory domain, baseline TMTV, and number of previous lines of treatment retained a significant impact (Table 3).

3.6 | Impact of radiomics and conventional PET-based features on safety

Finally, we explored the association between the radiomics score and conventional PET-based features on significant adverse events after CAR T-cell therapy. Neither the radiomics score TMTV nor SUV_{max} was associated with a higher incidence of grade ≥ 2 CRS or ICANS in our cohort ($p > 0.05$). Similarly, we could not identify risk factors for grade ≥ 2 CRS and/or ICANS among the clinical variables including age, ECOG ≥ 1 , number of previous lines of treatment, and costimulatory domain ($p > 0.05$).

4 | DISCUSSION

The advent of CAR T-cell therapy has significantly improved the prognosis of patients with R/R DLBCL. However, only a subset of patients achieves durable responses after therapy. The identification of predictive biomarkers to aid patient selection and prognostic stratification is an unmet need in the field. In this study, we developed a PET-based radiomics signature capable of predicting clinical benefit after CAR T-cell therapy with an accuracy of 75%. This radiomics signature combined metabolic consumption (SUV), tumor size, and metabolic activity distribution, indicating that patients with smaller lesions and lower SUV_{max} or negatively skewed distribution are more likely to benefit from CAR T-cell therapy. The radiomics signature outperformed other established PET parameters, indicating that a more comprehensive evaluation of PET images including radiomic features could, more accurately capture, which DLBCL patients are more likely to respond to this novel cellular therapy.

Conventional PET-based biomarkers such as TMTV and SUV_{max} have been analyzed as prognostic biomarkers in lymphoma patients treated with CAR T cells [9–11], showing that higher baseline values are associated with a lower probability of response and long-term survival. Moreover, PET-based radiomic signatures have been associated with response to treatment in DLBCL patients treated with conventional chemotherapy, showing a better performance than standard clinical scoring systems, such as the IPI score [16]. However, few studies have implemented PET-based radiomics for modeling tumor response to CAR T-cell therapy. Zhou et al. evaluated the prognostic value of different PET-based radiomic features and explored the association between baseline radiomics and cytokine release syndrome

TABLE 3 Univariate and multivariate Cox for progression-free survival (PFS) and overall survival (OS).

	Univariate and multivariate analysis for PFS				Univariate and multivariate analysis for OS			
	Univariate analysis		Multivariate analysis		Univariate analysis		Multivariate analysis	
	HR [95% CI]	<i>p</i>	HR [95% CI]	<i>p</i>	HR [95% CI]	<i>p</i>	HR [95% CI]	<i>p</i>
Age	1.01 [0.99–1.04]	0.223			1.01 [0.99–1.04]	0.299		
ECOG ≥ 1	1.91 [1.06–3.44]	0.031			2.39 [1.14–5.03]	0.021	1.83 [0.84–3.97]	0.127
Number of previous lines of treatment	1.71 [1.29–2.27]	<0.001			1.89 [1.36–2.63]	<0.001	1.85 [1.27–2.69]	0.001
Costimulatory domain 4-1BB	4.17 [2.04–8.53]	<0.001	3.90 [1.9–8.00]	<0.001	2.81 [1.24–6.35]	0.013	1.89 [0.81–4.4]	0.138
TMTV (dL)	2.01 [1.27–3.17]	0.003			2.33 [1.41–3.86]	0.001	1.44 [0.79–2.61]	0.236
SUV _{max}	1.03 [1.01–1.05]	0.022			1.03 [1–1.06]	0.024		
Low RAD score	2.97 [1.47–5.10]	<0.001	2.72 [1.59–4.67]	<0.001	2.83 [1.52–5.27]	<0.001	2.33 [1.2–4.51]	<0.001

Note: All parameters were assessed from the pretreatment PET imaging.
In bold *p*-value below 0.05 (i.e., statistically significant).

in a small cohort of lymphoma patients. Similar to our results, Zhou et al. found that some textural features had a better performance than conventional PET parameters. Reinert et al. identified temporal associations between radiomic textural features and serologic markers of response such as C-reactive protein, LDH and leukocyte count. Noteworthy, these studies were performed in limited cohorts (less than 30 patients each) and only carried out univariate analysis to associate PET-based radiomic features to outcome. In our study, a multivariate analysis combining different tumor characteristics was performed to define responsive phenotypes. Furthermore, the availability of a larger cohort of patients treated with CAR T cells allowed us to test our model using internal validation.

In our study, tumor heterogeneity by radiomics evaluation was associated with a lower probability of response after CAR T-cell therapy. This could be explained by a diverse composition of the tumor cells and microenvironment and, potentially, heterogeneous expression of the target antigen, which have been described as predictive biomarkers of response in the setting of adoptive T-cell therapies [25, 26].

Several limitations should be taken into account in this study. Despite being one of the largest studies on radiomics in CAR T-cell patients, our results are limited by the relatively small sample size of the validation cohort. Nevertheless, the results are encouraging to explore further in larger, more heterogeneous datasets, allowing an improvement of the model and additional validation. Of note, our pipeline incorporates early preprocessing steps including SUV standardization, making our model easier to be computed from images acquired with different scanners and facilitating its implementation in clinical practice. Furthermore, we have provided the preprocessing methods as well as the equation to obtain the final score. This will facilitate the validation in of our signature in other external cohorts without needing the data access. Moreover, further analysis of these biomarkers, integrating clinical, imaging and molecular data, could lend insight into the biological mechanisms that lead to higher response rates in these patients. Thus, the indication of CAR T-cell therapy should not be

made based on a single variable but integrating radiomics with other prognostic factors.

In conclusion, in this study, we trained and validated a proof-of-concept PET-based radiomics signature to predict response to CAR T cells. This model outperformed conventional PET parameters and showed an independent predictive value after adjusting for well-known clinical factors.

AUTHOR CONTRIBUTIONS

Study design and development: RPL, PB, MS, GI, ML, MB. Data acquisition, management, and retrieval: MS, MB, CC, PB, GI, VN, MSz, SB, AM-N, JI, PA. Image manual segmentation: MS, MB. Data preprocessing, analysis, and inference: ML, MB, VN. Interpretation and validation of results: ML, RPL, PB, GI, RSL, VN. All of the authors reviewed and approved the manuscript for its publication and certify the integrity of the presented work.

CONFLICT OF INTEREST STATEMENT

PB receives honoraria from Allogene, Amgen, BMS, Gilead, Miltenyi Biomedicine, Pierre Fabre, Nektar, Novartis, and Pfizer. GI receives consultancy and honoraria from Novartis, Roche, Kite/Gilead, Bristol-Myers Squibb, Abbvie, Janssen, Sandoz, Miltenyi, and AstraZeneca. SB receives honoraria from AstraZeneca, ROCHE, Janssen, and Abbvie. This work was partially presented at the 64th Annual Meeting of the American Society of Hematology, 2022.

FUNDING INFORMATION

This research received no external funding. RPL is supported by LaCaixa Foundation, a CRIS Foundation Talent Award (TALENT19-05), the FERO Foundation, the Instituto de Salud Carlos III-Investigacion en Salud (PI18/01395 and PI21/01019), and the Prostate Cancer Foundation (18YOUN19). ML is supported by the PERIS PIF-Salut Grant (SLT017/20/000080).

ETHICS STATEMENT

The study was approved by the ethical committee of the Vall d'Hebron University Hospital (PR[AG]404/2020) and conducted in accordance with the Declaration of Helsinki.

PATIENT CONSENT STATEMENT

All patients included in the study provided informed consent

DATA AVAILABILITY STATEMENT

The raw data of imaging scans analyzed in this study are not publicly available due to their containing information, as this would compromise the privacy of research participants. Any queries for data access used in this study should be directed to the corresponding author.

ORCID

Marta Ligeró  <https://orcid.org/0000-0001-9824-7316>

Cecilia Carpio  <https://orcid.org/0000-0001-9346-0134>

Raquel Perez-Lopez  <https://orcid.org/0000-0002-9176-0130>

REFERENCES

1. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375–90.
2. Sant M, Miniccozzi P, Mounier M, Anderson LA, Brenner H, Hollecsek B, et al. EURO-CARE-5 Working Group: survival for haematological malignancies in Europe between 1997 and 2008 by region and age: results of EURO-CARE-5, a population-based study. *Lancet Oncol*. 2014;15:931–42.
3. Howlader N, Noone A-M, Krapcho M, et al. SEER cancer statistics review, 1975–2010. Bethesda, MD: National Cancer Institute. 2013;21:12.
4. Susanibar-Adaniya S, Barta SK. 2021 Update on Diffuse large B cell lymphoma: a review of current data and potential applications on risk stratification and management. *Am J Hematol*. 2021;96(5):617–29.
5. Neelapu SS, Locke FL, Bartlett NL, Lekakis LJ, Miklos DB, Jacobson CA, et al. Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma. *N Engl J Med*. 2017;377(26):2531–44.
6. Schuster SJ, Svoboda J, Chong EA, Nasta SD, Mato AR, Anak Ö, et al. Chimeric antigen receptor T cells in refractory B-cell lymphomas. *N Engl J Med*. 2017;377(26):2545–54.
7. Locke FL, Miklos DB, Jacobson CA, Perales M-A, Kersten M-J, Oluwole OO, et al. Axicabtagene ciloleucel as second-line therapy for large B-cell lymphoma. *N Engl J Med*. 2022;386(7):640–54.
8. Kamdar M, Solomon SR, Arnason J, Johnston PB, Glass B, Bachanova V, et al. Lisocabtagene maraleucel versus standard of care with salvage chemotherapy followed by autologous stem cell transplantation as second-line treatment in patients with relapsed or refractory large B-cell lymphoma (TRANSFORM): results from an interim analysis of an open-label, randomised, phase 3 trial. *Lancet*. 2022;399:2294–308.
9. Iacoboni G, Simó M, Villacampa G, Catalá E, Carpio C, Díaz-Lagares C, et al. Prognostic impact of total metabolic tumor volume in large B-cell lymphoma patients receiving CAR T-cell therapy. *Ann Hematol*. 2021;100(9):2303–10.
10. Vercellino L, Cottreau A-S, Casasnovas O, Tilly H, Feugier P, Chartier L, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. *Blood*. 2020;135(16):1396–405.
11. Dean EA, Mhaskar RS, Lu H, Mousa MS, Krivenko GS, Lazaryan A, et al. High metabolic tumor volume is associated with decreased efficacy of axicabtagene ciloleucel in large B-cell lymphoma. *Blood Adv*. 2020;4(14):3268–76.
12. Cottreau A-S, Nioche C, Dirand A-S, Clerc J, Morschhauser F, Casasnovas O, et al. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med*. 2020;61(1):40–5.
13. Wang J, Hu Y, Yang S, Wei G, Zhao X, Wu W, et al. Role of fluorodeoxyglucose positron emission tomography/computed tomography in predicting the adverse effects of chimeric antigen receptor T cell therapy in patients with non-Hodgkin lymphoma. *Biol Blood Marrow Transpl*. 2019;25(6):1092–8.
14. Schuster SJ, Bishop MR, Tam CS, Waller EK, Borchmann P, McGuirk JP, et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. *N Engl J Med*. 2019;380(1):45–56.
15. Vercellino L, Di Blasi R, Kanoun S, Tessoulin B, Rossi C, D'aveni-Piney M, et al. Predictive factors of early progression after CAR T-cell therapy in relapsed/refractory diffuse large B-cell lymphoma. *Blood Adv*. 2020;4(22):5607–15.
16. Eertink JJ, Van De Brug T, Wiegers SE, Zwezerijnen GJC, Pfähler EAG, Lugtenburg PJ, et al. 18F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma. *Eur J Nucl Med Mol Imaging*. 2022;49(3):932–42.
17. Zhou Y, Hong Z, Zhou M, Sang S, Zhang B, Li J, et al. Prognostic value of baseline 18 F-FDG PET/CT metabolic parameters in paediatric lymphoma. *J Med Imaging Radiat Oncol*. 2020;64(1):87–95.
18. Reinert CP, Perl RM, Faul C, Lengerke C, Nikolaou K, Dittmann H, et al. Value of CT-textural features and volume-based PET parameters in comparison to serologic markers for response prediction in patients with diffuse large B-cell lymphoma undergoing CD19-CAR-T cell therapy. *J Clin Med Res*. 2022;11(6):1522.
19. Lee DW, Santomaso BD, Locke FL, Ghobadi A, Turtle CJ, Brudno JN, et al. ASTCT consensus grading for cytokine release syndrome and neurologic toxicity associated with immune effector cells. *Biol Blood Marrow Transplant*. 2019;25(4):625–38.
20. Hayden PJ, Roddie C, Bader P, Basak GW, Bonig H, Bonini C, et al. Management of adults and children receiving CAR T-cell therapy: 2021 best practice recommendations of the European Society for Blood and Marrow Transplantation (EBMT) and the Joint Accreditation Committee of ISCT and EBMT (JACIE) and the European Haematology Association (EHA). *Ann Oncol*. 2022;33(3):259–75.
21. Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol*. 2014;32(27):3059–67.
22. Ulaner G. Fundamentals of Oncologic PET/CT E-Book. Elsevier Health Sciences; 2018.
23. Ha S, Choi H, Paeng JC, Cheon GJ. Radiomics in oncological PET/CT: a methodological overview. *Nucl Med Mol Imaging*. 2019;53(1):14–29.
24. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–7.
25. Guedan S, Luu M, Ammar D, Barba P, Bonini C, Bousso P, et al. Time 2EVOLVE: predicting efficacy of engineered T-cells—how far is the bench from the bedside? *J Immunother Cancer*. 2022;10(5):e003487.
26. Majzner RG, Mackall CL. Clinical lessons learned from the first leg of the CAR T cell journey. *Nat Med*. 2019;25(9):1341–55.

How to cite this article: Ligeró M, Simó M, Carpio C, Iacoboni G, Balaguer-Montero M, Navarro V, et al. PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma. *eJHaem*. 2023;4:1081–1088. <https://doi.org/10.1002/jha2.757>

4.4 Study 4

Radiomics and outcome prediction to anti-angiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumours: findings from the phase II TALENT trial

This study corresponds to **objective 3**. In this research, we implemented radiomics pipelines to predict response to anti-angiogenics treatment in a population of patients with neuroendocrine tumors, characterized by a distinct imaging phenotype attributed to their highly vascularized nature. Additionally, the unique attributes of this tumor type, which is usually scanned in multiple CT-acquisition phases, facilitated the exploration of multi-modal radiomics analysis employing data from portal and arterial phases.

Reference:

Ligero M., Hernando J., Delgado E. et al. Radiomics and outcome prediction to anti-angiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumours: findings from the phase II TALENT trial. BJC Reports 1, 9 (2023).

ARTICLE OPEN



Radiomics and outcome prediction to antiangiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumours: findings from the phase II TALENT trial

Marta Liger^{1,16}, Jorge Hernando^{2,16}, Eric Delgado¹, Alonso Garcia-Ruiz¹, Xavier Merino-Casabiel³, Toni Ibrahim⁴, Nicola Fazio⁵, Carlos Lopez⁶, Alexandre Teulé⁷, Juan W. Valle⁸, Salvatore Tafuto⁹, Ana Custodio¹⁰, Nicholas Reed¹¹, Markus Raderer¹², Enrique Grande¹³, Rocio Garcia-Carbonero¹⁴, Paula Jimenez-Fonseca¹⁵, Alejandro Garcia-Alvarez², Manuel Escobar³, Oriol Casanovas⁷, Jaume Capdevila^{2,17} and Raquel Perez-Lopez^{1,3,17}✉

© The Author(s) 2023

BACKGROUND: More accurate predictive biomarkers in patients with gastroenteropancreatic neuroendocrine tumours (GEP-NETs) are needed. This study aims to investigate radiomics-based tumour phenotypes as a surrogate biomarker of the tumour vasculature and response prediction to antiangiogenic targeted agents in patients with GEP-NETs.

METHODS: In this retrospective study, a radiomics signature was developed in patients with GEP-NETs and liver metastases receiving lenvatinib. Patients were selected from the multicentre phase II TALENT trial (NCT02678780) (development cohort). Radiomics variables were extracted from liver metastases in the pre-treatment CT-scans and selected using LASSO regression and minimum redundancy maximum relevance (mRMR). Logistic regression and Cox proportional-hazards models for radiomics and combined radiomics with clinical data were explored. The performance of the models was tested in an external cohort of patients treated with sunitinib (test cohort). Associations between the radiomics score and vascularisation factors in plasma were studied using hierarchical clustering and Mann–Whitney *U* test.

RESULTS: A total of 89 patients were included in the study, 408 liver metastases were analysed. The CT-based radiomics signature was associated with clinical benefit in the development (training and validation sets) and test cohorts (AUC 0.75 [0.66–0.90], 0.67 [0.49–0.92] and 0.67 [0.43–0.91], respectively). The combined radiomics-clinical signature (including the radiomics score, Ki-67 index and primary tumour site) improved on radiomics-only signature performance (AUC 0.79 [95% CI 0.64–0.93]; *p* < 0.001). A higher radiomics score indicated longer progression-free survival (hazard ratio of 0.11 [0.03–0.45]; *p* = 0.002) and was associated with vascularisation factors (*p* = 0.01).

CONCLUSIONS: Radiomics-based phenotypes can provide valuable information about tumour characteristics, including the vasculature, that are associated with response to antiangiogenics.

CLINICAL TRIAL REGISTRATION: This is a study of the Lenvatinib Efficacy in Metastatic Neuroendocrine Tumours (TALENT) phase II clinical trial (NCT02678780).

BJC Reports; <https://doi.org/10.1038/s44276-023-00010-0>

BACKGROUND

Neuroendocrine tumours (NETs) are considered heterogeneous and complex to treat malignancies. Nevertheless, in the last two decades the treatment landscape for gastroenteropancreatic neuroendocrine tumours (GEP-NETs) has improved considerably, with an increase in available treatment strategies, making patient stratification and treatment selection more challenging. There is a

wide range of effective therapies including somatostatin analogues, radiolabelled somatostatin analogues, molecular-targeted agents or chemotherapy [1–3]. Among them, antiangiogenic targeted therapies such as lenvatinib have demonstrated a high radiological response rate in these patients [4, 5]. Several clinical indicators have been considered as prognostic factors for NETs such as the Ki-67 proliferative index and disease staging [6].

¹Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. ²Oncology Department, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. ³Radiology Department, Vall d'Hebron University Hospital (VHUH), Barcelona, Spain. ⁴Osteoncology and Rare Tumours Centre, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST), Meldola, Italy. ⁵Units of Gastrointestinal and Neuroendocrine Tumours, European Institute of Oncology, Milan, Italy. ⁶Oncology Department, Marques de Valdecilla University Hospital (IDIVAL), Santander, Spain. ⁷Oncology Department, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat (Barcelona), Spain. ⁸University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. ⁹S.C. Sarcomi e Tumori Rari, Istituto Nazionale Tumori, IRCCS, Fondazione "G. Pascale", Naples, Italy. ¹⁰Oncology Department, La Paz University Hospital, Madrid, Spain. ¹¹Gartnavel Hospital, Beatson Oncology Centre, Glasgow, UK. ¹²Department of Oncology and Internal Medicine, Medical University of Vienna, Vienna, Austria. ¹³Oncology Department, MD Anderson Cancer Center, Madrid, Spain. ¹⁴Oncology Department, Hospital Universitario 12 de Octubre, Ima12, UCM, Madrid, Spain. ¹⁵Oncology Department, Central de Asturias University Hospital, Oviedo, Spain. ¹⁶These authors contributed equally: Marta Liger, Jorge Hernando. ¹⁷These authors jointly supervised this work: Jaume Capdevila, Raquel Perez-Lopez. ✉email: rperez@vhio.net

Received: 2 March 2023 Revised: 6 July 2023 Accepted: 10 July 2023

Published online: 02 August 2023

However, there remains a clinical need to find predictive biomarkers of response to novel targeted therapies for achieving more precise patient selection [7].

Radiomics analysis allows for the extraction of quantitative data from routinely acquired medical images and correlation of imaging features with the underlying tumour characteristics including the tumour vasculature [8]. Radiomics opens a window of opportunity to develop new tools for improved prediction and response evaluation to novel treatment options. Current studies have been focused on applying radiomics for NET staging, grade characterisation and determination of alternative prognostic factors [9–19], but there is scarce data about radiomics to predict response to antiangiogenics in NETs [19]. Accounting for the emerging treatment strategies for NET patients, radiomics could also contribute in determining the best therapeutic approach for each patient.

The aim of this study was to develop and evaluate the performance of a computed tomography (CT)-based radiomics signature for tumour response prediction to antiangiogenic agents prior to treatment. Patients with advanced GEP-NETs enrolled in a multicentre phase II clinical trial conducted with the antiangiogenic targeted agent lenvatinib were included in the study [5]. To validate the signature, an external cohort of patients with pancreatic NET treated with another multikinase inhibitor with also proved antiangiogenic activity used in routine clinical practice, sunitinib, was studied. Secondly, a multiphase model, that included information of both arterial and venous CT-image acquisitions, was explored to see whether this improved the performance for response prediction. Moreover, integration of the CT-based radiomic model and clinical data was also investigated in an attempt to improve the predictive value of this tool. We hypothesised that a CT-based radiomics phenotype can provide meaningful information about biology of NETs (including the tumour vasculature) and its susceptibility to respond to antiangiogenic treatment. The CT-based radiomics signature could be used to stratify patients by identifying those that are more likely to benefit from antiangiogenic targeted agents.

METHODS

The institutional review board approved this retrospective study. Informed consent for computational image analysis was waived.

All patients included in the clinical trial Lenvatinib Efficacy in Metastatic Neuroendocrine Tumours (TALENT) NCT02678780 [5] provided written informed consent.

Study sample

The development cohort consisted of patients with GEP-NETs treated with the multikinase inhibitor lenvatinib in a multicentre, international, phase II clinical trial conducted from October 2015 to August 2020, identified as NCT02678780 (Supplementary Table 1 participant centres). The test cohort consisted of patients with pancreatic NETs treated with the multikinase inhibitor sunitinib as standard of care at the Vall d'Hebron Institute of Oncology (Barcelona, Spain) from October 2011 to September 2020.

Patients with GEP-NET liver metastasis and intravenous contrast-enhanced CT scans at baseline were included. Patients with artifacts at baseline CT scans and patients in which the clinical outcome could not be assessed due to toxicity or non-disease related death were excluded. A total of 408 liver metastases from 89 patients were included (46 men [52%] and 43 women [48%]), mean age 62 years (range 33–86).

Image acquisition and radiomics analysis

All CT scans were acquired within 28-days before the treatment starting day. Contrast-enhanced images were obtained with 16- or 64-channel CT scanners (Siemens, Philips, GE Medical Systems, Toshiba, Agfa), 1–5 mm slice thickness and 100–120 kV of voltage (Supplementary Table 2). Up to six liver metastases per patient of at least 1 cm diameter were segmented using the semi-automatic segmentation tool of 3D Slicer (version 4.11.0; www.slicer.org; RRID:SCR_005619) [20] by a radiologist with more than 10 years of experience in oncological imaging (RPL). The tumours were

segmented in the CT arterial or venous phase depending on the tumour contrast enhancement, selecting the phase in which tumours were better depicted (Supplementary Fig. 1). Combined models including information from both CT contrast-enhanced phases (arterial and venous) were also explored. Images were resampled to isotropic voxels of $1 \times 1 \times 1 \text{ mm}^3$ by using spline interpolation. Hounsfield units were binarized to discrete values of 25 HU. CT-based radiomics features including first-order, shape and five gray-level texture matrices (Gray Level Co-occurrence Matrix [GLCM], Gray Level Dependence Matrix [GLDM], Gray Level Run Length Matrix [GLRLM], Gray Level Size Zone Matrix [GLSZM], and Neighbouring Gray Tone Different Matrix [NGTDM]) were calculated in three dimensions using the Pyradiomics package (version 3.0.1) for Python (version 3.6.13 Python Software Foundation, Delaware, USA; RRID:SCR_008394), compliant with the Image Biomarker Standardisation Initiative guidelines [21]. To investigate the prediction power of a multiphase radiomics model, the arterial and venous CT-images were co-registered and the radiomics features from both acquisition phases were extracted (Supplementary Methods).

Clinical data

Clinical data from the development cohort were collected from the TALENT clinical trial database where age, sex, primary tumour site, Ki-67 index, tumour burden, tumour grade and pre-treatment were registered. Vascular endothelial growth factor receptor 2 (VEGFR2) and angiopoietin 2 (ANG2) quantifications were obtained from patients with plasma samples. VEGFR2 and ANG2 plasma levels were determined by multiplex ELISA with a custom-made glass-slide sandwich Quantibody Array (RayBioTech, GA, USA). Regarding the test cohort, data were obtained from the electronic patient records. Clinical benefit was defined as achieving either a complete response (CR), partial response (PR) or stable disease (SD) by RECIST 1.1 for a duration exceeding the median progression-free survival (PFS) in the populations according to the phase II and III clinical trials [5, 22]. Therefore, patients treated with lenvatinib who progressed before 15.7 months and those treated with sunitinib who progressed before 11.4 months were considered as not achieving clinical benefit. Clinical data were used to develop a clinical only and a combined clinical-radiomics model.

Modelling and statistical analysis

The development cohort was divided into training and validation sets (70–30%) balanced for outcome in both sets and then tested in an external cohort of patients with GEP-NETs (test cohort). As a standard method in machine learning modelling, the training set corresponded to the group of patients where the model was initially trained and the hyperparameters were fitted using cross-validation; in the validation set, we explored the model performance robustness; the final model generalisability was evaluated in the test cohort (i.e., the external validation cohort).

The median radiomics value of all the evaluated liver tumours per patient was implemented as feature aggregation method. We performed a two-step procedure for feature selection based on LASSO and minimum Redundancy Maximum Relevance (mRMR), these two have been shown to benchmark other feature selection methods [23]. Three-fold cross-validation was performed for LASSO hyperparameter tuning, choosing the model with one standard error from the best area under the curve (AUC). To avoid multicollinearity in the model features, mRMR was implemented to reduce the number of variables controlling for a variance inflation factor (VIF) < 2 and a Pearson square $R < 0.5$ [24]. Logistic regression model was performed including the selected features. The area under the curve (AUC) and 95% confidence interval (CI) (DeLong method) were computed from the receiver operating characteristic (ROC) curve and p values were assessed from Mann–Whitney U test. The decision threshold to compute sensitivity and specificity was defined by Youden's index.

We explored the combination of the radiomics score with established prognostic factors in NETs, including Ki-67 index and primary tumour site. Since tumour grade is correlated with Ki-67 expression, it was not included as a separate factor in our analysis. Furthermore, the assessment of tumour burden, which can indicate different prognoses, was already incorporated into the radiomics model analysis. A logistic regression model including the radiomics score and non-correlated clinical variables (Ki-67 index, primary tumour site) was developed. Clinical data imputation was done using random forest. PFS associations with clinical variables and predictive radiomics scores were investigated using Cox Proportional-Hazard regression and Log-rank test. A multiphase model combining radiomics features from arterial and venous phases was also investigated (Supplementary Methods).

Patients with both imaging and liquid biopsy were clustered based on vascularisation factors (including VEGFR2 and ANG2) using hierarchical clustering. Associations between radiomics score and vascularisation factors were studied using Mann–Whitney *U* test. The study workflow is summarised in Fig. 1.

RESULTS

Study patient characteristics

From a total of 128 patients (88 from the development and 40 from the test cohorts), 15 patients without baseline contrast-enhanced CT scans or with artifacts in the area of interest were excluded, 19 without liver metastases or measurable lesions per RECIST 1.1 and 5 patients who presented toxicity before clinical outcome evaluation were also excluded. Eighty-nine patients were included in the radiomics analysis, 65 patients to build up the model (development cohort) and 24 for external validation (test cohort). Forty-four patients from the TALENT clinical trial had concomitant CT images and liquid biopsy samples.

For the exploratory analysis of combined information from both CT contrast-enhanced phases (arterial and venous), 65 patients (41 from the development and 24 from the test cohorts) had CT scans with both arterial and venous phases. A total of 282 tumours (5 [1–6] mean[range] lesion/patient) from the 65 patients of the development cohort; 34 women and 31 men were included in the final analysis. The median PFS was 14.9 [IQR 3.04–46.72] months; 46% (30/65) of patients presented clinical benefit (i.e., CR, PR or SD over the median PFS) and 54% (35/65) did not.

The model was validated in the test cohort including a total of 126 tumours from 24 patients (5 [1–6] mean[range] lesion/patient); 12 women and 12 men. The median PFS was 9.75 [IQR 2.90–29.47] months; 50% (12/24) of patients presented clinical benefit and 50% (12/24) did not.

In the development cohort, clinical benefit was defined as the combination of CR, PR or SD after 15.7 months, corresponding to the median PFS time in the TALENT clinical trial [5]. In the test cohort the cut-off for clinical benefit was defined at 11.4 months, corresponding to the median PFS time in the cohort [22].

Flow chart of the study population selection is shown in Supplementary Fig. 2. The population description is reported in Table 1.

Predictive model development and testing

LASSO regression lambda hyperparameter was set to 0.03 after cross-validation analysis with 13 radiomics features with non-zero coefficients. Six radiomics features were selected using mRMR method as the maximum number of variables to avoid multicollinearity in logistic regression. The final model included variables from first order, shape, and GLCM and GLRLM texture matrices (Table 2). The radiomics model combining the selected features predicted clinical benefit with an AUC of 0.75 [95% CI 0.60–0.90; $p = 0.001$] and 0.67 [95% CI 0.41–0.92; $p = 0.115$] in the training and validation sets, respectively. In the test cohort, the radiomics predicted response with an AUC of 0.67 [95% CI 0.43–0.91; $p = 0.060$] (Fig. 2a). Sensitivity and specificity for classifying patients with clinical benefit with an optimal Youden's cut-off of 0.49 are described in Table 3. Internal cross-validation showed a mean AUC of 0.67 [0.50–0.87] (Supplementary Fig. 3). The radiomics model showed that tumour sphericity, heterogeneity assessed by GLCM Informational Measure of Correlation (imc), and enhancement were associated with clinical benefit (Fig. 2d). No significant differences in radiomics scores were found between patients previously treated with antiangiogenics and non-pre-treated patients ($p = 0.33$) (Supplementary Fig. 4). The radiomics score for predicting clinical benefit showed significant association with continuous PFS (HR 0.11[0.03–0.45]; $p = 0.002$) (Fig. 2b).

The exploratory multiphase predictive model (combining information from the arterial and venous CT phases) did not improve the prediction capacity (AUC in test set 0.63 [0.39–0.87]) (Supplementary Fig. 5).

Integrating radiomics and clinical data

The clinical model combining Ki-67 and primary tumour site showed poor performance in both the development (training 0.57 [0.40–0.74] and validation 0.56 [0.30–0.82]) and test sets (0.23 [0–0.46]). When combined with the radiomics score, the radiomics-clinical model predicted response with an AUC of 0.79 [0.64–0.93; $p < 0.001$] and 0.72 [0.48–0.95; $p = 0.06$] in the training and validation sets, respectively. In the test cohort, the combined radiomics-clinical score associated with response with an AUC of 0.58 [0.33–0.82; $p = 0.27$]. The sensitivity and specificity for classifying patients with clinical benefit with an optimal cut-off of 0.36 are described in Table 3. The ROC curves of the combined signature integrating radiomics-score and clinical variables were also computed (Fig. 2a). Tumour burden was not included in this integrative analysis because it was quantitatively included in the radiomics analysis. Tumour grade was also excluded due to correlation with Ki-67 index. The combined clinical-radiomics score for predicting clinical benefit showed significant association with continuous PFS (HR 0.12[0.04–0.43]; $p = 0.001$) (Fig. 2b). However, none of the clinical variables including Ki67, primary tumour site or tumour burden showed significant associations with PFS ($p < 0.05$) (Supplementary Fig. 6).

Correlation of the CT-based radiomics score and tumour vasculature

The study population was clustered according to proangiogenic profiling factors, VEGFR2 and ANG2, by hierarchical clustering. Cluster A included 21 patients grouped as highly vascularised tumours showing higher values of VEGFR2 [median 18,784 (IQR 17,579–21,842) pg/ml] and ANG2 [2062 (1296–3328) pg/ml] expression. Cluster B included 23 patients with poorly vascularised tumours showing lower VEGFR2 [2289 (1577–9776) pg/ml] and ANG2 level expression [709 (4079–1204) pg/ml]. Significant associations were found between the radiomics score and vascularisation clusters showing higher radiomics score in the vascularised cluster A [0.51 (0.37–0.66)] and lower radiomics score in the poorly vascularised cluster B [0.29 (0.21–0.445)] ($p = 0.01$) (Fig. 2c).

DISCUSSION

The increasing development of targeted therapies for NETs has pointed out the unmet need to define predictive biomarkers of response for improving patient selection. The neoangiogenesis process is a key feature in NET carcinogenesis and the field of new antiangiogenic compound development is still active after sunitinib, the only tyrosine kinase inhibitor utilisable in clinical practice. In this study we explored an imaging phenotype based on radiomics features from baseline CT scans that allows the identification of patients most likely to benefit from antiangiogenic treatment.

The developed radiomics signature outperforms clinical features (including Ki-67 and primary tumour type) for predicting response to treatment (AUC 0.75 vs. 0.57). The radiomics signature consisted of five radiomics features, indicating that patients with more spherical and heterogeneous tumours are more likely to respond to antiangiogenic treatment. Also, more intravenous contrast-enhancing tumours are associated with better response. Therefore, we investigated the association of tumour vascularisation (by means of plasma VEGFR2 and ANG2) and the predictive radiomics score, showing that in the TALENT cohort, the CT-radiomics score was associated with VEGFR2 and ANG2 expression. This suggests that radiomics quantification from baseline CT

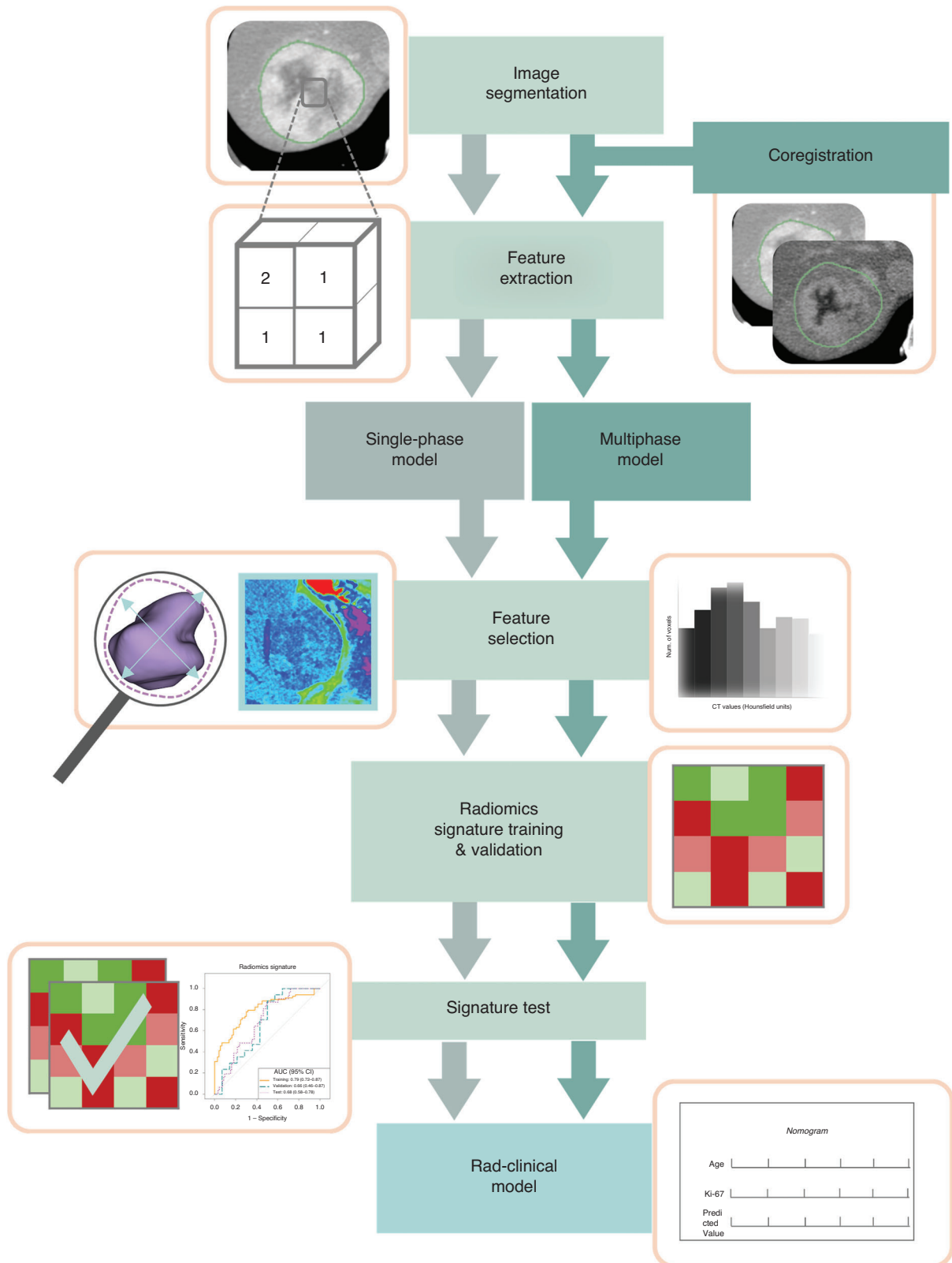


Fig. 1 Radiomics analysis flowchart. The single-phase model is obtained from one of both acquisition CT phases (arterial or venous) of the development cohort, whereas the multiphase model included the co-registered images of the arterial and venous phases. Feature selection and model fitting was performed for both approaches. The model is validated in the test cohort and combined with clinical data to improve the predictive capacity.

scans can non-invasively provide valuable information about the tumour vascularisation. The TALENT research group has previously shown that in antiangiogenic pre-treated patients, the plasma proangiogenic profiling (VEGFR2 and ANG2 levels) significantly predicts response in NETs [25]. In the subset of non-pre-treated patients, different mechanisms of vascularisation and activation of alternative signalling pathways could sustain the lack of association between proangiogenic profiling and tumour response [26]. Interestingly, our signature did not show significant accuracy differences between patients who had received previously antiangiogenics or not, which may indicate that the radiomics

phenotype captures the tumour characteristics that make it susceptible to antiangiogenic response that includes, but is not limited to, tumour vascularisation.

The capacity of the radiomics signature to predict response was tested in an independent population of patients with pancreatic NETs treated with sunitinib, another multikinase inhibitor with antiangiogenic effects but with different targets and affinity than lenvatinib, with a stable performance of the radiomics signature. The integration of prognostic factors such as Ki-67 index and primary tumour site improved the performance of the radiomics-only signature. However, the combined model (clinical-radiomics) presented a modest performance in the sunitinib population that can be related to the clinical differences between patients receiving standard of care treatment and those within clinical trials.

In previous research, a radiomics model based on the analysis of the entire liver was reported, demonstrating predictive value for response to everolimus in advanced NETs [19]. However, to our knowledge, no previous studies have been reported exploring the role of CT-radiomics in predicting response to antiangiogenic multikinase inhibitors in NETs. Several studies have shown that high-grade NETs present lower density (likely related to a lower intratumoral microvascular density) than low grade NETs [18, 27]. Our study population included mostly grade 1 and 2 NETs, and no significant differences in the tumour enhancement was identified between them (Supplementary Fig. 7). Therefore, the relevance of tumour intensity in the CT-radiomics signature represents most likely a true specific predictive biomarker rather than an indicator of aggressiveness or tumour grade.

Although NETs are often highly-dense tumours in contrast-enhanced CT, usually more conspicuous in the arterial phase of the CT scan, some tumours present different patterns and, as shown in the study population, some tumours are better depicted in the venous phase [28]. In this study, an expert radiologist selected the phase (arterial or venous) of the CT scan where the tumour was best depicted. We also investigated the integration of a multiphase model (combining radiomics data from arterial and venous phases) to address this concern. We demonstrated that including radiomics features from multiple phases did not improve the prediction in our population. Furthermore, we correlated the radiomics score with vascularisation factors (VEGFR2 and ANG2) in the population with concomitant imaging and plasma samples and showed significant associations between the radiomics-score and vascularisation-factor expression.

There are some limitations encountered in this study. First, the population used for the training dataset was tied to the clinical trial population. The test cohort was treated with a different agent, although both are multikinase inhibitors with an antiangiogenic effect, and the population slightly differs clinically from the development cohort (i.e., primary site and grade). This limitation is mainly affecting the performance of the combined radiomics-clinical model in the test dataset. Second, we developed a model using only one acquisition phase selected by an expert radiologist, which can influence the radiomics features. However, we explored the performance of including both phases and it did not improve the prediction

Table 1. Population characteristics of the development and test cohorts.

Parameter	Development cohort (n = 65)	Test cohort (n = 24)
Mean age (years) ^a	60 ± 12 (33–85)	63 ± 12 (38–86)
Sex		
Female	34 (52)	12 (50)
Male	31 (48)	12 (50)
Primary tumour site		
PanNET	24 (37)	23 (96)
GI-NET	33 (51)	1 (4)
Unknown	8 (12)	0 (0)
Pre-treatment		
Sunitinib	10 (15)	NA
Everolimus	19 (29)	NA
None	36 (55)	NA
Best response		
Partial response	24 (37)	6 (25)
Stable disease	40 (62)	15 (62)
Progressive disease	1 (1)	3 (13)
Mean Ki-67 ^{a,b}	7.13 ± 5.87 (1.00–18.00)	21.96 ± 18.98 (1.30–55.10)
Grade ^b		
Grade 1	22 (34)	3 (12)
Grade 2	40 (61)	12 (50)
Grade 3	...	9 (38)
Unknown	3 (5)	
PFS (months) ^c	14.9 [3.04–46.72]	9.75 [2.90–29.47]
OS (months) ^c	33.7 [9.21–46.94]	12.1 [4.00–47.00]
Number of lesions	4 (1–6)	5 (1–6)
Mean hepatic tumour burden (dl) ^a	0.037 (0.001–2.209)	1.086 (0.007–1.054)
CT segmentation phase		
Arterial	23 (35)	7 (29)
Pancreatic NET	14 (21)	6 (25)
Gastrointestinal NET	9 (14)	1 (4)
Venous	42 (65)	17 (71)
Pancreatic NET	15 (23)	17 (71)
Gastrointestinal NET	27 (42)	...

Data in parentheses are percentages.

^aData are ± standard deviation and parentheses are range.

^bData presented has no information from three patients.

^cData are median [interquartile range].

Table 2. Radiomics model coefficients.

Radiomics features	Coefficient [95% CI]
Shape flatness	−0.82 [−1.6–−0.05]
GLCM lmc1	0.31 [−0.47–1.09]
First order 90th percentile	0.37 [−0.29–1.04]
GLRLM Long Run Low Gray Level Emphasis	0.16 [−0.58–0.91]
First order skewness	0.34 [−0.35–1.03]
Shape sphericity	0.84 [0.05–1.63]

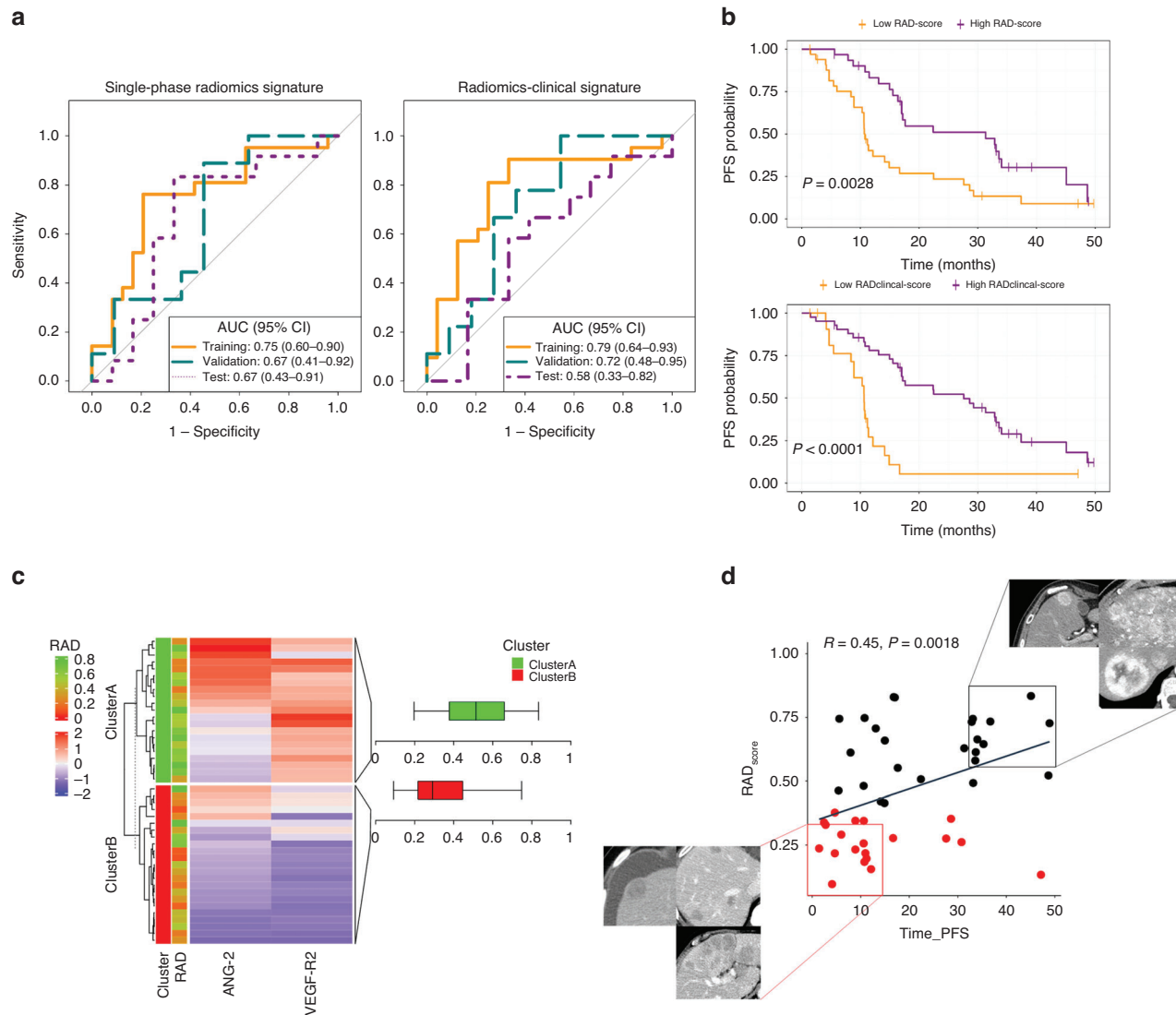


Fig. 2 Response prediction performance and explainability. Performance of the radiomics and radiomics-clinical models to predict response to treatment using: **a** Receiver operating characteristic (ROC) curves for median progression-free survival (PFS) prediction and **b** Kaplan–Meier curves for PFS probability. Explainability of the radiomics signature: **c** Associations between vascular factor expression and radiomics score. The cluster A corresponds to highly vascularised tumours with high-radiomics score, while cluster B includes tumours with poor vascularisation and low-radiomics score. **d** Visualisation of tumours with high-radiomics score (i.e., more spherical, heterogenous and highly enhancing) and long PFS vs. low-radiomics score and short PFS. Spearman's rank correlation analysis ($p < 0.01$).

Table 3. Radiomics and radiomics-clinical model performance evaluation.

	AUC [95% CI]	<i>p</i> value	Accuracy (%)	Sensitivity (%)	Specificity (%)
Radiomics model					
Training	0.75 [0.60–0.90]	<0.001	78	76	79
Validation	0.67 [0.41–0.92]	0.114	55	55	56
Test	0.67 [0.43–0.91]	0.09	63	67	58
Radiomics clinical model					
Training	0.79 [0.64–0.93]	<0.001	78	90	67
Validation	0.72 [0.48–0.95]	0.12	70	45	100
Test	0.58 [0.33–0.82]	0.276	54	17	92

Table containing the area under the curve (AUC), accuracy, sensitivity and specificity for the Youden's cut-off 0.49 for the radiomics-only model and 0.36 for the radiomics-clinical model.

capacity and showed higher model overfitting. Third, the images in this study underwent evaluation by a consistent expert radiologist. Nevertheless, it is important to acknowledge that variations in segmentation techniques and observers may introduce biases that could impact the reliability of first-order and textural radiomics features, as well as shape and size metrics [29]. Finally, while we recognise the need for homogeneity in radiomics modelling, achieved by focusing solely on liver metastases, we acknowledge that further exploration and potential validation in larger cohorts encompassing primary tumours and other metastatic sites should be pursued. This will contribute to a more comprehensive understanding of the radiomics-based phenotypes in GEP-NETs.

In conclusion, the GEP-NET phenotype evaluated by means of CT-based radiomics can be a useful non-invasive surrogate of the tumour vasculature and has predictive value of tumour response to antiangiogenic targeted agents. The improved prediction of response to antiangiogenic therapy achieved by combining radiomics with clinical prognostic factors can facilitate medical decision making and optimise treatment outcomes for patients with GEP-NETs. Nevertheless, despite the promising results of this study, further research is necessary in larger, prospective trials for implementation of this tool in clinical practice.

DATA AVAILABILITY

The raw data of imaging scans analysed in this study are not publicly available due to their containing information, as this would compromise the privacy of research participants. Any queries for data access used in this study should be directed to the corresponding author. The codes can be publicly accessed at https://github.com/radiomicsgroup/TALENT_project. We relied on the open-source software dcm2niix (<https://github.com/rordenlab/dcm2niix/>) for DICOM conversion and 3D Slicer (www.slicer.org; RRID:SCR_005619) [20] for image annotations and computing.

REFERENCES

- Cives M, Pelle E, Strosberg J. Emerging treatment options for gastroenteropancreatic neuroendocrine tumors. *J Clin Med*. 2020;9:3655.
- Ito T, Masui T, Komoto I, Doi R, Osamura RY, Sakurai A, et al. JNETS clinical practice guidelines for gastroenteropancreatic neuroendocrine neoplasms: diagnosis, treatment, and follow-up: a synopsis. *J Gastroenterol*. 2021;56:1033–44.
- Cives M, Strosberg JR. Gastroenteropancreatic neuroendocrine tumors. *CA Cancer J Clin*. 2018;68:471–87.
- Megdanova-Chipeva VG, Lamarca A, Backen A, McNamara MG, Barriuso J, Sergieva S, et al. Systemic treatment selection for patients with advanced pancreatic neuroendocrine tumours (PanNETs). *Cancers*. 2020;12:1988.
- Capdevila J, Fazio N, Lopez C, Teulé A, Valle JW, Tafuto S, et al. Lenvatinib in patients with advanced grade 1/2 pancreatic and gastrointestinal neuroendocrine tumors: results of the phase II TALENT Trial (GETNE1509). *J Clin Oncol*. 2021;39:2304–12.
- García-Carbonero R, Capdevila J, Crespo-Herrero G, Díaz-Pérez JA, Martínez Del Prado MP, Alonso Orduña V, et al. Incidence, patterns of care and prognostic factors for outcome of gastroenteropancreatic neuroendocrine tumors (GEP-NETs): results from the National Cancer Registry of Spain (RGETNE). *Ann Oncol*. 2010;21:1794–803.
- Auernhammer CJ, Spitzweg C, Angele MK, Boeck S, Grossman A, Nölting S, et al. Advanced neuroendocrine tumours of the small intestine and pancreas: clinical developments, controversies, and future strategies. *Lancet Diabetes Endocrinol*. 2018;6:404–15.
- Braman N, Prasanna P, Bera K, Alilou M, Khorrami M, Leo P, et al. Novel radiomic measurements of tumor-associated vasculature morphology on clinical imaging as a biomarker of treatment response in multiple cancers. *Clin Cancer Res*. 2022;28:4410–24.
- Canellas R, Burk KS, Parakh A, Sahani DV. Prediction of pancreatic neuroendocrine tumor grade based on CT features and texture analysis. *Am J Roentgenol*. 2018;210:341–6.
- D'Onofrio M, Ciaravino V, Cardobi N, De Robertis R, Cingarlini S, Landoni L, et al. CT enhancement and 3D texture analysis of pancreatic neuroendocrine neoplasms. *Sci Rep*. 2019;9:2176.
- Gu D, Hu Y, Ding H, Wei J, Chen K, Liu H, et al. CT radiomics may predict the grade of pancreatic neuroendocrine tumors: a multicenter study. *Eur Radiol*. 2019;29:6880–90.
- Wang H, Lin Z, Li G, Zhang D, Yu D, Lin Q, et al. Validation and modification of staging systems for poorly differentiated pancreatic neuroendocrine carcinoma. *BMC Cancer*. 2020;20:188.
- Benedetti G, Mori M, Panzeri MM, Barbera M, Palumbo D, Sini C, et al. CT-derived radiomic features to discriminate histologic characteristics of pancreatic neuroendocrine tumors. *Radiol Med*. 2021;126:745–60.
- Guo C, Zhuge X, Wang Q, Xiao W, Wang Z, Wang Z, et al. The differentiation of pancreatic neuroendocrine carcinoma from pancreatic ductal adenocarcinoma: the values of CT imaging features and texture analysis. *Cancer Imaging*. 2018;18:37.
- Loi S, Mori M, Benedetti G, Partelli S, Broggi S, Cattaneo GM, et al. Robustness of CT radiomic features against image discretization and interpolation in characterizing pancreatic neuroendocrine neoplasms. *Phys Med*. 2020;76:125–33.
- Martini I, Polici M, Zerunian M, Panzuto F, Rinzivillo M, Landolfi F, et al. CT texture analysis of liver metastases in PNETs versus NPNETs: correlation with histopathological findings. *Eur J Radiol*. 2020;124:108812.
- Okabe H, Hashimoto D, Chikamoto A, Yoshida M, Taki K, Arima K, et al. Shape and enhancement characteristics of pancreatic neuroendocrine tumor on pre-operative contrast-enhanced computed tomography may be prognostic indicators. *Ann Surg Oncol*. 2017;24:1399–405.
- Rodallec M, Vilgrain V, Couvelard A, Rufat P, O'Toole D, Barrau V, et al. Endocrine pancreatic tumours and helical CT: contrast enhancement is correlated with microvascular density, histoprognostic factors and survival. *Pancreatol*. 2006;6:77–85.
- Caruso D, Polici M, Rinzivillo M, Zerunian M, Nacci I, Marasco M, et al. CT-based radiomics for prediction of therapeutic response to Everolimus in metastatic neuroendocrine tumors. *Radiol Med*. 2022;127:691–701.
- Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30:1323–41.
- Zwanenburg A, Vallières M, Abdallah MA, Aerts H, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295:328–38.
- Raymond E, Dahan L, Raoul JL, Bang YJ, Borbath I, Lombard-Bohas C, et al. Sunitinib malate for the treatment of pancreatic neuroendocrine tumors. *N Engl J Med*. 2011;364:501–13.
- Demircioglu A. Benchmarking feature selection methods in radiomics. *Invest Radiol*. 2022;57:433–43.
- Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology*. 2016;6:227.
- Capdevila J, Jimenez-Valerio G, Martinez A, Hernandez J, Ibrahim T, Fazio N, et al. Plasma biomarker study of lenvatinib in gastroenteropancreatic neuroendocrine tumors reveals Ang2 and FGF2 as predictors of treatment response: results from the international phase II TALENT trial (GETNE 1509). *J Clin Oncol*. 2021. https://doi.org/10.1200/JCO.2021.39.15_suppl.4113.
- Haibe Y, Kreidieh M, El Hajj H, Khalifeh I, Mukherji D, Temraz S, et al. Resistance mechanisms to anti-angiogenic therapies in cancer. *Front Oncol*. 2020;10:221.
- Pulvirenti A, Yamashita R, Chakraborty J, Horvat N, Seier K, McIntyre CA, et al. Quantitative computed tomography image analysis to predict pancreatic neuroendocrine tumor grade. *JCO Clin Cancer Inform*. 2021;5:679–94.
- Sahani DV, Bonaffini PA, Fernandez-Del Castillo C, Blake MA. Gastroenteropancreatic neuroendocrine tumors: role of imaging in diagnosis and management. *Radiology*. 2013;266:38–61.
- Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102:1143–58.

ACKNOWLEDGEMENTS

This research received no external funding. RPL is supported by LaCaixa Foundation, a CRIS Foundation Talent Award (TALENT19-05), the FERO Foundation, the Instituto de Salud Carlos III-Investigación en Salud (PI18/01395 and PI21/01019) and the Prostate Cancer Foundation (18YOUN19). ML is supported by the PERIS PIF-Salut Grant (SLT017/20/000080).

AUTHOR CONTRIBUTIONS

The authors of this submission have made significant contributions to the work presented herein. Specifically, each author has played a critical role in conceiving and/or designing the study, acquiring and analysing the data, and interpreting the results. Additionally, each author has actively participated in drafting and/or revising the manuscript to ensure its accuracy and clarity. Furthermore, all authors have given their approval for the final version of the manuscript and have agreed to be

accountable for all aspects of the work. This includes a commitment to investigate and resolve any questions related to the accuracy or integrity of any part of the work to ensure its validity and credibility.

COMPETING INTERESTS

JH has received honoraria for speaker engagements, advisory roles or funding of continuous medical education from Eisai, Ipsen, Novartis, AAA, Angelini, Roche, Pfizer, LEO Pharma. EG has received honoraria for speaker engagements, advisory roles or funding of continuous medical education from Adacap, AMGEN, Angelini, Astellas, AstraZeneca, Bayer, Blueprint, Bristol Myers Squibb, Caris Life Sciences, Celgene, Clovis-Oncology, Eisai, Eusa Pharma, Genetracer, Guardant Health, HRA-Pharma, IPSEN, ITM-Radiopharma, Janssen, Lexicon, Lilly, Merck KGaA, MSD, Nanostring Technologies, Natera, Novartis, ONCODNA (Biosequence), Palex, Pharmamar, Pierre Fabre, Pfizer, Roche, Sanofi-Genzyme, Servier, Taiho, and Thermo Fisher Scientific. EG has received research grants from Pfizer, AstraZeneca, Astellas, and Lexicon Pharmaceuticals. RPL receives grant support from AstraZeneca and Roche Pharma and has received honoraria for speaker engagements or advisory roles from Astellas and Roche Pharma.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The Institutional Review Board approved this retrospective study. Informed consent for computational image analysis was waived. All patients included in the clinical trial Lenvatinib Efficacy in Metastatic Neuroendocrine Tumours (TALENT) NCT02678780 provided written informed consent.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44276-023-00010-0>.

Correspondence and requests for materials should be addressed to Raquel Perez-Lopez.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

4.5 Study 5

Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression

This study corresponds to **objective 4**. It was developed in the framework of exploring different medical imaging fields beyond radiology with the aim of learning new techniques that could be translated to the radiology field and developing multimodal models integrating radiology and pathology. In this publication, we trained a weakly supervised deep learning multiple to classified PD-L1 stained IHC and assess its performance to predict response to immunotherapy.

Reference:

Ligero M., Serna G. El Nahhas Omar S.M., et al. Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression. Cancer Research Communications (2023) - *InPress*.

Title Page

Title: Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression

Running title: DL Predicts Immunotherapy Response through PD-L1 Expression

Marta Ligeró (1), Garazi Serna (2), Omar S.M. El Nahhas (3), Irene Sansano(4), Siarhei Mauchanski (2), Cristina Viaplana (5), Julien Calderaro (6), Rodrigo Toledo (7), Rodrigo Dienstmann (5), Rami S. Vanguri (8), Jennifer L. Sauter (9), Francisco Sanchez-Vega (10), Sohrab P. Shah (10), Santiago Ramón y Cajal (4), Elena Garralda (11), Paolo Nuciforo (2), Raquel Perez-Lopez (1)*, Jakob Nikolas Kather (3,12,13,14)#*

* equal contribution

- (1) Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain
- (2) Molecular Oncology Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain
- (3) Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany
- (4) Pathology Department, Vall d'Hebron University Hospital (VHUH), Barcelona, Spain
- (5) Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain
- (6) Assistance Publique-Hôpitaux de Paris, Département de Pathologie, CHU Henri Mondor, F-94000 Créteil, France; Université Paris-Est Créteil, Faculté de Médecine, Créteil, France
- (7) Biomarkers and Clonal Dynamics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain
- (8) Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA
- (9) Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
- (10) Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
- (11) Department of Medical Oncology, Vall d'Hebron University Hospital and Institute of Oncology (VHIO), Barcelona, Spain
- (12) Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom
- (13) Department of Medicine I, University Hospital Dresden, Dresden, Germany
- (14) Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Corresponding Author (#)

Prof. Jakob Nikolas Kather, MD, MSc
Professor of Clinical Artificial Intelligence
Else Kröner Fresenius Center for Digital Health
Technical University Dresden
Tel.: +49 351 458 15812
Mail: jakob_nikolas.kather@tu-dresden.de

Conflicts of interest

JNK declares consulting services for Owkin, France; DoMore Diagnostics, Norway and Panakeia, UK and has received honoraria for lectures by Eisai, Roche, MSD and Fresenius. RPL reports grants from Roche and AstraZeneca, and a consulting or advisory role at Roche. IS declares to have received honoraria in the last three years for participating as a speaker in meetings sponsored by Astrazeneca, Boehringer-Ingelheim, Roche, Pfizer and Takeda. She has also received financial aid from Takeda for congress attendance. RD declares advisory role for Roche, Boehringer Ingelheim, received a speaker's fee from Roche, Boehringer Ingelheim, Ipsen, Amgen, Servier, Sanofi, Libbs, Merck Sharp & Dohme, Lilly, AstraZeneca, Janssen, Takeda and research grants from Merck, Novartis, Daiichi-Sankyo, Pierre Fabre and AstraZeneca. SPS declares shareholder Canexia Imagia Health Inc and consulting services for AstraZeneca Inc. EG declares research grants from Novartis, Roche, Thermo Fisher, AstraZeneca, Taiho, BeiGene, Janssen. EG also reports consultant or advisory role for Roche, Ellipses Pharma, Boehringer Ingelheim, Janssen Global Services, Seattle Genetics, Thermo Fisher, MabDiscovery, Anaveon, F-Star Therapeutics, Hengrui, Sanofi, Incyte, Medscape and speakers bureau from Merck Sharp & Dohme, Roche, Thermo Fisher, Lilly, Novartis, SeaGen. PN declares advisory or consultant role for MSD ONCOLOGY, BAYER and speaker's fee from Novartis. No other competing interests are disclosed by any author.

Authors' contributions

ML, RPL and JNK design the study. ML and OSMEN developed the software. ML, GS, CV, IS, SM, PN, RT, SRC, EG contributed to data acquisition, management, interpretation and retrieval from VHIO. RSV, JLS, FSV and SPS contributed to data acquisition, management, interpretation and retrieval from MSKCC. ML, GS, OSMEN, PN, RPL, JNK interpreted and analyzed the data. All of the authors reviewed and approved the manuscript for its publication and certify the integrity of the presented work.

List of Abbreviations

AUC - Area under the receiver operating characteristic

CPS - Combined positive score

DL - Deep learning

FFPE - formalin-fixed paraffin-embedded

IHC - Immunohistochemistry

ICIs - Immune checkpoint Inhibitors

MIL - Multiple instance learning

NSCLC - Non-small cell lung cancer

PD-L1 - Programmed death-ligand 1

PFS - Progression-free survival

ROI - Regions of interest

TPS - Tumor proportion score

WSI - Whole slide image

Abstract

Programmed death-ligand 1 (PD-L1) immunohistochemistry (IHC) is the most commonly used biomarker for immunotherapy response. However, quantification of PD-L1 status in pathology slides is challenging. Neither manual quantification nor a computer-based mimicking of manual readouts are perfectly reproducible, and the predictive performance of both approaches regarding immunotherapy response is limited. In this study, we developed a deep learning (DL) method to predict PD-L1 status directly from raw IHC image data, without explicit intermediary steps such as cell detection or pigment quantification. We trained the weakly supervised model on PD-L1-stained slides from the NSCLC-MSK cohort (N=233) and validated it on the pancancer-VHIO cohort (N=108). We also investigated the performance of the model to predict response to immune checkpoint inhibitors (ICIs) in terms of progression-free survival (PFS). In the pancancer-VHIO cohort, the performance was compared with Tumor proportion score (TPS) and Combined Positive Score (CPS). The DL-model showed good performance in predicting PD-L1 expression (TPS $\geq$ 1%) in both NSCLC-MSK and pancancer-VHIO cohort (AUC 0.88 $\pm$ 0.06 and 0.80 $\pm$ 0.03, respectively). The predicted PD-L1 status showed an improved association with response to ICIs (hazard ratio (HR) 1.5[95%CI 1-2.3], p-value=0.049) compared to TPS (HR 1.4[0.96-2.2], p-value=0.082) and CPS (HR 1.2[0.79-1.9], p-value=0.386). Notably, our explainability analysis showed that the model does not just look at the amount of brown pigment in the IHC slides, but also considers morphological factors such as lymphocyte conglomerates. Overall, end-to-end weakly supervised DL shows potential for improving patient stratification for cancer immunotherapy by analyzing PD-L1 immunohistochemistry, holistically integrating morphology and PD-L1-staining intensity.

Significance

The weakly supervised deep learning model to predict PD-L1 status from raw immunohistochemistry data, integrating tumor staining intensity and morphology, enables enhanced patient stratification in cancer immunotherapy compared to traditional pathologist assessment.

Introduction

Over the last decade, tumor immunotherapy has irrevocably altered the landscape of oncology. Checkpoint inhibitors, in particular, are now employed in practically all tumor types (1–3). These treatments are increasingly being employed in early stages, palliative lines of therapy in advanced cancer, as well as adjuvant or neoadjuvant therapy. However, the identification of suitable patients remains a critical issue, as only a subset of individuals reacts to checkpoint inhibition. The immunotherapy biomarker which is most commonly used for patient selection in clinical practice is the programmed death-ligand 1 (PD-L1) status, which is assessed using immunohistochemistry (IHC) assays (4–6). Pathologists typically assess these stains by visual inspection of the sample. The number of positive tumor cells and positive immune cells are visually estimated. Nevertheless, manual estimation of a biomarker is obviously not desirable for such a decisive therapy recommendation, and multiple studies have demonstrated that pathologists exhibit a considerable level of variability in this assessment (7–10). Hence, a more precise and reliable approach is imperative to ensure accurate diagnosis and treatment.

As a result, dozens of studies have been conducted to build image analysis tools that assess the PD-L1 status of a tissue sample (11). The majority of these systems mimic the human workflow explicitly: First, they are recognizing tissue on a slide. Second, all individual cells are recognized. These cells are classified into relevant categories, such as tumor cells or immune cells. Finally, the color intensity of each individual cell is evaluated, and the cell is classified as positive or negative (12–16). Although this procedure is theoretically relatively well-defined, there are issues with the approach of such a multistep pipeline in practice. If the tissue slice differs from the slices in the training data set, for example due to artifacts or a difference in sample handling, and even if only one of the intermediary steps of the pipeline is confused here, the final result is untrustworthy. Consequently, another study used deep learning (DL) to combine PanCytokeratin and PD-L1 stained images to differentiate tumor regions from epithelial tissue and assess the proportion of PD-L1 positive inside the tumor areas as a multi-label task (17). Other studies investigated as an alternative the implementation of deep-learning (DL) methods to predict PD-L1 status from other image types such as H&E histology images (18,19) or from radiology images as a surrogate for immunotherapy response (20).

An alternative to such multistep image analysis pipelines is a weakly supervised end-to-end approach. In such an approach, a DL network is trained to predict a biomarker directly from the digitized whole slide images without intermediate steps such as the detection of individual cells being explicitly modeled (21–23). Through the implementation of attention-based methods, the DL model is trained to identify distinct patterns across the different regions of the sample that are contributing to the PD-L1 status. As a result, the model can learn to focus on the areas containing tumor cells and avoiding nonspecific tissue from the biopsy. Weakly supervised end-to-end methods have made enormous progress in the last five years and have been used successfully in numerous tumor entities (21,24,25).

Today, the first commercial products using such methods are already on the market, including a DL system for predicting clinical outcome and genetic alterations such as microsatellite instability (MSI) status from colorectal cancer H&E slides (26,27).

In contrast to previous studies, we developed a weakly supervised DL-based quantification of PD-L1 status in IHC images from non-small cell lung cancer (NSCLC) patients and evaluated it in a cohort of patients with solid tumors, all of them treated with immunotherapy. Thereafter, we investigated the performance of the DL-based PD-L1 scoring to stratify patient response to immunotherapy.

Methods

Ethics statement

This study was performed in accordance with the Declaration of Helsinki. The overall analysis was approved by the Ethics commission of the Medical Faculty of Technical University of Dresden (BO-EK-444102022). The Vall d'Hebron Institute of Oncology (VHIO) institutional review board approved this study for retrospective data [PR(AG)70/2018] and the prospective PREDICT study [PR(AG)29/2020]. All patients included in the clinical trials provided written informed consent. Need for informed consent for the computational analysis of the images in the retrospective study was waived.

Cohort description

We included PD-L1 stained IHC digitalized slides of NSCLC patients from the publicly available NSCLC-MSK MIND cohort (28) with PD-L1 Tumor Proportion Score (TPS) assessed by a pathologist. All patients were treated with immune check-point inhibitors (ICIs) in monotherapy or in combination with other immunotherapies. Progression-free survival (PFS) was assessed as the time between therapy starting date and the progression date.

As an external validation set, we collected data from patients with different tumor types (including NSCLC, gastrointestinal cancer and melanoma among others) treated with ICIs at Vall d'Hebron Institute of Oncology between 2012 and 2021. Only patients with tissue available from the biopsy prior to treatment and with PD-L1 TPS assessed by a pathologist were included in the analysis. In the external cohort, the Combined Positive Score (CPS) was also collected when available. Biopsies at multiple time points during ICIs treatment from the same patient were collected and included in the PD-L1 status prediction performance. For those patients with multiple biopsies, only the samples closest to the start of treatment with ICIs were considered for treatment response prediction. An overview of the included population is described in Figure 1A.

PD-L1 immunohistochemistry

In the NSCLC-MSK cohort, PD-L1 IHC staining of the slides was performed as described by Vanguri et al. (28). Briefly, the IHC was performed using the Bond III, Leica automated immunostaining platform

using the PD-L1 antibody (E1L3N; Cell Signaling Technologies). In the pan-cancer-VHIO cohort, 3.5 μm -thick slides stained with PD-L1 IHC were obtained from formalin-fixed paraffin-embedded (FFPE) tumor blocks of the VHIO cohort. The staining was performed with PD-L1 (SP263) Rabbit Monoclonal Primary Antibody (#790-4905, VENTANA) on the Benchmark ULTRA automated immunostaining platform. The OptiView DAB IHC Detection Kit (#760-700; VENTANA) was used for detecting the primary antibody. Placental and tonsil tissues were used as positive controls for all the stainings. The slides were digitalized using NanoZoomer 2.0-HT slide scanner by Hamamatsu Photonics, with a magnification of 20x obtaining which corresponds to a pixel size of $0.46\mu\text{m} \times 0.46\mu\text{m}$. A trained pathologist (S.M.) interpreted the slides and provided the TPS and CPS for each sample. A negative score was defined as staining in less than 1% of tumor cells or the absence of staining in tumor cells. Slides with less than 100 tumor cells were excluded from PD-L1 CPS/TPS assessment. Investigators were blinded to patient response during the PD-L1 assessment.

Image preprocessing

In the NSCLC-MSK cohort, regions of interest (ROI) of the patient sample tissue were provided from the HALO software (Indica Labs). In the VHIO cohort, samples were annotated in polygonal regions of interest using the open source QuPath software v0.3.2. These annotations were used to crop the image excluding any other tissue from the WSI that was not from the patient, i.e. reference samples. For both cohorts, PD-L1 stained whole-slide images (WSI) were tessellated into non-overlapping tiles of 224×224 pixels corresponding to an edge length of $256\mu\text{m}$ from the ROIs. Blurry tiles or tiles without tissue samples, defined as tiles containing less than 2% of edges in the image, were discarded using Canny edge detection (29–33).

PD-L1 status classification model

We developed weakly supervised DL methods to classify high ($\text{TPS} \geq 1\%$) and low ($\text{TPS} < 1\%$) PD-L1 status from PD-L1 stained IHC samples. The DL-approach consisted of two-step: first, feature vectors were extracted from every tile of the WSI, obtaining a tile-level feature vector. Thereafter, we performed weakly supervised feature aggregation and classification.

Feature extraction was performed using the Retrieval with Clustering-guided Contrastive Learning (RetCCL) model (34), an ImageNet-weighted ResNet50 pretrained in a self-supervised manner on a pathology dataset with 32,000 H&E-stained histology slides across various cancer types. After feature extraction, vectors of 2048 features were obtained from every 224×224 pixels tile image. The attention-based multiple instance learning (MIL) model (35) consisted of an attention-based mechanism, followed by a classification head. First, a trainable weight is assigned to each tile-level feature vector, referred to as attention. Then, the features and weights were aggregated by the attention scores of the tile-level features to obtain the attention-weighted embeddings per sample. Finally, the embeddings were given to the classification head to obtain the model's sample-level prediction scores.

The model was trained using 5-fold cross-validation, with each fold stratified to maintain a proportional representation of PD-L1 status in the training, validation and test set. The Area Under the Receiver Operating Characteristic (AUC) metric was used to evaluate the models. In the external validation cohort, we ran inference from the five models from cross-validation and the prediction scores were averaged to predict the PD-L1 status. For explainability analysis, the model with better performance on the training set was selected. We investigated the performance of the model when it is trained in the pan-cancer-VHIO cohort and validated in the NSCLC-MSK cohort following the same methods.

Response prediction analysis and explainability

Association between PFS and PD-L1 status were investigated using Cox-proportional Hazard Ratio and Kaplan-Meier curves with DeLong method in both NSCLC-MSK and the pan-cancer-VHIO cohort. For explainability and visualization purposes, we computed high resolution attention heatmaps on the WSI as reported previously (29). Attention heatmaps highlight the areas of the WSI where the model paid more attention for the decision-making. To compute high resolution heatmaps, the attention MIL model architecture was loaded into a fully convolutional equivalent with the trained weights. The equivalent convolutional neural network (CNN) with attention MIL weights is running inference with a kernel size of 32x32 across the WSI. This approach allows for higher resolution attention heatmaps than using the original tiles of 224x224 pixels used to train the model for visualization purposes. We compared the attention heatmaps with PanCytokeratin-stained slides that highlight tumor areas. Additionally, to assess whether the diversity of tumor types in the external cohort was affecting the performance of the model in predicting PD-L1, we studied the ratio of true and false positive and true and false negative of the model across tumor types. To have a better understanding of the model performance, an expert pathologist manually inspected the attention heatmaps to find which morphological factors the model is identifying to make predictions, including cases from false positive and false negative where the model finds confounding factors in the images. Finally, the predicted scores were compared as continuous variables with both TPS and CPS. Figure 1B shows the workflow of the methods.

Data Availability Statement

All data from the MSKCC cohort are publicly available at <https://www.synapse.org/#!/Synapse:syn26642505>. All other images are available from the respective study principal investigators upon reasonable request. All source code for image preprocessing is publicly available at <https://github.com/KatherLab>. The tessellation script is available at <https://github.com/KatherLab/preprocessing-ng>. Extraction of RetCCL features was carried out with scripts from <https://github.com/KatherLab/marugoto> as well as the weakly supervised Deep Learning model and the heatmaps for explainability used for the study.

Results

Cohort clinical data description

We included 233 patients from the NSCLC-MSK cohort treated with ICIs with available PD-L1 stained IHC slides and TPS score assessed by a pathologist. One IHC-slide was discarded due to scarce tissue sample for image analysis. The median PFS was 3.2 [95%CI 2.6-4.7] months with 30% (70/233) of patients responding to ICIs (i.e., achieved complete or partial response by RECIST 1.1). To test the PD-L1 status classifier in an external independent cohort, 122 PD-L1 stained slides from 108 patients from the pan-cancer-VHIO cohort were collected. From the all samples, we selected baseline samples from 106 patients treated with ICIs in monotherapy or in combinations with other ICIs that had evaluable response to predict response to immunotherapy. The pan-cancer-VHIO cohort had a median PFS of 2.73 [95%CI 1.84-3.98] months, with 15% (16/106) of patients with partial or complete response to ICIs. The cohort description is reported in Table 1.

Weakly-supervised deep learning approach to classify PD-L1 status predicts response to ICIs

Our attention MIL-based model was trained on PD-L1 stained IHC images from the NSCLC-MSK cohort to predict PD-L1 status TPS>1%, reaching a high performance with a cross-validated mean AUC of 0.88 ± 0.06 (Figure 2A). Subsequently, we explored how the model generalizes in an external cohort (pan-cancer-VHIO), which consisted of different tumor types, including NSCLC. We deployed all the models from cross-validation in the IHC images of the external pan-cancer-VHIO achieving a mean AUC of 0.80 ± 0.03 (Figure 2B). The detailed performance statistics are available in Suppl. Table S1 (accuracy, sensitivity, specificity, positive predictive value [PPV], negative predicted value [NPV]). Overall, these results show that the PD-L1 classifier trained on the NSCLC can generalize to broader and more diverse datasets, even when stained with different protocols. Similar results were obtained when training on the pan-cancer-VHIO cohort and validating on the NSCLC-MSK. (Suppl. Figure S1)

We investigated whether the predicted PD-L1 status associated with response to immunotherapy in terms of PFS endpoint. A strong association between PD-L1 predicted scores and response to ICIs is observed in both NSCLC-MSK cohort (HR: 1.4 [95%CI 1-1.8], p-value = 0.03) and pan-cancer-VHIO cohort (HR: 1.5 [95%CI 1-2.3], p-value = 0.049), as measured by Cox-regression analysis. In the external validation cohort, the predicted PD-L1 status showed more significant association with response to immunotherapy compared to TPS (HR 1.4 [95% 0.96-2.2], p-value=0.082) and CPS (HR 1.2 [95% 0.79-1.9], p-value = 0.386). Additionally, patients classified as high PD-L1 status showed significant longer median PFS compared to patients classified as low PD-L1 status in both the training and test cohorts by Kaplan-Meier curves and long-rank test; median PFS of 4.70 [IQR 2.7-9.0] vs 2.7 [IQR 2.1-3.5] months, p-value<0.05 and 3.06 [IQR 1.84-5.46] vs 2.63 [IQR 1.71-3.94] months, p-value<0.05; respectively (Figure 2C-D). In summary, these results suggest that the predicted PD-L1 status can discriminate which patients are more likely to benefit from ICI treatment.

Explainability analysis: PD-L1 predicted status differentiates tumor cells from surrounding tissue

The assessment of PD-L1 status can be performed considering the PD-L1 expression only from the tumor cells (TPS) or including both tumor cells and surrounding tissue, such as the stroma component (CPS). The classifier was trained on a NSCLC cohort where PD-L1 was assessed only from tumor cells (TPS) as standard of care. Then, the model was validated in a pan-cancer cohort, in which PD-L1 status was assessed with both TPS and CPS. For explainability analysis of the classifier, we investigated the distribution of attention scores in some of the patients that were correctly classified as high and low PD-L1 status in both NSCLC-MSK and pan-cancerVHIO cohort. A pathologist review identified the model's capability to correctly identify tumor cells with and without expression of PD-L1 in the membrane and that it can differentiate tumor cells expressing PD-L1 from other cellular structures such as the stromal tissue, adipose tissue or immune cells in the tumor microenvironment (Figure 3). This visual analysis was further validated through a comparative examination of attention heatmaps and PanCytokeratin-stained images (Suppl. Figure S2). Taken together, these results suggest that even when the PD-L1 is expressed in the immune component, the model will recognize that tissue as non-tumoral and will not consider it as relevant for the final classification, which means that our DL-model is more sensitive to TPS than CPS.

Furthermore, the model showed an overall accuracy of 77% to correctly classify samples with high and low PD-L1 expression in the tumor cells (Figure 4A, Suppl. Table S1). However, as shown in Figure 4, when testing the model in an heterogeneous cohort such as the pan-cancer-VHIO dataset, some potential factors may limit the overall performance of the model. As can be seen in Figure 4B and Suppl. Table S1, the model's classification performance is weaker in tumor histologies such as gastrointestinal tumor, melanoma and sarcoma. These results are likely to be related to the histological and molecular differences between these tumor types and the NSCLC, which is the tumor type used to train the model.

Samples with inconsistent PD-L1 expression between the tumor cells and the stromal tissue (e.g. high CPS and low TPS) also pose a challenge to the model. In some of these cases, a pathologist review identified that the DL-model recognizes the PD-L1 expression from the stromal component as high PD-L1 status (Figure 4C-D). These findings may also be linked to histological differences in population. Higher false positive rates correspond to differentiated tumor types such as melanoma, head and neck and other tumor types. Figure 4E shows magnification of a PD-L1 stained IHC sample from a melanoma biopsy. In this case, the model fails to recognize the tumor cells from skin (blue magnification) and sees the immune dense component (red magnification) as tumor area, classifying the patient as high PD-L1 when the ground truth based on TPS is low PD-L1. Other misclassification cases correspond to samples with TPS falling near the classification threshold ($TPS < 1\%$) as shown in Figure 4C (77% of the false negative cases had a low value of TPS, ranging between 1 to 7%) or samples with low cellularity as shown in Figure 4F, where the model fails to classify as low due to the large amount of sample without PD-L1 expression. Together, these results indicate that the trained

model on the NSCLC-MSK cohort generalizes to a more diverse cohort and pays attention to relevant tissue components, i.e., focused on tumor cells.

Discussion

Immunohistochemical analysis of PD-L1 expression in tumor samples is the most commonly used biomarker in clinical practice and currently employed as patient selection for some of the standard ICI therapies (11,36). However, the manual evaluation of this biomarker results in inter-observer variability, which may lead to patient misclassification in terms of treatment options (7). Therefore, there is a need for quantitative and standardized tools to evaluate PD-L1 status in clinical practice.

In this study, we implemented and validated a weakly supervised DL method that can predict PD-L1 status in patients with several tumor types. The current study found that a DL-model trained with weak labels for predicting TPS in a NSCLC cohort is capable of differentiating between tumor cells and immune cells in most of the cases and correctly quantifying the expression of PD-L1 in the specific tumor cells. This is a particularly encouraging result that supports the idea of using weak labels to classify WSI without the need of delineating structures, as done in previous studies (14,15). Interestingly, the model proves generalization to other cohorts across different tumor types, not limited to lung, and across two different PD-L1 antibodies. Furthermore, we proved that the predicted PD-L1 status from our model can help to identify patients that would benefit from immunotherapy.

Despite the promising findings of our study, several limitations should be acknowledged. First, the model was trained and tested in a relatively small cohort, which may limit the accuracy and generalizability of the model. Additionally, the evaluation of the model in different histological tumor types might influence the performance. The model showed lower performance in the tumor types such as melanoma, gastrointestinal, sarcomas, head and neck, which may be related to the different biology and histology of these tumor types, compared to NSCLC (37,38). Moreover, differences in staining protocols between laboratories and tumor types might contribute to the higher rate of misclassifications when evaluating the model (10,38). For example, the presence of non-specific staining such as melanin in the case of melanoma, tumor necrosis or dense lymphoid aggregates could be potential sources of biases for the model. Finally, the features used in the DL model were trained using H&E images and applied to IHC images. Even though the model shows a good performance using RetCCL features, this could be limiting the model to achieve enhanced classification. Nonetheless, despite all the limitations encountered in the heterogeneous external cohort, our study demonstrates that a weakly supervised model, trained on NSCLC, can accurately classify PD-L1 status in most cases.

Furthermore, the differences in standard methods for assessing PD-L1 based on tumor type and treatment regime could influence the performance of DL-model as an immunotherapy response biomarker when applied to a diverse cohort of tumor types. Although the model was originally

developed in a NSCLC cohort, where PD-L1 status is a gold standard, a widely accepted biomarker used in the clinical routine for treatment-decision, it was tested on a pan-cancer cohort with advanced patients from clinical trials. In this setting, the expression of PD-L1 in the tumor cells (TPS) as a response biomarker remains unclear, and it is often evaluated using CPS, which included both stromal and tumoral PD-L1 expression. The use of CPS in more advanced patients allows for broader inclusion of patients into clinical trials (39,40). Interestingly, our model showed acceptable performance for predicting response to immunotherapy in both NSCLC and pan-cancer cohort.

Nevertheless, in order to address the previous constraints related to cohort characteristics, we investigated the performance of the model when trained on the pan-cancer-VHIO cohort and validated on the MSK-NSCLC cohort. However, the limitations of the pan-cancer-VHIO cohort, characterized by a small sample size, substantial tumor type heterogeneity, and the inclusion of predominantly advanced tumors accessible only via biopsy, may potentially reduce the reliability of the model performance.

One additional limitation we acknowledge is the potential impact of inter-observer variability in model development and its bias towards subjective assessment. It is worth noting, however, that the PD-L1 measurements were sourced from clinical reports, which were assessed by multiple pathologists in routine clinical practice. This property enables our model to get insights from a wide range of pathologist evaluations, thereby mitigating bias from a single evaluation and accounting for inter-observer variability. In terms of potential future work, these models could benefit from the integration of uncertainty modeling methods that consider the inherent variability of the ground truth during modeling.

To conclude, our findings differ from the previous studies mainly in terms of outcome prediction and methods implementation. Several studies have reported promising results and similar variability between the AI tool and the pathologist performance for assessing PD-L1 status (15). However, most of these methods require either manual or automatic annotation of the tumor tissue, as well as the detection and classification of the different cells in the sample. Despite other studies trying to address this limitations by proposing tissue classification as an annotation solution instead of using cell detection, our proposed method predicts PD-L1 status directly from the WSI, without any physician input or need for further stainings such as pan-cytokeratins. This one-step approach reduces the risk of errors, while yielding similar results to previously developed methods and providing explainability of the model by terms of attention. Additionally, as a distinguishing characteristic of this study, it investigates the potential value of the DL-tool to predict response to immunotherapy, which is the main objective of PD-L1 quantification.

This study supports the idea that weakly supervised DL methods are able to distinguish between the different histological patterns in immunohistochemistry and learn how to quantify expression of PD-L1 in tumor samples without any additional input from the physician. However, further research is needed

in larger and more diverse cohorts for the model to gain a better understanding of the different tumor histologies, so it can properly generalize to any tumor type.

Acknowledgments

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312) and the National Institute for Health and Care Research (NIHR, NIHR213331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. RPL is supported by LaCaixa Foundation, a CRIS Foundation Talent Award (TALENT19-05), the FERRO Foundation, the Instituto de Salud Carlos III-Investigacion en Salud (PI18/01395 and PI21/01019) and the Prostate Cancer Foundation (18YOUN19). ML is supported by the PERIS PIF-Salut Grant. As per the ICMJE guidelines of April 2023, we hereby disclose that the following artificial intelligence tools were used to write this article: ChatGPT-4 for checking and correcting spelling and grammar.

References

1. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*. 2016;375:1823–33.
2. Ribas A, Puzanov I, Dummer R, Schadendorf D, Hamid O, Robert C, et al. Pembrolizumab versus investigator-choice chemotherapy for ipilimumab-refractory melanoma (KEYNOTE-002): a randomised, controlled, phase 2 trial. *Lancet Oncol*. 2015;16:908–18.
3. Bellmunt J, de Wit R, Vaughn DJ, Fradet Y, Lee J-L, Fong L, et al. Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. *N Engl J Med*. 2017;376:1015–26.
4. Roach C, Zhang N, Corigliano E. Development of a companion diagnostic PD-L1 immunohistochemistry assay for pembrolizumab therapy in non-small-cell lung cancer. *Applied* [Internet]. ncbi.nlm.nih.gov; 2016; Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4957959/>
5. Cyprian FS, Akhtar S, Gatalica Z, Vranic S. Targeted immunotherapy with a checkpoint inhibitor in combination with chemotherapy: A new clinical paradigm in the treatment of triple-negative breast cancer. *Bosn J Basic Med Sci*. 2019;19:227–33.
6. Daud AI, Wolchok JD, Robert C, Hwu W-J, Weber JS, Ribas A, et al. Programmed Death-Ligand 1 Expression and Response to the Anti-Programmed Death 1 Antibody Pembrolizumab in Melanoma. *J Clin Oncol*. 2016;34:4102–9.
7. Brunnström H, Johansson A, Westbom-Fremer S, Backman M, Djureinovic D, Patthey A, et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. *Mod Pathol*. 2017;30:1411–21.
8. Chang S, Park HK, Choi Y-L, Jang SJ. Interobserver Reproducibility of PD-L1 Biomarker in Non-small Cell Lung Cancer: A Multi-Institutional Study by 27 Pathologists. *J Pathol Transl Med*. 2019;53:347–53.
9. Cooper WA, Russell PA, Cherian M, Duhig EE, Godbolt D, Jessup PJ, et al. Intra-and Interobserver Reproducibility Assessment of PD-L1 Biomarker in Non--Small Cell Lung Cancer. *Reproducibility of PD-L1 Biomarker Assessment in NSCLC*. *Clin Cancer Res*. AACR; 2017;23:4569–77.
10. Rimm DL, Han G, Taube JM, Yi ES, Bridge JA, Flieder DB, et al. A Prospective, Multi-institutional, Pathologist-Based Assessment of 4 Immunohistochemistry Assays for PD-L1 Expression in Non-Small Cell Lung Cancer. *JAMA Oncol*. 2017;3:1051–8.
11. Inge LJ, Dennis E. Development and applications of computer image analysis algorithms for scoring of PD-L1 immunohistochemistry. *Immunooncol Technol*. 2020;6:2–8.
12. Hondelink LM, Hüyük M, Postmus PE, Smit VTHBM, Blom S, von der Thüsen JH, et al. Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer. *Histopathology*. Wiley; 2022;80:635–47.
13. Puladi B, Ooms M, Kintsler S, Houschyar KS, Steib F, Modabber A, et al. Automated PD-L1 Scoring Using Artificial Intelligence in Head and Neck Squamous Cell Carcinoma. *Cancers* [Internet]. 2021;13. Available from: <http://dx.doi.org/10.3390/cancers13174409>
14. Wu J, Liu C, Liu X, Sun W, Li L, Gao N, et al. Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer. *Mod Pathol*. 2022;35:403–11.
15. Wang X, Wang L, Bu H, Zhang N, Yue M, Jia Z, et al. How can artificial intelligence models

- assist PD-L1 expression scoring in breast cancer: results of multi-institutional ring studies. *NPJ Breast Cancer*. 2021;7:61.
16. Naso JR, Povshedna T, Wang G, Banyi N, MacAulay C, Ionescu DN, et al. Automated PD-L1 Scoring for Non-Small Cell Lung Carcinoma Using Open-Source Software. *Pathol Oncol Res*. 2021;27:609717.
 17. Kapil A, Meier A, Steele K, Rebelatto M, Nekolla K, Haragan A, et al. Domain Adaptation-Based Deep Learning for Automated Tumor Cell (TC) Scoring and Survival Analysis on PD-L1 Stained Tissue Images. *IEEE Trans Med Imaging*. 2021;40:2513–23.
 18. Shamai G, Livne A, Polónia A, Sabo E, Cretu A, Bar-Sela G, et al. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. *Nat Commun*. 2022;13:6753.
 19. Sha L, Osinski BL, Ho IY, Tan TL, Willis C, Weiss H, et al. Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images. *J Pathol Inform*. 2019;10:24.
 20. Wang C, Ma J, Shao J, Zhang S, Li J, Yan J, et al. Non-Invasive Measurement Using Deep Learning Algorithm Based on Multi-Source Features Fusion to Predict PD-L1 Expression and Survival in NSCLC. *Front Immunol*. 2022;13:828560.
 21. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25:1054–6.
 22. Ghaffari Laleh N, Muti HS, Loeffler CML, Echle A, Saldanha OL, Mahmood F, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med Image Anal*. 2022;79:102474.
 23. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. *J Pathol* [Internet]. 2022; Available from: <http://dx.doi.org/10.1002/path.5898>
 24. Schrammen PL, Ghaffari Laleh N, Echle A, Truhn D, Schulz V, Brinker TJ, et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J Pathol*. Wiley; 2022;256:50–60.
 25. Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat Cancer*. 2022;3:1026–38.
 26. Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, et al. Blind validation of MSIntuit, an AI-based pre-screening tool for MSI detection from histology slides of colorectal cancer [Internet]. *medRxiv*. 2022 [cited 2023 Jan 29]. page 2022.11.17.22282460. Available from: <https://www.medrxiv.org/content/10.1101/2022.11.17.22282460v2>
 27. Kleppe A, Skrede O-J, De Raedt S, Hveem TS, Askautrud HA, Jacobsen JE, et al. A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and pathological staging markers: a development and validation study. *Lancet Oncol*. 2022;23:1221–32.
 28. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat Cancer* [Internet]. 2022; Available from: <http://dx.doi.org/10.1038/s43018-022-00416-8>
 29. El Nahhas OSM, Loeffler CML, Carrero ZI, van Treeck M, Kolbinger FR, Hewitt KJ, et al. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides [Internet]. *arXiv [cs.CV]*. 2023. Available from: <http://arxiv.org/abs/2304.05153>

30. Lavinia Loeffler CM, El Nahhas OSM, Muti HS, Seibel T, Cifci D, van Treeck M, et al. Direct prediction of Homologous Recombination Deficiency from routine histology in ten different tumor types with attention-based Multiple Instance Learning: a development and validation study. medRxiv [Internet]. 2023; Available from: <http://dx.doi.org/10.1101/2023.03.08.23286975>
31. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986;8:679–98.
32. Saldanha OL, Loeffler CML, Niehues JM, van Treeck M, Seraphin TP, Hewitt KJ, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. NPJ Precis Oncol. 2023;7:35.
33. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, et al. Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell. 2023;41:1650–61.e4.
34. Wang X, Du Y, Yang S, Zhang J, Wang M, Zhang J, et al. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. Med Image Anal. 2023;83:102645.
35. Ilse M, Tomczak J, Welling M. Attention-based Deep Multiple Instance Learning. In: Dy J, Krause A, editors. Proceedings of the 35th International Conference on Machine Learning. PMLR; 10--15 Jul 2018. page 2127–36.
36. Davis AA, Patel VG. The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. J Immunother Cancer. 2019;7:278.
37. Kluger HM, Zito CR, Turcu G, Baine MK, Zhang H, Adeniran A, et al. PD-L1 Studies Across Tumor Types, Its Differential Expression and Predictive Value in Patients Treated with Immune Checkpoint Inhibitors. Clin Cancer Res. 2017;23:4270–9.
38. Kintsler S, Cassataro MA, Drosch M, Holenya P, Knuechel R, Braunschweig T. Expression of programmed death ligand (PD-L1) in different tumors. Comparison of several current available antibody clones and antibody profiling. Ann Diagn Pathol. 2019;41:24–37.
39. Hagi T, Kurokawa Y, Kawabata R, Omori T, Matsuyama J, Fujitani K, et al. Multicentre biomarker cohort study on the efficacy of nivolumab treatment for gastric cancer. Br J Cancer. 2020;123:965–72.
40. Cohen EEW, Bell RB, Bifulco CB, Burtneess B, Gillison ML, Harrington KJ, et al. The Society for Immunotherapy of Cancer consensus statement on immunotherapy for the treatment of squamous cell carcinoma of the head and neck (HNSCC). J Immunother Cancer. BMJ; 2019;7:184.

Tables

Cohort	NSCLC-MSK	pan-cancer-VHIO
N	233	108 (106 treated with immunotherapy)
Age, median (range)	68 [30-93]	57 [29-77]
Sex		
Male	102	57
Female	131	51
Treatment		
Anti-PDL1	N/A	35
Anti-PD1	N/A	13
Anti-PD1 + other ICIs	N/A	30
Anti-PDL1 + other ICIs	N/A	8
Other ICIs	N/A	14
Anti-PD1 + chemotherapy	N/A	4
Other combinations	N/A	4
Tumor type		
Lung	233	27
Gastrointestinal (including gastric, colon, pancreas)		22
Melanoma		10
Gynecologic (including ovarian, uterine and endometrial)		9
Breast		8
Urinary		6
Head and Neck		8
Hepatobiliary		4
Sarcoma		3
Other		9
Biopsy location		
Primary	105	69
Metastatic	128	39
PD-L1 expression		
TPS < 1%	73	59
TPS ≥1%	163	49
Best response		
CR/PR	70	16
SD/PD	154	90
PFS in months (median [95% CI])	3.2 [2.6-4.7]	2.73 [1.84-3.98]

Table 1. Clinical population characteristics for NSCLC-MSK and pan-cancer-VHIO cohort. ICI (Immune Checkpoint Inhibitors), TPS (Tumor proportion score), CI (Confidence Interval), CR (Complete Response), PR (Partial Response), SD (Stable Disease), PD (Progressive Disease), *N/A - Not Available

Figures

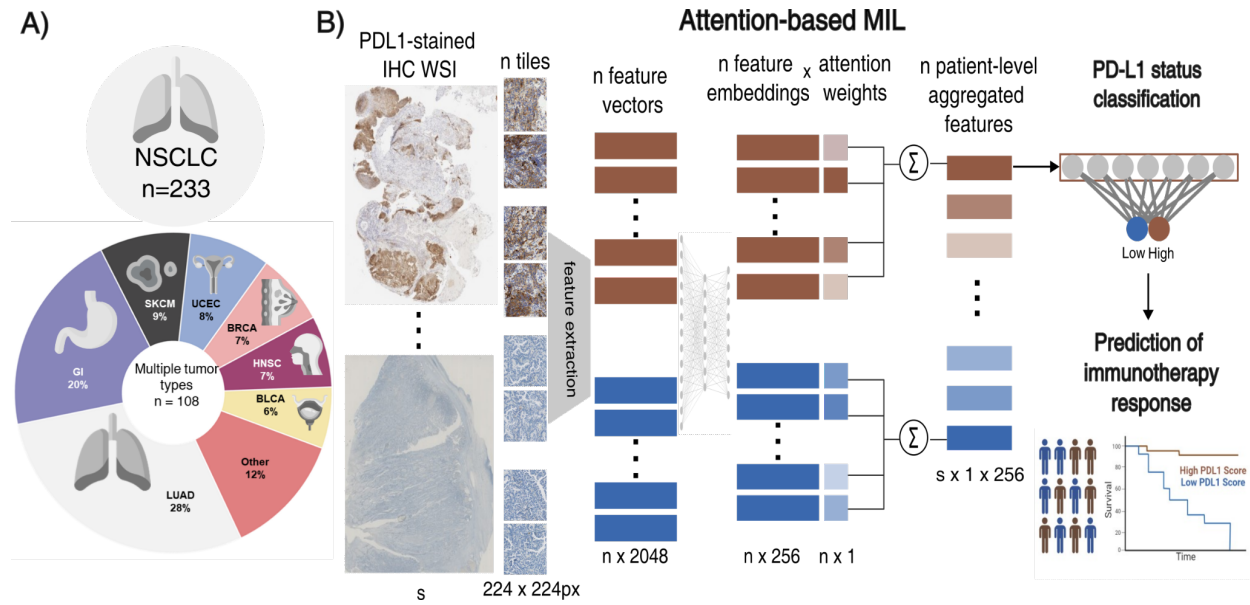


Figure 1. Overview of the study design: (A) Population description of the NSCLC-MSK (training) and Pan-cancer- VHIO (test) cohorts. (B) Workflow of the attention-based Multiple Instance Learning (MIL) pipeline to classify PD-L1 status on Immunohistochemistry (IHC) slides. The predicted PD-L1 status was investigated as predictive biomarker of response to immunotherapy.

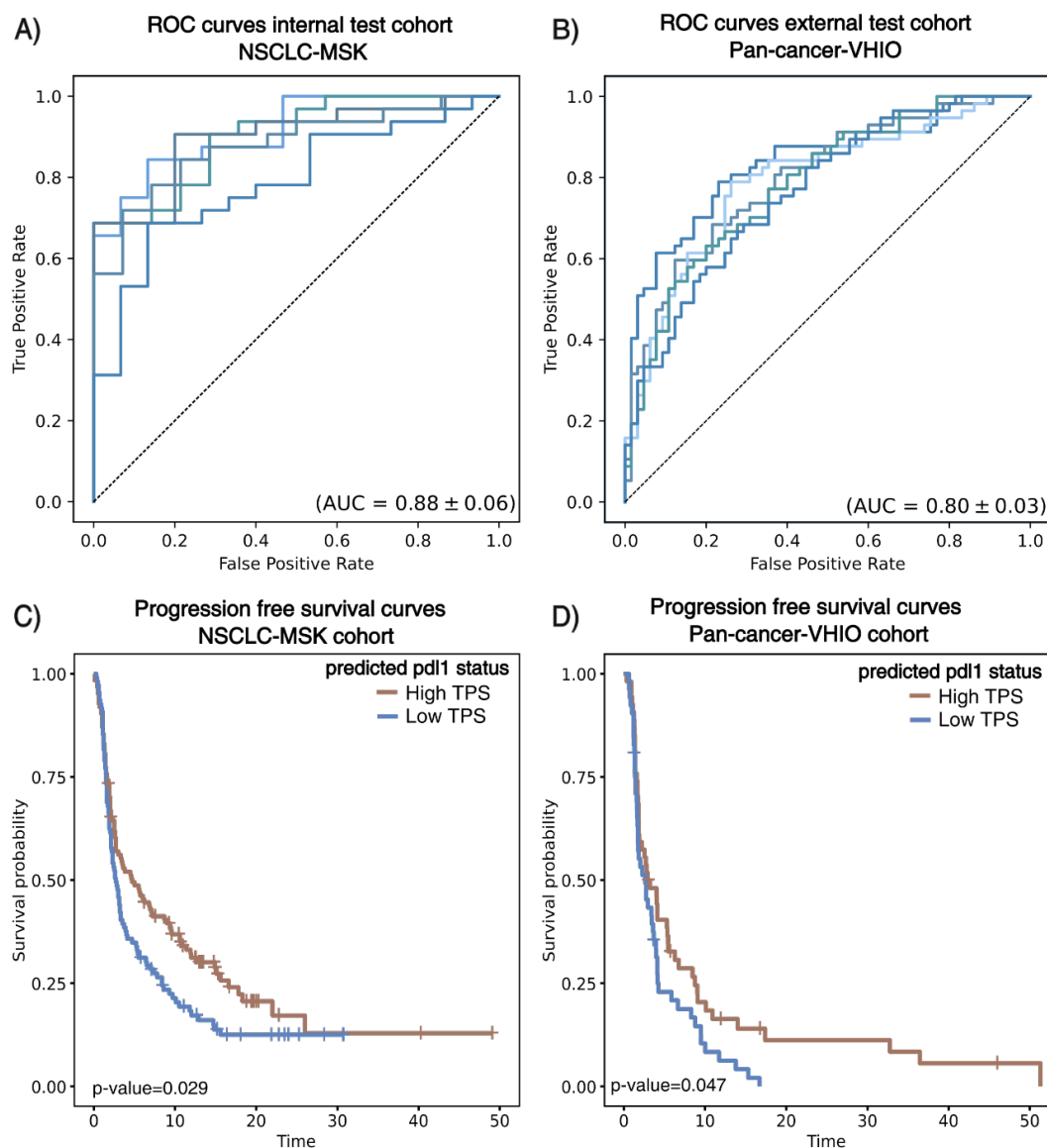


Figure 2. Performance overview of the model for predicting PD-L1 status and response to immunotherapy in NSCLC-MSK and pan-cancer-VHIO cohort. Area under the receiver operating characteristic (ROC) curves for the model to predict PD-L1 status ($\text{TPS} \geq 1\%$) in the training (NSCLC-MSK cohort) (A) and in the test cohort (pan-cancer-VHIO cohort) (B) for the 5-folds cross-validation. All trained models were deployed in the test cohort. Kaplan-Meier curves for the predicted PD-L1 status (high/low) from the model differentiates patients with longer Progression Free Survival to immunotherapy from patients with shorter survival in both NSCLC-MSK (C) and pan-cancer-VHIO cohort (D).

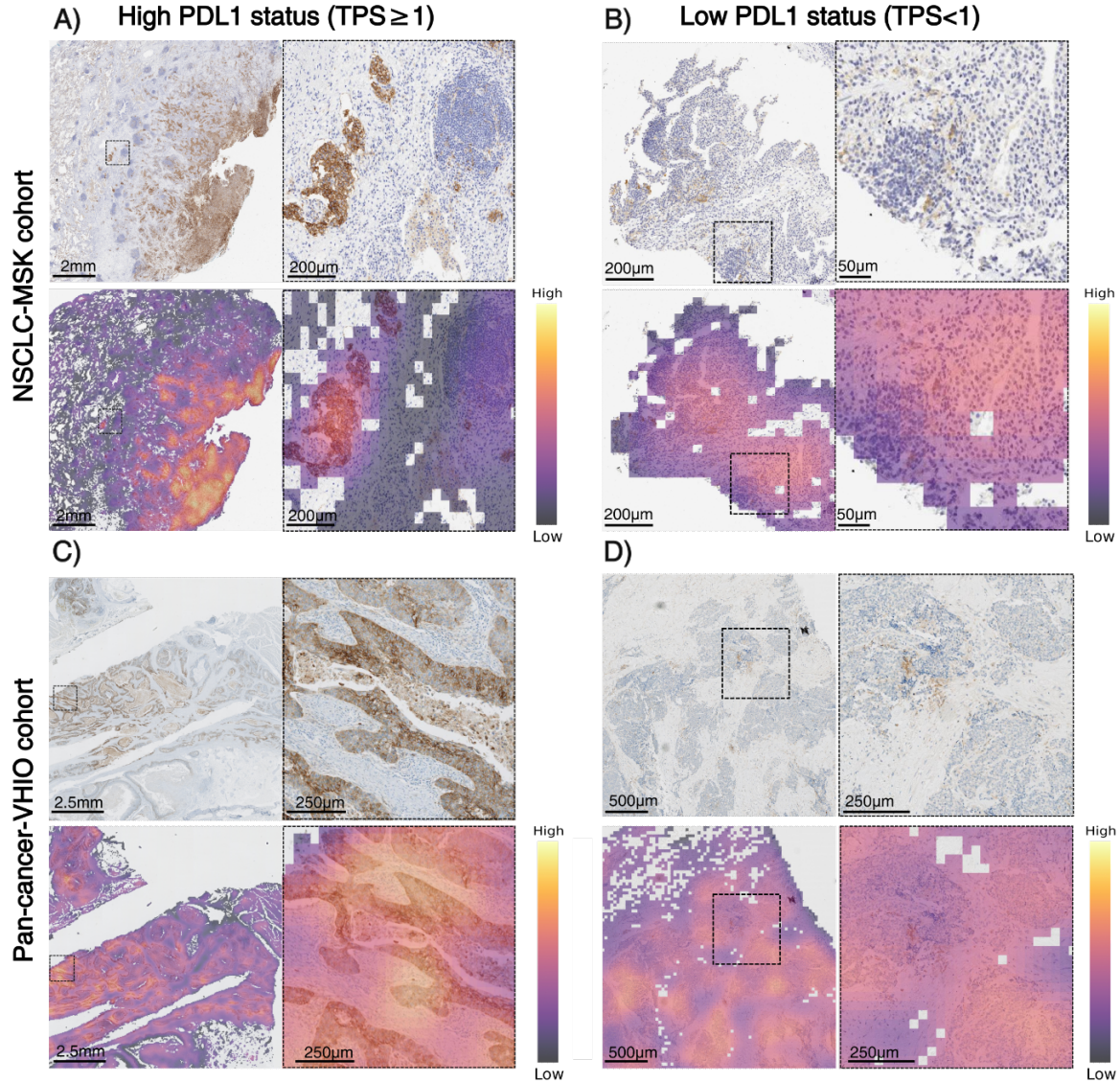


Figure 3. Visualization of high and low PD-L1 score patients from the NSCLC-MSK and pan-cancer-VHIO cohort. Magnification of tumor areas with the highest attention scores for both high PD-L1 (TPS $\geq 1\%$) and low PD-L1 (TPS $< 1\%$) samples from the training and validation cohort. Magnification shows that for both high and low scores the model gives more attention to tumor cells, ignoring areas of high lymphocyte density (A: tumor cells (high attention, left), lymphocytes (low attention, right); B and D: tumor cells (high attention, right), lymphocytes (low attention, left); C: tumor cells (high attention), stroma (low attention)). (Image magnification, A: 1.25x-6.12x, B: , C: 1.25x-2.5x, D: 2.5x-5.0x)

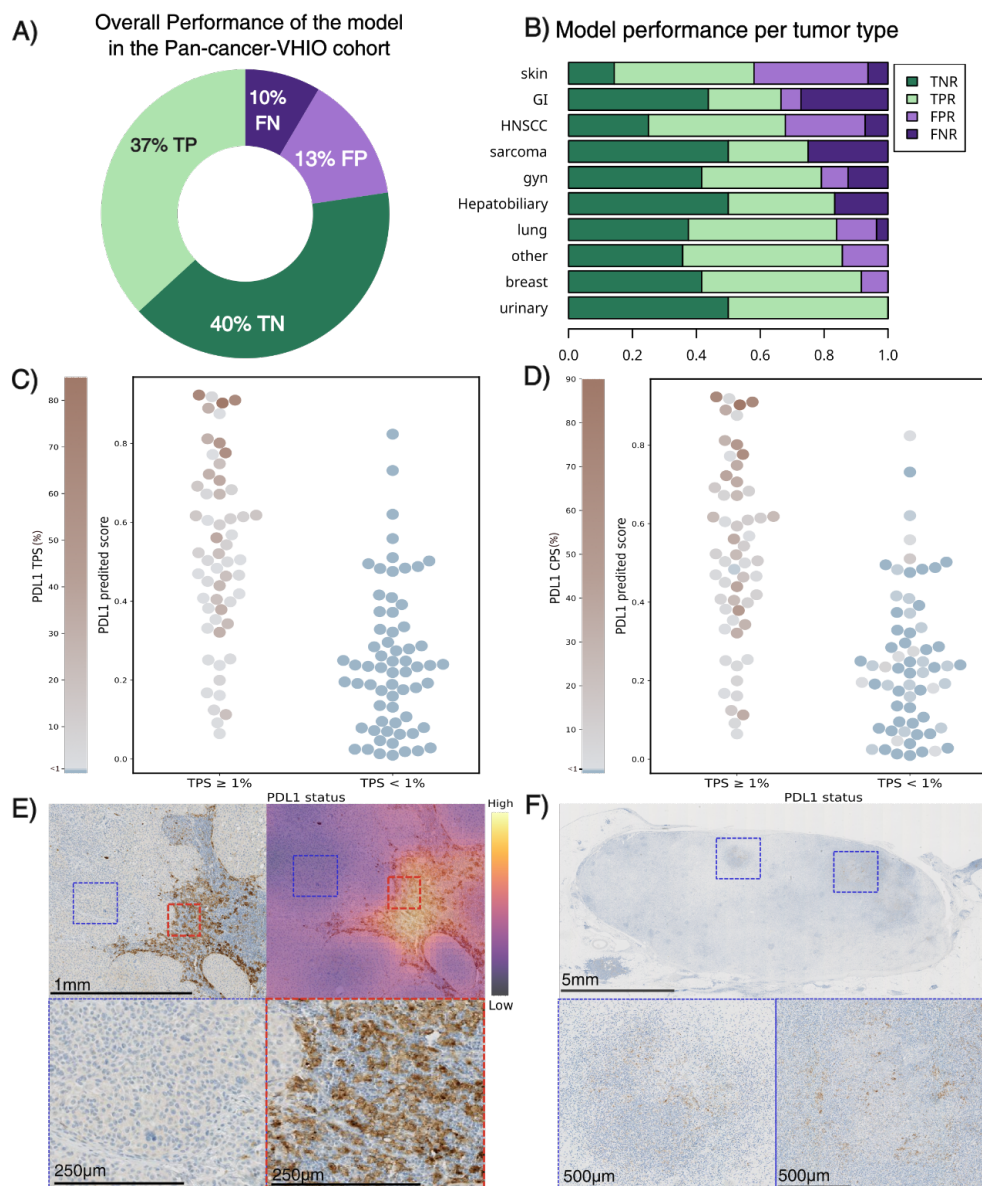


Figure 4. Explainability for model performance in the pan-cancer-VHIO validation cohort. Overall performance of the model in the validation cohort (A). Distribution of True Negative Ratio (TNR), True Positive Ratio (TPR), False Negative Ratio (FNR), and False Positive Ratio (FPR) per tumor type in the pan-cancer-VHIO validation cohort (B). Distribution of predicted scores in the pan-cancer-VHIO cohort compared with continuous Tumor Proportion Score (C) and Combined Positive Score (D). Examples of false positives due to histological differences from NSCLC: PD-L1 stained IHC image from melanoma tumor sample predicted as high PD-L1 by the model (predicted TPS $\geq$ 1%) with low PD-L1 expression in the tumor cells (TPS<1%) and high PD-L1 expression in the immune component (CPS $\geq$ 1%). Magnification of areas with tumor cells with no PD-L1 expression (blue) and immune cells with PD-L1 expression (red), image magnification: 5x and 20x (E). Examples of false negative due to low cellularity samples: PD-L1 stained IHC image from gastrointestinal tumor sample predicted as low PD-L1 by the model (predicted TPS<1%) with high PD-L1 expression in the tumor cells (TPS=20%) and high PD-L1 expression in the immune component (CPS $\geq$ 35%). Magnification of regions with tumor cells, image magnification, 0.35x and 5x (F).

4.6 Supervisors statement

This thesis in compendium of publications format, entitled “Machine Learning for predicting Immunotherapy response from routine Medical Imaging”, has been developed under our supervision. It consists of 5 published articles and one accepted article in peer review journals.

n	Title	Scientific Journal; DOI	Year	IF (Q)	Author Contribution
1	Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis	European Radiology; 10.1007/s00330-020-07174-0	2020	7.13 (Q1)	Literature research, Study conceptualization and design, Data acquisition, Data analysis and interpretation, Manuscript drafting and Revision
2	A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid	Radiology; 10.1148/radiol.2021200928	2021	29.146 (Q1)	
3	PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma.	eJHaem; 10.1002/jha2.757	2023	NA	
4	Radiomics and outcome prediction to anti-angiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumors: findings from the phase II TALENT trial.	BJC Rep; 10.1038/s44276-023-00010-0	2023	NA	
5	Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression.	Cancer Research Comm	2023	NA	

The author of the thesis declares equal contribution with Marc Simo in Study 3 and Jorge Hernando in Study 4. None of the studies in this thesis were used implicitly or explicitly in any other thesis.

Signed on the day **January 5th, 2024**



Raquel Perez-Lopez



Roser Sala-Llonch

CHAPTER 5

Summary of Results

Study 1 - Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis

Journal: European Radiology, **IF:** 7.13 (Q1)

This study corresponds to **objective 1** of the thesis. This research is aimed to identify CT-acquisition parameters accounting for radiomics variability and to develop a post-acquisition CT-image correction method to reduce variability. By correcting variability, this study aims to improve radiomics classification in both phantom and clinical applications.

For this study, fifty-three different CT-acquisition protocols were prospectively tested in a phantom containing 13 tissue densities. To validate the results in a real-world multicentric clinical setting, we retrospectively included 97 liver metastases from 43 patients with colorectal/renal cancer liver metastases scanned with CT machines from two different vendors (SIEMENS and General Electrics). Ninety-three radiomic features of first order and texture were extracted.

This study consist of a two-step framework, with the primary objective of **discerning the source of variability in CT image acquisition (Step A)**, followed by the implementation of strategies to **correct the introduced variability in the acquired dataset (Step B)**.

The findings in the first analysis (Step A) indicated that imaging acquisition parameters – including pixel size, reconstruction slice spacing, convolution kernel and acquisition slice thickness – were potential sources of variability. When these parameters differed between acquisitions, specially in extreme acquisitions, the number of robust features could be decreased to only 40%-50% of robust features, defining robust features as those with an

Intra-class Correlation Coefficient (ICC) greater than 0.8.

In the second part of the study (Step B), we explored different pre- and post-processing methods including voxel resize, ComBat correction, and singular value decomposition (SVD) compensation for reducing the radiomics variability. Interestingly, our results showed that when images were acquired with different pixel sizes (ranging from 0.39-1) and slice thickness (ranging from 1 to 5 mm), by only resampling images to isometric voxels in the preprocessing step, the number of robust features already increased significantly.

However, resampling techniques could not correct for all variability, specially regarding reconstruction convolution kernels and slice thickness. Further analysis where we implemented batch effect correction methods, revealed that ComBat harmonization method could correctly reduce features variability due to reconstruction kernels and slice thickness, significantly increasing the number of robust features and improving classification performance.

Overall, the results indicate that accounting for CT imaging acquisition and reconstruction parameters is crucial to get robust predictive models. This study proved that image preprocessing techniques can enhance reproducibility. Additionally, in datasets with defined batches such as different reconstruction kernels, ComBat harmonization method is a potent tool to harmonize the data.

Study 2 - A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid

Journal: Radiology, **IF:** 29.146 (Q1)

This study corresponds to objective 2. In this research, we explore the implementation of CT-based radiomic features to develop predictive signatures of response to immunotherapy in a population of advanced cancer patients with different tumor types and validate it in two separate cohorts of bladder and lung cancer.

In this retrospective study, a CT-based radiomics signature to predict response to immunotherapy was developed in 85 patients with advanced solid tumors with ICIs (cohort 1). The model was tested in 46 patients with bladder and 47 patients with lung cancer (cohorts 2 and 3) and biologically validated in 20 patients available contrast-enhanced CT and gene expression profiling (cohort 4).

Our findings showed that the radiomics signature was associated with clinical response to ICIs, area under the curve (AUC) of 0.70 (95% CI: 0.64-0.77) in the training set and

showed a moderate validation performance with AUC of 0.67 (95% CI: 0.58-0.76) and 0.67 (95% CI: 0.56-0.77) in cohorts 2 and 3, respectively. In concordance with state-of-the-art biomarkers, our interpretation of the signature coefficients indicate that homogeneous tumors were associated with a high probability of immune response.

Additionally, when integrating radiomic features with clinical data, including baseline albumin level and lymphocyte count, the signature outperformed radiomics-only performance with AUC of 0.74 (95% CI: 0.63-0.84). Finally, for the purpose of uncovering the biological explanation of the signature, we explored the correlation between the radiomics score and immune cytotoxic cells count, derived from gene expression profiling signatures. Our findings suggested that a higher radiomics score (0.29) appeared to correspond to an increased count of immune cytotoxic cells, although this correlation was not statistically significant (0.80 ± 0.12 vs. 0.62 ± 0.25 ; $P = 0.29$).

These results confirmed the potential of using CT radiomics to characterize responsive tumor phenotypes for immunotherapy, leading to the exploration of these tools in other settings in the following studies.

Study 3 - PET-based radiomics signature can predict durable responses to CAR T-cell therapy in patients with large B-cell lymphoma

Journal: eJHaem, **IF:** Not assigned

This study corresponds to objective 3. In this research, we investigated the generalizability of radiomic features to Positron Emission Tomography (PET) component in a setting of large B-cell lymphoma patients treated with CAR T-cell therapy.

In this study, we explored whether PET-based radiomics quantification of the SUV component could outperform conventional PET biomarkers such as maximum standardized uptake value (SUVmax) or total metabolic tumor volume (TMTV) for predicting response to CAR T-cell therapy. We conducted a single-center study including 93 consecutive R/R LBCL patient who received a CAR T-cell infusion from 2018 to 2021, split in training set (73 patients) and test set (20 patients).

Radiomic features were extracted from baseline PET scans and clinical benefit was defined based on median PFS. Our results showed that a radiomics signature combining 4 PET-based radiomic features, including tumor size and maximum SUV value, achieved an AUC of 0.73 (95% CI 0.48–0.98) for predicting clinical benefit in the test-set. Therefore, by

quantifying the SUV component using radiomics we can obtain enhanced predictive performance compared to the conventional PET biomarkers (TMTV AUC of 0.66 and SUVmax AUC of 0.59).

Additionally, Cox regression models were implemented to explore associations between the PET radiomic score and PFS or OS. Through multivariate analysis, these associations were compared with clinical variables (including age, ECOG, treatment type and number of previous lines). Univariate and multivariate cox regression showed that a high radiomic score was significantly associated with longer PFS and OS.

All together, we can conclude that radiomic quantification of the PET component adds value to conventional PET biomarkers limited to maximum values and tumor volume for predicting response to CAR T-cell therapy.

Study 4 - Radiomics and outcome prediction to anti-angiogenic treatment in advanced gastroenteropancreatic neuroendocrine tumors: findings from the phase II TALENT trial.

Journal: BJC reports, **IF:** Not assigned

This study corresponds to objective 3. In this research, we implemented radiomics pipelines to predict response to anti-angiogenics treatment in a population of patients with neuroendocrine tumors, characterized by a distinct imaging phenotype attributed to their highly vascularized nature. Additionally, the unique attributes of this tumor type, which is usually scanned in multiple CT-acquisition phases, facilitated the exploration of multi-modal radiomics analysis employing data from portal and arterial phases.

We developed the model in a population of 57 patients treated with Levantinib (cohort 1, split into training (80%) and validation (20%) sets) and test it in a cohort of 26 patients treated with Sunitinib. In the training and validation set, the model associated with response (PFS larger than the median) with an AUC of 0.75 (95% CI 0.66–0.90) and 0.67 (95% CI 0.49–0.92), respectively. However, in the test set the model showed moderate associations with response, AUC of 0.67 (95% CI 0.43–0.91).

Additionally, in this setting, patients were scanned with two contrast phases: arterial and venous. Accounting for that and the relevance of vasculature in such population treated with anti-angiogenics, we developed a multimodal model including radiomic features from the

two contrast-enhanced CT phases. Surprisingly, combining radiomic features from multiple phases did not improve the prediction capacity of the model (AUC = 0.59). On the other hand, combining radiomic features and clinical variables, including Ki-67 and primary tumor site, slightly improved the performance for predicting response in the three sets, AUC of 0.79 (95% CI 0.64–0.93), 0.72 (95% CI 0.48–0.95) and 0.58 (95% CI 0.33–0.82), respectively.

Regarding model interpretability, the coefficients of the model indicated that high radiomic score was presented in more spherical and hypervascularised tumors. Therefore, we explored the associations between radiomic score and proangiogenic profiling factors, including expression of VEGFR2 and ANG2. Hierarchical clustering was used to grouped patients based on their proangiogenic profiling, finding that patients in the high vascularization cluster showed significantly higher radiomic score ($P=0.030$).

The outcome of this study stresses the potential of contrast-enhanced image quantification for characterizing tumor vasculature towards better characterization of angiogenic properties of the tissue.

Study 5 - Weakly supervised deep learning predicts immunotherapy response in solid tumors based on PD-L1 expression

Journal: Cancer Research Communications (In Press), **IF:** Not assigned

This study corresponds to objective 4. It was developed in the framework of exploring different medical imaging fields beyond radiology with the aim of learning new techniques that could be translated to the radiology field and developing multimodal models integrating radiology and pathology. In this publication, we trained a weakly supervised deep learning multiple to classified PD-L1 stained IHC and assess its performance to predict response to immunotherapy.

In this study, we developed a deep learning (DL) method to predict PD-L1 status directly from raw IHC image data, without explicit intermediary steps such as cell detection or pigment quantification. We trained the weakly supervised model on PD-L1-stained slides from the NSCLC-MSK cohort (N=233) and validated it on the pancancer-VHIO cohort (N=108). We also investigated the performance of the model to predict response to immune checkpoint inhibitors (ICIs) in terms of PFS. In the pancancer-VHIO cohort, the performance was compared with TPS and CPS.

The DL-model showed good performance in predicting PD-L1 expression ($TPS \geq 1\%$)

in both NSCLC-MSK and pancancer-VHIO cohort (AUC 0.88 ± 0.06 and 0.80 ± 0.03 , respectively). The predicted PD-L1 status showed an improved association with response to ICIs (hazard ratio (HR) 1.5[95%CI 1-2.3], p-value=0.049) compared to TPS (HR 1.4[0.96-2.2], p-value=0.082) and CPS (HR 1.2[0.79-1.9], p-value=0.386). Notably, our explainability analysis showed that the model does not just look at the amount of brown pigment in the IHC slides, but also considers morphological factors such as lymphocyte conglomerates.

Overall, end-to-end weakly supervised DL shows potential for improving patient stratification for cancer immunotherapy by analyzing PD-L1 immunohistochemistry, holistically integrating morphology and PD-L1-staining intensity.

CHAPTER 6

Discussion

Immunotherapy, alongside other cutting-edge cancer treatments, has showcased remarkable efficacy among advanced cancer patients [2–4, 6]. With the increasing accessibility to these therapies and new combinations emerging, the demand for personalized medicine and precise treatment selection is already an urgent need in the current landscape [1]. Much of the discussion about personalized treatment has focused on molecular and genetic characterization of the tumor [26]. However, to date, there is no stand-alone biomarker suitable for accurately predicting response to immunotherapy or some other cancer treatments such as anti-angiogenics [14].

This thesis hypothesizes that routine medical images can translate meaningful information about tumor biology. Moreover, it proposes that the integration of machine learning and medical image analysis is key for identifying patterns linked to treatment response. Consequently, this thesis aimed to develop medical imaging based biomarkers that could predict response to treatment. In essence, this research explores the fundamental question of whether treatment response can be reliably predicted through medical imaging.

To address this question, the primary focus of the research is centered on CT-based radiomics analysis. The initial step involved a reproducibility study of the technique, driven by the hypothesis that identifying and addressing inherent data variability would enhance the model performance (**Study 1**). Subsequent studies delved into developing predictive models for immunotherapy response using CT-based radiomics (**Study 2**), followed by assessing the applicability of these methods when extended to other imaging techniques and targeted therapies, specifically to PET imaging (**Study 3**) and anti-angiogenic treatment (**Study 4**). Finally, we explored the potential of predicting immunotherapy outcomes from digital pathology imaging using deep learning (**Study 5**).

6.1 Challenges in the development of radiomics predictive signatures

Scarce data was published regarding radiomics implementation for cancer applications [83, 84] at the beginning of the thesis research. Specially regarding methods to handle multi-center studies and advanced metastatic patients. Most of the studies were focussed on finding the best combination of data dimensionality reduction and machine learning techniques [91, 105]. Despite that there were a few studies reporting the reproducibility of radiomic features [Jha2021-fg, 93], there was no solution proposed to deal with radiomics variability among multi-center acquisitions, feature aggregation methods and model selection. Therefore, in the studies corresponding to this thesis, we explored different methods regarding the study design.

6.1.1 Radiomics variability

The first study of this thesis played a pivotal role in shaping the subsequent research methodology. It allowed for identifying the potential sources of variability in CT data and set the basis for the radiomics methods implemented in the subsequent studies. By implementing pre-processing methods and batch-effect correction methods such as image resampling and ComBat harmonization, we demonstrated that it is possible to reduce data variability in radiomic features and improve the predictive performance of the models (Study 1). Despite the limited existing literature on batch effect correction methods for radiomic data, our findings were consistent with those of Orlhac et al [97]. Both studies demonstrated that after applying ComBat, the distribution of the radiomic data was harmonized among acquisitions. In addition, our study evidenced not only a harmonized data distribution but also an increase in the number of robust features, contributing to enhanced classification performance of the models.

6.1.2 Radiomic feature aggregation methods

Another relevant point in the study design for developing predictive signatures for treatment response is the definition of the target lesions. Many radiomics studies rely on delineating and extracting features from a single selected tumor, often the primary or largest lesion, arguing that the single chosen lesion can be representative of the whole disease [165, 166]. However, in the context of advanced cancer patients with multi-site metastatic disease, this approach may overlook relevant spatial information, miss tumor heterogeneity and, ultimately, limit the predictive capacity of the model [167]. To partially overcome this limitation and capture the

full spectrum of patient heterogeneity, feature aggregation methods such as average, volume-weighted average, or attention-based multiple instance learning (MIL) can be applied [168]. In Study 2, 3 and 4, where radiomic predictive signatures were developed, we selected specific aggregation methods according to the clinical setting of each study. These are described in more detail below.

In Study 2 we proposed to aggregate all the radiomic data from all the metastatic disease by including mixed models. These mixed models fit the patient distribution inside the whole population distribution, allowing to account for random effects that are intrinsically from the patient. This method will output a lesion-level prediction while considering the global patient response. We believed that including the whole burden of disease presents an advantage to single lesion methods. Some studies that focused on predicting surrogate biomarkers quantified on the biopsied lesion, simplified the radiomics analysis by using the same lesion where the sample was obtained [131, 169, 170]. Although this solution is methodologically correct in terms of the surrogate endpoint, it excludes relevant information from the patient status and heterogeneity that might influence the overall patient response [12].

A similar approach to our proposed solution was the one from Trebeschi et al., where they included the total burden of disease by predicting tumor shrinkage at a lesion-level and aggregating to a patient-level through dissociated response analysis [132]. Similarly, Himoto et al. presented a study to assess intra- and inter- lesion heterogeneity among all lesions of the patients [134]. Interestingly, these studies associated heterogeneity in tumors or in volume changes to response by means of PFS but failed to model all the data towards a final prediction. In our proposed solution, we implemented a method to account for all this heterogeneity while still providing a final response prediction. Therefore, despite being aware of the limitations of this modeling, which requires large data to be trained, we believe that including a representation of the whole disease in the model could improve the performance.

However, feature aggregation methods are still limited by the inherent properties of the data. For instance, in Study 4 we opted for the median of the features across all liver lesions as a feature aggregation method, given that the lesions' tissue properties were the same. On the other hand, in study 3, due to the diffuse nature of lymphoma throughout the entire body and the non-mesurable properties of the majority of the disease, we decided to perform radiomics analysis of the largest lesion. Therefore, we can conclude that information aggregation of metastatic disease poses a complex challenge in radiomics studies and that no definitive solution exists for study design. While numerous studies have attempted to compare aggregation methods, there is still a lack of consistent results [165–167, 171].

6.1.3 Definition of dimensionality reduction and classifier selection of radiomics studies

When formulating the study design for a radiomic model, a critical aspect is the incorporation of techniques for reducing data dimensionality and choosing an optimal model. Numerous radiomics studies implemented LASSO for feature selection, followed by logistic regression for classification [165, 172–174]. Similarly to LASSO, in Study 2 we implemented elastic-net as feature selection, a regularized regression method that adds an extra penalty to the LASSO regression. However, in the following studies (Study 3 and 4), we implemented LASSO and logistic regression as the standard pipeline, prioritizing the focus on the generalizability of the technique.

In terms of classification method, several studies have explored the performance of different classification algorithms for predicting response (such as SVM, RF, decision trees and kNN). All of them showed that logistic regression had similar or slightly better performance than other more complex classifiers [109–111, 165]. In the same lines, we have found the implementation of logistic regression applied to handcrafted features radiomics as the most straightforward option, given that more complex classifiers were not improving logistic regression performance in our study cases.

6.2 The translational impact of radiomics and AI for cancer treatment

The main research question addressed in the included studies in this thesis relies on the application of predictive models to improve patient selection and customize the most effective treatment strategy for each individual. However, in the development of predictive models several assumptions are done regarding population description. It is crucial to account for the implications of cohort definition and target endpoints in the design choices.

6.2.1 Development of predictive models across diverse tumor types

Different properties contribute to make a tumor phenotype more responsive to treatment. These properties can be tumor-specific or can be found across different tumor types [14, 175]. Moreover, the tumor type itself is a critical determinant of a patient’s likelihood to respond to certain treatments [12]. In the case of immunotherapy, for example, patients with tumor types like NSCLC and melanoma tend to be more responsive compared to other tumor types

such as breast or colorectal cancer [32]. The notable success of immunotherapy in NSCLC and melanoma has positioned it as a standard of care treatment, which directly influences on the availability and quality of the data for predictive models. For that reason, to date, the majority of radiomics studies on immunotherapy response prediction have focused on NSCLC, benefiting from the availability of larger datasets and a higher degree of treatment responses in this tumor type [170, 174, 176–179]. Other tumor types including melanoma, gastric, head and neck, bladder and kidney cancers have been investigated to a lesser extent [132, 135, 172, 180, 181] and only few studies have developed predictive radiomic models in pan-cancer settings [131]).

In Study 2 we used advanced patients with different tumor types treated in clinical trials, and we aimed to develop a generalizable predictive model for response to immunotherapy. We showed that the model could learn the tumor phenotypes across different tumor types and generalize to tumor specific cohorts from bladder and lung cancer. However, in Study 3 and 4, limited by the available cohorts and the specificity of the included treatment to a tumor type, anti-angiogenics for neuroendocrine tumors and CAR T-cells for lymphoma, we developed tumor type-specific radiomic signatures. Similarly, in Study 5, we trained the model on a NSCLC population treated with immunotherapy but then demonstrated its generalizability to pan-cancer cohorts. In that study, we also explore the performance when training in pan-cancer and validating in a single cohort. However, the limitations of the pan-cancer cohort, characterized by a small sample size, substantial tumor type heterogeneity, and the inclusion of predominantly advanced tumors accessible only via biopsy, could potentially reduce the reliability of the model performance in comparison to the other approach.

In conclusion, the development of tumor type-specific radiomic signatures allows finding radiomic features unique to that population and treatment pattern, however it reduces the generalization of the methods to other tumor types that are less common or rarely treated with that specific treatment. On the other hand, pan-cancer approaches require the use of larger cohorts for the model to comprehend the inherent heterogeneity of the population, thereby reducing the bias towards the response probability of each tumor type.

6.2.2 Endpoint definition for immunotherapy response predictive model

Another relevant consideration in the study design for developing a model with potential clinical translation is the definition of the target endpoint that will indicate response to treatment. Studies focusing on predictive signatures and, specially, in immunotherapy can be categorized into two types based on the endpoints: studies aiming to directly predict

clinical outcome and studies focused on predicting known surrogate biomarkers of response (i.e., biological and molecular markers used as predictors of a patient’s response such as PD-L1).

Defining when a tumor is responding is a challenge itself, given the subjectivity of the assessment. Consequently, the most relevant and robust measure for evaluating if a patient is benefitting from treatment is overall survival (OS). Although certain radiomic studies have considered OS as the clinical endpoint [135, 136, 174], most studies rely on tumor size changes by the Response Evaluation Criteria in Solid Tumors version 1.1 (RECIST 1.1) [77].

From the RECIST assessment, multiple measurements can be computed and used as endpoints, including PFS, disease control (which gathers CR and PR and SD) or objective response rate (ORR). These measures, despite being the gold standard to decide patient treatment discontinuation in clinical trials and standard of care, present some limitations: they are subjectively defined and highly variable. For that reason, these response measurements can be considered as surrogate endpoints for OS [76].

The same problem is found when surrogate biomarkers are defined as endpoints. Most of these surrogates of response are still not good predictors, and therefore, the performance of predictive models targeting these biomarkers will be inherently limited by the predictive capacity of the surrogate biomarker itself. This implies that the effectiveness of the model in predicting treatment response will be constrained by the predictive capabilities of the surrogate biomarker being used.

Furthermore, these endpoints, despite their subjectivity, are usually continuous variables and in the case of survival, they are also censored. Due to the complexity of the regression modeling, most of the studies, including the ones from this thesis, tend to discretize the endpoint. In all studies 2, 3 and 4, we opted for a definition of clinical benefit, which was defined as PFS larger than the median for that specific type of treatment or primary tumor type. This option allowed for simplifying the predictive model towards a classification problem. Similarly, when predicting molecular surrogate biomarkers (such as PD-L1 expression), many studies, including study 5, also tend to discretize the target variable and transform it into a classification problem. However, this limitation is tackled in all four studies by studying the associations of these models with the ultimate clinical endpoint of OS.

6.2.3 Multi-center studies to improve model generalizability

According to the presented data, we can infer that medical imaging has great potential for the development of new biomarkers of response. However, further work is needed to reach the implementation of image-based biomarkers into clinical practice. Some of the limitations of

our projects is the development of single-center studies like in Study 2 and 3. Even though we can correct variability from different centers using methods defined in Study 1, developing multi-center studies is essential to demonstrate the applicability of these methods across large and heterogeneous datasets. That will ensure reliability and fairness by encompassing diverse populations and scanners from various institutions. This limitation is tackled in this thesis by Study 4 where we developed a predictive signature in patients from a multi-centric clinical trial.

6.2.4 Integration of radiomics and clinical data

Furthermore, imaging biomarkers can also be seen as an extra value for integrative models that can capture information from different clinical sources. For that reason, in study 2, 3 and 4 we explored the potential of combining radiomic data to clinical data. In all those studies, we proved that the combination of radiomics with clinical variables always improves the response prediction.

Some studies have highlighted potential improvements of predictive models through multimodal approaches that combine radiomics with histopathology or genomics [Crispin-Ortuzar2023-rp, 157]. With that purpose, in Study 5 we explored the development of deep learning methods for histopathology to demonstrate its potential as biomarkers for immunotherapy but to also understand how to integrate them with radiomic models such as the one presented in Study 2. During this thesis, large efforts were done to obtain a big dataset of patients treated with immunotherapy that contained information from multiple data sources, including radiology and pathology. In line with other studies [155], preliminary results showed that the combination of PD-L1 and CT image data into a model improved the prediction performance compared to models based on a single image modality [182]. Still, this requires representative heterogeneous data, ideally from multiple centers including all sources of information, which has been a notable limitation thus far.

6.3 Biological validation of radiomic features

Integration of radiomics-based biomarkers into clinical practice hinges on the critical aspects of explainability and trustability. It is crucial that healthcare professionals can comprehend and rely on these complex data-driven insights to make informed patient care decisions. Therefore, the path to integrating radiomics into clinical practice still relies on biological translation of the predictive models [183].

In Study 2 and 4, we explored potential biological meaning to the developed radiomic

predictive models. In both studies, we demonstrated that radiomic features can capture biological differences of the tumor phenotype that are associated with response to specific treatment. For instance, in Study 2, we observed that radiomic features indicating homogeneity proved to be more strongly associated with response. This property of tumor, though well-established, is often challenging to quantify accurately. Through the application of radiomics, we provided a non-invasive method to quantify this relevant aspect of tumor phenotypes. Similarly, in Study 4, we found that radiomic features related to homogeneity in shape (sphericity) and enhancement were indicative of tumor response to anti-angiogenic treatment. One possible explanation for this result is that increased enhancement may be attributed to enhanced vascularization, making it more susceptible to therapies targeting tumor vascular growth, as is the case with anti-angiogenic treatment. In conclusion, our research demonstrates that radiomic features can capture aspects of tumor biology and that in combination with machine learning methods could potentially enhance patient stratification towards treatment selection options.

To validate this potential biological explanation of the radiomic features in both studies we correlated the predicted radiomic score with well known biomarkers for each treatment; cytotoxic immunophenotype enrichment in the tumor for immunotherapy (Study 2) and vasculature growth factors such as VGFR and ANG2 (Study 4). In both studies, we saw a relationship between the molecular biomarkers and the radiomic score. However, ongoing investigation in this direction is needed to enhance the reliability and applicability of these models for seamless integration into routine clinical practice.

6.4 Limitations and future perspectives in the development of AI-based clinical biomarkers

The main limitation in the studies included in the thesis is the use of real-world clinical data in two growing fields, such as immunotherapy and artificial intelligence. The limited sample size and data acquisition variability from clinical practice limited the implementation of more complex methods. However, it raised the opportunity to explore variability correction methods and to focus on more specific data populations.

It is noteworthy that both the fields of immunotherapy and artificial intelligence have experienced substantial growth. The recent advancements in immunotherapy may pose challenges for the development of AI tools, requiring methods that can deal with increasingly diverse and less standardized patient cohorts, specially regarding treatment type. What initially started with single-agent immunotherapies, such as anti-PD-L1 and anti-PD1, has

expanded to include new agents in combination, targeting new pathways related to immune response or other biological processes. However, increasing population diversity in immunotherapy is not solely attributed to the development of combinations with new targets. Clinical studies investigating the integration of immunotherapy with other treatments like anti-angiogenics, chemotherapy or radiotherapy, are yielding notable results, potentially influencing future implementations [184].

Even though radiomics has been proposed as a solution to extract quantitative representation of tumor phenotype from medical images [84], several limitations have been raised in the past years. Radiomics field is limited to the pre-defined features that were not designed specifically for medical images but more for general statistical heterogeneity quantification. Additionally, many studies have raised the concern about variability among image acquisition parameters and the need for labeled data to obtain representations [93, 98]. Therefore, future research questions that could be asked include how can we obtain more medical imaging specific feature representations that can account for the variability of image acquisition and represent the medical imaging without the need of manually image labeling.

New lines of research are pointing to the use of deep learning development of foundation models that can learn from large sets of images without any specific task [185]. By leveraging medical imaging knowledge, these models can enhance task-specific models used for specific medical problems with fewer data available. In the case of cancer treatment and personalized medicine, the development of foundation models, which can learn from previously unseen tasks, will support the development of predictive biomarkers for novel therapies tested in small subsets of patients. Some efforts have already been made in the field of radiology [186], digital pathology [128] and medical text domain-based language models [187] that are already showing this reality [188]. However, ongoing research is needed to obtain a clinical supporting tool for cancer treatment.

In conclusion, the new developments and efforts in the AI field are pointing towards the construction of AI models that will support precision oncology and will provide solutions for a better understanding of the complexity of cancer evolution and treatment response [18].

CHAPTER 7

Conclusion

1. The measurements in radiomics are susceptible to image acquisition and reconstruction parameters, including factors like slice thickness and reconstruction filters. In the context of multi-centric studies, addressing this variability through harmonization techniques can be beneficial. The incorporation of the ComBat batch effect correction method in radiomics analysis holds the potential to enhance the performance of classification methods, ultimately improving the robustness of the models.
2. Radiomics quantification enables the extraction of representative features from CT images, effectively capturing various tumor phenotypes. These identified imaging tumor phenotypes provide insights into tumor heterogeneity and immune phenotype, directly influencing immunotherapy response. Integrating these radiomic representations with clinical data could significantly improve the accuracy of predicting immunotherapy response.
3. Radiomic features can be extended to capture additional representations of tumor phenotypes, including tumor vasculature. This representation can be predictive of response to anti-angiogenic targeted agents. Furthermore, the application of radiomic features is not limited to CT imaging alone, it can also be extended to other imaging modalities, such as PET imaging. This application allows for the quantification of tumor metabolism, potentially achieving improved performance compared to traditional state-of-the-art metrics.
4. Pathology imaging can provide a more detailed cellular-level representation compared to radiology imaging. By extracting information from the tumor phenotype, these pathology images can be used to predict immunotherapy response through surrogate biomarker prediction, notably by predicting PD-L1 status from immunohistochemistry (IHC) images. These methods have demonstrated generalization to different tumor types, even when initially trained on specific tumors.

Bibliography

- [1] F. Ciardiello, D. Arnold, P. G. Casali, *et al.*, “Delivering precision medicine in oncology today and in future-the promise and challenges of personalised cancer medicine: A position paper by the european society for medical oncology (ESMO),” en, *Ann. Oncol.*, vol. 25, no. 9, pp. 1673–1678, Sep. 2014.
- [2] C. Robert, G. V. Long, B. Brady, *et al.*, “Nivolumab in previously untreated melanoma without BRAF mutation,” en, *N. Engl. J. Med.*, vol. 372, no. 4, pp. 320–330, Jan. 2015.
- [3] J. Brahmer, K. L. Reckamp, P. Baas, *et al.*, “Nivolumab versus docetaxel in advanced Squamous-Cell Non-Small-Cell lung cancer,” en, *N. Engl. J. Med.*, vol. 373, no. 2, pp. 123–135, Jul. 2015.
- [4] T. André, K.-K. Shiu, T. W. Kim, *et al.*, “Pembrolizumab in Microsatellite-Instability-High advanced colorectal cancer,” *N. Engl. J. Med.*, vol. 383, no. 23, pp. 2207–2218, Dec. 2020.
- [5] F. S. Hodi, S. J. O’Day, D. F. McDermott, *et al.*, “Improved survival with ipilimumab in patients with metastatic melanoma,” en, *N. Engl. J. Med.*, vol. 363, no. 8, pp. 711–723, Aug. 2010.
- [6] T. Powles, I. Durán, M. S. van der Heijden, *et al.*, “Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): A multicentre, open-label, phase 3 randomised controlled trial,” en, *Lancet*, vol. 391, no. 10122, pp. 748–757, Feb. 2018.
- [7] J. Bellmunt, R. de Wit, D. J. Vaughn, *et al.*, “Pembrolizumab as Second-Line therapy for advanced urothelial carcinoma,” en, *N. Engl. J. Med.*, vol. 376, no. 11, pp. 1015–1026, Mar. 2017.
- [8] R. J. Motzer, B. Escudier, D. F. McDermott, *et al.*, “Nivolumab versus everolimus in advanced Renal-Cell carcinoma,” en, *N. Engl. J. Med.*, vol. 373, no. 19, pp. 1803–1813, Nov. 2015.

-
- [9] M. Binnewies, E. W. Roberts, K. Kersten, *et al.*, “Understanding the tumor immune microenvironment (TIME) for effective therapy,” en, *Nat. Med.*, vol. 24, no. 5, pp. 541–550, May 2018.
 - [10] I. J. Fidler, “Tumor heterogeneity and the biology of cancer invasion and metastasis,” en, *Cancer Res.*, vol. 38, no. 9, pp. 2651–2660, Sep. 1978.
 - [11] C. Swanton, “Intratumor heterogeneity: Evolution through space and time,” en, *Cancer Res.*, vol. 72, no. 19, pp. 4875–4882, Oct. 2012.
 - [12] A. Marusyk, M. Janiszewska, and K. Polyak, “Intratumor heterogeneity: The rosetta stone of therapy resistance,” en, *Cancer Cell*, vol. 37, no. 4, pp. 471–484, Apr. 2020.
 - [13] N. McGranahan and C. Swanton, “Biological and therapeutic impact of intratumor heterogeneity in cancer evolution,” en, *Cancer Cell*, vol. 27, no. 1, pp. 15–26, Jan. 2015.
 - [14] J. S. Lee and E. Ruppin, “Multiomics prediction of response rates to therapies to inhibit programmed cell death 1 and programmed cell death 1 ligand 1,” en, *JAMA Oncol*, vol. 5, no. 11, pp. 1614–1618, Nov. 2019.
 - [15] H. L. Duvivier, M. Rothe, P. K. Mangat, *et al.*, “Pembrolizumab in patients with tumors with high tumor mutational burden: Results from the targeted agent and profiling utilization registry study,” en, *J. Clin. Oncol.*, JCO2300702, Aug. 2023.
 - [16] T. A. Chan, M. Yarchoan, E. Jaffee, *et al.*, “Development of tumor mutation burden as an immunotherapy biomarker: Utility for the oncology clinic,” en, *Ann. Oncol.*, vol. 30, no. 1, pp. 44–56, Jan. 2019.
 - [17] Food and U. S. Drug Administration, “FDA approves pembrolizumab for adults and children with TMB-H solid tumors,” *FDA Med. Bull.*,
 - [18] P. J. Ballester and J. Carmona, “Artificial intelligence for the next generation of precision oncology,” en, *NPJ Precis Oncol*, vol. 5, no. 1, p. 79, Aug. 2021.
 - [19] N. Ghaffari Laleh, H. S. Muti, C. M. L. Loeffler, *et al.*, “Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology,” en, *Med. Image Anal.*, vol. 79, p. 102474, Jul. 2022.
 - [20] G. V. Long, J. S. Weber, J. Larkin, *et al.*, “Nivolumab for patients with advanced melanoma treated beyond progression: Analysis of 2 phase 3 clinical trials,” en, *JAMA Oncol*, vol. 3, no. 11, pp. 1511–1519, Nov. 2017.

BIBLIOGRAPHY

- [21] M. A. Postow, J. Chesney, A. C. Pavlick, *et al.*, “Nivolumab and ipilimumab versus ipilimumab in untreated melanoma,” en, *N. Engl. J. Med.*, vol. 372, no. 21, pp. 2006–2017, May 2015.
- [22] D. L. Jardim, D. de Melo Gagliato, F. J. Giles, *et al.*, “Analysis of drug development paradigms for immune checkpoint inhibitors,” en, *Clin. Cancer Res.*, vol. 24, no. 8, pp. 1785–1794, Apr. 2018.
- [23] C. Pilard, M. Ancion, P. Delvenne, *et al.*, “Cancer immunotherapy: It’s time to better predict patients’ response,” en, *Br. J. Cancer*, vol. 125, no. 7, pp. 927–938, Sep. 2021.
- [24] C. L. Ventola, “Cancer immunotherapy, part 3: Challenges and future trends,” en, *P T*, vol. 42, no. 8, pp. 514–521, Aug. 2017.
- [25] E. B. Garon, “Cancer immunotherapy trials not immune from imprecise selection of patients,” en, *N. Engl. J. Med.*, vol. 376, no. 25, pp. 2483–2485, Jun. 2017.
- [26] J. J. Havel, D. Chowell, and T. A. Chan, “The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy,” en, *Nat. Rev. Cancer*, vol. 19, no. 3, pp. 133–150, Feb. 2019.
- [27] J. H. Strickler, B. A. Hanks, and M. Khasraw, “Tumor mutational burden as a predictor of immunotherapy response: Is more always better?” en, *Clin. Cancer Res.*, vol. 27, no. 5, pp. 1236–1241, Mar. 2021.
- [28] M. Yarchoan, A. Hopkins, and E. M. Jaffee, “Tumor mutational burden and response rate to PD-1 inhibition,” en, *N. Engl. J. Med.*, vol. 377, no. 25, pp. 2500–2501, Dec. 2017.
- [29] J. E. Rosenberg, J. Hoffman-Censits, T. Powles, *et al.*, “Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: A single-arm, multicentre, phase 2 trial,” en, *Lancet*, vol. 387, no. 10031, pp. 1909–1920, May 2016.
- [30] A. Snyder, V. Makarov, T. Merghoub, *et al.*, “Genetic basis for clinical response to CTLA-4 blockade in melanoma,” en, *N. Engl. J. Med.*, vol. 371, no. 23, pp. 2189–2199, Dec. 2014.
- [31] A. Marabelle, M. Fakih, J. Lopez, *et al.*, “Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: Prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study,” en, *Lancet Oncol.*, vol. 21, no. 10, pp. 1353–1365, Oct. 2020.

- [32] N. A. Rizvi, M. D. Hellmann, A. Snyder, *et al.*, “Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” en, *Science*, vol. 348, no. 6230, pp. 124–128, Apr. 2015.
- [33] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, *et al.*, “Signatures of mutational processes in human cancer,” en, *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013.
- [34] R. M. Samstein, C.-H. Lee, A. N. Shoushtari, *et al.*, “Tumor mutational load predicts survival after immunotherapy across multiple cancer types,” en, *Nat. Genet.*, vol. 51, no. 2, pp. 202–206, Feb. 2019.
- [35] N. McGranahan, A. J. S. Furness, R. Rosenthal, *et al.*, “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade,” en, *Science*, vol. 351, no. 6280, pp. 1463–1469, Mar. 2016.
- [36] N. Andor, T. A. Graham, M. Jansen, *et al.*, “Pan-cancer analysis of the extent and consequences of intratumor heterogeneity,” en, *Nat. Med.*, vol. 22, no. 1, pp. 105–113, Jan. 2016.
- [37] L. G. T. Morris, N. Riaz, A. Desrichard, *et al.*, “Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival,” en, *Oncotarget*, vol. 7, no. 9, pp. 10 051–10 063, Mar. 2016.
- [38] D. T. Le, J. N. Durham, K. N. Smith, *et al.*, “Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade,” en, *Science*, vol. 357, no. 6349, pp. 409–413, Jul. 2017.
- [39] D. T. Le, J. N. Uram, H. Wang, *et al.*, “PD-1 blockade in tumors with Mismatch-Repair deficiency,” en, *N. Engl. J. Med.*, vol. 372, no. 26, pp. 2509–2520, Jun. 2015.
- [40] G. Germano, S. Lamba, G. Rospo, *et al.*, “Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth,” en, *Nature*, vol. 552, no. 7683, pp. 116–120, Dec. 2017.
- [41] S. L. Topalian, F. S. Hodi, J. R. Brahmer, *et al.*, “Safety, activity, and immune correlates of Anti-PD-1 antibody in cancer,” *N. Engl. J. Med.*, vol. 366, no. 26, pp. 2443–2454, Jun. 2012.
- [42] M. Reck, D. Rodríguez-Abreu, A. G. Robinson, *et al.*, “Pembrolizumab versus chemotherapy for PD-L1-Positive Non-Small-Cell lung cancer,” *N. Engl. J. Med.*, vol. 375, no. 19, pp. 1823–1833, Nov. 2016.
- [43] S. L. Topalian, J. M. Taube, R. A. Anders, *et al.*, “Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy,” en, *Nat. Rev. Cancer*, vol. 16, no. 5, pp. 275–287, Apr. 2016.

- [44] G. T. Gibney, L. M. Weiner, and M. B. Atkins, “Predictive biomarkers for checkpoint inhibitor-based immunotherapy,” *Lancet Oncol.*, vol. 17, no. 12, e542–e551, Dec. 2016.
- [45] A. I. Daud, J. D. Wolchok, C. Robert, *et al.*, “Programmed Death-Ligand 1 expression and response to the Anti-Programmed death 1 antibody pembrolizumab in melanoma,” en, *J. Clin. Oncol.*, vol. 34, no. 34, pp. 4102–4109, Dec. 2016.
- [46] M. Mino-Kenudson, “Programmed cell death ligand-1 (PD-L1) expression by immunohistochemistry: Could it be predictive and/or prognostic in non-small cell lung cancer?” en, *Cancer Biol Med*, vol. 13, no. 2, pp. 157–170, Jun. 2016.
- [47] H. Brunnström, A. Johansson, S. Westbom-Fremer, *et al.*, “PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: Inter-pathologist variability is higher than assay variability,” en, *Mod. Pathol.*, vol. 30, no. 10, pp. 1411–1421, Oct. 2017.
- [48] S. Chang, H. K. Park, Y.-L. Choi, *et al.*, “Interobserver reproducibility of PD-L1 biomarker in non-small cell lung cancer: A Multi-Institutional study by 27 pathologists,” en, *J Pathol Transl Med*, vol. 53, no. 6, pp. 347–353, Nov. 2019.
- [49] K. Kulangara, D. A. Hanks, S. Waldroup, *et al.*, “Development of the combined positive score (CPS) for the evaluation of PD-L1 in solid tumors with the immunohistochemistry assay PD-L1 IHC 22C3 pharmdx,” *J. Clin. Orthod.*, vol. 35, no. 15-suppl, e14589–e14589, May 2017.
- [50] W. H. Fridman, F. Pagès, C. Sautès-Fridman, *et al.*, “The immune contexture in human tumours: Impact on clinical outcome,” en, *Nat. Rev. Cancer*, vol. 12, no. 4, pp. 298–306, Mar. 2012.
- [51] J. Galon, A. Costes, F. Sanchez-Cabo, *et al.*, “Type, density, and location of immune cells within human colorectal tumors predict clinical outcome,” en, *Science*, vol. 313, no. 5795, pp. 1960–1964, Sep. 2006.
- [52] B. Mlecnik, G. Bindea, H. K. Angell, *et al.*, “Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability,” en, *Immunity*, vol. 44, no. 3, pp. 698–711, Mar. 2016.
- [53] F. Pagès, B. Mlecnik, F. Marliot, *et al.*, “International validation of the consensus immunoscore for the classification of colon cancer: A prognostic and accuracy study,” en, *Lancet*, vol. 391, no. 10135, pp. 2128–2139, May 2018.
- [54] M. Sade-Feldman, K. Yizhak, S. L. Bjorgaard, *et al.*, “Defining T cell states associated with response to checkpoint immunotherapy in melanoma,” en, *Cell*, vol. 176, no. 1-2, p. 404, Jan. 2019.

- [55] D. S. Chen and I. Mellman, “Elements of cancer immunity and the cancer-immune set point,” en, *Nature*, vol. 541, no. 7637, pp. 321–330, Jan. 2017.
- [56] R. S. Herbst, J.-C. Soria, M. Kowanetz, *et al.*, “Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients,” en, *Nature*, vol. 515, no. 7528, pp. 563–567, Nov. 2014.
- [57] R. Cristescu, R. Mogg, M. Ayers, *et al.*, “Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy,” en, *Science*, vol. 362, no. 6411, Oct. 2018.
- [58] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer,” en, *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.
- [59] G. C. Jayson, R. Kerbel, L. M. Ellis, *et al.*, “Antiangiogenic therapy in oncology: Current status and future directions,” en, *Lancet*, vol. 388, no. 10043, pp. 518–529, Jul. 2016.
- [60] J. Folkman, “Tumor angiogenesis: Therapeutic implications,” en, *N. Engl. J. Med.*, vol. 285, no. 21, pp. 1182–1186, Nov. 1971.
- [61] N. Ferrara and R. S. Kerbel, “Angiogenesis as a therapeutic target,” en, *Nature*, vol. 438, no. 7070, pp. 967–974, Dec. 2005.
- [62] B. Escudier, C. Szczylik, C. Porta, *et al.*, “Treatment selection in metastatic renal cell carcinoma: Expert consensus,” en, *Nat. Rev. Clin. Oncol.*, vol. 9, no. 6, pp. 327–337, Apr. 2012.
- [63] H. Hurwitz, L. Fehrenbacher, W. Novotny, *et al.*, “Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer,” en, *N. Engl. J. Med.*, vol. 350, no. 23, pp. 2335–2342, Jun. 2004.
- [64] B. J. Monk, A. Poveda, I. Vergote, *et al.*, “Anti-angiopoietin therapy with trebananib for recurrent ovarian cancer (TRINOVA-1): A randomised, multicentre, double-blind, placebo-controlled phase 3 trial,” en, *Lancet Oncol.*, vol. 15, no. 8, pp. 799–808, Jul. 2014.
- [65] R. A. Burger, M. F. Brady, M. A. Bookman, *et al.*, “Incorporation of bevacizumab in the primary treatment of ovarian cancer,” en, *N. Engl. J. Med.*, vol. 365, no. 26, pp. 2473–2483, Dec. 2011.
- [66] J. M. Llovet, S. Ricci, V. Mazzaferro, *et al.*, “Sorafenib in advanced hepatocellular carcinoma,” en, *N. Engl. J. Med.*, vol. 359, no. 4, pp. 378–390, Jul. 2008.

- [67] A.-L. Cheng, Y.-K. Kang, Z. Chen, *et al.*, “Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: A phase III randomised, double-blind, placebo-controlled trial,” en, *Lancet Oncol.*, vol. 10, no. 1, pp. 25–34, Jan. 2009.
- [68] E. Raymond, L. Dahan, J.-L. Raoul, *et al.*, “Sunitinib malate for the treatment of pancreatic neuroendocrine tumors,” en, *N. Engl. J. Med.*, vol. 364, no. 6, pp. 501–513, Feb. 2011.
- [69] J. Capdevila, N. Fazio, C. Lopez, *et al.*, “Lenvatinib in patients with advanced grade 1/2 pancreatic and gastrointestinal neuroendocrine tumors: Results of the phase II TALENT trial (GETNE1509),” en, *J. Clin. Oncol.*, vol. 39, no. 20, pp. 2304–2312, Jul. 2021.
- [70] M. Hilmi, C. Neuzillet, J. Calderaro, *et al.*, “Angiogenesis and immune checkpoint inhibitors as therapies for hepatocellular carcinoma: Current knowledge and future research directions,” en, *J Immunother Cancer*, vol. 7, no. 1, p. 333, Nov. 2019.
- [71] M. Manzoni, B. Rovati, M. Ronzoni, *et al.*, “Immunological effects of bevacizumab-based treatment in metastatic colorectal cancer,” en, *Oncology*, vol. 79, no. 3-4, pp. 187–196, 2010.
- [72] E. C. Martino, G. Misso, P. Pastina, *et al.*, “Immune-modulating effects of bevacizumab in metastatic non-small-cell lung cancer patients,” en, *Cell Death Discov*, vol. 2, p. 16 025, Oct. 2016.
- [73] J. J. Wallin, J. C. Bendell, R. Funke, *et al.*, “Atezolizumab in combination with bevacizumab enhances antigen-specific t-cell migration in metastatic renal cell carcinoma,” en, *Nat. Commun.*, vol. 7, p. 12 624, Aug. 2016.
- [74] A. Hosny, C. Parmar, J. Quackenbush, *et al.*, “Artificial intelligence in radiology,” en, *Nat. Rev. Cancer*, vol. 18, no. 8, pp. 500–510, Aug. 2018.
- [75] J. Hsieh, *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*, en. SPIE Press, 2003.
- [76] C. K. Kuhl, Y. Alparslan, J. Schmoe, *et al.*, “Validity of RECIST version 1.1 for response assessment in metastatic cancer: A prospective, multireader study,” en, *Radiology*, vol. 290, no. 2, pp. 349–356, Feb. 2019.
- [77] E. A. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, “New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1),” en, *Eur. J. Cancer*, vol. 45, no. 2, pp. 228–247, Jan. 2009.

-
- [78] D. L. Bailey, D. W. Townsend, P. E. Valk, *et al.*, *Positron Emission Tomography: Basic Sciences*, en. Springer Science & Business Media, Jul. 2006.
- [79] G. Iacoboni, M. Simó, G. Villacampa, *et al.*, “Prognostic impact of total metabolic tumor volume in large b-cell lymphoma patients receiving CAR t-cell therapy,” en, *Ann. Hematol.*, vol. 100, no. 9, pp. 2303–2310, Sep. 2021.
- [80] L. Vercellino, A.-S. Cottreau, O. Casasnovas, *et al.*, “High total metabolic tumor volume at baseline predicts survival independent of response to therapy,” en, *Blood*, vol. 135, no. 16, pp. 1396–1405, Apr. 2020.
- [81] A.-S. Cottreau, H. Lanic, S. Mareschal, *et al.*, “Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-Cell lymphoma,” en, *Clin. Cancer Res.*, vol. 22, no. 15, pp. 3801–3809, Aug. 2016.
- [82] J. F. Gomes Marin, P. S. Duarte, J. Willegaignon de Amorim de Carvalho, *et al.*, “Comparison of the variability of SUV normalized by skeletal volume with the variability of SUV normalized by body weight in 18F-Fluoride PET/CT,” en, *J. Nucl. Med. Technol.*, vol. 47, no. 1, pp. 60–63, Mar. 2019.
- [83] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” en, *Nat. Commun.*, vol. 5, p. 4006, Jun. 2014.
- [84] P. Lambin, R. T. H. Leijenaar, T. M. Deist, *et al.*, “Radiomics: The bridge between medical imaging and personalized medicine,” en, *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, Dec. 2017.
- [85] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, “Radiomics: Extracting more information from medical images using advanced feature analysis,” en, *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012.
- [86] V. Kumar, Y. Gu, S. Basu, *et al.*, “Radiomics: The process and the challenges,” en, *Magn. Reson. Imaging*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012.
- [87] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images are more than pictures, they are data,” en, *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016.
- [88] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

- [89] N. Ghaffari Laleh, M. Ligerio, R. Perez-Lopez, *et al.*, “Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer,” en, *Clin. Cancer Res.*, Sep. 2022.
- [90] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, *et al.*, “Computational radiomics system to decode the radiographic phenotype,” en, *Cancer Res.*, vol. 77, no. 21, e104–e107, Nov. 2017.
- [91] C. Parmar, P. Grossmann, J. Bussink, *et al.*, “Machine learning methods for quantitative radiomic biomarkers,” en, *Sci. Rep.*, vol. 5, p. 13 087, Aug. 2015.
- [92] A. Zwanenburg, M. Vallières, M. A. Abdalah, *et al.*, “The image biomarker standardization initiative: Standardized quantitative radiomics for High-Throughput image-based phenotyping,” en, *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020.
- [93] R. Berenguer, M. D. R. Pastor-Juan, J. Canales-Vázquez, *et al.*, “Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters,” en, *Radiology*, vol. 288, no. 2, pp. 407–415, Aug. 2018.
- [94] D. Mackin, X. Fave, L. Zhang, *et al.*, “Measuring computed tomography scanner variability of radiomics features,” en, *Invest. Radiol.*, vol. 50, no. 11, pp. 757–765, Nov. 2015.
- [95] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical bayes methods,” en, *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007.
- [96] C. Lazar, S. Meganck, J. Taminiau, *et al.*, “Batch effect removal methods for microarray gene expression data integration: A survey,” en, *Brief. Bioinform.*, vol. 14, no. 4, pp. 469–490, Jul. 2013.
- [97] F. Orlhac, F. Frouin, C. Nioche, *et al.*, “Validation of a method to compensate multicenter effects affecting CT radiomics,” en, *Radiology*, vol. 291, no. 1, pp. 53–59, Apr. 2019.
- [98] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, *et al.*, “Radiomics in medical imaging-“how-to” guide and critical reflection,” en, *Insights Imaging*, vol. 11, no. 1, p. 91, Aug. 2020.
- [99] B. Kocak, B. Baessler, S. Bakas, *et al.*, “CheckList for EvaluAtion of radiomics research (CLEAR): A step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII,” en, *Insights Imaging*, vol. 14, no. 1, p. 75, May 2023.
- [100] H. Hotelling, *Analysis of a Complex of Statistical Variables Into Principal Components*, en. Warwick & York, 1933.

-
- [101] G. Ge and J. Zhang, “Feature selection methods and predictive models in CT lung cancer radiomics,” en, *J. Appl. Clin. Med. Phys.*, vol. 24, no. 1, e13869, Jan. 2023.
- [102] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” en, *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.
- [103] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
- [104] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” en, *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Mar. 2005.
- [105] A. Demircioğlu, “Benchmarking feature selection methods in radiomics,” en, *Invest. Radiol.*, vol. 57, no. 7, pp. 433–443, Jul. 2022.
- [106] D. R. Cox, “Regression models and life-tables,” en, *J. R. Stat. Soc.*, vol. 34, no. 2, pp. 187–202, Jan. 1972.
- [107] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [108] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [109] G. Yuan, Y. Song, Q. Li, *et al.*, “Development and validation of a contrast-enhanced CT-based radiomics nomogram for prediction of therapeutic efficacy of anti-PD-1 antibodies in advanced HCC patients,” en, *Front. Immunol.*, vol. 11, p. 613 946, 2020.
- [110] G. Ungan, A.-F. Lavandier, J. Rouanet, *et al.*, “Metastatic melanoma treated by immunotherapy: Discovering prognostic markers from radiomics analysis of pretreatment CT with feature selection and classification,” en, *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 10, pp. 1867–1877, Oct. 2022.
- [111] Z. Jiang, Y. Dong, L. Yang, *et al.*, “CT-Based hand-crafted radiomic signatures can predict PD-L1 expression levels in non-small cell lung cancer: A Two-Center study,” en, *J. Digit. Imaging*, vol. 34, no. 5, pp. 1073–1085, Oct. 2021.
- [112] C. Liu, J. Gong, H. Yu, *et al.*, “A CT-Based radiomics approach to predict nivolumab response in advanced Non-Small-Cell lung cancer,” en, *Front. Oncol.*, vol. 11, p. 544 339, Feb. 2021.
- [113] P. Afshar, A. Mohammadi, K. N. Plataniotis, *et al.*, “From Hand-Crafted to deep learning-based cancer radiomics: Challenges and opportunities,” Aug. 2018. arXiv: [1808.07954](https://arxiv.org/abs/1808.07954) [cs.CV].

- [114] P. M. Cheng, E. Montagnon, R. Yamashita, *et al.*, “Deep learning: An update for radiologists,” en, *Radiographics*, vol. 41, no. 5, pp. 1427–1445, 2021.
- [115] K. Lång, V. Josefsson, A.-M. Larsson, *et al.*, “Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study,” en, *Lancet Oncol.*, vol. 24, no. 8, pp. 936–944, Aug. 2023.
- [116] A. Sreekumar, K. R. Nair, S. Sudheer, *et al.*, “Malignant lung nodule detection using deep learning,” in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, Jul. 2020, pp. 0209–0212.
- [117] J. N. Kather, C. Hörner, C.-A. Weis, *et al.*, “CD163+ immune cell infiltrates and presence of CD54+ microvessels are prognostic markers for patients with embryonal rhabdomyosarcoma,” en, *Sci. Rep.*, vol. 9, no. 1, p. 9211, Jun. 2019.
- [118] J. N. Kather, M. Suarez-Carmona, P. Charoentong, *et al.*, “Topography of cancer-associated immune cells in human solid tumors,” en, *Elife*, vol. 7, Sep. 2018.
- [119] A. Echle, N. T. Rindtorff, T. J. Brinker, *et al.*, “Deep learning in cancer pathology: A new generation of clinical biomarkers,” en, *Br. J. Cancer*, vol. 124, no. 4, pp. 686–696, Feb. 2021.
- [120] O.-J. Skrede, S. De Raedt, A. Kleppe, *et al.*, “Deep learning for prediction of colorectal cancer outcome: A discovery and validation study,” en, *Lancet*, vol. 395, no. 10221, pp. 350–360, Feb. 2020.
- [121] D. Cifci, S. Foersch, and J. N. Kather, “Artificial intelligence to identify genetic alterations in conventional histopathology,” en, *J. Pathol.*, vol. 257, no. 4, pp. 430–444, Jul. 2022.
- [122] *Cancer Genomics: What Does It Mean for You? : the Cancer Genome Atlas*, en. National Cancer Institute, 2010.
- [123] Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, *et al.*, “The cancer genome atlas Pan-Cancer analysis project,” en, *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.
- [124] J. N. Kather, A. T. Pearson, N. Halama, *et al.*, “Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer,” en, *Nat. Med.*, vol. 25, no. 7, pp. 1054–1056, Jul. 2019.

-
- [125] O. L. Saldanha, P. Quirke, N. P. West, *et al.*, “Swarm learning for decentralized artificial intelligence in cancer histopathology,” en, *Nat. Med.*, vol. 28, no. 6, pp. 1232–1239, Jun. 2022.
- [126] R. Balestrieri, M. Ibrahim, V. Sobal, *et al.*, “A cookbook of Self-Supervised learning,” Apr. 2023. arXiv: [2304.12210 \[cs.LG\]](#).
- [127] X. Wang, Y. Du, S. Yang, *et al.*, “RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval,” en, *Med. Image Anal.*, vol. 83, p. 102645, Jan. 2023.
- [128] X. Wang, S. Yang, J. Zhang, *et al.*, “Transformer-based unsupervised contrastive learning for histopathological image classification,” en, *Med. Image Anal.*, vol. 81, p. 102559, Oct. 2022.
- [129] M. Ilse, J. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2127–2136.
- [130] S. J. Wagner, D. Reisenbüchler, N. P. West, *et al.*, “Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study,” en, *Cancer Cell*, vol. 41, no. 9, 1650–1661.e4, Sep. 2023.
- [131] R. Sun, E. J. Limkin, M. Vakalopoulou, *et al.*, “A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: An imaging biomarker, retrospective multicohort study,” *Lancet Oncol.*, vol. 19, no. 9, pp. 1180–1191, Sep. 2018.
- [132] S. Trebeschi, S. G. Drago, N. J. Birkbak, *et al.*, “Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers,” en, *Ann. Oncol.*, vol. 30, no. 6, pp. 998–1004, Jun. 2019.
- [133] M. Khorrami, P. Prasanna, A. Gupta, *et al.*, “Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in Non-Small cell lung cancer,” en, *Cancer Immunol Res*, vol. 8, no. 1, pp. 108–119, Jan. 2020.
- [134] Y. Himoto, H. Veeraraghavan, J. Zheng, *et al.*, “Computed Tomography-Derived radiomic metrics can identify responders to immunotherapy in ovarian cancer,” en, *JCO Precis Oncol*, vol. 3, Aug. 2019.
- [135] L. Dercle, B. Zhao, M. Gönen, *et al.*, “Early readout on overall survival of patients with melanoma treated with immunotherapy using a novel imaging analysis,” en, *JAMA Oncol*, vol. 8, no. 3, pp. 385–392, Mar. 2022.

BIBLIOGRAPHY

- [136] S. Trebeschi, Z. Bodalal, T. N. Boellaard, *et al.*, “Prognostic value of deep Learning-Mediated treatment monitoring in lung cancer patients receiving immunotherapy,” en, *Front. Oncol.*, vol. 11, p. 609 054, Mar. 2021.
- [137] S. Trebeschi, Z. Bodalal, N. van Dijk, *et al.*, “Development of a prognostic AI-Monitor for metastatic urothelial cancer patients receiving immunotherapy,” en, *Front. Oncol.*, vol. 11, p. 637 804, Apr. 2021.
- [138] L. M. Hondelink, M. Hüyük, P. E. Postmus, *et al.*, “Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer,” en, *Histopathology*, vol. 80, no. 4, pp. 635–647, Mar. 2022.
- [139] J. Liu, Q. Zheng, X. Mu, *et al.*, “Automated tumor proportion score analysis for PD-L1 (22c3) expression in lung squamous cell carcinoma,” en, *Sci. Rep.*, vol. 11, no. 1, p. 15 907, Aug. 2021.
- [140] A. Kapil, A. Meier, A. Zuraw, *et al.*, “Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies,” en, *Sci. Rep.*, vol. 8, no. 1, p. 17 343, Nov. 2018.
- [141] J. Wu, C. Liu, X. Liu, *et al.*, “Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer,” en, *Mod. Pathol.*, vol. 35, no. 3, pp. 403–411, Mar. 2022.
- [142] M. Shaban, S. A. Khurram, M. M. Fraz, *et al.*, “A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma,” en, *Sci. Rep.*, vol. 9, no. 1, p. 13 341, Sep. 2019.
- [143] Y. Schirris, M. Engelaer, A. Panteli, *et al.*, “WeakSTIL: Weak whole-slide image level stromal tumor infiltrating lymphocyte scores are all you need,” en, in *Medical Imaging 2022: Digital and Computational Pathology*, vol. 12039, SPIE, Apr. 2022, pp. 55–59.
- [144] R. Yamashita, J. Long, T. Longacre, *et al.*, “Deep learning model for the prediction of microsatellite instability in colorectal cancer: A diagnostic study,” en, *Lancet Oncol.*, vol. 22, no. 1, pp. 132–141, Jan. 2021.
- [145] J. N. Kather, L. R. Heij, H. I. Grabsch, *et al.*, “Pan-cancer image-based detection of clinically actionable genetic alterations,” en, *Nat Cancer*, vol. 1, no. 8, pp. 789–799, Aug. 2020.
- [146] M. S. Jain and T. F. Massoud, “Predicting tumour mutational burden from histopathological images using multiscale deep learning,” en, Jun. 2020.

-
- [147] S. Chen, J. Xiang, X. Wang, *et al.*, “Pan-cancer computational histopathology reveals tumor mutational burden status through weakly-supervised deep learning,” *arXiv e-prints*, arXiv-2204, 2022.
- [148] L. Sha, B. L. Osinski, I. Y. Ho, *et al.*, “Multi-Field-of-View deep learning model predicts nonsmall cell lung cancer programmed Death-Ligand 1 status from Whole-Slide hematoxylin and eosin images,” *en, J. Pathol. Inform.*, vol. 10, p. 24, Jul. 2019.
- [149] Q. Zeng, C. Klein, S. Caruso, *et al.*, “Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology,” *en, J. Hepatol.*, vol. 77, no. 1, pp. 116–127, Jul. 2022.
- [150] C. Shen, C. Schlager, D. Rajan, *et al.*, “Abstract 1922: Application of an interpretable graph neural network to predict gene expression signatures associated with tertiary lymphoid structures in histopathological images,” *en, Cancer Res.*, vol. 82, no. 12_Supplement, pp. 1922–1922, Jun. 2022.
- [151] J. Hu, C. Cui, W. Yang, *et al.*, “Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images,” *en, Transl. Oncol.*, vol. 14, no. 1, p. 100921, Jan. 2021.
- [152] C. Xie, C. Vanderbilt, C. Feng, *et al.*, “Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E,” *en, Jan. 2022.*
- [153] R. J. Chen, M. Y. Lu, D. F. K. Williamson, *et al.*, “Pan-cancer integrative histology-genomic analysis via multimodal deep learning,” *en, Cancer Cell*, vol. 40, no. 8, 865–878.e6, Aug. 2022.
- [154] K. M. Boehm, E. A. Aherne, L. Ellenson, *et al.*, “Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer,” *en, Nat Cancer*, vol. 3, no. 6, pp. 723–733, Jun. 2022.
- [155] R. S. Vanguri, J. Luo, A. T. Aukerman, *et al.*, “Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer,” *en, Nat Cancer*, Aug. 2022.
- [156] Z. Chen, Y. Du, J. Hu, *et al.*, “Multi-modal masked autoencoders for medical Vision-and-Language pre-training,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Springer Nature Switzerland, 2022, pp. 679–689.
- [157] K. M. Boehm, P. Khosravi, R. Vanguri, *et al.*, “Harnessing multimodal data integration to advance precision oncology,” *en, Nat. Rev. Cancer*, vol. 22, no. 2, pp. 114–126, Feb. 2022.

BIBLIOGRAPHY

- [158] P. Ohorodnyk, E. A. Eisenhauer, and C. M. Booth, “Clinical benefit in oncology trials: Is this a patient-centred or tumour-centred end-point?” en, *Eur. J. Cancer*, vol. 45, no. 13, pp. 2249–2252, Sep. 2009.
- [159] X. Li, P. S. Morgan, J. Ashburner, *et al.*, “The first step for neuroimaging data analysis: DICOM to NIfTI conversion,” en, *J. Neurosci. Methods*, vol. 264, pp. 47–56, May 2016.
- [160] G. Ulaner, *Fundamentals of Oncologic PET/CT E-Book*, en. Elsevier Health Sciences, Jun. 2018.
- [161] S. Ha, H. Choi, J. C. Paeng, *et al.*, “Radiomics in oncological PET/CT: A methodological overview,” *Nucl. Med. Mol. Imaging*, vol. 53, no. 1, pp. 14–29, Feb. 2019.
- [162] C. Parmar, J. D. Barry, A. Hosny, *et al.*, “Data analysis strategies in medical imaging,” en, *Clin. Cancer Res.*, vol. 24, no. 15, pp. 3492–3499, Aug. 2018.
- [163] J. Canny, “A computational approach to edge detection,” en, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Jun. 1986.
- [164] T. K. Koo and M. Y. Li, “A guideline of selecting and reporting intraclass correlation coefficients for reliability research,” en, *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016.
- [165] Y. Liu, M. Wu, Y. Zhang, *et al.*, “Imaging biomarkers to predict and evaluate the effectiveness of immunotherapy in advanced Non-Small-Cell lung cancer,” en, *Front. Oncol.*, vol. 11, p. 657615, Mar. 2021.
- [166] X. Chen, M. Zhou, Z. Wang, *et al.*, “Immunotherapy treatment outcome prediction in metastatic melanoma through an automated multi-objective delta-radiomics model,” en, *Comput. Biol. Med.*, vol. 138, p. 104916, Nov. 2021.
- [167] E. Chang, M. Z. Joel, H. Y. Chang, *et al.*, “Comparison of radiomic feature aggregation methods for patients with multiple tumors,” en, *Sci. Rep.*, vol. 11, no. 1, p. 9758, May 2021.
- [168] R. Sun, T. Henry, A. Laville, *et al.*, “Imaging approaches and radiomics: Toward a new era of ultraprecision radioimmunotherapy?” en, *J Immunother Cancer*, vol. 10, no. 7, Jul. 2022.
- [169] B. Varghese, S. Cen, H. Zahoor, *et al.*, “Feasibility of using CT radiomic signatures for predicting CD8-T cell infiltration and PD-L1 expression in renal cell carcinoma,” en, *Eur J Radiol Open*, vol. 9, p. 100440, Sep. 2022.

-
- [170] L. Dercle, M. Fronheiser, L. Lu, *et al.*, “Identification of Non–Small cell lung cancer sensitive to systemic cancer therapies using radiomics,” en, *Clin. Cancer Res.*, vol. 26, no. 9, pp. 2151–2162, May 2020.
- [171] R. Sun, N. Sundahl, M. Hecht, *et al.*, “Radiomics to predict outcomes and abscopal response of patients with cancer treated with immunotherapy combined with radiotherapy using a validated signature of CD8 cells,” en, *J Immunother Cancer*, vol. 8, no. 2, Nov. 2020.
- [172] Y.-M. Zheng, M.-G. Yuan, R.-Q. Zhou, *et al.*, “A computed tomography-based radiomics signature for predicting expression of programmed death ligand 1 in head and neck squamous cell carcinoma,” en, *Eur. Radiol.*, vol. 32, no. 8, pp. 5362–5370, Aug. 2022.
- [173] K. J. Park, J.-L. Lee, S.-K. Yoon, *et al.*, “Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma,” en, *Eur. Radiol.*, vol. 30, no. 10, pp. 5392–5403, Oct. 2020.
- [174] K. Jazieh, M. Khorrami, A. Saad, *et al.*, “Novel imaging biomarkers predict outcomes in stage III unresectable non-small cell lung cancer treated with chemoradiation and durvalumab,” en, *J Immunother Cancer*, vol. 10, no. 3, Mar. 2022.
- [175] L. Paré, T. Pascual, E. Seguí, *et al.*, “Association between PD1 mRNA and response to anti-PD1 monotherapy across multiple cancer types,” en, *Ann. Oncol.*, vol. 29, no. 10, pp. 2121–2128, Oct. 2018.
- [176] A. Singh, H. Horng, L. Roshkovan, *et al.*, “Development of a robust radiomic biomarker of progression-free survival in advanced non-small cell lung cancer patients treated with first-line immunotherapy,” en, *Sci. Rep.*, vol. 12, no. 1, p. 9993, Jun. 2022.
- [177] M. Nishino, F. Hong, B. Ricciuti, *et al.*, “Tumor response dynamics during First-Line pembrolizumab therapy in patients with advanced Non-Small-Cell lung cancer,” en, *JCO Precis Oncol*, vol. 5, Mar. 2021.
- [178] B. He, D. Dong, Y. She, *et al.*, “Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker,” en, *J Immunother Cancer*, vol. 8, no. 2, Jul. 2020.
- [179] C. Tang, B. Hobbs, A. Amer, *et al.*, “Development of an Immune-Pathology informed radiomics model for Non-Small cell lung cancer,” en, *Sci. Rep.*, vol. 8, no. 1, p. 1922, Jan. 2018.

BIBLIOGRAPHY

- [180] W. Huang, Y. Jiang, W. Xiong, *et al.*, “Noninvasive imaging of the tumor immune microenvironment correlates with response to immunotherapy in gastric cancer,” en, *Nat. Commun.*, vol. 13, no. 1, p. 5095, Aug. 2022.
- [181] J. R. Mains, F. Donskov, E. M. Pedersen, *et al.*, “Dynamic Contrast-Enhanced computed Tomography–Derived blood volume and blood flow correlate with patient outcome in metastatic renal cell carcinoma,” *Invest. Radiol.*, vol. 52, no. 2, p. 103, Feb. 2017.
- [182] M. Ligerio, O. S. M. El Nahhas, O. P. Palomares, *et al.*, “¹⁷⁶P enhancing immunotherapy response prediction via multimodal integration of radiology and pathology deep learning models,” *Ann. Oncol.*, vol. 34, S251–S252, 2023.
- [183] M. R. Tomaszewski and R. J. Gillies, “The biological meaning of radiomic features,” en, *Radiology*, vol. 299, no. 2, E256, May 2021.
- [184] S. Rutella, M. A. Cannarile, S. Gnjatic, *et al.*, “Society for immunotherapy of cancer clinical and biomarkers data sharing resource document: Volume i-conceptual challenges,” en, *J Immunother Cancer*, vol. 8, no. 2, Oct. 2020.
- [185] M. Moor, O. Banerjee, Z. S. H. Abad, *et al.*, “Foundation models for generalist medical artificial intelligence,” en, *Nature*, vol. 616, no. 7956, pp. 259–265, Apr. 2023.
- [186] S. Pai, D. Bontempi, V. Prudente, *et al.*, “Foundation models for quantitative biomarker discovery in cancer imaging,” en, *medRxiv*, Sep. 2023.
- [187] Z. Chen, A. H. Cano, A. Romanou, *et al.*, “MEDITRON-70B: Scaling medical pre-training for large language models,” Nov. 2023. arXiv: [2311.16079 \[cs.CL\]](#).
- [188] C. Li, C. Wong, S. Zhang, *et al.*, “LLaVA-Med: Training a large Language-and-Vision assistant for biomedicine in one day,” Jun. 2023. arXiv: [2306.00890 \[cs.CV\]](#).

APPENDIX A

A whirl of radiomics-based biomarkers in cancer immunotherapy, why is large scale validation still lacking?

Reference:

Ligero M., Gielen B., Navarro V, et al.; A whirl of radiomics-based biomarkers in cancer
immunotherapy, why is large scale validation still lacking?
NPJ Precision Oncology (Accepted).

A whirl of radiomics-based biomarkers in cancer immunotherapy, why is large scale validation still lacking?

Marta Ligeró* (1), Bente Gielen* (1), Victor Navarro (2), Pablo Cresta (2,3), Olivia Prior (1), Rodrigo Dienstmann (2), Paolo Nuciforo (4), Stefano Trebeschi (5, 6), Regina Beets-Tan (5,6,7), Evis Sala (8,9), Elena Garraïda (3), Raquel Perez-Lopez+ (1)

*equal contribution

(1) Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain

(2) Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain

(3) Department of Medical Oncology, Vall d'Hebron University Hospital and Institute of Oncology (VHIO), Barcelona, Spain

(4) Molecular Oncology Group, Vall d'Hebron University Hospital and Institute of Oncology (VHIO), Barcelona, Spain

(5) Department of Radiology, Netherlands Cancer Institute, Amsterdam, The Netherlands

(6) GROW School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands

(7) Faculty of Health Sciences, University of Southern Denmark, Denmark

(8) Dipartimento Diagnostica per Immagini, Radioterapia Oncologica ed Ematologia, Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

(9) Dipartimento di Scienze Radiologiche ed Ematologiche, Università Cattolica del Sacro Cuore, Rome, Italy

+ Corresponding author

Dr. Raquel Perez-Lopez

Team Leader Radiomics Group

Vall d'Hebron Institute of Oncology (VHIO)

Cellex Center, Natzaret 115-117. Barcelona 08035, Spain.

+34 932543450 ext. 8682

rperez@vhio.net

twitter:@RaqPerezLopez

Abstract

The search for understanding immunotherapy response has sparked interest from diverse areas of oncology, with artificial intelligence (AI) and radiomics emerging as promising tools, capable of gathering large amounts of information to identify suitable patients for treatment. The application of AI in radiology has grown, driven by the hypothesis that radiology images capture tumor phenotypes and thus could provide valuable insights into immunotherapy response likelihood. However, despite the rapid growth of studies in the field, none of the developed algorithms have yet made it to clinical implementation, mainly due to the lack of standardized methods, hampering study comparisons and reproducibility across different datasets. In this review, we performed a comprehensive assessment of published data to identify sources of variability in radiomics study design that hinder the comparison of the different model performance and therefore, clinical implementation. Subsequently, we conducted a use-case meta-analysis using homogenous studies to assess the overall performance of radiomics in estimating the predictor biomarker of immunotherapy response, programmed death-ligand 1 (PD-L1) expression. Our findings indicate that, despite numerous attempts to predict immunotherapy response, only a limited number of studies share comparable methodologies and report sufficient data about cohorts and methods to be suitable for meta-analysis. Nevertheless, although only a few studies meet these criteria, their promising results underscore the importance of ongoing standardization and benchmarking efforts. This review highlights the importance of uniformity in study design and reporting. Such standardization is crucial to enable meaningful comparisons and demonstrate the validity of biomarkers across diverse populations, facilitating the implementation of these biomarkers into the immunotherapy patient selection process.

Introduction

Cancer immunotherapy, particularly immune checkpoint inhibitors (ICI), has emerged as the gold standard for treating various cancers, including lung, renal, and melanoma [1–4]. The remarkable success achieved with ICI has ignited optimism for its potential application in treating numerous other types of cancer. However, the variability in patient responses makes it necessary to identify biomarkers capable of predicting individual responses to ICI. This crucial step is instrumental in enhancing patient stratification, maximizing treatment efficacy, detecting treatment resistance and thus minimizing potential harm for those who may not benefit. Various tissue-based predictive biomarkers have been proposed, such as microsatellite instability (MSI) [5,6], tumor mutational burden (TMB) [7], programmed death-ligand 1 (PD-L1) expression [8], and tumor-infiltrating lymphocyte (TIL) count [9]. However, these biomarkers often require invasive procedures to obtain tumor tissue for analysis, and their accuracy in identifying suitable candidates for immunotherapy remains suboptimal [10]. Radiomics analysis, in combination with machine learning methods, efficiently extracts meaningful information from medical images, enabling three-dimensional evaluation of tumors throughout the entire body, and repeated assessments over the course of cancer treatment [11]. In particular, extracting radiomics features from standard-of-care CT images, a widely used imaging technique for cancer staging and follow-up, offers a valuable tool with potential for developing predictive biomarkers in the context of immunotherapy [12–15]. This is especially pertinent in cancer immunotherapy, where treatment may occur after the initial diagnosis, in pretreated patients with evolving tumors and non-reachable lesions [16,17]. The non-invasive nature of radiomics applications thus becomes highly valuable.

In fact, the emergence of encouraging radiomics signatures for predicting response to immunotherapy has sparked a surge in research endeavors within this domain. Nevertheless, the absence of standardized protocols and benchmarking studies of biological validation of such signatures poses a significant challenge for the application of these signatures in clinical practice. Despite numerous radiomics studies predicting response across various tumor types, inconsistencies persist in data selection, model construction, and outcome definition. To assess the reliability of predictive radiomics studies, standardization research criteria such as the Radiomics Quality Score (RQS)

[11] and the CLEAR checklist [18] have been introduced [19]. However, some studies have reported low RQS in most published radiomics studies, indicating poor reporting practices and limited reproducibility [20]. Furthermore, efforts are emerging to develop PRISMA-AI guidelines [21] that will define standardized frameworks, comprehensive method descriptions, and data sharing practices in radiomics-based studies, that will allow study comparison, validation, and meta-analysis efforts in this domain.

In this review, we provide an overview of the current state of radiomics-based biomarkers to guide the use of immunotherapy through a comprehensive examination, encompassing the potential biases and variations in the currently developed radiomics pipelines that challenge the development of comparison studies through meta-analysis. Additionally, we present a short case study featuring a meta-analysis of studies predicting PD-L1 status from CT imaging, comparing radiomics machine learning (ML) and deep learning (DL) models. By examining the existing literature and conducting a meta-analysis, we aim to offer valuable insights and perspectives on the efficacy and reliability of radiomics as immunotherapy biomarkers.

Uncovering potential sources of variability in radiomics study design

We conducted a systematic review encompassing all studies utilizing radiomics or deep learning techniques in CT imaging for predicting either direct response to immunotherapy or any biomarker surrogate of response. Supplementary material 1 provides a detailed description of our review methodology. Our findings highlight the significant diversity in study design among publications aiming to create similar predictive models (Figure 1). This variability in methodology presents a challenge when attempting to compare the performance of these models through meta-analysis. In this section, we aim to summarize all these studies and the differences among them.

Cohort setting

The characteristics that define the tumor phenotype and make it more responsive to immunotherapy can encompass tumor-specific aspects or those that can be expressed and captured across multiple tumor types. Consequently, researchers have pursued two approaches in the development and validation of radiomics-based biomarkers. One approach being tumor specific, aiming to create validated biomarkers within each type. While the other approach addresses this challenge by incorporating

multiple tumor types and considering the location of metastatic disease when feeding the models.

To date, the majority of radiomics studies on immunotherapy response prediction have focused on non-small cell lung cancer (NSCLC), benefiting from the availability of larger datasets and a higher degree of treatment responses in this tumor type. In fact, most of the studies exploring radiomics to predict direct response or surrogates of response (i.e., biological and molecular markers used as predictors of a patient's response such as PD-L1) to immunotherapy have been done in NSCLC populations. Other tumor types including melanoma, gastric, head and neck, bladder and kidney cancers have been investigated to a lesser extent and only few studies have developed predictive radiomics models in pan-cancer settings [12,14]. Despite the increasing number of lung cancer and melanoma patients receiving immunotherapy as part of standard care, it is noteworthy that only around 30% of the articles included cohorts larger than 200 patients, and merely 22% reported the utilization of external validation cohorts (Supplementary Table 1).

The development of tumor type-specific radiomics signatures allows finding radiomics features unique to that population, however it reduces the generalization of the methods to other tumor types that are less common or rarely treated with immunotherapy. On the other hand, pan-cancer approaches require the use of larger cohorts for the model to comprehend the inherent heterogeneity of the population, thereby reducing the bias towards the response probability of each tumor type.

Outcome evaluation

Studies focusing on predictive radiomics signatures and immunotherapy can be categorized into two types: those aiming to directly predict clinical outcome and those focused on predicting known surrogate biomarkers. However, the lack of standardization regarding outcome definition poses a significant challenge, making it difficult to compare and assess the predictive capabilities of the resulting radiomics signatures.

One major challenge is the wide range of clinical endpoints used to assess treatment response. The most relevant measure for evaluating the benefit of immunotherapy

treatment in patients is overall survival (OS). While certain radiomics studies have considered OS as the clinical endpoint [13,15,22–33], most studies rely on tumor size changes by the Response Evaluation Criteria in Solid Tumors version 1.1 (RECIST 1.1) [16]. From the RECIST assessment, multiple measurements can be computed and used as endpoints, including progression-free survival (PFS) [29–32,34–38], disease control (which gathers complete response (CR), partial response (PR), and stable disease (SD)) [14,22,28,34,35,39–46] or objective response rate (ORR) [23,47–50]. However, it is important to note that these response evaluations are considered surrogate endpoints for OS, and their reliability is hindered by their inherent subjectivity and variability, challenging the development of reproducible models [51,52]. Furthermore, the wide range of response evaluation criteria derived from RECIST [53,54] (e.g., PFS, ORR, disease control) also limits the direct comparison of radiomics signatures across studies.

Similarly, when predicting molecular surrogate biomarkers (such as PD-L1 expression), many studies tend to discretize the target variable and transform it into a classification problem. However, these biomarker cutoffs are subjected to the primary tumor biology or the type of treatment. Therefore, the lack of standardized cutoff values further complicates the comparison of radiomics signatures for predicting surrogate biomarkers in immunotherapy. In addition to the heterogeneity in endpoint definitions, it is important to consider that the performance of radiomics signatures predicting surrogate biomarkers will be inherently limited by the predictive capacity of the surrogate biomarker itself. This implies that the effectiveness of the radiomics signatures in predicting treatment response will be constrained by the predictive capabilities of the surrogate biomarker being used.

Study design regarding number of lesions, region of interest and time-points

Another relevant point in the study design for immunotherapy radiomics signatures is the selection of target lesions for analysis. Many radiomics studies rely on delineating and extracting features from a single selected tumor (~63% of the studies found in the review), often the primary or largest lesion, arguing that the single chosen lesion can represent the whole disease. However, in patients with metastases at multiple sites, heterogeneous immunophenotypes can drive different immune responses [55,56]. Therefore, analyzing only one lesion per patient may not fully capture the tumor

heterogeneity and limit the predictive capacity of the model. To partially overcome this limitation and capture the full spectrum of patient heterogeneity, feature aggregation methods such as average, volume-weighted average, or attention-based multiple instance learning (MIL) are commonly used [14,41,49,57]. Additionally, the analysis of inter- and intra-lesion heterogeneity through radiomics studies to capture the whole metastatic disease has also been considered as a potential indicator of immunotherapy response [58].

Moreover, with the aim of providing the model with all the potential relevant data and knowing the effect of surrounding tumor microenvironment for immunotherapy response, certain studies have also explored the value of incorporating peritumoral area information into predictive models for predicting response to immunotherapy [22,50]. Nevertheless, the models obtained more relevant information from the intratumoral features. Some studies have also shown that intratumoral 3D radiomics features provide more informative insights compared to using only 2D radiomics features [28].

Finally, regarding the imaging time-points, the majority of studies in radiomics research have focused on the development of predictive biomarkers using baseline scans, which refer to the scans obtained just before initiating treatment. This approach facilitates improved patient selection for treatment decision-making. However, some studies have demonstrated enhanced outcomes by analyzing changes in the radiomics tumor phenotype between baseline and early follow-up time points, commonly known as early readouts or delta radiomics signatures [13,15,22,29,43,47,59]. Such approaches enable the capture of response or progression patterns that may go unnoticed by radiologists, thereby potentially preventing patients to stay longer under ineffective treatment. Some of these studies have shown that tracking these changes in CT scans provides better prognostic value compared to the current standard of care, RECIST [13,60]. It is important to note, however, that these early readouts do not represent true predictive biomarkers *per se*, but rather serve as indicators of early response, and should be thought of as alternative response criteria themselves, rather than predictive biomarkers. This is because at the time these early readouts are assessed, treatment decisions have already been made, and the patient is already receiving immunotherapy.

Radiomics feature selection and model implementation

Fifty percent of the pipelines implemented for hand-crafted radiomics analysis correspond to Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection (implemented in 40% of the studies), followed by a logistic regression for classification (implemented in 25% of the studies). Multiple studies have highlighted the benefits of utilizing LASSO as the feature selection method due to its efficacy in high-dimensional data regression, thereby mitigating the risk of overfitting [30,49]. In terms of classification method, several studies have explored the performance of different classification algorithms for predicting response (such as support vector machine (SVM), Random Forest (RF), decision tree and k-nearest neighbor). All of them showed that logistic regression had similar or slightly better performance than other more complex classifiers [28,38,50,61].

Only a few studies have used more advanced DL methods to predict response to immunotherapy [13,32,45,46]. These methods are data-hungry and need large cohorts of patients, as well as reliable and objective annotations, to achieve good performance. However, gathering this amount of data regarding immunotherapy treatment response is still challenging. For that reason, most of the CT-based DL models currently developed are focused on predicting surrogates of response such as PD-L1 status [32,62–64].

Case-study: meta-analysis for predicting PD-L1 status from CT imaging comparing DL vs radiomics with classical ML

In order to get a better understanding of the overall performance of the radiomics signatures as predicting biomarkers for immunotherapy, we conducted a meta-analysis of all the studies that implemented CT-based radiomics with classical ML or DL to predict PD-L1 status. Detailed description of the methodology is reported in Supplementary 1 (Methods of the systematic review). Figure 2 shows a flow chart illustrating the systematic review conducted in PubMed, outlining the predefined inclusion and exclusion criteria.

We identified a total of 56 articles developing CT-based predictive signatures in patients treated with immunotherapy; 34 for predicting direct response and 22

predicting surrogate molecular biomarkers (Supplementary 2. Detailed results systematic review). In Supplementary Table 1, all included papers are listed. We reviewed the CLEAR guidelines for all these studies (Supplementary Table 2). However, we could not include harmonized image preprocessing techniques or feature selection methods. Accounting for the previously described variability in the methods of radiomics signatures and with the aim of investigating the most standardized models, we found seven comparable studies to perform the meta-analysis. All of them predicted PD-L1 expression assessed as tumor proportion score (TPS) $\geq 1\%$, using the area under the curve (AUC) as the evaluation metric and implementing either logistic regression (n=4) [61,64–66] or DL (n=3) [62,64,67] as the predictive model. External validation performance was also explored in three studies applying logistic regression methods and one DL modeling.

The included papers showed varying performance in predicting PD-L1 expression, with AUROCs ranging from 0.76 to 0.96 in both logistic regression and DL methods. In the internal validation, the logistic regression models showed a pooled AUC of 0.86 (95%CI 0.77-0.94, $i^2= 94\%$) while the DL method exhibited a pooled AUROC of 0.86 (95%CI 0.79-0.92, $i^2= 89\%$), using random effects model (Figure 3). Interestingly, our findings revealed that the performance across different studies for logistic regression remained comparable in the external validation set, yielding an estimated AUROC of 0.80 (95%CI 0.78-0.82, $i^2= 0\%$) (Figure 4). These findings indicate low heterogeneity between studies in the external validation performance in contrast to the higher heterogeneity in the internal set. There was not enough data from DL studies to evaluate the heterogeneity in the external set. Notably, studies utilizing logistic regression and DL methods demonstrated similar results in the internal set, with a combined estimated AUROC of 0.86 (95% CI 0.80-0.91), despite DL models having access to a larger dataset compared to logistic regression studies.

Discussion

The application of artificial intelligence (AI) to improve patient stratification towards better treatment selection is of growing interest in both radiology and oncology fields.

Numerous studies have focused on uncovering radiological features of tumors that could predict response patterns to immunotherapy. However, the lack of standardized and homogeneous frameworks employed in these studies, as well as scarce data sharing, present challenges when comparing results and validating radiomics models, ultimately, hindering their integration into clinical practice. In this review, we aimed to shed light on the factors that contribute to the variability in radiomics studies, rendering the available models incomparable. Additionally, we conducted a comprehensive literature review specifically targeting studies that investigated CT-based radiomics signatures for predicting response to immune checkpoint inhibitors (ICI) and surrogate biomarkers of response to ICI.

Remarkably, our findings revealed that despite the abundance of studies predicting direct response to immunotherapy, only a limited number of these studies employed similar methods, making them unsuitable for meta-analysis. This significant variability in methodology poses challenges in terms of study comparisons and reproducibility across different datasets. Similar challenges have arisen in the development of other potential predictive biomarkers based on biological samples such as PD-L1 expression or TMB, highlighted in debates surrounding the heterogeneous distribution of these markers in tumor samples, variations in staining techniques, and establishment of appropriate thresholds, among other issues [68–70]. Addressing these challenges requires benchmarking studies that facilitate the comparison of established methods with novel techniques across diverse cohorts, thereby promoting advancement and standardization in the field.

Nonetheless, some studies investigating the development of radiomics signatures for predicting programmed death-ligand 1 (PD-L1) expression in tumors met the necessary criteria for meaningful pooling and meta-analysis. Consequently, our meta-analysis exclusively focused on studies examining CT-based radiomics for predicting PD-L1 status. This case study highlighted the promising performance of the reported models in predicting PD-L1 expression, with area under the receiver operating characteristic curve (AUROC) values ranging from 0.7 to 0.9. While these models have demonstrated positive outcomes and exhibited limited heterogeneity in accuracy during external validation, questions persist regarding the lack of widespread adoption in clinical practice. One potential contributing factor could be the absence of a reported

correlation between PD-L1 prediction and treatment response. Furthermore, it is important to acknowledge the potential influence of publication bias, which may result in a prevalence of positive results, potentially overshadowing scientifically crucial findings from studies that may not achieve high accuracy despite employing sound methodologies.

Developing multi-center studies is essential to demonstrate the applicability of these methods across large and heterogeneous datasets, ensuring reliability and fairness by encompassing diverse populations and machines from various institutions. Concerns regarding data privacy and patient data monetization have slowed down the development of large-scale multi-center models. Nevertheless, efforts have been made in this field to provide more secure methods of data sharing and decentralized model training, such as federated learning [71,72], where models can be trained on multi-institutional data without leaving the respective institutions, thus safeguarding data privacy. Moreover, some studies have highlighted potential improvements of predictive models through multimodal approaches that combine radiomics with histopathology or genomics [73,74]. Still, this requires representative heterogeneous data ideally from multiple centers including all sources of information, which has been a notable limitation thus far. Finally, integration of radiomics-based biomarkers into clinical practice hinges on the critical aspects of explainability and trustability, ensuring that healthcare professionals can comprehend and rely on these complex data-driven insights to make informed patient care decisions.

Moreover, the path to integrating radiomics into clinical practice, even when all the previous limitations are considered, still relies on biological translation of the predictive models. Certain studies have made substantial progress in this direction by correlating radiomics predictions with biological and molecular markers like PD-L1 [30], cellular pathways [39] or cytotoxic immunophenotype [14]. Other studies have focused on developing models that aim to predict directly the molecular properties of the tumor from surgical resections or biopsies [75]. However, ongoing investigation in this direction is needed to enhance the reliability and applicability of these models for seamless integration into routine clinical practice.

In conclusion, the journey towards establishing radiomics-based biomarkers is challenging, requiring technical development of imaging assays and computational methods, validation encompassing sensitivity, specificity, and reproducibility evaluations, biological validation, as well as proving clinical relevance ideally through embedding them in prospective clinical trials. Despite the considerable interest and expectations from the scientific community, as well as the abundance of papers exploring imaging phenotypes derived from radiomics as potential biomarkers of response to immunotherapy, these tools have yet to be implemented in clinical practice. To make a substantial impact on clinical trials and medical practice, larger prospective studies with appropriate external validation datasets, focusing on the clinical applicability of these signatures, are crucial.

Fortunately, changes are underway in the field that should facilitate the exploration of these novel biomarkers and their potential applicability in the clinic. The imaging scientific community, through collaborative efforts and consortia supported by the EU commissioner, is working to bridge the gap between research and real-world application. Among the most significant initiatives is the EUCAIM project, which is dedicated to establish an infrastructure for over 60 million cancer images from over 100,000 cancer patients with the goal to develop and benchmark trustworthy AI tools. Together, we strive to pave the way for the true integration of radiomics-based biomarkers into clinical decision-making, ultimately improving the care of cancer patients.

Conflicts of Interest

ES, has received speakers fees from GE healthcare and is a co-founder and shareholder of Lucida Medical Ltd. EG declares research funding from Novartis, Roche, Thermo Fisher, AstraZeneca, Taiho, BeiGene, Janssen. EG also reports consultant or advisor role for Roche, Ellipses Pharma, Boehringer Ingelheim, Janssen Global Services, Seattle Genetics, Thermo Fisher, MabDiscovery, Anaveon, F-Star Therapeutics, Hengrui, Sanofi, Incyte, Medscape and speakers bureau from Merck Sharp & Dohme, Roche, Thermo Fisher, Lilly, Novartis, SeaGen. RD declares advisory role for Roche, Foundation Medicine, received a speaker's fee from Roche, Ipsen, Amgen, Servier, Sanofi, Libbs, Merck Sharp & Dohme, Lilly, AstraZeneca, Janssen, Takeda, Bristol Myers Squibb, GlaxoSmithKline, Gilead and research grants from Merck, Novartis, Daiichi-Sankyo, GlaxoSmithKline and AstraZeneca. PN declares advisory or consultant role for MSD ONCOLOGY, BAYER and speaker's fee from Novartis. No other competing interests are disclosed by any author.

Acknowledgements

This research has been funded by the Comprehensive Program of Cancer Immunotherapy & Immunology II (CAIMI-II) supported by the BBVA Foundation (grant 53/2021). RPL is supported by LaCaixa Foundation, a CRIS Foundation Talent Award (TALENT19-05), the FERO Foundation, the Instituto de Salud Carlos III-Investigación en Salud (PI18/01395 and PI21/01019) and the Prostate Cancer Foundation (18YOUN19). ML is supported by the PERIS PIF-Salut Grant. OP is supported by a La Caixa INPhINIT Fellowship.

Authors' contributions

ML and RPL designed the study. ML and BG contributed to data collection and assembly. ML, BG, OP interpreted and analyzed the data. VN and PC performed statistical revision. All authors wrote and reviewed the report and approved the final version for submission.

References

1. Long GV, Weber JS, Larkin J, Atkinson V, Grob J-J, Schadendorf D, et al. Nivolumab for Patients With Advanced Melanoma Treated Beyond Progression: Analysis of 2 Phase 3 Clinical Trials. *JAMA Oncol.* 2017;3: 1511–1519.
2. Postow MA, Chesney J, Pavlick AC, Robert C, Grossmann K, McDermott D, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *N Engl J Med.* 2015;372: 2006–2017.
3. Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, et al. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *N Engl J Med.* 2015;373: 1803–1813.
4. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med.* 2015;373: 123–135.
5. Le DT, Kim TW, Van Cutsem E, Geva R, Jäger D, Hara H, et al. Phase II Open-Label Study of Pembrolizumab in Treatment-Refractory, Microsatellite Instability-High/Mismatch Repair-Deficient Metastatic Colorectal Cancer: KEYNOTE-164. *J Clin Oncol.* 2020;38: 11–19.
6. Marabelle A, Le DT, Ascierto PA, Di Giacomo AM, De Jesus-Acosta A, Delord J-P, et al. Efficacy of Pembrolizumab in Patients With Noncolorectal High Microsatellite Instability/Mismatch Repair-Deficient Cancer: Results From the Phase II KEYNOTE-158 Study. *J Clin Oncol.* 2020;38: 1–10.
7. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann Oncol.* 2019;30: 44–56.
8. Daud AI, Wolchok JD, Robert C, Hwu W-J, Weber JS, Ribas A, et al. Programmed Death-Ligand 1 Expression and Response to the Anti-Programmed Death 1 Antibody Pembrolizumab in Melanoma. *J Clin Oncol.* 2016;34: 4102–4109.
9. Lee JS, Ruppin E. Multiomics Prediction of Response Rates to Therapies to Inhibit Programmed Cell Death 1 and Programmed Cell Death 1 Ligand 1. *JAMA Oncol.* 2019;5: 1614–1618.
10. Pilard C, Ancion M, Delvenne P, Jerusalem G, Hubert P, Herfs M. Cancer immunotherapy: it's time to better predict patients' response. *Br J Cancer.* 2021;125: 927–938.
11. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14: 749–762.
12. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-

PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol.* 2018;19: 1180–1191.

13. Trebeschi S, Bodalal Z, Boellaard TN, Tareco Bucho TM, Drago SG, Kurilova I, et al. Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy. *Front Oncol.* 2021;11: 609054.
14. Liger M, Garcia-Ruiz A, Viaplana C, Villacampa G, Raciti MV, Landa J, et al. A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors. *Radiology.* 2021;299: 109–119.
15. Dercle L, Zhao B, Gönen M, Moskowitz CS, Firas A, Beylertgil V, et al. Early Readout on Overall Survival of Patients With Melanoma Treated With Immunotherapy Using a Novel Imaging Analysis. *JAMA Oncol.* 2022;8: 385–392.
16. Jiménez-Sánchez A, Cybulska P, Mager KL, Koplev S, Cast O, Couturier D-L, et al. Unraveling tumor-immune heterogeneity in advanced ovarian cancer uncovers immunogenic effect of chemotherapy. *Nat Genet.* 2020;52: 582–593.
17. Nguyen PHD, Ma S, Phua CZJ, Kaya NA, Lai HLH, Lim CJ, et al. Intratumoural immune heterogeneity as a hallmark of tumour evolution and progression in hepatocellular carcinoma. *Nat Commun.* 2021;12: 227.
18. Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging.* 2023;14: 75.
19. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—“how-to” guide and critical reflection. *Insights Imaging.* 2020;11: 1–16.
20. Ramlee S, Hulse D, Bernatowicz K, Pérez-López R, Sala E, Aloj L. Radiomic Signatures Associated with CD8+ Tumour-Infiltrating Lymphocytes: A Systematic Review and Quality Assessment Study. *Cancers* . 2022;14. doi:10.3390/cancers14153656
21. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nat Med.* 2023;29: 14–15.
22. Khorrami M, Prasanna P, Gupta A, Patil P, Velu PD, Thawani R, et al. Changes in CT Radiomic Features Associated with Lymphocyte Distribution Predict Overall Survival and Response to Immunotherapy in Non-Small Cell Lung Cancer. *Cancer Immunol Res.* 2020;8: 108–119.
23. Peisen F, Hänsch A, Hering A, Brendlin AS, Afat S, Nikolaou K, et al. Combination of Whole-Body Baseline CT Radiomics and Clinical Parameters to

Predict Response and Survival in a Stage-IV Melanoma Cohort Undergoing Immunotherapy. *Cancers* . 2022;14. doi:10.3390/cancers14122992

24. Tunali I, Tan Y, Gray JE, Katsoulakis E, Eschrich SA, Saller J, et al. Hypoxia-Related Radiomics and Immunotherapy Response: A Multicohort Study of Non-Small Cell Lung Cancer. *JNCI Cancer Spectr*. 2021;5. doi:10.1093/jncics/pkab048
25. Schraag A, Klumpp B, Afat S, Gatidis S, Nikolaou K, Eigentler TK, et al. Baseline clinical and imaging predictors of treatment response and overall survival of patients with metastatic melanoma undergoing immunotherapy. *Eur J Radiol*. 2019;121: 108688.
26. Corino VDA, Bologna M, Calareso G, Licitra L, Ghi M, Rinaldi G, et al. A CT-Based Radiomic Signature Can Be Prognostic for 10-Months Overall Survival in Metastatic Tumors Treated with Nivolumab: An Exploratory Study. *Diagnostics (Basel)*. 2021;11. doi:10.3390/diagnostics11060979
27. Zerunian M, Caruso D, Zucchelli A, Polici M, Capalbo C, Filetti M, et al. CT based radiomic approach on first line pembrolizumab in lung cancer. *Sci Rep*. 2021;11: 6633.
28. Ungan G, Lavandier A-F, Rouanet J, Hordonneau C, Chauveau B, Pereira B, et al. Metastatic melanoma treated by immunotherapy: discovering prognostic markers from radiomics analysis of pretreatment CT with feature selection and classification. *Int J Comput Assist Radiol Surg*. 2022;17: 1867–1877.
29. Guerrisi A, Russillo M, Loi E, Ganeshan B, Ungania S, Desiderio F, et al. Exploring CT Texture Parameters as Predictive and Response Imaging Biomarkers of Survival in Patients With Metastatic Melanoma Treated With PD-1 Inhibitor Nivolumab: A Pilot Study Using a Delta-Radiomics Approach. *Front Oncol*. 2021;11: 704607.
30. Jazieh K, Khorrami M, Saad A, Gad M, Gupta A, Patil P, et al. Novel imaging biomarkers predict outcomes in stage III unresectable non-small cell lung cancer treated with chemoradiation and durvalumab. *J Immunother Cancer*. 2022;10. doi:10.1136/jitc-2021-003778
31. Nardone V, Tini P, Pastina P, Botta C, Reginelli A, Carbone SF, et al. Radiomics predicts survival of patients with advanced non-small cell lung cancer undergoing PD-1 blockade using Nivolumab. *Oncol Lett*. 2020;19: 1559–1566.
32. He B-X, Zhong Y-F, Zhu Y-B, Deng J-J, Fang M-J, She Y-L, et al. Deep learning for predicting immunotherapeutic efficacy in advanced non-small cell lung cancer patients: a retrospective study combining progression-free survival risk and overall survival risk. *Transl Lung Cancer Res*. 2022;11: 670–685.
33. Mazzaschi G, Milanese G, Pagano P, Madeddu D, Gnetti L, Trentini F, et al. Integrated CT imaging and tissue immune features disclose a radio-immune signature with high prognostic impact on surgically resected NSCLC. *Lung Cancer*. 2020;144: 30–39.

34. Yang Y, Yang J, Shen L, Chen J, Xia L, Ni B, et al. A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer. *Am J Transl Res.* 2021;13: 743–756.
35. Yang B, Zhou L, Zhong J, Lv T, Li A, Ma L, et al. Combination of computed tomography imaging-based radiomics and clinicopathological characteristics for predicting the clinical benefits of immune checkpoint inhibitors in lung cancer. *Respir Res.* 2021;22: 189.
36. Dercle L, Fronheiser M, Lu L, Du S, Hayes W, Leung DK, et al. Identification of Non–Small Cell Lung Cancer Sensitive to Systemic Cancer Therapies Using Radiomics. *Clin Cancer Res.* 2020;26: 2151–2162.
37. Ladwa R, Roberts KE, O’Leary C, Maggacis N, O’Byrne KJ, Miles K. Computed tomography texture analysis of response to second-line nivolumab in metastatic non-small cell lung cancer. *Lung Cancer Manag.* 2020;9: LMT38.
38. Liu C, Gong J, Yu H, Liu Q, Wang S, Wang J. A CT-Based Radiomics Approach to Predict Nivolumab Response in Advanced Non-Small-Cell Lung Cancer. *Front Oncol.* 2021;11: 544339.
39. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Delli Pizzi A, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol.* 2019;30: 998–1004.
40. Liu Y, Wu M, Zhang Y, Luo Y, He S, Wang Y, et al. Imaging Biomarkers to Predict and Evaluate the Effectiveness of Immunotherapy in Advanced Non-Small-Cell Lung Cancer. *Front Oncol.* 2021;11: 657615.
41. Wu M, Zhang Y, Zhang J, Zhang Y, Wang Y, Chen F, et al. A Combined-Radiomics Approach of CT Images to Predict Response to Anti-PD-1 Immunotherapy in NSCLC: A Retrospective Multicenter Study. *Front Oncol.* 2021;11: 688679.
42. Ji Z, Cui Y, Peng Z, Gong J, Zhu H-T, Zhang X, et al. Use of Radiomics to Predict Response to Immunotherapy of Malignant Tumors of the Digestive System. *Med Sci Monit.* 2020;26: e924671.
43. Wang Z-L, Mao L-L, Zhou Z-G, Si L, Zhu H-T, Chen X, et al. Pilot Study of CT-Based Radiomics Model for Early Evaluation of Response to Immunotherapy in Patients With Metastatic Melanoma. *Front Oncol.* 2020;10: 1524.
44. Malone ER, Sim H-W, Stundzia A, Pierre S, Metser U, O’Malley M, et al. Predictive radiomics signature for treatment response to nivolumab in patients with advanced renal cell carcinoma. *Can Urol Assoc J.* 2022;16: E94–E101.
45. Ren Q, Xiong F, Zhu P, Chang X, Wang G, He N, et al. Assessing the robustness of radiomics/deep learning approach in the identification of efficacy of anti–PD-1 treatment in advanced or metastatic non-small cell lung carcinoma patients. *Front Oncol.* 2022;12. doi:10.3389/fonc.2022.952749

46. Rundo F, Bersanelli M, Urzia V, Friedlaender A, Cantale O, Calcara G, et al. Three-Dimensional Deep Noninvasive Radiomics for the Prediction of Disease Control in Patients With Metastatic Urothelial Carcinoma treated With Immunotherapy. *Clin Genitourin Cancer*. 2021;19: 396–404.
47. Gong J, Bao X, Wang T, Liu J, Peng W, Shi J, et al. A short-term follow-up CT based radiomics approach to predict response to immunotherapy in advanced non-small-cell lung cancer. *Oncoimmunology*. 2022;11: 2028962.
48. Liang Z, Huang A, Wang L, Bi J, Kuang B, Xiao Y, et al. A radiomics model predicts the response of patients with advanced gastric cancer to PD-1 inhibitor treatment. *Aging* . 2022;14: 907–922.
49. Park KJ, Lee J-L, Yoon S-K, Heo C, Park BW, Kim JK. Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma. *Eur Radiol*. 2020;30: 5392–5403.
50. Yuan G, Song Y, Li Q, Hu X, Zang M, Dai W, et al. Development and validation of a contrast-enhanced CT-based radiomics nomogram for prediction of therapeutic efficacy of anti-PD-1 antibodies in advanced HCC patients. *Front Immunol*. 2020;11: 613946.
51. Kuhl CK, Alparslan Y, Schmoe J, Sequeira B, Keulers A, Brümmendorf TH, et al. Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study. *Radiology*. 2019;290: 349–356.
52. Garralda E, Laurie SA, Seymour L, de Vries EGE. Towards evidence-based response criteria for cancer immunotherapy. *Nat Commun*. 2023;14: 3001.
53. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol*. 2017;18: e143–e152.
54. Hodi FS, Ballinger M, Lyons B, Soria J-C, Nishino M, Tabernero J, et al. Immune-Modified Response Evaluation Criteria In Solid Tumors (imRECIST): Refining Guidelines to Assess the Clinical Benefit of Cancer Immunotherapy. *J Clin Oncol*. 2018;36: 850–858.
55. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med*. 2016;22: 105–113.
56. McGranahan N, Furness AJS, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 2016;351: 1463–1469.
57. Sun R, Henry T, Laville A, Carré A, Hamaoui A, Bockel S, et al. Imaging approaches and radiomics: toward a new era of ultraprecision radioimmunotherapy? *J Immunother Cancer*. 2022;10. doi:10.1136/jitc-2022-004848
58. Himoto Y, Veeraraghavan H, Zheng J, Zamarin D, Snyder A, Capanu M, et al.

Computed Tomography-Derived Radiomic Metrics Can Identify Responders to Immunotherapy in Ovarian Cancer. *JCO Precis Oncol.* 2019;3. doi:10.1200/PO.19.00038

59. Tunali I, Gray JE, Qi J, Abdalah M, Jeong DK, Guvenis A, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. *Lung Cancer.* 2019;129: 75–79.
60. Trebeschi S, Bodalal Z, van Dijk N, Boellaard TN, Apfaltrer P, Tareco Bucho TM, et al. Development of a Prognostic AI-Monitor for Metastatic Urothelial Cancer Patients Receiving Immunotherapy. *Front Oncol.* 2021;11: 637804.
61. Jiang Z, Dong Y, Yang L, Lv Y, Dong S, Yuan S, et al. CT-Based Hand-crafted Radiomic Signatures Can Predict PD-L1 Expression Levels in Non-small Cell Lung Cancer: a Two-Center Study. *J Digit Imaging.* 2021;34: 1073–1085.
62. Wang C, Ma J, Shao J, Zhang S, Li J, Yan J, et al. Non-Invasive Measurement Using Deep Learning Algorithm Based on Multi-Source Features Fusion to Predict PD-L1 Expression and Survival in NSCLC. *Front Immunol.* 2022;13: 828560.
63. Wang C, Xu X, Shao J, Zhou K, Zhao K, He Y, et al. Deep Learning to Predict EGFR Mutation and PD-L1 Expression Status in Non-Small-Cell Lung Cancer on Computed Tomography Images. *J Oncol.* 2021;2021: 5499385.
64. Wang C, Ma J, Shao J, Zhang S, Liu Z, Yu Y, et al. Predicting EGFR and PD-L1 Status in NSCLC Patients Using Multitask AI System Based on CT Images. *Front Immunol.* 2022;13: 813072.
65. Zheng Y-M, Zhan J-F, Yuan M-G, Hou F, Jiang G, Wu Z-J, et al. A CT-based radiomics signature for preoperative discrimination between high and low expression of programmed death ligand 1 in head and neck squamous cell carcinoma. *Eur J Radiol.* 2022;146: 110093.
66. Bracci S, Dolciemi M, Trobiani C, Izzo A, Pernazza A, D'Amati G, et al. Quantitative CT texture analysis in predicting PD-L1 expression in locally advanced or metastatic NSCLC patients. *Radiol Med.* 2021;126: 1425–1433.
67. Zhu Y, Liu Y-L, Feng Y, Yang X-Y, Zhang J, Chang D-D, et al. A CT-derived deep neural network predicts for programmed death ligand-1 expression status in advanced lung adenocarcinomas. *Ann Transl Med.* 2020;8: 930.
68. Jardim DL, Goodman A, de Melo Gagliato D, Kurzrock R. The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell.* 2021;39: 154–173.
69. Wang M, Wang S, Trapani JA, Neeson PJ. Challenges of PD-L1 testing in non-small cell lung cancer and beyond. *J Thorac Dis.* 2020;12: 4541–4548.
70. Duvivier HL, Rothe M, Mangat PK, Garrett-Mayer E, Ahn ER, Al Baghdadi T, et al. Pembrolizumab in Patients With Tumors With High Tumor Mutational

Burden: Results From the Targeted Agent and Profiling Utilization Registry Study. *J Clin Oncol*. 2023; JCO2300702.

71. Lu MY, Chen RJ, Kong D, Lipkova J, Singh R, Williamson DFK, et al. Federated learning for computational pathology on gigapixel whole slide images. *Med Image Anal*. 2022;76: 102298.
72. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3: 119.
73. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat Cancer*. 2022. doi:10.1038/s43018-022-00416-8
74. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. 2022;3: 723–733.
75. Jiang Y, Zhang Z, Wang W, Huang W, Chen C, Xi S, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. *Nat Commun*. 2023;14: 5135.

Figures

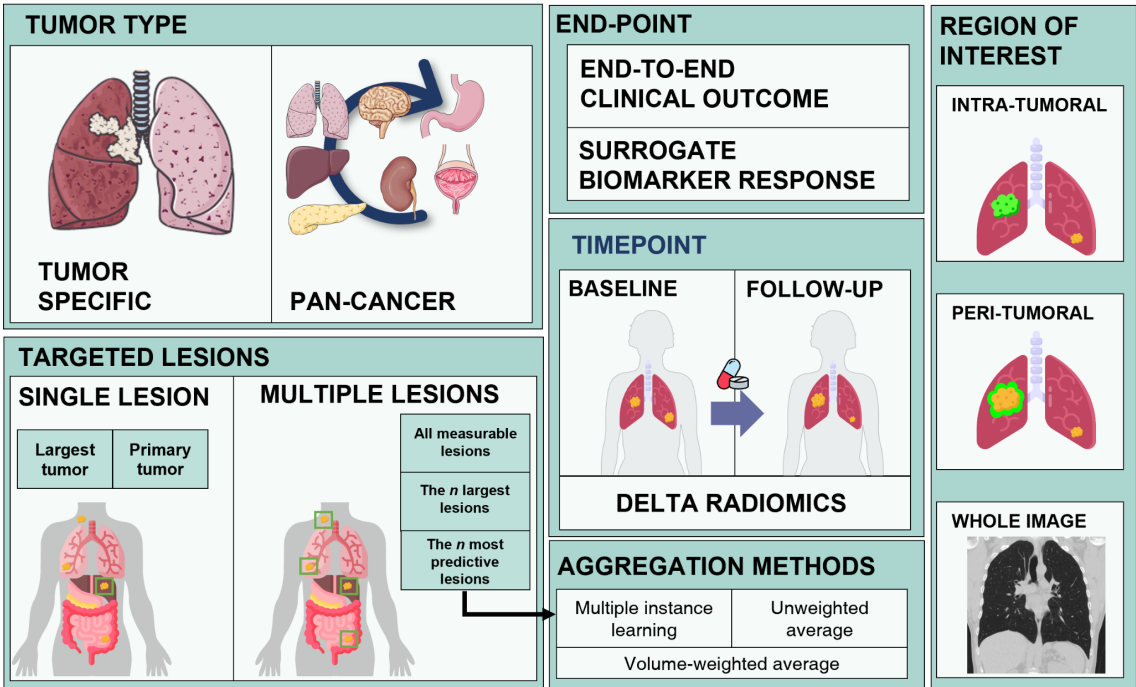


Figure 1. Potential sources of variability in radiomics study design including features related to the cohort setting (specific signature for a single tumor type or pan-cancer), end-point (for clinical outcome such as response yes/no or for predicting molecular surrogates biomarkers such as PD-L1 expression), number of lesions, imaging timepoints and region of interest.

n: number of lesions

PD-L1: programmed death-ligand 1

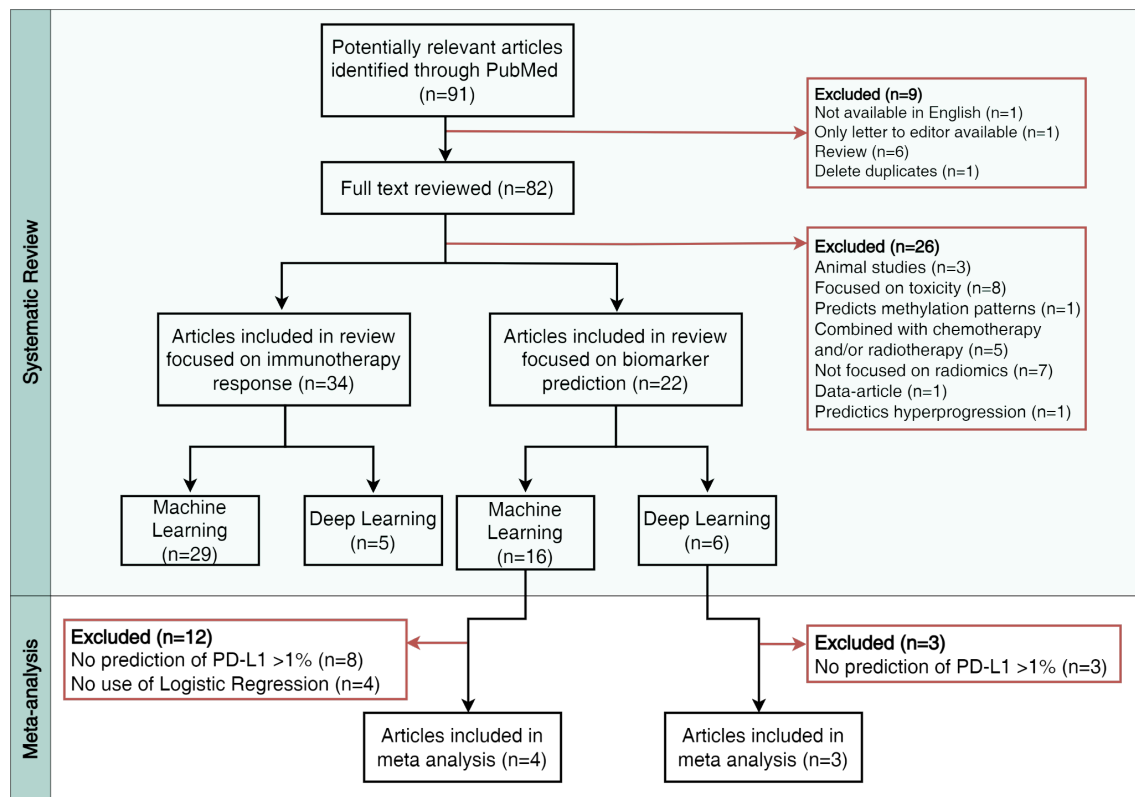


Figure 2. Referred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram for study illustrating the number of records screened in the review and articles included and excluded, outlining the predefined inclusion and exclusion criteria. In total, 56 articles were included in the review and 35 articles were excluded, reasons for exclusion were reported. Seven studies exploring CT-based radiomics models for predicting programmed death-ligand 1 (PD-L1) expression were included in the meta-analysis.

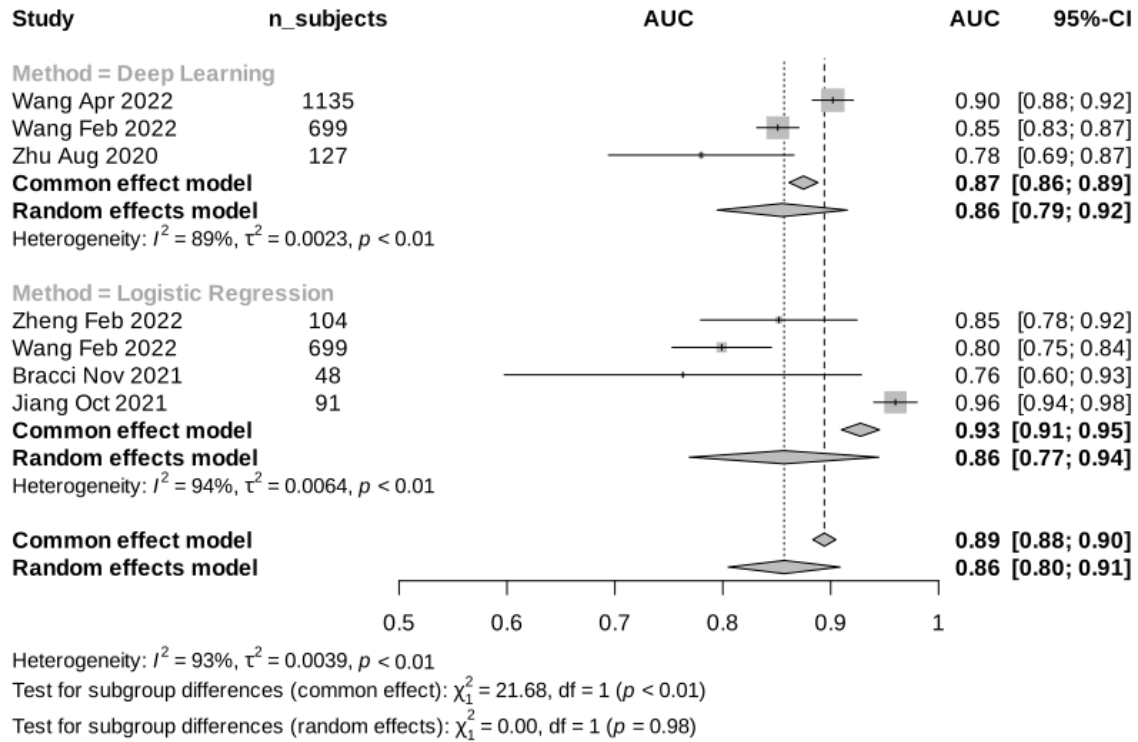


Figure 3. Meta-analysis results: Internal validation performance of the reported studies that implemented CT-based radiomics with classical machine learning (ML) or deep learning (DL) for predicting programmed death-ligand 1 (PD-L1) expression.

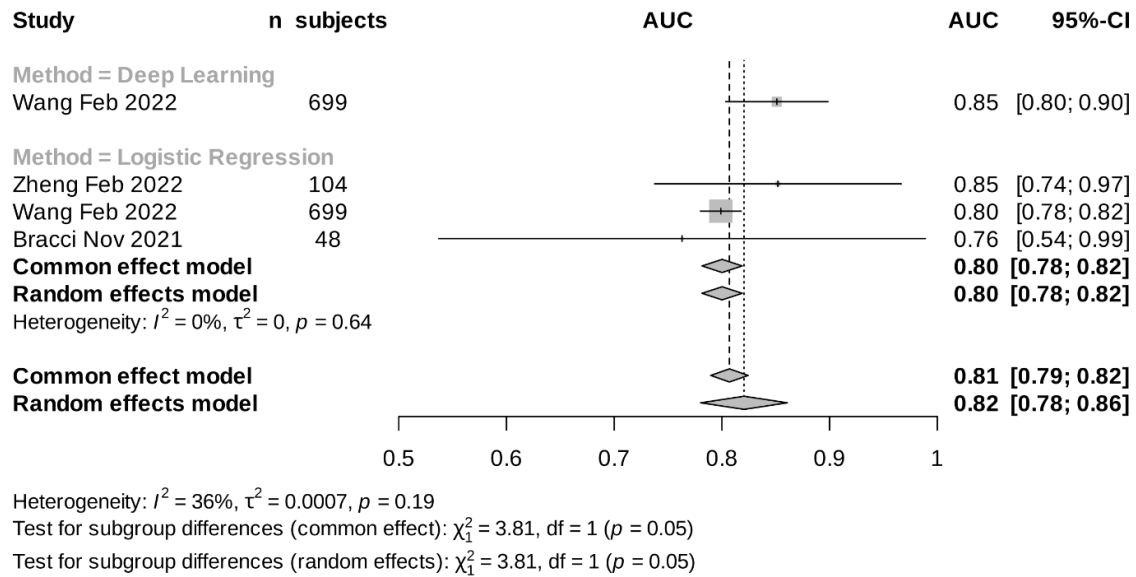


Figure 4. Meta-analysis results: External validation performance of the reported studies that implemented CT-based radiomics with machine learning (ML) or deep learning (DL) for predicting programmed death-ligand 1 (PD-L1) expression.

APPENDIX B

Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer

Reference:

Ghaffari Laleh N., Ligerio M., Perez-Lopez R., Kather JN.; Facts and Hopes on the Use of
Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer.
Clin Cancer Res 15 January 2023; 29 (2): 316–323.

Facts and Hopes on the Use of Artificial Intelligence for Predictive Immunotherapy Biomarkers in Cancer

Narmin Ghaffari Laleh¹, Marta Ligeró², Raquel Perez-Lopez^{2,3}, and Jakob Nikolas Kather^{1,4,5,6}



ABSTRACT

Immunotherapy by immune checkpoint inhibitors has become a standard treatment strategy for many types of solid tumors. However, the majority of patients with cancer will not respond, and predicting response to this therapy is still a challenge. Artificial intelligence (AI) methods can extract meaningful information from complex data, such as image data. In clinical routine, radiology or histopathology images are ubiquitously available. AI has been used to predict the response to immunotherapy from radiology or histopathology images, either directly or indirectly via surrogate

markers. While none of these methods are currently used in clinical routine, academic and commercial developments are pointing toward potential clinical adoption in the near future. Here, we summarize the state of the art in AI-based image biomarkers for immunotherapy response based on radiology and histopathology images. We point out limitations, caveats, and pitfalls, including biases, generalizability, and explainability, which are relevant for researchers and health care providers alike, and outline key clinical use cases of this new class of predictive biomarkers.

Introduction

Response prediction in immunotherapy—one of the major challenges of oncology

The unprecedented results of immune checkpoint inhibitors (ICI) in melanoma patients led to the extensive use of immunotherapy in many tumor types (1). However, while some patients achieve excellent responses, many others do not respond, but can still suffer from serious toxicity. Therefore, accurate predictive biomarkers towards a more precise patient stratification for ICI are needed to optimize patient selection and to minimize undesirable side effects. Several molecular biomarkers for ICI response prediction have been introduced. These include microsatellite instability (MSI), the tumor mutational burden (TMB), the expression of programmed death ligand-1 (PD-L1), as well as the number of tumor-infiltrating lymphocytes (TIL; refs. 2, 3). However, the performance for predicting the response of each biomarker individually or when combined is suboptimal, with some tumors showing resistance despite the presence of the biomarker and vice versa (4). Hence, there is a need for improved predictive

biomarkers for supporting clinical decisions, ideally with high reproducibility, high accuracy, and low costs.

Artificial intelligence in medical imaging

Artificial intelligence (AI) is a branch of computer science that integrates different technologies able to replicate human brain functions such as learning and problem solving. Machine learning (ML) is a subset of AI techniques capable of teaching a computer to detect patterns in datasets. ML has been used in oncology for many years, for example to define gene signatures of treatment response (5). A particularly powerful method within ML is artificial neural networks (ANN). An ANN is an interconnected group of units which collectively can perform complex computations, and is loosely modeled after biological neural networks. Multi-layered ANNs are particularly powerful, and using them in ML is called “Deep Learning” (DL). In the last decade, DL has provided strong performance gains in many scientific domains. One of its most common fields of application is computer-based image processing. In image processing, ANNs typically use a mathematical operation called convolutions to reduce the raw pixel information to relevant concepts. ANNs relying on convolutions are called convolutional neural networks (CNN). For decades, teaching computer programs to automatically read and interpret digital images was a very hard task, but in the last decade, the use of CNNs coupled with better hardware and larger data collections have resulted in methodologic breakthroughs. Today, DL with CNNs is the state-of-the-art method for almost any image processing task in nonmedical and medical applications.

Medical applications of AI

As of 2022, dozens of AI-based systems have received regulatory approval to be used in clinical routine (6). Many of these systems are tools to analyze digital medical images, including radiology, histopathology, dermoscopy, and endoscopy images. In oncology, image data is abundant and instrumental for medical decisions: the suspicion of a malignant tumor is usually confirmed with radiology imaging, which also serves to determine the spread of the disease. The diagnosis and most components of the stage of a malignant tumor are usually obtained through histopathology, i.e., visual examination of a piece of tumor tissue under a microscope by an expert pathologist. Hence, for almost every single patient with a solid tumor, radiology and histopathology images are available. The central paradigm of

¹Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany. ²Radiomics Group, Vall d'Hebron Institute of Oncology, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain. ³Department of Radiology, Vall d'Hebron University Hospital, Barcelona, Spain. ⁴Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom. ⁵Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. ⁶Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany.

N. Ghaffari Laleh, M. Ligeró, R. Perez-Lopez, and J.N. Kather contributed equally to this article.

Corresponding Authors: Jakob Nikolas Kather, Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Fetscherstrasse 74, Dresden 01307, Germany. E-mail: jakob-nikolas.kather@alumni.dkfz.de; and Raquel Perez-Lopez, Vall d'Hebron Institute of Oncology (VHIO), Cellex Center, Natzaré 115-117, Barcelona 08035, Spain. Phone: 349-3254-3450, ext. 8682; E-mail: rperez@vhio.net

Clin Cancer Res 2023;29:316-23

doi: 10.1158/1078-0432.CCR-22-0390

©2022 American Association for Cancer Research

image-based biomarkers in oncology is that these routinely available images contain much more information than is currently being used in clinical care and that this information can be extracted by AI.

Image-based biomarkers for cancer immunotherapy response prediction

In this work, we will review applications of AI to extract biomarkers for ICI response from radiology and histopathology images. The methods can be classified either as a “surrogate” strategy or an “end-to-end” strategy. The “surrogate” strategy means that AI is used to predict TMB, MSI status, TILs or PD-L1 expression. The “end-to-end” strategy means that AI is used to directly predict treatment response (Fig. 1). Together, these approaches have the potential to yield new diagnostic methods for better treatment decisions in clinical routine. However, before clinical adoption, it is important to discuss potential limitations and enable clinicians to understand and interpret the output of these systems.

Response Prediction from Radiology Imaging

Classical radiomics and DL radiomics

Radiology imaging modalities, such as CT, MRI, and PET are powerful tools for cancer detection, characterization, and follow-up,

providing a comprehensive view of the entire tumor repeatedly throughout the course of the disease. As early as 2014, ML has been used to quantify tumor features in CT images, which are linked to clinical outcome (7). This method has been termed “radiomics” (7). Classical radiomics follow a two-step approach. First, to derive a quantifiable set of handcrafted features from imaging data. Second, to train ML methods for predicting clinically relevant categories from these features. Handcrafted features provide information about the intensity, shape, and texture of the tumor phenotype. However, handcrafting sets of features limit the type of information that ML models can learn, potentially reducing the performance. To overcome this, a more recent approach is “Deep Radiomics”, which uses DL (mostly CNNs) to predict a target category directly from image data, learning features and their combination in one go. CNNs can learn a larger number of features at different levels of abstraction (Fig. 2A). Thereby, the extracted features are selected and weighted on the basis of the task at hand, which provides more flexibility and can improve performance.

Radiomics to predict surrogate biomarkers for immunotherapy response

Several studies have used AI to predict surrogate markers of immunotherapy response from radiology images. These markers include MSI (8–10), TMB (11–13), TILs (14–16), and PD-L1 (Table 1;

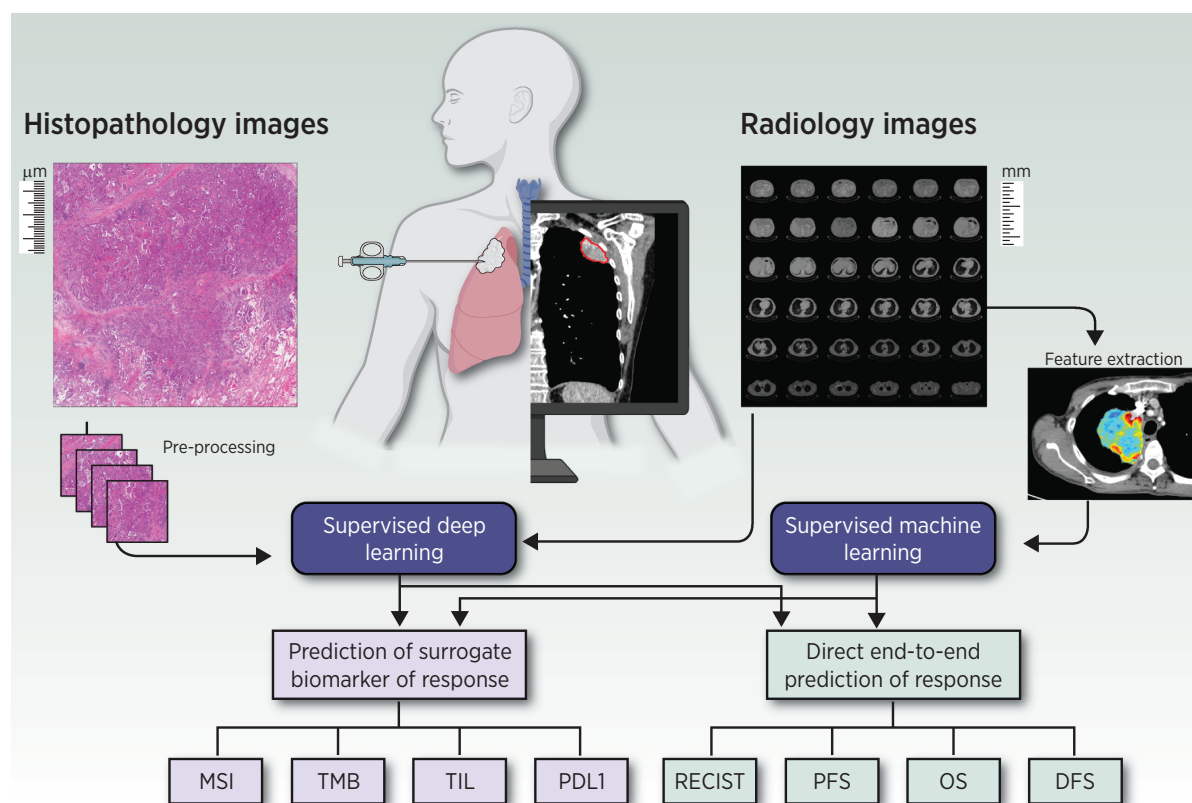


Figure 1.

Supervised learning workflows for prediction of immunotherapy response from clinical image data. Pathology (left) and radiology (right) image data are available for almost any patient with a solid tumor. Both types of image data have been used to predict surrogate markers of immunotherapy response or to directly predict immunotherapy response. PFS, progression-free survival; DFS, disease-free survival. (Adapted from an image created with BioRender.com.)

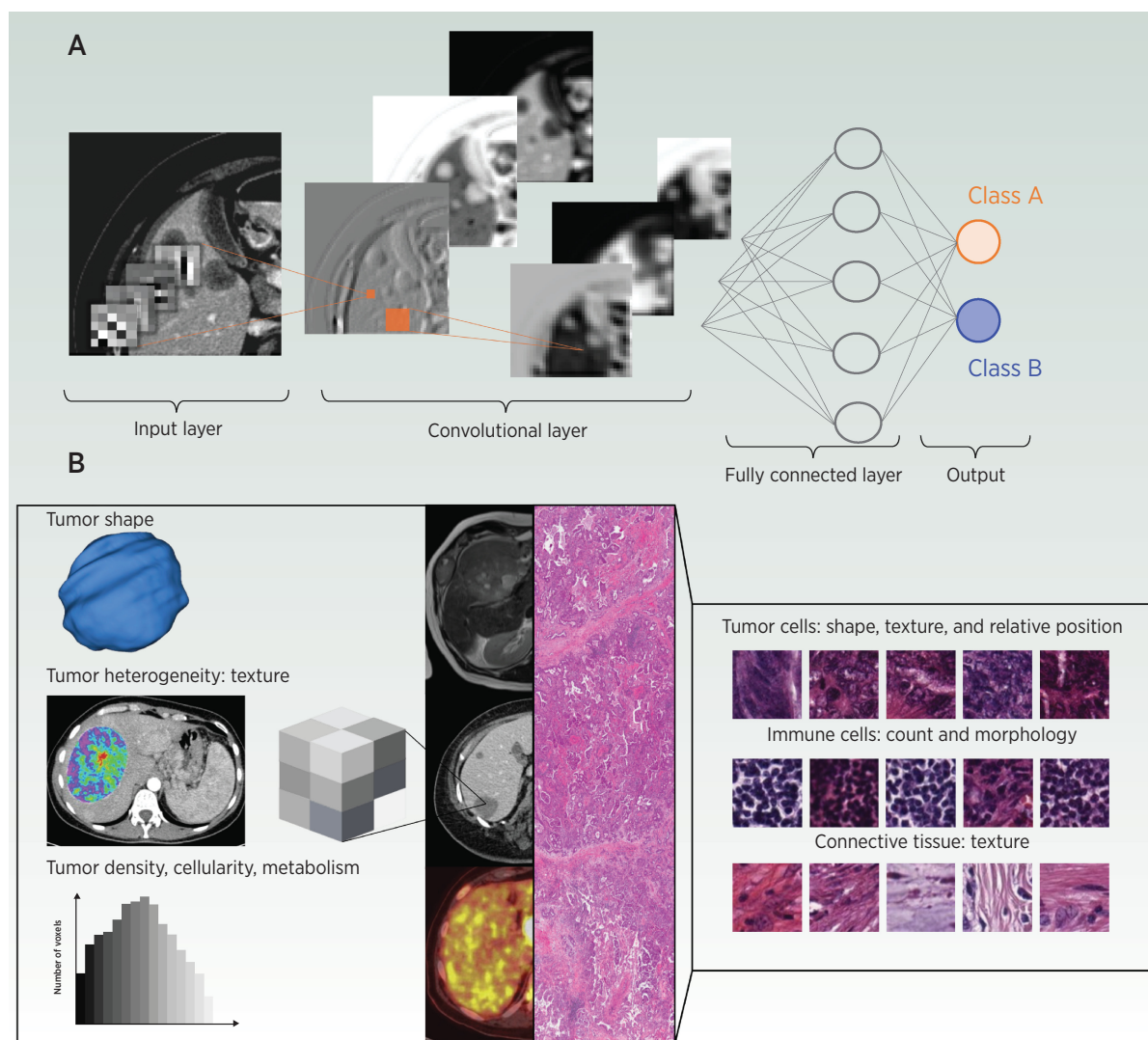


Figure 2.

Visual features in image data, which have been implicated in clinical outcome. **A**, A commonly used method to extract complex features from medical imaging data is CNNs. In this technology, artificial neurons are stacked in layers, which recognize visual patterns at different levels of abstraction. **B**, In radiology and pathology image data, neural networks can recognize a wide range of visual patterns that are related to immunotherapy response in patients.

refs. 17–19). The basic idea behind these studies is that the molecular properties of tumors change the tumor's phenotype, which can be observed in radiology images (Fig. 2B). Many of these studies use classical radiomics to classify a discrete surrogate biomarker, such as microsatellite stable versus MSI-high, high versus low TMB, high versus low CD8-cell infiltration, or PD-L1 positive versus negative. These studies showed promising results: the area under the receiver operating characteristic curve (AUC) usually falls within the range of 0.70 to 0.90. Other studies have used DL methods, showing similar performances for MSI and PD-L1 prediction (17, 20). The ability to predict the status of such biomarkers from radiology images opens exciting opportunities for clinical practice. For example, monitoring changes of these biomarkers during treatment, analyzing the whole tumor and not just a small part of tumor tissue, and the noninvasive-

ness are clear advantages compared with invasive tissue-based methods. However, the performance of these imaging-based biomarkers is not perfect, and it remains to be shown if it is sufficient for clinical decision-making.

Radiomics for end-to-end prediction of immunotherapy response

While “surrogate” approaches predict an established biomarker from image data, “end-to-end” approaches can directly predict immunotherapy response, as measured by the response evaluation criteria in solid tumors (RECIST), progression-free survival, or overall survival (OS) as endpoints (Table 1; refs. 21–23). Successful applications of this approach have been shown in melanoma (23), lung cancer (21, 22), and bladder cancer, among other tumor types, and are generally

Table 1. An overview of indirect and end-to-end biomarkers in radiology and pathology image analysis.

Target	References	Summary	Data type	Tumor type	Technology	Validation AUROC
RADIOLOGY	(8–10, 20)	Prediction of MSI status	CT	Colorectal	Radiomics features + ML model	0.73–0.89 ^a
	(11–13)	Prediction of TMB	CT/MRI	Lung, endometrial, glioma	Radiomics features + ML model	0.81–0.87 ^a
	(14–16)	Prediction of TIL number	CT/MRI	Head and neck, lung, hepatocellular, bladder, endothelial	Radiomics features + ML model	0.67–0.90 ^a
	(17–19)	Prediction of PD-L1 expression	CT/PET	Lung, pancreatic	Radiomics features + ML model / CNN	0.68–0.82 ^a
	(23)	Direct prediction of OS under immunotherapy	CT	Melanoma	Radiomics features + RF	0.92 (95% CI, 0.89–0.95) ^b
	(22)	Direct prediction of response to immunotherapy	CT	Multiple solid tumors	Radiomic features, Elastic-net, GLMM	0.67 (95% CI, 0.58–0.76) ^b
	(21)			Melanoma, lung	Radiomics features + RF	0.64–0.83 ^a
	(41–50)	Prediction of MSI status		Colorectal, gastric, endometrial	CNN	0.60–0.92 ^a
	(45, 61, 80)	Prediction of EBV or HPV status	H&E	Gastric, head and neck	CNN	0.67–0.87 ^a
	(51–54)	Prediction of TMB	H&E	Lung, urogenital, colorectal, gastric, head and neck, and others	CNN	0.64–0.92 ^a
PATHOLOGY	(57, 58, 81)	Prediction of immune-related gene expression signatures	H&E	Hepatocellular, colorectal	CNN, MIL	0.81–0.92 ^a
	(55, 56)	Prediction of PD-L1 status from H&E	H&E	Lung	CNN	0.63–0.93 ^a
	(37, 38)	Prediction of TIL number from H&E	H&E	Breast, oral	SSL, MIL, CNN	0.88–0.89 ^a
	(64, 65)	Direct prediction of response to immunotherapy	H&E	Lung, melanoma	CNN, GNN	0.62–0.78 ^a
	(29–33)	PD-L1 expression scoring assistance	IHC	Lung	GAN, CNN	0.90–0.99

Note: Representative studies are referenced, and the list may not be exhaustive in all categories.

Abbreviations: AUROC, area under the receiver operating characteristic curve; CI, confidence interval; MIL, multiple instance learning; RF, random forest; FS, feature selection; GAN, generative adversarial network; CPS, combined positivity score; IHC, immunohistochemistry; H&E, hematoxylin and eosin.

^aWhenever multiple studies are reported, the range of AUROC from the validation/test set of the studies are listed (lowest and highest value).

^bWhenever only a single study is reported (or only a single study includes detailed performance statistics), the AUROC and 95% CI from the validation/test sets are listed as given in the publication.

based on “classical” handcrafted radiomics, not DL radiomics. In general, these approaches have achieved a good performance with AUC values greater than 0.70 for predicting ICI response. Furthermore, in most studies, these predictive radiomics scores are significantly associated with other surrogate biomarkers with the aim of explaining the underlying biology of the radiomics responsive phenotypes (21, 22). OS is the less common endpoint of response, given that it is highly dependent on other factors, such as previous treatments. However, it is the most relevant endpoint in model development for reaching clinical practice (24). In summary, end-to-end imaging biomarkers may play a key role in complementing surrogate biomarkers to improve the understanding of the tumor phenotype.

Response Prediction from Digital Histopathology Images

The rise of computational pathology

Histopathology images reflect properties of tumors, which are associated with immunotherapy response. Compared with radiology, the spatial scale of histopathology is much lower, such that phenotypes at cellular and subcellular levels can be directly observed (Fig. 2B). Even routine tissue slides stained with hematoxylin and eosin (H&E) are sufficient to identify many different types of immune cells in the lymphoid and myeloid spectrum. A key disadvantage of histopathology, however, is that it requires invasive procedures. Also, while radiomics biomarkers can be recomputed in serial imaging during

the course of the disease, histopathology image biomarkers are usually only measured in the initial tumor sample. In the last 5 to 10 years, AI approaches have been widely used to extract biomarkers from H&E slides of solid tumors (25). AI can extract diagnostic (26), prognostic (27), and predictive (28) information from H&E slides. This use of AI methods in pathology image analysis is referred to as “computational pathology”.

Quantification of IHC

One of the applications of DL methods is to facilitate the quantification of established biomarkers in IHC slides. For example, the expression level of the PD-L1 protein on tumor cells and immune cells (combined positivity score), is assessed by manual observation of IHC. Several studies have used DL to automate the subjective scoring of PD-L1 status (29–33). Similarly, the number of TILs in H&E or IHC slides is associated with survival and immunotherapy response in many tumor types (34, 35). As early as 2006, several years before DL was used as a tool in biomedicine, handcrafted image analysis pipelines with classical ML methods were used to count TILs (36). More recently, several studies have used DL for TIL quantification (37, 38). In addition to quantifying lymphocytes, ML methods were also used to quantify other types of innate and lymphoid immune cells, demonstrating that cell counts are prognostic of survival in multiple tumor types (39, 40).

Extracting molecular biomarkers from H&E slides

However, the abilities of DL are not limited to simple quantification tasks such as counting cells. DL can solve complex visual pattern recognition tasks and can piece together subtle visual cues related to cell numbers, cellular shape and textures, relative cell positions and phenotypes of connective tissue. For example, by training DL systems on raw histopathology slides, it is possible to infer the MSI status in colorectal, gastric, and endometrial cancer (41–50). This is particularly relevant for immunotherapy because MSI is a clinically approved biomarker to select patients for immunotherapy, independent of the tumor type (46). MSI prediction systems have been made explainable. When they are queried for relevant visual features driving the classifications, immune cell-rich tumor regions are highlighted (50). Similarly, DL has been used to predict TMB directly from H&E in multiple major tumor types, including lung, breast, and colorectal cancer (51–54). Also, PD-L1 expression levels have been inferred from H&E directly, without the need for a dedicated IHC (55, 56). In addition, the expression level of gene signatures which predict ICI response can be inferred by DL from H&E slides, in hepatocellular carcinoma (57), lung cancer (58), and multiple other tumor types (48). This could help to solve some of the practical problems associated with widespread clinical use of these signatures. Likewise, DL can predict immunotherapy-sensitive tumor subtypes. For example, the luminal subtype of urothelial carcinoma which conveys a better anti-PD-L1 response (59) has been predicted from H&E slides alone (60). Relatedly, virus positivity in head and neck and gastric cancer—which associates with better ICI response—can be inferred from H&E slides with DL (45, 61). Finally, some mutations in clinically relevant driver genes have been shown to increase the probability of immunotherapy response and these mutations can be inferred from H&E. Prominent examples are detection of mutant BRAF in melanoma (62) which predicts benefit to anti-PD-1 response (63). In summary, routinely available H&E slides of solid tumors seem to harbor a wealth of information, some of which is related to ICI response, and which can be extracted by DL.

Computational pathology for end-to-end prediction of immunotherapy response

All of these above-mentioned approaches use DL to predict a known biological marker from H&E images. An alternative strategy is to train DL directly on response or outcome data. This has been attempted by at least two studies (64, 65) that used CNNs or graph neural networks (GNN) to predict immunotherapy responses. These studies reported an AUC of 0.778 for prediction of responders in melanoma and an AUC of 0.69 for predicting response in lung cancer. Another report about the prediction of response from the morphology of cancer cell nuclei is publicly available (66). In general, such end-to-end prediction studies are quite difficult to perform for practical reasons: they require the DL system to be trained on clinical outcomes, ideally RECIST data. In practice, it is very difficult to collect a sufficient number of pathology tissue specimens of ICI-treated patients with matched response data. In addition, cancer immunotherapy is often administered in late lines of therapy, after patients have received one or more regimens of chemotherapy and other treatments. However, tissue for histopathology is typically acquired at the initial diagnosis, and re-biopsies at later time points are not commonly performed in most tumor types. This means that DL systems are trained on treatment-naïve tumor tissue while patients might start immunotherapy only months later after multiple previous lines of treatment have failed. This is a strong conceptual limitation that could only be resolved by a repetitive sampling of tissue, or a move of immunotherapies towards earlier lines of treatment.

Limitations and Outlook

Key limitations

AI biomarkers have a number of conceptual limitations. The first limitation is data quality. Beyond the large amount of data that AI models require for achieving accurate, generalizable results, this data must be of high quality (67). If we train a model with noisy or artefactual images, many more cases will be necessary for the model to converge and achieve a good performance. The second limitation is generalization. If the training data is not representative of real-world populations, DL models can fail to generalize. This is especially an issue in medical contexts, where data distributions vary markedly between different countries or even different hospitals. Without adequate precautions, such batch effects can inflate performance statistics (68). Mitigation strategies are to train on diverse datasets (42) or augmenting data (69). The third limitation is biases. AI models can be biased, which means that the performance can be dependent on patient characteristics like age, gender, or ethnicity (70). For AI models to be deployed in medicine, large-scale validation studies with predefined performance metrics are required to guarantee model performance in the real world (26). The fourth limitation is the quality of the ground truth. When developing a model using molecular biomarkers as a surrogate of response to immunotherapy, the performance of the model will be limited to the established molecular biomarker predictive capacity. This is a motivation for clear definition of ground truth for AI biomarker training, but also a motivation for end-to-end training of biomarkers on clinical outcome data. In all of these efforts, data standardization and quality control systems are paramount before application in clinical routine. Best practice guidelines for such quality aspects are formalized and collected in the Equator Network (<https://www.equator-network.org/>), which also includes AI-specific aspects, such as in the STARD-AI (71) and TRIPOD-AI (72) guidelines. In addition to radiology, an Image Biomarker Standardization Initiative (73) as well as a radiomics quality score (74) have been reported

with the aim of ensuring reproducible and trustable predictive models. In addition to these research-focused guidelines, the FDA has published a list of guiding principles to promote the safe, effective, and high-quality application of AI and ML in the medical field (75).

Multimodal models

One way to combine the benefits of radiology and histology, as well as other data types, would be to develop multimodal AI models, which can integrate multiple data types. Multimodality also helps with interpretability of the resulting models because many image features only make sense in the light of specific host factors, such as age, immune status, comorbidities/disease, genomics. ML models have been used to extract immunotherapy biomarkers from non-image data such as serum profiles from liquid biopsies of patients with cancer (76–78). Combining such data with image data could further improve the predictive performance. On the technical side, transformer neural networks have achieved remarkable performance in non-medical tasks, especially for combinations of different types of image data. However, evaluating such multimodal models exacerbated the practical problems associated with data collection: for most researchers at academic institutions, it is currently very difficult if not impossible to collect radiology, histopathology, and clinical data for a set of patients. Systematic data collection in clinical trials could be a solution to this problem if the data is made available to researchers with low barriers.

Explainability

The increasing use of AI for developing immune surrogate biomarkers and the potential impact on clinical decision-making has raised the need for humans to understand these algorithms. This has been termed “Explainable AI” (XAI). Many classical ML algorithms are intrinsically explainable by their structure, e.g., logistic regression or decision trees. However, models with higher complexity, such as DL models, tend to lack explainability. To solve the trade-off between performance and explainability, several *post hoc* techniques have been developed for understanding the decision-making procedure of these so-called black-box models. XAI aims to comprehend model predictions and explain them in human and understandable terms. This could increase the trustability of all the stakeholders: medical doctors and patients to rely on fair decisions, regulatory agencies for quality control, and developers for improving the product.

Regulatory approval

Routine clinical implementation of imaging AI-based tools should be driven by clear demonstration of clinical value and strict ethical and regulatory requirements. An added complexity for validating AI biomarkers, compared with other medical devices, is the ability for AI systems to learn from real-world data in real time. These evolving biomarkers require appropriately tailored regulatory frameworks. AI-based imaging biomarkers are tools embedded in software applications that are intended to be used, alone or in combination, for predicting or monitoring cancer response, therefore being considered

as Medical Device Software. The framework to develop, clinically qualify and implement a software tool as a medical device will depend on the local regulations where the device is planned to be used. In the United States, the FDA has defined a framework to enable developers, users and the Agency itself to evaluate and monitor Clinical Decision Support (CDS) software from its premarket development through post-market performance accounting for its iterative nature, while still ensuring its continued safety and effectiveness evaluation. In Europe, CDS software should be compliant with the European Union Medical Devices Regulation (MDR 2017/745), that sets the standards of performance, quality, safety, and efficacy (79).

Embedding in routine workflows

The main constraints of AI-based tool implementation in clinical practice are numerous, particularly related to a lack of clinical qualification, but also to the complexity of conducting prospective clinical trials that evaluate biomarkers. Moreover, advances in the field are sometimes hampered by the reluctance of a part of the medical community to embrace these technologies due to these potential users’ lack of confidence and their presumed resistance to the heavier workload that high-throughput imaging analysis may involve. Importantly, the implementation of AI-based tools should not increase physician workload but actually reduce and facilitate radiologists’ and pathologists’ workflows, enabling standardized data reporting. Therefore, efforts are necessary to integrate these software applications within the clinical routine analysis platforms of radiology and pathology departments. In this regard, all the stakeholders (i.e., manufacturers, researchers, clinicians, patients) involved in the development of imaging AI-based tools should work together to accelerate the validation and implementation of these tools in clinical routine and truly impact clinical practice. Imaging AI-based tools must be accessible, user-friendly, rapid to compute and able to promote equality in healthcare to be implemented in routine clinical practice. Finally, the results must be considered as a decision support tool to assist physicians, rather than a substitute for expert physician decision-making.

Authors’ Disclosures

R. Perez-Lopez reports grants from AstraZeneca and Roche Pharma outside the submitted work. J.N. Kather reports personal fees from Owkin, Panakeia, MSD, Eisai, and Bayer outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

R. Perez-Lopez is supported by La Caixa Foundation, a CRIS Foundation Talent Award (TALENT19–05), the FERO Foundation, the Instituto de Salud Carlos III–Investigacion en Salud (PI18/01395 and PI21/01019), and the Prostate Cancer Foundation (18YOUN19). M. Ligerio is supported by PERIS PIF-Salut Grant. J.N. Kather is supported by the German Federal Ministry of Health (DEEP LIVER, ZMV11–2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864).

Received June 7, 2022; revised July 26, 2022; accepted August 29, 2022; published first September 9, 2022.

References

1. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer* 2019;19:133–50.
2. Center for Drug Evaluation, Research. FDA approves pembrolizumab for adults and children with TMB-H solid tumors. In: U.S. Food and Drug Administration. 17 Jun 2020 [cited 25 Apr 2022]. Available from: <https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors>.
3. Planchard D, Popat S, Kerr K, Novello S, Smit EF, Faivre-Finn C, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment, and follow-up. *Ann Oncol* 2018;29:iv192–237.

4. Lee JS, Ruppini E. Multiomics prediction of response rates to therapies to inhibit programmed cell death 1 and programmed cell death 1 ligand 1. *JAMA Oncol* 2019;5:1614–8.
5. Wiesweg M, Mairinger F, Reis H, Goetz M, Kollmeier J, Misch D, et al. Machine learning reveals a PD-L1-independent prediction of response to immunotherapy of non-small cell lung cancer by gene expression context. *Eur J Cancer* 2020;140:76–85.
6. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118.
7. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumor phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014;5:4006.
8. Li Z, Zhong Q, Zhang L, Wang M, Xiao W, Cui F, et al. Computed tomography-based radiomics model to preoperatively predict microsatellite instability status in colorectal cancer: a multicenter study. *Front Oncol* 2021;11:666786.
9. Pei Q, Yi X, Chen C, Pang P, Fu Y, Lei G, et al. Pretreatment CT-based radiomics nomogram for predicting microsatellite instability status in colorectal cancer. *Eur Radiol* 2022;32:714–24.
10. Cao Y, Zhang G, Zhang J, Yang Y, Ren J, Yan X, et al. Predicting microsatellite instability status in colorectal cancer based on triphasic enhanced computed tomography radiomics signatures: a multicenter study. *Front Oncol* 2021;11:687771.
11. He B, Dong D, She Y, Zhou C, Fang M, Zhu Y, et al. Predicting response to immunotherapy in advanced non-small cell lung cancer using tumor mutational burden radiomic biomarker. *J Immunother Cancer* 2020;8:e000550.
12. Veeraraghavan H, Friedman CF, DeLair DF, Ninčević J, Himoto Y, Bruni SG, et al. Machine learning-based prediction of microsatellite instability and high tumor mutation burden from contrast-enhanced computed tomography in endometrial cancers. *Sci Rep* 2020;10:17769.
13. Liu E-T, Zhou S, Li Y, Zhang S, Ma Z, Guo J, et al. Development and validation of an MRI-based nomogram for the preoperative prediction of tumor mutational burden in lower-grade gliomas. *Quant Imaging Med Surg* 2022;12:1684–97.
14. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumor-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. *Lancet Oncol* 2018;19:1180–91.
15. Chen S, Feng S, Wei J, Liu F, Li B, Li X, et al. Pretreatment prediction of immunoscore in hepatocellular cancer: a radiomics-based clinical model based on Gd-EOB-DTPA-enhanced MRI imaging. *Eur Radiol* 2019;29:4177–87.
16. Liao H, Zhang Z, Chen J, Liao M, Xu L, Wu Z, et al. Preoperative radiomic approach to evaluate tumor-infiltrating CD8⁺ T cells in hepatocellular carcinoma patients using contrast-enhanced computed tomography. *Ann Surg Oncol* 2019;26:4537–47.
17. Mu W, Jiang L, Shi Y, Tunali I, Gray JE, Katsoulakis E, et al. Noninvasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images. *J Immunother Cancer* 2021;9:e002118.
18. Iwatate Y, Hoshino I, Yokota H, Ishige F, Itami M, Mori Y, et al. Radiogenomics for predicting p53 status, PD-L1 expression, and prognosis with machine learning in pancreatic cancer. *Br J Cancer* 2020;123:1253–61.
19. Tang C, Hobbs B, Amer A, Li X, Behrens C, Canales JR, et al. Development of an immune-pathology informed radiomics model for non-small cell lung cancer. *Sci Rep* 2018;8:1922.
20. Zhang W, Yin H, Huang Z, Zhao J, Zheng H, He D, et al. Development and validation of MRI-based deep learning models for prediction of microsatellite instability in rectal cancer. *Cancer Med* 2021;10:4164–73.
21. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Delli Pizzi A, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol* 2019;30:998–1004.
22. Ligerio M, Garcia-Ruiz A, Viaplana C, Villacampa G, Raciti MV, Landa J, et al. A CT-based radiomics signature is associated with response to immune checkpoint inhibitors in advanced solid tumors. *Radiology* 2021;299:109–19.
23. Dercle L, Zhao B, Gönen M, Moskowitz CS, Firas A, Beylertgil V, et al. Early readout on overall survival of patients with melanoma treated with immunotherapy using a novel imaging analysis. *JAMA Oncol* 2022;8:385–92.
24. Alban TJ, Chan TA. Immunotherapy biomarkers: the long and winding road. *Nature reviews. Nat Rev Clin Oncol* 2021;18:323–4.
25. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* 2021;124:686–96.
26. Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* 2021;21:199–211.
27. Skrede O-J, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* 2020;395:350–60.
28. Cifci D, Foersch S, Kather JN. Artificial intelligence to identify genetic alterations in conventional histopathology. *J Pathol* 2022;257:430–44.
29. Hondelink LM, Hüyük M, Postmus PE, Smit VTHBM, Blom S, von der Thüsen JH, et al. Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumor proportion score assessment in non-small cell lung cancer. *Histopathology* 2022;80:635–47.
30. Liu J, Zheng Q, Mu X, Zuo Y, Xu B, Jin Y, et al. Automated tumor proportion score analysis for PD-L1 (22C3) expression in lung squamous cell carcinoma. *Sci Rep* 2021;11:15907.
31. Kapil A, Meier A, Zuraw A, Steele KE, Rebelatto MC, Schmidt G, et al. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. *Sci Rep* 2018;8:17343.
32. Wu J, Liu C, Liu X, Sun W, Li L, Gao N, et al. Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer. *Mod Pathol* 2022;35:403–11.
33. Wang X, Chen P, Ding G, Xing Y, Tang R, Peng C, et al. Dual-scale categorization based deep learning to evaluate programmed cell death ligand 1 expression in non-small cell lung cancer. *Medicine* 2021;100:e25994.
34. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing tumor-infiltrating lymphocytes in solid tumors: A practical review for pathologists and proposal for a standardized method from the International Immunology biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Adv Anat Pathol* 2017;24:311–35.
35. Yu Y, Zeng D, Ou Q, Liu S, Li A, Chen Y, et al. Association of survival and immune-related biomarkers with immunotherapy in patients with non-small cell lung cancer: a meta-analysis and individual patient-level analysis. *JAMA Netw Open* 2019;2:e196879.
36. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006;313:1960–4.
37. Schirris Y, Engelaer M, Panteli A, Horlings HM, Gavves E, WeakSTIL TJ, et al. Weak whole-slide image level stromal tumor-infiltrating lymphocyte scores are all you need. *arXiv [eess.IV]*. 2021. Available from: <http://arxiv.org/abs/2109.05892>.
38. Shaban M, Khurram SA, Fraz MM, Alsubaie N, Masood I, Mushtaq S, et al. A novel digital score for abundance of tumor infiltrating lymphocytes predicts disease-free survival in oral squamous cell carcinoma. *Sci Rep* 2019;9:13341.
39. Kather JN, Hörner C, Weis C-A, Aung T, Vokuhl C, Weiss C, et al. CD163⁺ immune cell infiltrates and presence of CD54⁺ microvessels are prognostic markers for patients with embryonal rhabdomyosarcoma. *Sci Rep* 2019;9:9211.
40. Kather JN, Suarez-Carmona M, Charoentong P, Weis C-A, Hirsch D, Bankhead P, et al. Topography of cancer-associated immune cells in human solid tumors. *eLife* 2018;7:36967.
41. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019;25:1054–6.
42. Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology* 2020;159:1406–16.
43. Bilal M, Raza SEA, Azam A, Graham S, Ilyas M, Cree IA, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* 2021;3:e763–72.
44. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol* 2021;22:132–41.
45. Muti HS, Heij LR, Keller G, Kohlruß M, Langer R, Dislich B, et al. Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicenter cohort study. *The Lancet Digital Health* 2021;3:e654–64.

46. Echle A, Laleh NG, Schrammen PL, West NP, Trautwein C, Brinker TJ, et al. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review. *Immunoinformatics* 2021;3–4:100008.
47. Schrammen PL, Laleh NG, Echle A, Truhn D, Schulz V, Brinker TJ, et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J Pathol* 2022;256:50–60.
48. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* 2020;1:789–99.
49. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-seq expression of tumors from whole slide images. *Nat Commun* 2020;11:3877.
50. Echle A, Laleh NG, Quirke P, Grabsch HI, Muti HS, Saldanha OL, et al. Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a prescreening tool for clinical application. *ESMO Open* 2022;7:100400.
51. Jain MS, Massoud TF. Predicting tumor mutational burden from histopathological images using multiscale deep learning. *bioRxiv* 2020;.2020.06.15.153379.
52. Xu H, Park S, Lee SH, Hwang TH. Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients. *bioRxiv* 2019;554527.
53. Chen S, Xiang J, Wang X, Zhang J, Yang S, Huang J, et al. Pan-cancer computational histopathology reveals tumor mutational burden status through weakly supervised deep learning. *arXiv [cs.CV]*. 2022. Available from: <http://arxiv.org/abs/2204.03257>.
54. Niu Y, Wang L, Zhang X, Han Y, Yang C, Bai H, et al. Predicting tumor mutational burden from lung adenocarcinoma histopathological images using deep learning. *Front Oncol* 2022;12:927426.
55. Sha L, Osinski BL, Ho IY, Tan TL, Willis C, Weiss H, et al. Multi-field-of-view deep learning model predicts non-small cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. *J Pathol Inform* 2019;10:24.
56. Ebert MP, Meindl-Beinker NM, Gutting T, Maenz M, Betge J, Schulte N, et al. Second-line therapy with nivolumab plus ipilimumab for older patients with esophageal squamous cell cancer (RAMONA): a multicenter, open-label, phase II trial. *The Lancet Healthy Longevity* 2022;3:e417–27.
57. Zeng Q, Klein C, Caruso S, Maillé P, Laleh NG, Sommacale D, et al. Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology. *J Hepatol* 2022;77:116–27.
58. Shen C, Schlager C, Rajan D, Pourayhya M, Lin M, Mountain V, et al. Abstract 1922: Application of an interpretable graph neural network to predict gene expression signatures associated with tertiary lymphoid structures in histopathological images. *Cancer Res* 2022;82:1922.
59. Rosenberg JE, Hoffman-Censits J, Powles T, van der Heijden MS, Balar AV, Necchi A, et al. Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicenter, phase II trial. *Lancet* 2016;387:1909–20.
60. Woerl A-C, Eckstein M, Geiger J, Wagner DC, Daher T, Stenzel P, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol* 2020;78:256–64.
61. Kather JN, Schulte J, Grabsch HI, Loeffler C, Muti H, Dolezal J, et al. Deep learning detects virus presence in cancer histology. *bioRxiv* 2019;690206.
62. Kim RH, Nomikou S, Coudray N, Jour G, Dawood Z, Hong R, et al. Deep learning and pathomics analyses reveal cell nuclei as important features for mutation prediction of BRAF-mutated melanomas. *J Invest Dermatol* 2022;142:1650–8.
63. Wolchok JD, Chiarion-Sileni V, Gonzalez R, Rutkowski P, Grob J-J, Cowey CL, et al. Overall survival with combined nivolumab and ipilimumab in advanced melanoma. *N Engl J Med* 2017;377:1345–56.
64. Hu J, Cui C, Yang W, Huang L, Yu R, Liu S, et al. Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images. *Transl Oncol* 2021;14:100921.
65. Xie C, Vanderbilt C, Feng C, Ho D, Campanella G, Egger J, et al. Computational biomarker predicts lung ICI response via deep learning-driven hierarchical spatial modelling from H&E. *Research Square*. 2022.
66. Madabhushi A, Wang X, Barrera C. Predicting response to immunotherapy using computer extracted features of cancer nuclei from hematoxylin and eosin (HandE) stained images of non-small cell lung cancer (NSCLC). US Patent 11,055,844. 2021. Available: <https://patents.google.com/patent/US11055844B2/en>.
67. Schömig-Markieka B, Prylukhin A, Hulla W, Bychkov A, Fukuoka J, Madabhushi A, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod Pathol* 2021;34:2098–108.
68. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun* 2021;12:4423.
69. Yamashita R, Long J, Banda S, Shen J, Rubin DL. Learning Domain-agnostic visual representation for computational pathology using medically irrelevant style transfer augmentation. *IEEE Trans Med Imaging* 2021;40:3945–54.
70. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
71. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centered diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
72. Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
73. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295:328–38.
74. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14:749–62.
75. U.S. Food and Drug Administration. Good machine learning practice for medical device development: guiding principles. Available from: <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>.
76. Park Y, Kim MJ, Choi Y, Kim NH, Kim L, Hong SPD, et al. Role of mass spectrometry-based serum proteomics signatures in predicting clinical outcomes and toxicity in patients with cancer treated with immunotherapy. *J Immunother Cancer* 2022;10:e003566.
77. Wei C, Wang M, Gao Q, Yuan S, Deng W, Bie L, et al. Dynamic peripheral blood immune cell markers for predicting the response of patients with metastatic cancer to immune checkpoint inhibitors. *Cancer Immunol Immunother* 2022.
78. Kato S, Li B, Adashek JJ, Cha SW, Bianchi-Frias D, Qian D, et al. Serial changes in liquid biopsy-derived variant allele frequency predict immune checkpoint inhibitor responsiveness in the pan-cancer setting. *Oncoimmunology* 2022;11:2052410.
79. European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance). Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745>.
80. Klein S, Quaas A, Quantius J, Löser H, Meinel J, Peifer M, et al. Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains. *Clin Cancer Res* 2021;27:1131–8.
81. Sirinukunwattana K, Domingo E, Richman SD, Redmond KL, Blake A, Verrill C, et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* 2021;70:544–54.

APPENDIX C

Enhancing immunotherapy response prediction via multimodal integration of radiology and pathology deep learning models

Reference:

Ligero M., El Nahhas O., Prior O. et al. Enhancing immunotherapy response prediction via
multimodal integration of radiology and pathology deep learning models
Annals of Oncology (ESMO Abstract) 2023

Interests, Institutional, Advisory Board: BMS, Takeda, Novartis; Financial Interests, Personal, Invited Speaker: Roche, Incyte; Financial Interests, Institutional, Research Grant: Bayer, Chugai, BMS, Incyte. K. Pantel: Financial Interests, Personal, Advisory Board: Menarini/Silicon Biosystems. All other authors have declared no conflicts of interest.

<https://doi.org/10.1016/j.annonc.2023.09.2898>

175P Radiomic biomarker of vessel tortuosity for monitoring treatment change: Preliminary findings in prospective evaluation of ECOG-ACRIN EA5163

P. Mutha¹, F. Duan², V.S. Viswanathan¹, M. Khorrami¹, N. An³, V. Velcheti⁴, A. Gupta⁵, S.S. Ramalingam⁶, K. Kelly⁷, K. Sandler⁸, H. Borghaei⁹, A. Madabhushi¹

¹Wallace H. Coulter Department of Biomedical Engineering at Emory University and Georgia Tech, Emory University, Atlanta, GA, USA; ²Department of Biostatistics, Brown University, Providence, RI, USA; ³Center for Statistical Sciences, Brown University, Providence, RI, USA; ⁴Thoracic Oncology, NYU Langone Medical Center and School of Medicine, New York, NY, USA; ⁵Department of Radiology, Case Western Reserve University / University Hospitals, Cleveland, OH, USA; ⁶Department of Hematology and Medical Oncology, Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA, USA; ⁷Hematology Oncology Department, UC Davis Comprehensive Cancer Center, Sacramento, CA, USA; ⁸Department of Radiology, Vanderbilt University Medical Center, Nashville, TN, USA; ⁹Hematology/Oncology Department, Fox Chase Cancer Center - Main Campus, Philadelphia, PA, USA

Background: EA5163 is a phase 3 randomized trial investigating first-line mono immunotherapy (IO) or IO with chemotherapy in advanced non-small cell lung cancer (NSCLC). Previous work has retrospectively validated a CT based radiomic signature of tumor associated quantitative vessel tortuosity (QVT) for prognosticating outcomes in IO treated NSCLC (Alilou et al. Sci Adv 2022). In this study, we report initial results of evaluating QVT in assessing IO related treatment changes in EA5163, specifically associating changes in QVT (Δ -QVT) with changes in tumor volume (Δ -TV) between baseline (B) and 6 week post-treatment scans (TP1).

Methods: CT scans from 29 patients enrolled in EA5163 with measurable lung lesions at B and TP1 were included. Vascularity was automatically segmented in a 25mm region around manually annotated tumors in CT using MATLAB R2022b. 84 QVT features measuring tortuosity, curvature and angle statistics were extracted. No information on outcomes, survival, or treatment arm assignment was available due to the ongoing nature of the trial. Spearman correlation and false discovery rate corrected p-values were computed to associate Δ -QVT with Δ -TV in an unsupervised setting. A risk score (QRS) was generated as the output of a classifier previously trained on 5 Δ -QVT features prognostic of RECIST v1.1 response.

Results: 14 Δ -QVT features were significantly ($p < 0.05$) correlated with Δ -TV with an absolute range of 0.45 and 0.76. 4 features in the QRS signature were significantly correlated with Δ -TV (Table). QRS based stratification showed net decrease in TV in low risk patients, whereas a net increase in high-risk patients at TP 1 ($p < 0.001$). The Δ -QVT features were not associated with baseline tumor volume.

Table: 175P

Feature	Correlation	p-value
Vessel volume	0.71	<0.001
1 st bin of tortuosity histogram	0.65	0.002
12 th bin of angles histogram	0.61	0.004
2 nd bin of curvature histogram	0.65	0.002
Vessel volume to tumor volume ratio	-0.48	0.053

Conclusions: This study demonstrated an association between CT derived radiomic biomarker Δ -QVT and Δ -TV in an ongoing prospective clinical trial. Validation of these results upon interim analyses or trial completion is warranted.

Clinical trial identification: NCT03793179.

Legal entity responsible for the study: The authors.

Funding: This study was coordinated by the ECOG-ACRIN Cancer Research Group (Peter J. O'Dwyer, MD and Mitchell D. Schnall, MD, PhD, Group Co-Chairs) and supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under the following award numbers: U10CA180820, U10CA180794, UG1CA233234, UG1CA233270, and UG1CA233247. This study was also supported by an NCI grant NIH 1 R01CA257612-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure: V. Velcheti: Financial Interests, Personal, Advisory Board: ITeos Therapeutics, Bristol Myers Squibb, Merck, AstraZeneca/MedImmune, GSK, Amgen, Elevation Oncology, Merus, Taiho Oncology; Financial Interests, Institutional, Research Funding: Genentech, Trovogene, Eisai, OncoPlex Diagnostics, Alkermes, NantWorks, Genoptix, Altos BioScience, Merck, Bristol Myers Squibb, Atreca, Heat Biologics, Leap Therapeutics, RSIP Vision, GSK. A. Gupta: Financial Interests, Personal, Other, Consultancy for future development of AI tools on chest radiographs: GE Healthcare; Financial Interests, Personal, Stocks/Shares, Own shares of the company presently valued at 56 dollars: Picture Health; Financial Interests, Institutional, Funding, Research support for ongoing AI projects around chest radiography: GE Healthcare. S.S. Ramalingam: Financial Interests, Personal, Other, Editor in Chief, Cancer journal: American Cancer Society; Financial Interests, Research Grant: Merck, AstraZeneca, Advaxis, BMS, Amgen, Takeda, Genmab, GSK. K. Kelly: Financial Interests, Personal, Royalties, Primary reviewer for the small cell lung cancer chapter: UPTODATE. H. Borghaei: Financial

Interests, Personal, Advisory Board: BMS, Genentech, Eli Lilly, Merck, EMD-serono, AstraZeneca, Novartis, Genmab, Regeneron, Amgen, Takeda, PharmaMar, Jazz Pharma, Mirati, Daiichi, Guardant, Natera, Oncocyte, BeiGene, iTeo, Boehringer Ingelheim, Puma, BerGenbio, Janssen; Financial Interests, Personal, Other, Training discussion: Pfizer; Financial Interests, Personal, Other, DSMB: Novartis; Financial Interests, Institutional, Other, Clinical trial support: BMS, Amgen; Financial Interests, Personal, Stocks/Shares, Options for scientific advisory role: Sonnetbio; Financial Interests, Personal, Stocks/Shares, Options for scientific advisory board: Nucleai, Inspira (Rgenix); Financial Interests, Institutional, Trial Chair, Investigator initiated trial support: BMS; Financial Interests, Institutional, Coordinating PI, Investigator initiated trial support: Amgen, Lilly; Financial Interests, Personal and Institutional, Trial Chair, Chair steering committee: Mirati; Financial Interests, Personal and Institutional, Steering Committee Member: Amgen, AstraZeneca; Financial Interests, Personal, Steering Committee Member, Also trial support: BMS; Other, DSMB: Novartis, Takeda, Incyte, Springworks. A. Madabhushi: Financial Interests, Personal, Advisory Board, Serve on SAB and consult: SimbioSys; Financial Interests, Personal, Advisory Board: Aliforia, Picture Health; Financial Interests, Personal, Full or part-time Employment: Picture Health; Financial Interests, Personal, Ownership Interest: Picture Health, Elucid Bioimaging, Inspirata Inc; Financial Interests, Personal, Royalties: Picture Health, Elucid Bioimaging; Financial Interests, Institutional, Funding: AstraZeneca, Bristol Myers Squibb, Boehringer Ingelheim, Eli Lilly. All other authors have declared no conflicts of interest.

<https://doi.org/10.1016/j.annonc.2023.09.2899>

176P Enhancing immunotherapy response prediction via multimodal integration of radiology and pathology deep learning models

M. Ligerio¹, O.S.M. El Nahhas², O. Prior Palomares¹, V. Navarro³, I. Sansano⁴, G. Serna⁵, S. Mauchanski⁵, R.D.A. Toledo⁶, R. Dienstmann⁷, S. Ramon y Cajal⁴, E. Garralda⁸, P.G. Nuciforo⁹, J.N. Kather⁹, R. Perez Lopez²

¹Radiomics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain; ²Clinical AI, EKfZ - Else Kröner Fresenius Zentrum für Digitale Gesundheit, Dresden, Germany; ³Oncology Data Science (ODysSey) Group, Vall d'Hebron Institute of Oncology (VHIO)-Cellex Center, Barcelona, Spain; ⁴Pathology Department, Vall d'Hebron University Hospital, Barcelona, Spain; ⁵Molecular Oncology Group, Vall d'Hebron Institute of Oncology (VHIO)-Cellex Center, Barcelona, Spain; ⁶Medical Oncology Department, Vall d'Hebron Institute of Oncology (VHIO)-Cellex Center, Barcelona, Spain; ⁷Oncology Data Science (ODysSey) Group group, Vall d'Hebron Institute of Oncology, Barcelona, Spain; ⁸Early Drug Development Unit, Hospital Universitario Vall d'Hebron, Barcelona, Spain; ⁹Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO)-Cellex Center, Barcelona, Spain

Background: The growing need for better predictive biomarkers in immunotherapy has driven the development of AI models across various fields. Despite this progress, most studies focus on single modalities, such as pathology or radiology. This study emphasizes the importance of multimodal biomarkers, combining radiology and pathology data, to optimize treatment outcomes and advance personalized medicine in immunotherapy.

Methods: We developed weakly-supervised deep learning (DL) models for CT and PDL1-stained immunohistochemistry (IHC) images. We trained a CT-based model to predict immunotherapy clinical benefit (PFS > 5 months) in a pan-cancer cohort from VHIO (n=369). Simultaneously, we trained a DL method to predict PD-L1 status from IHC images in an open source NSCLC cohort (n=233), validated in a pan-cancer cohort from VHIO (n=106). The model predictions of both modalities were aggregated and utilized for patient stratification in immunotherapy response. In the overlapping subset with both modalities' data (n=75), we explored the association between PFS and the combination of predicted clinical benefit by CT and predicted PDL1 status by IHC using Cox regression and log-rank test.

Results: The CT-based and IHC-based DL models exhibited effective performance in predicting PFS for immunotherapy response in the pan-cancer cohort, with hazard ratios (HR) of 0.62 [95% CI 0.5-0.78], $p < 0.05$ and 0.66 [95% CI 0.44-1], $p < 0.05$, respectively. In the multi-modal data subsets, patients classified as clinical benefit prediction by the CT-based model and PDL1 high by the IHC-based model (n = 20) exhibited better immunotherapy response (HR 0.44 [0.25-0.78], $p < 0.05$) with longer PFS (median PFS 4.08 [2.70-10.91] months) compared to the rest (median PFS 1.81 [1.68-2.96] months).

Conclusions: Integrating radiology and pathology data improves response prediction compared to single modality models, emphasizing the potential of a multimodal approach for better patient stratification in immunotherapy.

Legal entity responsible for the study: The authors.

Funding: CRIS Foundation Talent Award (TALENT19-05).

Disclosure: R. Dienstmann: Financial Interests, Personal, Invited Speaker: Amgen, AstraZeneca, Boehringer Ingelheim, Ipsen, Libbs, Lilly, Merck Sharp & Dohme, Roche, Sanofi, Servier, GSK, Takeda, Janssen, Foundation Medicine; Financial Interests, Personal, Advisory Board: Bayer, Roche; Financial Interests, Personal, Full or part-time Employment, Oncoclinics is a private healthcare provider in Brazil. I work part time as Medical Director of the Precision Medicine and Big Data Initiative. We develop molecular tests (pathology and genomics) that are offered to patients treated in the organisation as part of support programs sponsored by pharmaceutical companies and I coordinate research activities with real-world clinico-genomics cohorts: Oncoclinics; Financial Interests, Personal, Stocks/Shares: Trialing; Financial Interests, Personal, Research Grant: Merck. E. Garralda: Financial Interests, Personal, Advisory Board: Roche, Ellipses Pharma, Boehringer Ingelheim, Janssen Global Services, Seattle Genetics, Alkermes, Thermo Fisher, MabDiscovery, Anaveon, Hengrui, F-Star Therapeutics, Sanofi, Incyte; Financial Interests, Personal, Invited Speaker: MSD, Lilly, Roche, Thermo Fisher, Novartis, Seagen; Financial Interests, Personal, Full or part-time Employment: NEXT Oncology; Financial Interests, Institutional, Funding: Novartis, Roche, Thermo Fisher, AstraZeneca, Taiho; Financial Interests, Institutional, Research Grant: BeiGene, Janssen. P.G. Nuciforo: Financial Interests,