Artificial intelligence for imaging in immunotherapy

Doctoral thesis

ARTIFICIAL INTELLIGENCE FOR IMAGING IN IMMUNOTHERAPY

Stefano Trebeschi

2021

ARTIFICIAL INTELLIGENCE FOR IMAGING IN IMMUNOTHERAPY

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. R.M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public
on Monday 22nd of March, at 16.00 hours

by

Stefano Trebeschi

Promotor

Prof. Regina G.H. Beets-Tan, MD PhD

Prof. Hugo Aerts, PhD

Assessment Committee

Prof. dr. Dirk de Ruysscher (chairman)

Prof. dr. Paul Hofman

Dr. Nicky Peters

Prof. dr. Paul Baas
The Netherlands Cancer Institute & Leiden University

Prof. dr. Klaus Maier-Hein
German Cancer Research Center

The publication of this thesis was financially supported by the Netherlands Cancer Institute and Maastricht University.

Cover	Elmar van Zyl | *www.elmar.design*
Production	Gildeprint Enschede | *www.gildeprint.nl*
ISBN	978-94-6419-161-5

Contents

1
Introduction

Immune checkpoint inhibitors have provided improved response rates and prolonged overall survival in advanced metastatic cancer patients. Several clinical trials have demonstrated more favourable outcomes compared to standard therapy, with the largest body of evidence being in melanoma [Lar+18; Web+15], non-small cell lung cancer [Vok+18; Bra+15; Car+17], renal cell carcinoma [Alb+20; Mot+20], and head and neck carcinoma [Sab+19; Coc+19; Pai+19; Fer+18]. Additionally, the FDA has authorized the use of pembrolizumab in microsatellite instability high tumours (MSI-H) [Zha+19b; Luc+19] — i.e. tumours that had a genetic predisposition to mutation resulting from impaired DNA repair mechanisms [Ion+93; TBS93].

Despite the success of these immunotherapies, there is still a substantial number of patients who do not benefit from the treatment [Lar+18; Web+15; Vok+18; Bra+15; Alb+20; Mot+20; Fer+18]. Biomarkers able to identify these responders would ultimately improve treatment outcomes — both in terms of prolonged overall survival, as well as reduced therapy-induced toxicity — while simultaneously helping to contain the costs of these expensive anti-cancer therapies [GWA16; HCC19a; Voo+17; Tar+16].

Immunotherapy biomarker research has been focusing so far on biological markers, often extracted from invasive tumour-tissue biopsies. These include levels of infiltration of lymphocytes (i.e. white blood cells) in the tumour [Zit+17; He+17], or more general markers of inflammation [Aye+17], genetic mutations [McG+16], among others [Ma+16; Men+15; Ker+15]. Their values, however, depend on the biopsied lesion, and on how generalizable they can be to the total tumour burden [Ram+20]. In advanced stage metastatic patients (the most likely target population for immunotherapy), multiple lesions across the body are likely to have developed distinctively [Cun+15; TS16] biological profiles, potentially as a result of their intrinsic microenvironment or the milieu of the organ where they are situated [Cun+15; TS16]. Due to the invasiveness of the biopsy procedure [Ove+13] in an already fragile patient cohort of advanced-stage cancer, it is not always feasible to biopsy multiple locations to average out inter-lesional het-

erogeneity, nor is it possible to perform multiple biopsies of the same location over time to evaluate response to treatment.

Different alternatives are being explored to overcome these limitations, including whole-body imaging. Routine radiological imaging has already emerged as a fundamental tool in the clinics for detection, characterisation, and monitoring of disease [Hri11]. It is non-invasive, and it is broad — i.e. able to capture whole-body information in one single measurement. These characteristics led imaging to become the default tool for the diagnosis [Mek+18] and follow-up of advanced-stage cancer patients receiving immunotherapy [Sey+17a; The+00], playing an irreplaceable role in treatment planning [Ner+20]. Currently, the read-out and interpretation of radiological imaging is done by a medical professional via visual assessment [Sey+17a; The+00]. They report the presence and location of cancer lesions, their extent or size, their evolution in time [Sey+17a; The+00; Sch+16; Eis+09; Org+79; Hay+77], and the presence of other clinically-relevant non-cancer conditions [Mek+18]. The most common tool used in the clinics is the *Response Evaluation Criteria in Solid Tumour* (RECIST) [Sey+17a; Sch+16]. This prescribes the follow-up of the cumulative in-plane diameter of maximum 5 target lesions. Depending mainly on the changes in the cumulative diameter, the response is classified as complete or partial, stable, or progressive disease; with minor changes made to accomodate for patients receiving immunotherapy. While this method is time-efficient (required aspect in the busy radiological practice), it lacks to address any prognostic factor that is not related to total tumor growth or shrinkage — among other limitations [VS13].

There is a growing body of evidence that more clinically-valuable information can be extracted from radiological imaging [Aer+14; Hos+18; Bi+19a; AH16]. It is hypothesized that underlying biological processes (some of which regulate response to therapy or survival) are reflected in imaging patterns on radiological scans [Aer+14; Aer16; Par+15; OCo+17]. Computational methods, able to analyze and identify patterns in high dimensional imaging data, can serve this purpose. Two main techniques are relevant to this end: radiomics

[Aer+14] and artificial intelligence [Hos+18]. Radiomics entails all the computer algorithms for data characterisation in medical imaging that aim to extract imaging features, each feature describing anatomical or functional characteristics of, e.g. a cancer lesion [Aer+14; Par+15; GKH16; Gri+17a]. The main advantage relies on the usage of computer algorithms, which allow for the disentanglement and identification of imaging (pixel) patterns that would not be able to be assessed by the naked eye. Once imaging patterns have been extracted, these need to be linked to a clinically-relevant variable, e.g. the outcome of the treatment. Artificial intelligence (AI) is used to this end [Hos+18; Bi+19a]. AI is the collection of all the computational methods which, through a learning process on plain or processed data, are able to decipher patterns and link variables together, ultimately extrapolating knowledge for "reasoning", or inference [RN03].

This thesis aims to develop and assess the value of AI-based radiomics methods on imaging data as clinical support systems for response evaluation, prediction and prognostication of advanced-stage cancer patients receiving immune checkpoint inhibitors.

This was formulated in two main research questions:

- Can AI-radiomics methods on radiological imaging detect and extract quantitative imaging patterns of single lesions that are related to lesion biology and lesion response to checkpoint inhibitors?
- Can AI-radiomics methods be developed to decipher prognostic morphological patterns at whole-body level in advanced stage metastatic patients receiving checkpoint inhibitors?

The outline of this thesis follows the research questions. It begins with the question of imaging features of single lesions for lesion assessment and response prediction. Chapter 2 evaluates predictive lesion-level radiomic features for the prediction of individual lesion response. Chapter 3 evaluates the diagnostic performance of radiomic features

for the diagnostic profiling of lung lesions, and gives an overview on how single lesion profiling can be correlated with long term outcome. The research question is then amplified to encompass the development of whole-body AI-based methods for the prognostication of metastatic patients receiving immunotherapy, overcoming in this manner the limitations of single lesion analysis. Chapter 4 introduces the concept of prognostication through AI-based image monitoring of chest imaging in NSCLC. Chapter 5 expands to thoracoabdominal imaging in urothelial cancer patients and formalizes it as the Prognostic AI-Monitor (PAM). Chapter 6 investigates PAM in brain metastasis, extending our findings to the whole body. This thesis concludes with Chapters 7 and 8, stating the future developments of artificial intelligence research in the clinics, and in cancer immunotherapy trials.

2

Lesion response prediction to immunotherapy

Stefano Trebeschi et al. "Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers". In: *Ann. Oncol.* 30.6 (2019), pp. 998–1004.

Abstract

Introduction Immunotherapy is regarded one of the major breakthroughs in cancer treatment. Despite its success, only a subset of patients responds — urging the quest for predictive biomarkers. We hypothesize that Artificial Intelligence (AI) algorithms can automatically quantify radiographic characteristics that are related to and may therefore act as non-invasive radiomic biomarkers for immunotherapy response.

Patients and Methods In this study, we analyzed 1055 primary and metastatic lesions from 203 patients with advanced melanoma and non-small cell lung cancer (NSCLC) undergoing anti-PD1 therapy. We performed a AI-based characterization of each lesion on the pretreatment contrast-enhanced CT imaging data to develop and validate a non-invasive machine learning biomarker capable of distinguishing between immunotherapy responding and non-responding. To define the biological basis of the radiographic biomarker, we performed gene-set enrichment analysis in an independent dataset of 262 NSCLC patients.

Results The biomarker reached significant performance on NSCLC lesions (up to 0.83 AUC, $p<0.001$) and borderline significant for melanoma lymph nodes (0.64 AUC, $p=0.05$). Combining these lesion-wide predictions on a patient level, immunotherapy response could be predicted with an AUC of up to 0.76 for both cancer types ($p<0.001$), resulting in a one year survival difference of 24% ($p=0.02$). We found highly significant associations with pathways involved in mitosis, indicating a relationship between increased proliferative potential and preferential response to immunotherapy.

Conclusions These results indicate that radiographic characteristics of lesions on standard-of-care imaging may function as non-invasive biomarkers for response to immunotherapy, and may show utility for improved patient stratification in both neoadjuvant and palliative settings.

2.1 Introduction

Cancer immunotherapy has made promising strides as a result of improved understanding of biological interactions between tumor cells and the immune system. Both the EMA and the FDA have approved anti-PD1 antibodies to treat melanoma or non-small cell lung cancer (NSCLC) patients with unresectable or metastatic disease, which progressed under platinum-based chemotherapy or display high expression of PD-L1 [U Sb; U Sa; Eurb; Eura] — with overall response rates of 44% and 32% in first and second line in melanoma [Wol+15; Web+15] and 19% in second line in lung cancer [Bor+15; Bra+15; Her+16]. Unlike traditional cancer treatment, anti-PD1 antibodies potentiate the anti-tumor immune response.

Despite their remarkable success, clinical benefit remains limited to only a subset of patients [Hod+10]. As immunotherapy is expensive and could bring toxicity, there is a need to stratify patients according to likely benefit prior to therapy. Different biomarkers have been investigated with variable success, such as levels of PD-L1 [Ma+16; Men+15; Ker+15], presence of tumor infiltrating lymphocytes [Zit+17; He+17], genetic mutations [McG+16; Riz+15; Hel+18a], and inflammatory cytokines [Aye+17].

Recent emergence of quantitative imaging biomarkers provide promising opportunities. Unlike traditional biopsy-based assays that represent only a sample of the tumor, images reflect the entire tumor burden, providing information on each cancer lesion with a single non-invasive examination. This is of particular importance in immunotherapy, where different lesions can have different microenvironments potentially leading to heterogeneous response patterns [Whi08]. Previously, radiolabeled anti-PD1 antibodies were used to visualize specific immunological expressions [Wu09].

Computational imaging approaches originating from Artificial Intelligence (AI) have achieved impressive successes in automatically quantifying radiographic characteristics of tumors [Hos+18].

AI-based characterization on radiology is referred to as "radiomics" and can provide more detailed characterization than possible by eye [Hos+18; Aer+14; AH16]. Radiomics-based biomarkers have shown success in different tumor types [Cor+16; Kir+17; Fav+17; Par+15; Kic+16; Pra+16; Li+16]; but to the best of our knowledge, there is no evidence yet in immunotherapy. Tumor morphology, visualized on imaging, is likely influenced by several aspects of tumor biology. We hypothesize that a set of morphological characteristics, quantified by radiomics, are related to and may therefore act as predictive markers.

In this study, we analyzed all visible cancer lesions to evaluate the potential predictive value of CT-derived radiomic biomarkers in metastatic NSCLC and melanoma patients receiving immunotherapy. A biologic evaluation was performed in an independent validation set of surgical NSCLC patients with imaging and gene-expression data.

2.2 Material and methods

2.2.1 Immunotherapy dataset

Patients with metastatic melanoma or NSCLC receiving 3mg/kg/2weeks of anti-PD1 at the Netherlands Cancer Institute (NKI) between 2014 and 2016 were retrospectively analyzed. Contrast-enhanced computed tomography (CE-CT) scans were acquired before (baseline) and around 12 weeks after start of treatment (follow-up). The study protocol was approved by the Medical Ethics Committee and Board of Directors of the NKI and informed consent was waived.

Image acquisition protocol. The CT scans were performed by either covering the chest (n=86) or covering the chest and abdomen (n=117) using multi-slice CT equipment (Toshiba Aquilion CX, Minato, Tokyo, Japan; Siemens Somatom Sensation Open, Erlangen, Germany) with

a tube voltage of 120 kVp, slice thickness of 1 mm, and in-plane resolution of 0.75 x 0.75 mm. The bolus injection was performed at 3 ml/s (Omnipaque 300, GE Healthcare, Chicago, Illinois, US) not prewarmed, with a total amount based on the patient weight + 40 cc (minimum of 90 cc and maximum of 130 cc) followed by a saline flush of 30 cc. The chest CT examinations were performed 40 seconds after contrast injection, whereas the chest and abdomen examinations were performed at 70 seconds.

2.2.2 Genomics dataset

To provide biological validation, we evaluated an independent, dataset of surgical NSCLC patients between 2006 and 2009 treated at the H. Lee Moffitt Cancer Center. Pre-surgical CE-CT (within 60 days of diagnosis) and gene expression data was available for 262 patients. The University of South Florida IRB approved and waived informed consent (IRB#16069); in accordance with HIPAA (more information in the original publication [Gro+17]).

Image acquisition protocol. Contrast-enhanced CT scans were acquired 60 days within diagnosis, as part of the Thoracic Oncology Program protocol, of the L. Lee Moffitt Cancer Center (Tampa, Florida, USA). Gene expression of 60,607 probes was measured on a custom Rosetta/Merk Affymetrix 2.0 microarray chipset (HuRSTA˙2a520709.CDF, GEO accession number GPL15048) by the Moffitt. The University of South Florida IRB institutional review board approved and waived the informed consent requirement (IRB#16069); data were collected and handled in accordance with the Health Insurance Portability and Accountability Act. Informed consent for gene expression collection was written and oral. For acquisition of imaging and clinical data USF IRB approved protocol (IRB#108426) provided a waiver of informed consent.

2.2.3 Chemotherapy dataset

To study the specificity of the radiomic biomarker for immunotherapeutic response prediction, we retrospectively collected a cohort of 39 patients with stage IV NSCLC treated with neoadjuvant chemoradiotherapy at NKI between 2012 and 2016 (IRBd18079).

Image acquisition protocol. The CT scans were performed covering the chest and abdomen (n=39) using multi-slice CT equipment (Toshiba Aquilion CX, Minato, Tokyo, Japan; Siemens Somatom Sensation Open, Erlangen, Germany) with a tube voltage of 120 kVp, slice thickness of 1 mm, and in-plane resolution of 0.75 x 0.75 mm. Specific of the scanning protocols were identical to the immunotherapy dataset.

2.2.4 Imaging data and lesion segmentations

Experienced readers manually delineated lesions on baseline and follow-up scans. Target lesions were defined as any tumor that was well-demarcated on both baseline and follow-up with diameter $\geq$5 mm. The inclusion criteria were: availability of CE-CT BL and FU and, presence of measurable target lesions at baseline. Measurable lesions were defined as any tumor lesions (primary or metastatic lesions) whose entire border could be identified on both BL and FU scans, as our radiomic feature extraction pipeline requires segmented region of interest to extract features Lesions that disappeared in the FU were flagged as complete response. Lesions that could not be accurately discriminated from surrounding tissues (e.g. lung nodule within atelectasis), with ill-defined borders (e.g. lung lesions adjacent to atelectasis) and lesions which could not be tracked down from other adjacent tumour lesions at baseline or follow-up CTs (e.g. confluent metastases) were not delineated and excluded. Lesions poorly visualized because of the presence of imaging artefacts (e.g. scattering, motion or breathing artefacts) were excluded as well. Examples are shown in Figures 2.1a-b.

2.2.5 Response kinetics

To assess the effects of mixed response, we performed a lesion-per-lesion assessment of relative change in diameter between baseline and follow-up, using RECIST criteria. Furthermore, in patients with >1 lesion, we classified response patterns on a patient basis as mixed for patients presenting both responding and progressive lesions and uniform for patients presenting only responding or progressive lesions (irrespective of stable lesions). This setup allows for the characterization of overall tumor burden.

2.2.6 Radiographic differences between responding and progressive lesions

To generate radiomic sequences for each lesion at baseline, a set of radiomic features was defined [G217] (see Figure 2.1e). Radiomic features of responding responding and progressive lesions were directly compared to identify differences. To reduce redundancy, ten complementary features were selected using unsupervised feature selection [Yin08]. Statistical significance was assessed using generalized mixed-effect models — controlling for patient, tumor type and organ. False discovery rate (FDR) was at 10% to correct for multiple comparisons.

2.2.7 Radiomic biomarkers to predict immunotherapy response of cancer lesions

To assess the performance of the radiomic biomarker, we developed a machine learning model [Cox58; B201; Ols+16]. We trained the model on all lesions (i.e. progressive, stable and responding) to discern progressive disease. The dataset was divided into training, tuning, and testing sets based on patient identifiers. The training set was used to model data distributions. The tuning set was used during training to control for overfitting. The test set was used for independent evaluation (see Figure 2.1f). Mann-Whitney-U test was used for statistical

testing of AUC curves, one-sided McNeils test was used to test if the radiomic biomarker was outperforming volume and maximum diameter, and log-rank test was used for statistical testing of survival performance.

To test for radiomic association with molecular pathways, Spearman's rank correlation coefficient was used. Pathways were then ranked by $-log10(p)$, where p is the correlation p-value, and put into a pre-ranked gene set enrichment analysis (GSEA) algorithm [Sub+05] version 2.0.14 on the C2 collection version Molecular Signature Database (MSigDB)[Lib+11].

Feature extraction. To reduce the influence of outlier intensity values in the image, the volume was clipped between -1000 HU and 3000 HU. Radiomic features were extracted from original images as well as from different image transformations including five Laplacian of Gaussian filters (σ = 1.0, 2.5, 5.0, 7.5, 10.0 mm), eight wavelets decompositions, and four non-linearities (exponential, square, square root and logarithm). We also repeated the extraction over three different scales, each defined by a set of radiomic parameters: (1) a fine scale with 1 mm isotropic resolution and 1HU bin width, (2) a medium scale with 3 mm isotropic resolution and bin width of 5HU and (3) coarse scale with 5 mm isotropic resolution and bin width of 25 HU. In this way, the algorithm can choose the best radiomic extraction parameters and/or their combination. Features which resulted in invalid values for more than one lesion were dropped.

Dataset preparation. The entire dataset was divided into train, validation and test set based on patient identification numbers (pid). Patients whose pid was divisible by three were assigned to the train set, those whose (pid - 1) was divisible by three were assigned to the validation set, and those whose (pid - 2) was divisible by three were assigned to the test set.

Classifier pool. The first group is composed by three linear classifiers based on logistic regression (LR) models [Cox58], each differentiated by a different feature selection method: (1) unsupervised result-

ing from PCA, (2) supervised resulting from wrapper feature selection (WFS), or (3) no feature selection. Similarly, we defined a second group of non-linear classifiers based on random forests (RF) [B201]. Finally, we generated two additional classifiers via genetic evolution (GEN-1 and GEN-2). Each classifier was trained using 2-fold cross validation and optimized via sequential model based optimization.

Training strategy. Each classifier is trained on the training set using a 2-fold cross validation procedure. To prevent the model from learning to recognize patients rather than the actual lesion-wise classification task, we enforced cross validation at a patient level, avoiding the distribution of lesions of the same patient across different folds. Once trained, the model is evaluated in on the test set to check for under- or overfitting, and model selection.

Classifier optimization. Each classifier comes with a set of tunable parameters, i.e. hyperparameters. We made use of a machine learning procedure, a.k.a. sequential model based optimization (SMBO), to tune the hyperparameters of each classifier. SMBO procedure is an iterative procedure, where at each iteration the performance is modelled as a function f of the hyperparameters. The search of the optimal hyperparameters is achieved via optimization of a criterion on f. We chose the commonly used Expected Improvement (EI). Parzen estimators were used to approximate the function f.

Hyperparameter space. Logistic regressions had only one tunable hyperparameter representing the weight of the L_2 regularization coefficient. Random forests had four hyperparameters: the max depth of the trees (d), the minimum number of samples in each leaf (mL), the minimum number of samples (mS) and minimum Gini impurity in each split (G). The hyperparameters of the genetic classifiers depend on the specific search result. Finally, wrapper feature selectors had one hyperparameter k, indicating the number of top-performant features selected. For each classifier, we selected the set of hyperparameters resulting in the highest AUC.

Model selection. Once completed, the optimization procedure results in a set of eight trained classifiers. We selected the final classifier by comparing their performance on the validation set. The classifier that achieved the highest AUC score was selected as candidate solution. All algorithms, except for wrapper random forests and the second genetic evolution classifier, reported a certain degree of overfitting quantified by a lower accuracy on the validation set w.r.t the one reported on the training set. During training, all algorithms perform similarly between the two folds of cross validation, except second genetic evolution classifier which showed higher variance. Our choice of using wrapper random forests as candidate classifier was motivated by the fact that this configuration reached the highest performance with the least amount of overfitting.

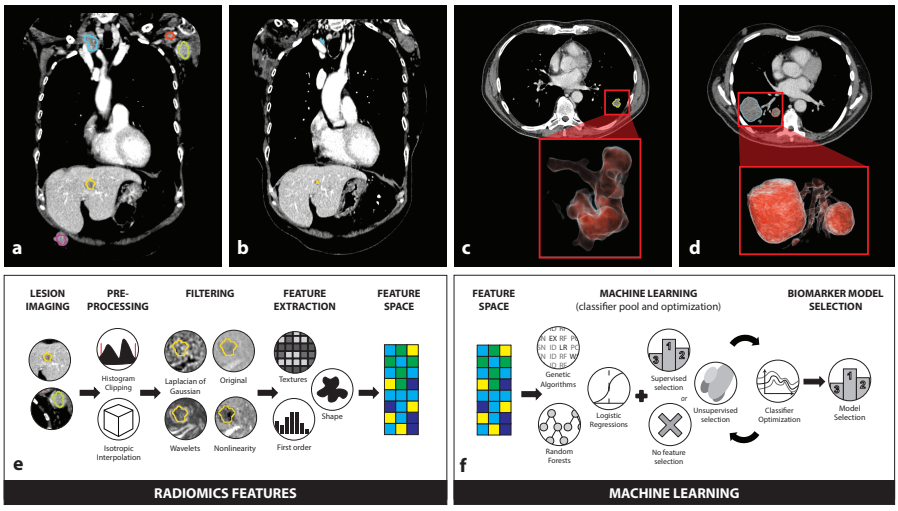

Figure 2.1: **(a)** Baseline contrast-enhanced CT scan of melanoma patient presenting with metastases in the liver and lymph nodes in the axilla and subclavicular area **(b)** Follow-up scan of the same patient showing complete response in the axillary region and partial response of the lesions in the liver and neck **(c)** Baseline CT scan of a NSCLC patient presenting lesion in the left lung, that showed progression at a later FU CT (not shown) **(d)** Baseline CT scan of a melanoma patient presenting lesions in the right lung, that showed response at a later FU CT (not shown) **(e)** Schematic representation of the radiomics feature extraction process **(f)** Schematic of the machine learning process.

2.3 Results

2.3.1 Immunotherapy response kinetics

To assess immunotherapy response kinetics, 203 (123 NSCLC, 80 melanoma) patients were analyzed with a total of 1055 target lesions. Lesions were similarly distributed between NSCLC (n=572, 54%) and melanoma (n=483, 46%). The most common lesion sites were lung (n=359, 34%), lymph nodes (n=312, 30%), and liver (n=212, 20%). Most lesions (n=746 vs 309, chi-square-test $p<0.01$) showed either stable (n=395) or partial response (n=351).

Melanoma lesions showed better overall response than NSCLC (40% vs 27% responding, $p<0.01$; 23% vs 34% progression, $p<0.01$). This trend was more evident for lung lesions, where we observed progression in NSCLC (39% vs 14%, $p<0.01$) and response in melanoma (48% vs 26%, $p<0.01$). Hepatic melanoma lesions showed response in comparison with NSCLC (22% vs 36%, $p=0.04$). Examples are shown in Figures 2.1c-d.

Comparing per-patient response patterns in both cancer types, we observed that 23% (n=47) showed uniform response, 27% (n=55) uniform progression, and 22% (n=45) mixed response. The remaining 28% (n=56) of the patients did not have multiple target lesions or presented only stable lesions. Significantly higher survival rates were seen in uniform response (log-rank test, $p<0.01$). This was evident in melanoma (log-rank-test, $p<0.01$), while in NSCLC, despite similar trends, did not reach significance ($p=0.08$). Per-patient response kinetics are shown in Figure 2.2a. Kaplan-Meier curves are shown in Figures 2.2b-d.

2.3.2 Radiographic differences between responding and progressive lesions

To investigate radiographic differences between responding and progressing lesions, we compared their radiomic features (see Table 2.1). Among the most common locations (lung, lymph nodes,

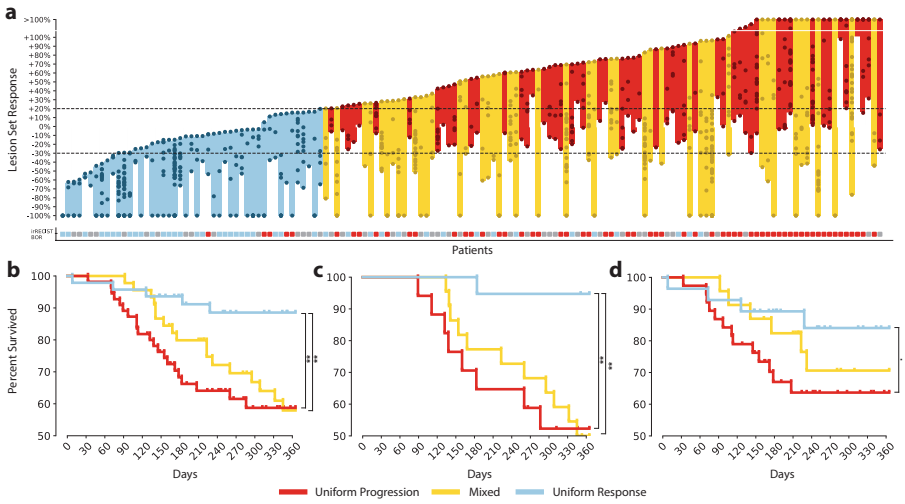

Figure 2.2: **(a)** Response kinetics curve depicting individual lesion responses (as dots) on a patient-to-patient basis **(b)** 1 year survival plot for all analyzed patients, (c) for melanoma patients only, **(d)** for NSCLC patients only.

liver and adrenal gland), responding lesions presented higher levels of irregular patterns (Wavelet.HLH_GLSZM_ZoneEntropy, Kenward-Roger-test p=0.007) with more compact, spherical profiles (SurfaceVolumeRatio, p=0.01). Subanalysis on location revealed increased values of morphological heterogeneity in hepatic, nodal, and splenic lesions associated with response (p<0.02). Of the most common NSCLC lesions, similar trends for morphological heterogeneity were seen at organ level in pulmonary and hepatic lesions, as well as lymph nodes also characterized by the presence of hypodense regions (p=0.007). No significance was observed in primary NSCLC tumors. Among most common melanoma lesions greater morphological heterogeneity showed association with response (GLCM˙DifferenceEntropy, p=0.006). Similar trends for morphological heterogeneity were seen but lower sample numbers did not allow to pass the patient correction.

2.3.3 Radiomic biomarker to predict immunotherapy responding and stable lesions

To assess the performance of radiomics to rule out progression, we used machine learning to develop a single radiomic biomarker with 133 patients in the discovery set and 70 patients in test (see Table 2.2). A random forest with wrapper feature selection was used to develop radiomic biomarkers based on the performance in the discovery set and were validated on the independent test set.

In NSCLC, radiomic biomarker from pulmonary (0.83 AUC, Mann-Whitney-U-test $p<0.001$) and nodal metastases (0.78 AUC, $p<0.001$) showed significant performance. Satisfactory performance was observed in NSCLC primary tumors (0.79 AUC, $p=0.05$), hepatic (0.75 AUC, $p=0.13$) and adrenal lesions (0.70 AUC, $p=0.18$) but did not reach significance mostly due to the low number of samples.The model performed poorly on both pulmonary and hepatic melanoma lesions (0.55 AUC). Despite these results, a trend toward significance is shown in nodal metastases (0.64 AUC, $p=0.05$) (see Figure 2.3a). Evaluation of the radiomic biomarker on all 303 lesions within the test dataset resulted in significant predictive performance (0.66 AUC, $p<0.01$; see Table 2.2).

By combining predictions made on individual lesions, it is possible to do a pre-treatment patient-wise prediction of immunotherapy response (see Methods). Significant performances were observed to predict OS for both tumour types (0.76 AUC for all patients, $p<0.01$; 0.76 AUC for NSCLC patients, $p<0.01$; 0.77 AUC for melanoma patients, $p<0.01$; see Figure 2.3b), with a significant survival difference at 1-year of 25% (77% vs 52%, log-rank-test, $p=0.02$; see Figure 2.3c). Interestingly, in melanoma patients, we observed significant performance to predict OS and response, despite the lower performance on a lesion level.

This radiomic immunotherapy response biomarker could not significant predict overall survival in patients treated with

neoadjuvant chemoradiotherapy (p=0.07), nor in terms of overall patient response (AUC=0.63; p=0.09). In terms of lesion response, the biomarker was inversely correlated to response of lung lesions in non-immunotherapy patients (n=61, AUC=0.70, p=0.04), but did not show any significant predictive value in the remaining nodal (n=61, AUC=0.59, p=0.24) and liver lesions (n=12, AUC=0.65, p=0.29).

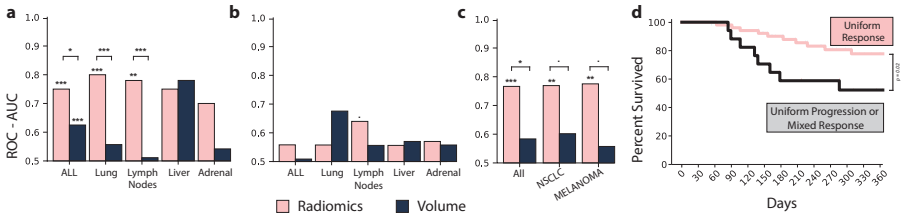

Figure 2.3: Performance of the selected classifier on the independent test set for **(a)** NSCLC lesions and **(b)** melanoma lesions. **(c)** Patient level response at first follow-up, and **(d)** Prognostic performance of the imaging biomarker on a patient level.

2.3.4 Biological validation of the radiomic biomarker

To evaluate the biological basis of the radiomic biomarker, we evaluated it in an independent dataset of 262 NSCLC patients with matched array-based gene expression data [Gro+17]. Through ranked gene set enrichment analysis, we found that the top gene-sets showing significant association with the radiomic biomarker were involved in cell cycle progression and mitosis. This indicates that a link between high tumor proliferation and improved response to immunotherapy may exist, and provides rationale for early-application immunotherapy as a therapeutic option for aggressive rapidly-expanding cancers.

2.4 Discussion

Our aim was to evaluate radiomics-based models and their potential to predict treatment response in metastatic melanoma and NSCLC patients receiving anti-PD-1 antibodies.

We found that lesions with more heterogeneous morphological profiles with non-uniform density patterns and compact borders are more likely to respond to immunotherapy - irrespective of organ and/or cancer type. Higher levels of SurfaceVolumeRatio in nonresponding lesions in both cancers suggest that more compact and spherical profiles are associated with better response.

Based on these results, it would be prudent to point out that morphological heterogeneity does not necessarily correspond to genetic heterogeneity: infiltration, inflammation, neovascularization, and necrosis could also be associated with morphological irregularities. Assuming that a well-vascularized monoclonal tumour growing in the absence of an immune system would expand uniformly in all directions, any deviation could suggest a fault of one of aforementioned characteristics. If we were to relax one of these conditions, e.g. by adding an immune system, we would observe infiltration and inflammatory microenvironment [McG+16] affecting the tumor morphology — now comprising more than solely tumor cells. Irregular vascularization might cause non-homogeneous growth patterns while hampering T-cell infiltration [Hua+13]. The role of the other compartments need to be taken into account in order to explain the overall tumor growth.

Overall results of machine learning model show good predictive performance for NSCLC metastases. In melanoma the same model performed poorly. Besides the smaller melanoma cohort, the heterogeneous therapeutic backgrounds likely played a role in the morphological characterisation. While NSCLC patients received chemotherapy as first-line, melanoma patients received a variety of different treatments prior to immunotherapy. This could potentially have led to standardization of defined genetic profiles and tumour microenvironments in

NSCLC [GM12; Aer+14; Gro+17; Rio+17]. In melanoma patients the diversity of therapeutic backgrounds might have induced different genetic profiles and microenvironments. Despite the lower performance on individual melanoma lesions, we still see a correlation with response and overall survival at a patient level, suggesting a relationship between individual lesion response and overall tumor burden.

GSEA on an external cohort revealed associations of the radiomic biomarker to proliferative potential in NSCLC, suggesting that highly proliferative tumors may show preferential response to immunotherapy. While standard of care for patients with aggressive cancer showing rapid expansion is platinum-doublet chemotherapy, these results provides the biological rationale for previous work demonstrating why combination therapy is a viable option in first-line metastatic settings, [Gan+18b].

We designed the study using a lesion-based approach, reflecting the metastatic condition characterizing patients receiving immunotherapy. This enabled us to investigate lesions individually while avoiding the issue of mixed response. Whenever possible, we limited selection biases and tried to avoid overfitting. Further validation in larger cohorts is warranted.

As imaging can provide information of the total tumor burden which allows the analysis of each lesion individually, its value lies complementary to currently known biomarkers (limited to single lesion samples). Despite the correlations found to overall patient survival and molecular pathways, further studies are needed to investigate the interaction between single (or clusters of) lesions, tumor biology and clinical status. Only a multidisciplinary aimed to integrate data from different disciplines can create a fully integrated solution that can be implemented into the clinical workflow.

2.5 Conclusions

Our findings suggest associations between radiomics characteristics and immunotherapy response showing consistent trends across cancer types and anatomical location. Lesions that are more likely to respond to immunotherapy typically present with more heterogeneous morphological profiles with non-uniform density patterns and compact borders. Moreover, we provide a predictive machine learning model that could be used within the context of lesion response to treatment, patient treatment response, and response pattern characterization. Furthermore, we evaluated the biological basis of the proposed biomarker and found to be correlated with cell proliferative potential. Motivated by the results and the wide availability of routine clinical CT scans for cancer immunotherapy patients, we aim to expand this research further to the design of clinically applicable automatic computer models that could support the oncological decision-making process.

Radiographic Feature					Responding vs Progressive (p-value)						
B	**R**	**Filter**	**Class**	**Feature**	All	A	H	LN	P	Sq	S
1	1.0	LoG.5.0	FirstOrd	Minimum	0.31	**<0.01**	0.28	0.26	0.09	0.33	**<0.01**
5	3.0	LoG.2.5	GLCM	Homoge1	0.47	0.80	0.52	**0.02**	0.43	1.00	0.80
5	3.0	-	GLCM	Homoge1	0.44	0.84	0.91	0.25	0.67	0.64	0.84
25	5.0	Wvl.HHH	GLCM	Homoge1	0.73	0.92	0.42	0.12	0.04	0.44	0.92
1	1.0	Wvl.HLH	GLSZM	ZoneEntrp	**<0.01**	0.99	**<0.01**	0.73	0.13	0.52	0.99
5	3.0	Square	FirstOrd	Entropy	0.22	N/A	0.16	**<0.01**	0.59	0.72	N/A
5	3.0	LoG.5.0	GLCM	DiffEntrp	0.43	0.95	0.24	0.17	0.26	0.97	0.95
5	3.0	Sqrt	GLRLM	LoGLREmp	0.88	0.76	**0.01**	0.37	0.53	0.90	0.76
25	5.0	-	Shape	SrfVolRtio	**0.01**	0.90	**0.01**	**0.01**	0.97	0.54	0.90
25	5.0	Wvl.LLL	GLCM	MaxProp	0.14	0.93	**0.02**	0.56	0.43	0.37	0.93

Table 2.1: Summary of radiographic differences in different metastatic locations: adrenal (A), hepatic (H), lymph nodes (LN), pulmonary (P), subcutaneous (SUBq) and spleen lesions (S). Feature settings of filter, class, feature name, binning (B) and resampling (R) are given. Association with response are shown by means of mixed model p-values. Significance after FDR is marked in bold. Failure of model convergence is reported as N/A.

Cancer	Organ	Discovery Set			Test Set			AUC	p-value
		Pts	N+	N-	Pts	N+	N-		
NSCLC	-	81	135	266	42	62	109	0.75	**<0.001**
	Lung	61	61	124	34	46	43	0.80	**<0.001**
	Lung (primary)	29	10	21	16	4	12	0.79	0.05
	Lung (metastases)	43	51	102	25	42	31	0.83	**<0.001**
	Lymph Nodes	47	37	88	22	9	48	0.78	**<0.01**
	Liver	16	30	38	3	4	5	0.75	0.14
	Adrenal	15	6	13	8	3	10	0.70	0.18
	Spleen	2	0	3	0	0	0	N/A	
	Subcutaneous	1	1	0	1	0	3	N/A	
Melanoma	-	52	77	274	28	35	97	0.55	0.20
	Lung	22	6	51	12	6	22	0.55	0.37
	Lymph Nodes	25	14	56	22	17	43	0.64	0.05
	Liver	16	20	88	7	7	20	0.55	0.35
	Adrenal	12	10	8	5	4	4	0.58	0.43
	Spleen	4	1	12	2	1	1	N/A	
	Subcutaneous	21	26	59	4	0	7	N/A	
All	-	133	212	540	70	97	206	0.66	**<0.001**

Table 2.2: Prediction performance of the chosen machine learning classifier on independent validation set. Size of both discovery and validation sets are reported in terms of number of patients (Pts), number of positive samples i.e. non-responding lesions (N+), and number of negative samples (N-).

3

Lesion diagnosis of therapy-induced lung disease

Stefano Trebeschi et al. "Deep learning distinguishing pulmonary progression from pulmonary sarcoid-like lesions in immunotherapy-treated melanoma patients". In: *British Journal of Cancer, accepted for publication* (2020).

Abstract

Background. Immunotherapy is being used in an increasingly variety of cancer types. As a result of its mechanism of action, immune-related side-effects may occur that are important to be distinguished from tumor progression - emphasizing the need for timely detection. In this study, we used deep learning applied to routine clinical CTs for diagnosis of intrapulmonary sarcoid-like granulomatous lesions subsequent to antiCTLA-4 monotherapy.

Methods. A deep learning network developed for lung cancer screening was used and fine-tuned on a cohort of 4579 lung nodules of 138 melanoma patients, of which 1679 lung nodules of 69 patients were used for independent testing(6 diagnostic outcomes).

Results. The network reached 0.68AUC (p<0.001) for histologically-proven sarcoid-like granulomatous lesions. Its performance could be improved for pulmonary metastases compared to the original screening network (0.76 versus 0.61AUC, p<0.001). These results suggest the presence of treatment-induced morphological changes, not present in the original treatment-naïve dataset. We found significant differences in the network's ability to distinguish between sarcoid-like and post-infection granulomas (0.71AUC, p<0.001), suggesting reliance on inflammation-associated features. The diagnostic score prognostic for 1-year OS (0.70AUC, p<0.002).

Conclusion. Artificial intelligence can improve the diagnosis of sarcoid-like granulomatous lesions. If validated, these findings could enhance the current diagnostic and treatment workup for patients receiving immunotherapy.

3.1 Introduction

Since the approval of the first checkpoint inhibitors in 2011 for stage-IV melanoma patients [Man11; Hod+10; US ; Eurc], cancer immunotherapy rapidly grew to include a larger variety of cancer types. This culminated in 2017 with the first ever approval of a cancer treatment in any solid tumor with high microsatellite instability or mismatch repair deficiency [Mar+19].

Due to the nature of immunotherapeutic agents, inflammations and autoimmune-like disorders are among the most common side-effects, termed immune-related adverse effects. These include immune-related toxicities of the skin, endocrinopathies, hepatotoxicity, and pneumonitis [FPP16]. While some of them pose little risk to the patient (e.g. skin rash) [Gol+16], others could lead to more serious, life-threatening conditions. Pneumonitis, for example, may represent a life-threatening situation [Fra+18], often affecting treatment continuation or (in more critical situations) trigger the administration of corticosteroids and immunosuppressants [LG16; SDL16], with the danger of developing irreversible interstitial lung disease that might limit patient outcome. Timely detection of these adverse effects is therefore essential for the management of patients undergoing immunotherapy [Dim+18].

While an effort has been made to adapt imaging follow-up schemes [Sey+17b], treatment-specific diagnosis of immune-related adverse effects is still limited due to the lack of familiarity of radiologists with this novel treatment modality [Nis+15], and the limitations of routine clinical imaging unable to clearly differentiate immunecompartments (i.e. the cellular composition of the microenvironment) and tissue immune-infiltrates. Due to the unique nature of the therapy and the rapidly increasing number of patients receiving checkpoint inhibitors, the need for novel diagnostic tools to fit treatment specific needs is clear.

We examined lung “sarcoid-like” granulomatous disease as a side

effect of CTLA-4 checkpoint inhibitors 11, which currently lacks accurate, non-invasive, treatment-specific diagnostic tools. Studies have raised the awareness of increasing incidence of sarcoid-like lesions in malignancies with an incidence rate of 4.3% [Ask+99; Bon+15; CK07; PGW95; Rei06; Egg+19]. These lesions arise from the checkpoint inhibitor mediated activation of the immune system, and present as nodular inflammations. Granulomatous lesions are not easily distinguishable from metastatic lesions on routine radiological scans, often being mistaken for progression of malignancy [Sid+17]. The reason for this is because they represent with similar imaging morphology on computed tomography (CT) as well as may show activity on positron emission tomography – computed tomography (PET-CT) resulting in a lack of diagnostic specificity of radiological appearance [Egg+19]. In addition, different nodule types have different treatment options. While guidelines have been established for the management of the incidental finding of solitary pulmonary nodule in patients without pre-existing conditions [NO19; Gra+16], the management of lung nodules still remains challenging in these patients [Nai+18] as well as in cancer patients [Gre+17]. Because of this, further imaging followup examinations during cancer immunotherapy are often needed to observe the development of a suspected possible granulomatous lesion, and even then, biopsy often remains the only option for accurate differential diagnosis [Egg+19; Ohs+17].

As standard radiological diagnostic tools do not provide accurate and non-invasive solutions, current evidence in the literature suggests that superior results could be achieved through the usage of artificial intelligence (AI) [Bi+19b] — more specifically, deep learning. Unlike standard radiological reporting based on qualitative evaluation of visible features, AI methods applied on routine clinical imaging allow for the quantitative evaluation of recurrent imaging patterns — anatomical structures and morphologies, possibly invisible to the human eye — that can be linked to the presence of a medical condition. AI effectively allows radiological images to be used as a source of quantita-

tive, minable data for diagnosis and prognostication. This study aims to evaluate the performance of deep learning methods in the differential diagnosis of (biopsy proven) lung sarcoid-like granulomatous disease of advanced melanoma patients undergoing CTLA-4 checkpoint inhibitors.

3.2 Materials and methods

3.2.1 Study cohort

For this study, we retrospectively included consecutive patients with stage IIIA-IV melanoma treated with anti-CTLA4 monotherapy (3 or 10 mg/kg) every 3 weeks (until disease progression, discontinuation due to toxicity or patient withdrawal) within the Department of Dermatology of the University Hospital of Zurich (USZ; Zurich, Switzerland) from 2012 - 2018. Of those n=72 patients were treated within clinical trials (clinicaltrials.gov NTC00636168, NTC01844505, NTC02388906). The remaining number of patients were treated with anti-CTLA4 as standard of clinical care. These patients were included according the ethical approval for the monocentric biodatabank (including imaging assessment and patient outcome) in melanoma patients treated with targeted and/or immunotherapy (EK 647). All patients underwent standardized imaging-based tumor response assessment with contrast-enhanced computed tomography (CT) with a follow up (FU) interval of 8-10 weeks during the treatment. Only patients with diagnosed intrapulmonary nodules and with a pre-existing CT examination at least 1 year prior to the time point of melanoma diagnosis were included. Patient history and clinical data were collected from electronic medical records. Patients with a known history of immunodeficiency and/or autoimmune disease were excluded. Data was collected according to the approval of the local ethics commission (EK 647 and KEK-ZH 2014-0193) and following the guidelines of the Helsinki Declaration on Human Rights with signed patient consent. Data analysis was carried out at the Netherlands

Cancer Institute (NKI; Amsterdam, The Netherlands) according to the approval of the local institutional review board (IRBd19-083).

3.2.2 Imaging Acquisition

CT images were acquired in a supine position in full inspiration. CT scans were performed with 3 different types of scanners: a 128 slice multidetector CT scanner (Somatom Flash, Siemens, Erlangen, Germany), a 128 slice multidetector CT scanner (Somatom Edge Plus, Siemens, Erlangen, Germany), and 64 slice multidetector CT scanner (Somatom AS, Siemens, Erlangen, Germany). The following parameters were used: tube voltage 120 kV, automated attenuation-based tube current modulation with a reference tube current-time product of 320mAs/rotation; pitch 3.2; gantry rotation time 0.25s, collimation 128×0.6 mm and 64×0.6 mm, respectively. Images were reconstructed with a slice thickness of 2 mm, an increment of 1.6 mm, a soft tissue kernel B36, and slice thickness 1.5, an increment of 1, a hard tissue kernel B57, respectively. For contrast enhanced CT scans 65 ml iopromide (Ultravist 300, 300mg/ml; Bayer Schering Parma, Berlin, Germany) was injected at a flow rate of 2.1 ml/s followed by 60 ml of saline solution at the same flow rate. Bolus tracking in the descending aorta was performed with a signal-attenuation threshold of 120 HU and a delay of 25 seconds for the arteriovenous contrast phase.

3.2.3 Imaging assessment, data preparation, and preprocessing

Read-out of imaging data was performed by two radiologists (LT 4 years of experience, TN 11 years of experience), consulting follow-up scans and clinical data when required. The readout assessment of the radiologists was performed in concordance, meaning unclear lesions were assessed in consensus. Because of this methodologic approach, no inter-observer assessment was evaluated. Each lung nodule was

marked with a 3-dimensional spherical region of interest (ROI) positioned in the center of the lesion, allowing to analyze the nodule itself and its surrounding anatomical information. Each nodule was assigned with a unique identifier that was kept consistent through its follow-ups for longitudinal tracking. An additional volumetric segmentation of lung nodules was provided only for lesions in the sarcoid-like granulomatous group, due to time constraints.

Recorded anatomic information included: side (left or right lung), lung lobe (upper, lower, middle or lingula), and subpleural versus intrapulmonary space. The intrapulmonary findings were categorized as following (also see Table 3.1) [Nai+18; CSA19]: metastasis, non-specific pulmonary perifissural lymph node, non-specific granuloma (generally post-infection), ground-glass lesion, scar tissue (post-infection or postoperative), focal infectious inflammation, and sarcoid-like granulomatous lesion. Scar tissue was excluded from the dataset, due to scarcity of samples (no samples in the discovery set). Non-specific granuloma were identified as known intrapulmonary lesions already documented on CT examination performed at least 1 year prior to the diagnosis of melanoma that remained stable during all follow up CTs, these lesions can be partially calcified.

To generate nodule-specific imaging data, each lung lesion was cropped to a cube of $64 \times 64 \times 64$ voxels (isotropic voxel resampling at 0.75 mm) according to the location recorded by the readers. Hounsfield units are clipped between -1000 (air density) and 1000 (cortical bone density) and normalized to zero mean and one standard deviation.

3.2.4 Deep learning

The dataset was divided into discovery and independent validation set based on patient identifiers. A publicly available 3-dimensional convolutional neural network was used on our dataset[1]. The network was

[1] url: github.com/LouisFoucard/DSB17_3d_lung_nodule_classifier

originally trained to classify malignant pulmonary lesions in a publicly available screening imaging dataset. We optimized the original network on our dataset via fine-tuning: the first six convolutional and batch normalization layers were frozen during fine-tuning, and the original multi-layer perceptron was replaced by a global average pooling followed by a softmax classifier. In other words, we employed the original screening network to extract quantitative imaging features, and replaced the original binary classifier (malignant vs benign) with a 6-class classifier, one class per diagnostic outcome. The improvement of the diagnostic performance was evaluated by comparing the performance of the original network on our dataset, to our optimized network for the classification of malignant lesions (as the original network was trained only to distinguish malignant lesions from benign ones with the use of the publicly available LUNA dataset[2]). The LUNA dataset is composed of 158 patients with different malignancies, including malignant and benign lesions. Sarcoid-like lesions labeling was not provided. The diagnostic parameters of the LUNA dataset were assessed by a team of expert readers. Biopsies were taken in a number of unclear cases. The LUNA dataset is more heterogeneous, as it contains contrast and non-contrast scans, as well as different voxel sizes and resolution parameters. To reduce overfitting and improve generalization, images were augmented with random rotation of 10°, shift and zoom of 10%, and random axis flipping. Additional regularization was implemented via dropout ($p = 0.5$) on the softmax classifier. To counter class imbalance, focal loss (β, α = 2.0, 1.0) was used during training [Lin+17]. Adam was used for stochastic optimization (lr = 0.001) [KB14], with batch size set to 32 samples. A detailed representation of the network is given in Figure 3.1. The trained network output consisted of a group of six probabilities, one per diagnostic outcome: non-specific granulomas, sarcoid-like granulomatous lesions, focal infections, perifissural lymph nodes, ground-glass lesions and metastases. Due to the lack of samples in the training set and/or test set, scars were excluded.

[2] url: luna16.grandchallenge.org

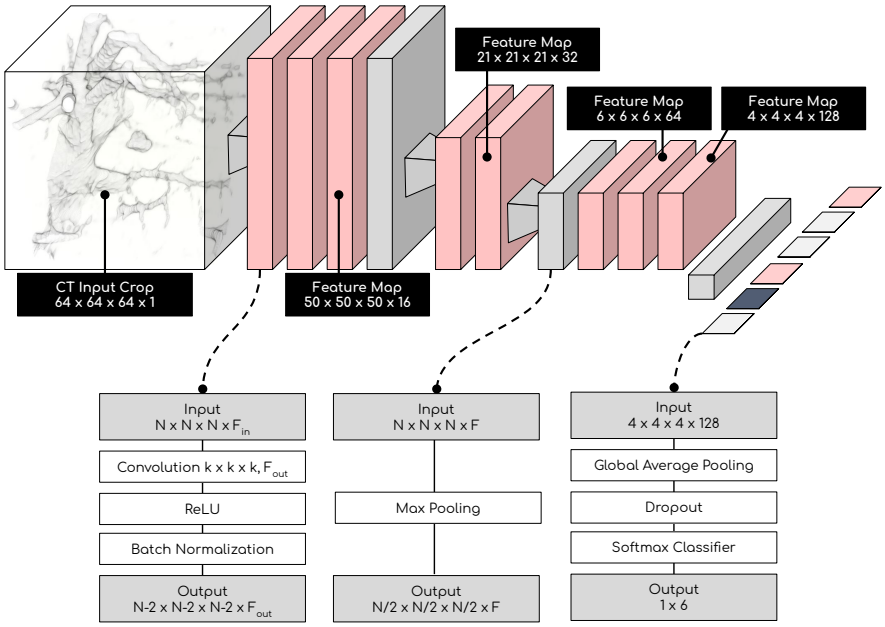

Figure 3.1: Deep learning architecture

3.2.5 Statistical analysis

To assess diagnostic performance, each score was analyzed individually against all other diagnostic outcomes (i.e. one-vs-all) using the area under the receiver operating curve (AUC). Confidence intervals were estimated via bootstrapping performed using repeated sampling with replacement (1000 times). Statistical significance was assessed via Mann-Whitney-U test. Further analysis was performed to the score associated with sarcoid-like granulomas lesions against all others diagnostic scores individually (i.e. one-vs-one) using the same statistical metrics. Lesions' malignancy scores generated by the original screening network, and lesions' volumes were used for diagnostic performance comparison. Significant differences in diagnostic performance

(i.e. between AUCs) were assessed via McNeil test.

Diagnostic scores can be used for both prognostication and stratification. To assess the prognostic performance of our AI-diagnostic signature at baseline (e.g. first available scan), a patient-wide imaging signature was created. In this imaging signature, the highest probability per outcome across all lesions in the scan was recorded. The resulting diagnostic signature described the probability of the patient having at least one lesion with a specific diagnostic outcome (e.g. metastasis, non-specific pulmonary perifissural lymph node, non-specific granuloma, ground-glass lesion, focal infectious inflammation, and sarcoid-like granulomatous lesion). Unsupervised principal component analysis was used to transform the patient-wise signature to a single-value prognostic score. Predictive performance of this score was estimated with Kaplan Meier curves. Statistical significance was assessed via log-rank test. Multivariate Cox regression was used to assess statistical significance against other known diagnostic and clinical parameters used for prognostication, tumor stage, presence of lung metastases and young age (<65 years).

3.2.6 Saliency Maps

Imaging features used by the network and associated with sarcoid-like granulomatous lesions were analyzed through saliency maps [SVZ13]. Saliency visualization is a technique which highlights sections of the original image that are deemed predictive by the network to estimate the diagnostic score. These highlighted sections are referred to as "salient regions". Saliency maps of sarcoid-like granulomas lesions were generated, along with metastases and non-specific granulomas (similar pathophysiology but different stages of infection) as controls. The qualitative (purely visual) analysis was limited to the top-8 lesions analyzed by the radiologists per class where the algorithm returned the highest probability of the classification. In other words, these were the eight lesions where the algorithm was most confident in its "diagnosis".

3.3 Results

3.3.1 Study cohort

From consecutive n=293 patients a total of n=165 melanoma patients were included in this study (62 female, median age 63 years, range 55 - 74 years). Twenty-seven patients were excluded due to the lack of visible lung nodules or insufficient image quality. In total, we included n=138 patients (87 males, 51 females, median age 64 years, melanoma cancer stage IIIA-C N=27, stage IV N=111), n=980 CT scans, and n=4579 lung nodules — n=1059 baseline nodules longitudinally followed-up with an average of 6 CT scans (IQR=6) every 8 to 12 weeks. Manifestation of granulomatous disease was confirmed through biopsy in n=8 patients (n=74 CT scans performed in these patients, see Table 3.2). In our study cohort all sarcoid-like lesions CT or PET/CT suggested a malignant cause of these pulmonary nodules, leading into biopsy or thoracic surgery. Also, retrospectively we could not find any imaging pattern that might be specific for sarcoid-like granulomatous disease in our study population. In Figure 3.2c morphologically the overlap between metastases and non-specific granuloma can be seen if compared to the introduced characteristics in Table 3.1. Out of n=317 lung nodules of the patients presenting sarcoid-like granulomas disease, 60.57% (n=192) nodules were sarcoid-like granulomatous lesions. The overall number of nodules per patient in this group does not significantly differ to the negative control (16 versus 17 nodules on average respectively, mann-whitney-u test $p = 0.23$). The majority of the nodules were either melanoma lung metastases (n=2772, 60.54%), or perifissural intrapulmonary lymph nodes (n=1150, 25.11%). Non-specific granulomas, sarcoid-like granulomatous lesions, and ground glass lesions were rare, accounting for 7.21% (n=330), 4.19% (n=192) and 1.64% (n=75) respectively. Because in all patients CT examinations prior to the diagnosis of melanoma were available, all metastases could be identified as new lesions occurring after melanoma diagnosis. All nonspecific lesions like non-specific

pulmonary perifissural lymph node and non-specific granuloma could be identified in the CT examination at least one year prior to first diagnosis of melanoma. Occurrences of scars, inflammations and lesions shifts (e.g. from perifissural lymph nodes to metastases) were each below 1%. Location-wise, nodules were unevenly distributed (chi-square test, $p<0.001$) with lower lobes accounting for half of the nodules (23.69% or n=1085, and 25.88% or n=1185, for left and right lungs respectively). Nodules in the lingula were the rarest (4.93% or n=226, see Table 3.3).

3.3.2 Diagnostic performance

The discovery set consisted of n=69 patients and n=2830 nodules, while the independent test set consisted of the remaining n=69 patients (n=5 with sarcoid-like granulomatous disease) and n=1679 nodules (n=35 sarcoid-like granulomatous lesions). On the independent test set, the network reached an average performance of 0.69 AUC. Higher performances were reached for ground-glass lesions (0.82 AUC, CI: 0.67 — 0.9, $p=0.002$), nonspecific granulomas (0.79 AUC, CI: 0.77 — 0.82, $p<0.001$), and metastases (0.76 AUC, CI: 0.74 — 0.78, $p<0.001$). Perifissural lymph nodes and sarcoid-like granulomas lesions reached diagnostic performance of 0.69 AUC (CI: 0.67 — 0.72, $p<0.001$) and 0.68 AUC (CI: 0.61 — 0.76, $p<0.001$), respectively. The network performed poorly on focal infectious inflammations (0.40 AUC, CI: 0.27 — 0.56, $p=0.26$). No significant difference in diagnostic performance has been detected between nodules at first appearance and nodules at follow-ups (McNeil-test, all $p>0.48$). Overall, the proposed fine-tuned network significantly outperformed the original one in the classification of metastatic lesions in our test set (0.76 AUC vs 0.61 AUC; McNeil-test, $p<0.001$) and the volumetric measure within the sarcoid-like granulomatous group (0.76 AUC vs 0.53 AUC; McNeil-test, $p<0.001$). Further analysis on diagnostic outcome pairs revealed higher diagnostic performance in distinguishing between sarcoid-like granulomatous lesions versus

ground-glass lesions (0.82 AUC, CI: 0.59 — 0.97, p=0.005), and sarcoid-like granulomatous lesions versus non-specific granulomas (0.71 AUC, CI: 0.63 — 0.80, $p<$ 0.001). Differential diagnosis of sarcoid-like granulomatous lesions versus perifissural lymph nodes and metastases performed similarly to the general case (0.69 AUC, CI: 0.62 — 0.76, p<0.001 and 0.68 AUC, CI: 0.60 — 0.74, p<0.001, respectively). The network appears to be unable to distinguish between sarcoid-like granulomatous lesions and focal infection inflammations (0.37 AUC, CI: 0.11 — 0.71, p=0.22). No significant difference in diagnostic performance has been detected between nodules at first appearance and nodules at follow-ups (mcneil-test, all p>0.57, see Figure 3.2a). To check that the different scan properties did not interfere with the deep learning model, a sub-analysis within the scans of the granulomatous patients only (N=63 lesions that were not granulomatous) was performed. The result kept being significant (p=0.01), and similar to the one reported for the entire dataset (0.64 AUC, CI: 0.55 — 0.73), which lead us to the conclusion that even when the scan properties are inhomogeneous, the result holds.

3.3.3 Predictive performance

Overall median survival of the cohort of patients in the independent test set was 16.6 months. Using the diagnostic score at baseline, we were able to identify two risk groups (i.e. high score and low score, according to score median) with a median survival difference of 12 months (11.3 vs 23.5 months, log-rank-test, p=0.03) and a predictive performance of 0.70 AUC for 1 year overall survival (CI: 0.59 — 0.81, p=0.002). In multivariate analysis, the score remained significant for overall survival (p=0.02) against tumor stage (>stage III, p=0.08), presence of lung metastases (p=0.43) and younger age (<65 years, p=0.87).

3.3.4 Saliency maps

The saliency maps generated for non-specific granulomas, sarcoid-like granulomatous disease and metastases (Figure 3.2c) revealed predictive features in the intralesional periphery, and in the perilesional parenchyma. The intralesional periphery was highlighted in case of sarcoid-like granulomatous-lesions and metastases, whereas imaging features in perilesional parenchyma were deemed predictive by the network in non-specific granulomas. Additionally, in non-specific granulomas a heterogeneous pattern could be depicted compared to the other cases. Salient regions on vasculature (within and around the lesions) could be found across all classes. Anatomical landmarks (e.g. ribs, intercostal muscles, surrounding organs) were not found to be predictive in any of the samples.

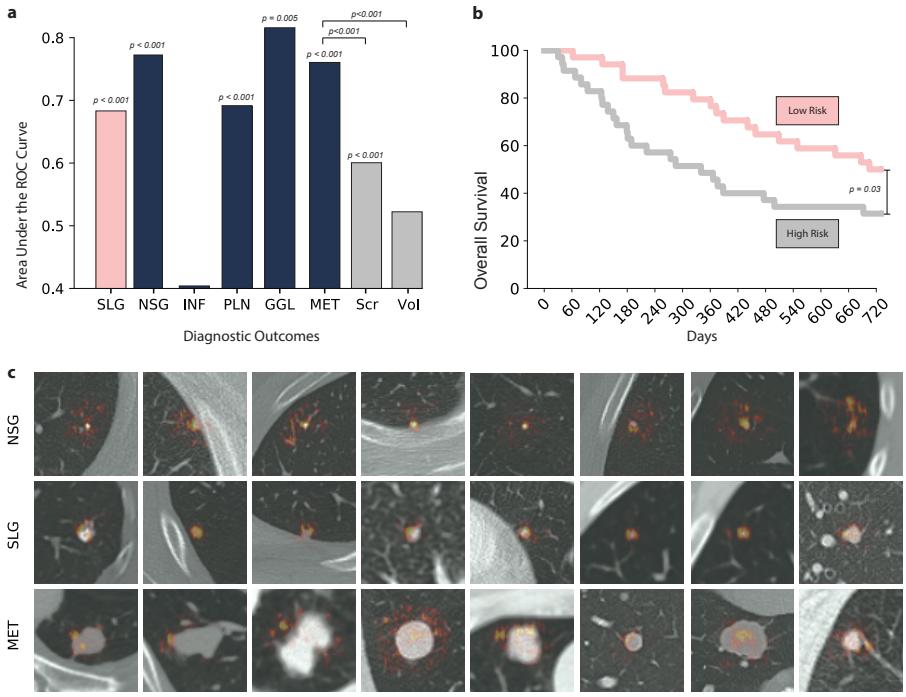

Figure 3.2: **(a)** Area under the receiver operating characteristic curve of each diagnostic outcomes versus all other diagnostic outcomes. **MET** metastases (**Scr** indicates the score of the original, non-fine tuned model trained on screening imaging data, **Vol** indicates the diagnostic performance of volume), **PLN** perifissural lymph nodes, **NSG** non-specific granuloma, **SLG** Sarcoid-like granulomatous lesions, **GGL** ground glass lesions, **SCA** scars, **INF** infections, **SH** shifts. **(b)** Kaplan Meier curves of high and low risk groups, stratified according to the diagnostic score at baseline **(c)** Saliency maps of sarcoidlike granulomatous lesions and respective controls (metastases and non-specific granulomas).

3.4 Discussion

Cancer immunotherapy checkpoint inhibitors changed the treatment landscape of various cancer types including melanoma. Due to the nature of immunotherapeutic agents, immune-related side effects are common. As these adverse effects could lead to life-threatening conditions, timely detection is essential. While efforts have been made to adapt imaging follow-up schemes, treatment-specific diagnosis is still limited partially due to the novelty of this therapy. Our aim was to evaluate the diagnostic performance of artificial intelligence (deep learning) and medical imaging in the diagnosis of sarcoid-like granulomatous lesions, which arose as an immune-related adverse effect to anti-CTLA4 checkpoint inhibitors in melanoma patients.

Our results show good performance of our AI model for differentiation of sarcoid-like granulomatous lesions against metastases, perifissural lymph nodes, non-specific granulomas, and ground-glass lesions with highly significant AUC-values. Despite the large data imbalance with very few sarcoid-like granulomatous lesions available for training, the careful design of the AI model and training procedure allowed to balance the diagnostic performance of each lesion outcome around 0.69 AUC. No significant difference has been observed in diagnostic performance at nodule baseline compared to their follow-ups, suggesting that the AI model was independent from treatment time-points.

Overall, our network significantly outperformed nodule volume, and visual features in distinguishing cancer lesions from benign lesions. While this difference might have been a direct result of unaccounted changes in the pre-processing steps, image acquisition (e.g. not all patients under screening receive contrast medium) or adjustments in the network architecture and training procedure, one should also account for the discrepancy of patient characteristics. The original network was trained on the LUNA dataset, a publicly available lung screening imaging dataset, containing mostly treatment-naive subjects. It has not yet been proven whether cancer lesions of patients undergoing immunotherapy are subject to morphological changes visible in

CT images. However, it would be reasonable to assume that the activation of the immune system driven by checkpoint inhibitors might cause changes in the visible (and potentially not-readily-visible) morphology of the tumor as a result of the inflammatory process.

Further analysis revealed higher diagnostic performance of the AI model (up to 0.82 AUC) in the distinction between sarcoid-like granulomatous lesions versus immune-related phenomena, such as non-specific granulomas, and ground-glass lesions. Ground-glass lesions are focal, diffuse, and (partially) solid abnormalities [Hen+02]. Their classification is purely based on radiological characteristics, while the histopathological type might include infection and drug toxicity, sarcoidosis, and malignancy [Hen+02; Kim+13; MS05], among others. The ability of the model in distinguishing between non-specific granulomas and sarcoid-like granulomatous lesions is particularly interesting, as these two nodule types share some similar underlying biological mechanisms. Their difference can be accounted for in two ways. First, non-specific granulomas are generally post-infection, while sarcoid-like granulomatous lesions present an ongoing immune-response. Second, non-specific granulomas are more likely to contain necrotic and fibrotic components, generally not associated with sarcoid-like lesions [IRT07]. Saliency maps of non-specific granulomas displayed predictive imaging features in the surrounding parenchyma and vasculature, whereas sarcoid-like granulomatous lesions showed salient regions in the intralesional periphery. One could hypothesize that the saliency map might depict a still active immune response in the intralesional periphery of the sarcoid-like granulomas lesion in comparison to a possible chronic enclosing immune reaction of the lung parenchyma surrounding the nonspecific granuloma. It is not possible to draw hypotheses, with the same level of certainty, that account for the difference between metastases and sarcoid-like lesions. While sarcoid-like lesions and non-specific granuloma share the same biological basis, cancer lesions lie far apart. Saliency maps clearly depict regions in the intralesional periphery, as well as parenchyma, to be predictive. This

could, however, be due to a number of factors, including potential differences in vascularization, immune- and tumor-infiltration, morphological features associated with the displacement of healthy tissue, and so on. Further validation is also needed to confirm its biological basis.

Prognostic analysis of the diagnostic score at baseline revealed significant correlation with overall survival, which remained significant even when compared to known negative prognostic factors, such as tumor stage, spread, and age. Diagnostic scores at baseline, such as the radiological tumor stage and pathological TNM classification of malignant tumors, have already been linked to overall survival. In this context, a more complete, quantitative descriptor of the patient status at start of treatment could potentially have prognostic

To the best of our knowledge, this is the first study to investigate the performance of artificial intelligence in the differential diagnosis of intrapulmonary sarcoid-like granulomatous disease in cancer patients undergoing immunotherapy. The earliest case-report of granulomatous disease associated with checkpoint inhibitors was published in 2016 by Danlos et al. [Dan+16] and presented a melanoma patient undergoing anti-PD1 checkpoint inhibitors [Dim+18] who developed a sarcoid-like granulomatous lesion in the mediastinal lymph node and skin after complete response. The authors acknowledged the link between the development of sarcoid-like lesions and the cell-mediated immunity induced by the treatment, while warning of possible adverse clinical implication for sarcoid-like lesions misdiagnosed as tumor progression. Similar reports have been published, including reports for patients receiving anti-PD1 + anti-CTLA4 combinations [Suo+16], and in lung cancer patients receiving anti-PD1 antibodies [Bir+17].

Regarding imaging, evidence in the literature is still scarce. A visual analysis on the morphology of granulomatous lesions in a cohort of 18 patients with common variable immune deficiency presented features, like generalized diffuse reticular pattern and lower lobe predom-

inance in 80% of non-specific granulomas positive patients [Par+05]. Molecular imaging, in the form of fluoro-D-glucose positron emission tomography (FDG-PET), is the only reported method for non-invasive, quantitative imaging assessment of sarcoid-like granulomatous lesions [Cap+16]. Activated leukocytes, macrophages and activated helper T-cell (CD4+) show increased FDG uptake, highlighting spots of ongoing inflammation. In these cases, FDG-PET has demonstrated high sensitivity (90%-100%). However, translating these results in the differential diagnosis of patients undergoing checkpoint inhibitors is not straightforward, as the method relies on the detection of activated inflammatory components which are likely to be present also in cancer lesions. Although these studies present insights into the viability of imaging for the assessment of sarcoid-like granulomatous lesions, none of them was performed with concurrent, immunotherapy-treated metastatic disease, where cancer lesions are likely to also have sizable active immune-compartments. Clinically, the prospect of routine imaging being able to provide a quantitative, non-invasive diagnostic profile of the whole tumor burden that can also be used in prognostication models is appealing. Such tools would enable clinicians to accurately monitor the treatment and steer it accordingly, while simultaneously avoiding additional invasive (and potentially harmful) examinations for the patient.

3.4.1 Limitations and future outlook

The lack of accurate non-invasive diagnostic methods can be explained by several factors. These include the rare (and often asymptomatic) nature of the disease, the novelty of the treatment, and the unfamiliarity of radiologists with treatment side-effects, among others. These factors make data collection complicated, leading to small-sized datasets. The diagnostic accuracy reported and the limited amount of diagnostic outcomes considered in our cohort, would not meet the requirements to use AI alone in the clinical workflow. Better performances are to be expected with increased availability of sarcoid-like granulomatous

lesion images for training and external validation cohorts. However, due to the rarity of the adverse effect, it is more likely to achieve better performance by leveraging more advanced AI methods, currently under research. For example, additional improvements in the diagnostic performance could be achieved by giving the model access to the clinical history of the patient. This however would require more complex, multi-modal approaches to be investigated separately. Because of the limited dataset in our study no clinical data was included to avoid an unwanted confounder in the evaluation of the performance solely based on imaging features. Additionally, further investigations should include biological validation where the imaging features learnt and leveraged by the AI model are linked back to micro-environmental and genetic quantitative features. While in this study we partially address the problem by leveraging the biological similarity between sarcoid-like granulomatous lesions and post-infectious granulomas, any further conclusion would need to be proven by adequate biological markers, possibly not relying on the generalizability assumption of single lesion biopsy. Finally, in this study we investigated sarcoid-like granulomatous lesions as a side-effect of CTLA-4 checkpoint inhibitors. Further research in the applicability of this method for treatments focusing on the PD1/PD-L1 axis and combination therapies are required.

3.5 Conclusions

Aim of this study was to apply novel technologies of artificial intelligence on routine medical imaging for the diagnosis of sarcoid-like granulomatous lesions induced by novel cancer immunotherapeutic agents. We found significant performance in the diagnosis of sarcoid-like granulomatous lesions, while simultaneously significantly improving the performance of the original screening network for the diagnosis of pulmonary metastases. Moreover, the network was able to distinguish between sarcoid-like granulomatous lesions and non-specific post-infection granulomas. Further investigation is needed to explore the links between the

imaging features and biological phenomena, and improvement of the diagnostic performance to clinical acceptability.

Nodule lesion	Semantic imaging characteristics
Metastases (MET)	Irregular borders (unsharp, polylobular, spiculated), solid with predominant soft tissue component, part solid with small nodular soft tissue component, homogenenous, inhomogenenous, partially centrally necrotic, unequifocal growth between follow up examinations, new intrapulmonary lesion during follow up examination with imaging, characteristics of metastases
Non-specific pulmonary perifissural lymph node (PLN)	No growth within the last 1-2 years, perifissural location, smooth borders, spheric shape, triangular shape, remains stable, no unequivocal growth during all follow up examinations
Non-specific granuloma (generally, post infection) (NSG)	No growth within last 1-2 years, central small calcification, calcified small nodule, remains stable, no unequivocal growth during all follow up examinations
Ground-glass lesion (GGL)	Pure ground-glass pattern, mixed ground-glass pattern with subsolid component
Scare tissue (post-infection or postoperative) (SCA)	Linear distribution, often subpleural, no change in shape during follow up examinations
Focal infectious inflammation (INF)	Linear distribution, not round shaped, focal consolidation, focal consolidation with perifocal ground-glass pattern, small clusters of intrapulmonary nodules, tree-in-bud phenomenon, all mentioned pattern resolve after antibiotic therapy within 1-4 weeks

Table 3.1: Imaging readout parameters for categorizing intrapulmonary nodule findings [Nai+18; CSA19]

Age	Gender	AJJC	Prior to ITx. (ipilimumab)	SLG appear (during ipi., months)	Histological documentation (non caeseating granuloma)	Corticost., Remission	Melanona Response
63	M	IV (M1c)	-	4	upper right lung	Yes, No	PD, death
57	F	IV (M1c)	-	1	left lower lung	-	SD
45	M	IV (M1b)	-	2	left lower lung	-	PD
78	M	IV (M1b)	-	3	hilar lymph node and left lower lung	Yes, Yes	PR
63	M	IV (M1c)	Nivolumab	1	right lower lung	-	SD
63	M	IV (M1c)	-	22	lymph node, diffuse in bronchial epithelia, in right lower lung	-	PR
47	M	IV (M1b)	-	4	left lower lung	-	PR
52	F	IV (M1c)	-	15	right upper lung)	-	PR

Table 3.2: Cases of sarcoid-like granulamtous disease: patients' characteristics, melanoma diagnosis and treatment, histological features, and patient outcome conforming to response evaluation criteria for solid tumors (RECIST 1.1) criteria. PD = progressive disease; SD = stable disease; PR = partial response. AJJC = American Joint Committee on Cancer (8th ed.) stage.

	MET	**PLN**	**NSG**	**SLG**	**GGL**	**SCA**	**INF**	***SH***
				Intrapulmonary Right Lung Nodules				
Upper Lobe	82 (300)	12 (58)	6 (35)	8 (25)	2 (17)	-	4 (7)	-
Middle Lobe	51 (169)	7 (44)	1 (6)	1 (3)	-	-	3 (6)	-
Lower Lobe	120 (492)	7 (27)	3 (27)	3 (8)	1 (14)	-	1 (3)	-
				Subpleural Right Lung Nodules				
Upper Lobe	42 (173)	41 (166)	2 (5)	-	-	1 (3)	-	-
Middle Lobe	42 (103)	25 (161)	3 (20)	-	-	-	1	-
Lower lobe	72 (225)	42 (261)	8 (87)	2 (25)	-	1 (16)	-	-
				Intrapulmonary Left Lung Nodules				
Upper Lobe	108 (403)	6 (15)	4 (32)	10 (50)	3 (36)	-	1	-
Lingula	24 (87)	-	1 (20)	2 (5)	-	1 (7)	1	-
Lower Lobe	118 (466)	7 (54)	2 (14)	4 (47)	2 (8)	-	1 (4)	1 (2)
Non-Specific	-	-	-	-	-	-	4 (6)	-
				Subpleural Left Lung Nodules				
Upper Lobe	24 (96)	10 (77)	7 (62)	2 (3)	-	-	-	-
Lingula	17 (63)	7 (31)	1 (6)	2 (6)	-	-	-	-
Lower Lobe	56 (195)	30 (256)	2 (16)	7 (20)	-	-	1 (3)	-

Table 3.3: Entry values indicate the number of unique nodules, and the number of nodules including their follow-ups. MET = metastases, PLN = perifissural lymph nodes, NSG = non-specific granuloma, SLG = Sarcoid-like granulomatous lesions, GGL = ground glass lesions, SCA = scars, INF = infections, SH = shifts.

	All nodule follow-ups							Only at first nodule appearance						
	N-	**N+**	**PR AUC**	**ROC AUC**	**SEN**	**SPE**	***p***	**N-**	**N+**	**PR AUC**	**ROC AUC**	**SEN**	**SPE**	***p***
General diagnostic performance														
NSG	1522	154	0.22	0.79	0.54	0.91	***.	372	18	0.13	0.76	0.52	0.89	***.
SLG	1641	35	0.04	0.68	0.50	0.75	***.	378	12	0.07	0.67	0.51	0.67	*
INF	1672	4	0.00	0.40	0.50	0.25	n.s.	388	2			-		
PLN	1349	327	0.35	0.69	0.56	0.73	***.	328	62	0.27	0.67	0.54	0.70	***.
GGL	1669	7	0.02	0.82	0.50	0.88	**.	388	2			-		
MET	527	1149	0.89	0.76	0.82	0.65	***.	96	294	0.90	0.74	0.83	0.61	***.
Sarcoid-like granulomatous versus														
NSG	154	35	0.53	0.71	0.55	0.71	***.	18	12	0.80	0.77	0.61	0.67	**
INF	4	35	0.89	0.37	0.25	0.46	n.s.	2	12			-		
PLN	327	35	0.24	0.69	0.53	0.74	***.	62	12	0.40	0.73	0.54	0.67	**
GGL	7	35	0.95	0.82	0.88	0.57	**	2	12			-		
MET	1149	35	0.05	0.68	0.51	0.75	***.	294	12	0.08	0.65	0.51	0.67	*

Table 3.4: Deep learning diagnostic performance. **MET** metastases, **PLN** perifissural lymph nodes, **NSG** non-specific granuloma, **SLG** Sarcoid-like granulomatous lesions, **GGL** ground glass lesions, **INF** infections. **AUC** = Area under the curve, **PR-AUC** = Precision-recall AUC, **ROC-AUC** = Receiver operating characteristic AUC, **SEN** = sensitivity, **SPE** = specificitiy, p = p-value. *** $p< 0.001$, ** $p< 0.01$, * $p< 0.05$, n.s. for not significant. Statistically significant after Bonferroni correction are indicated a (.) character. Not assessed due to the limited sample size are indicated by (-). Confidence intervals omitted for readability, avaialble in the original supplement.

4

Prognostic value of chest imaging monitoring

Stefano Trebeschi et al. "Prognostic value of deep learning mediated treatment monitoring in lung cancer patients receiving immunotherapy". In: *Frontiers in Oncology, accepted for publication* (2021).

Abstract

Background Checkpoint inhibitors provided sustained clinical benefit to metastatic lung cancer patients. Nonetheless, prognostic markers in metastatic settings are still under research. Imaging offers distinctive advantages, providing whole-body information non-invasively, while routinely available in most clinics. We hypothesized that more prognostic information can be extracted by employing artificial intelligence (AI) for treatment monitoring, superior to 2D tumor growth criteria.

Methods A cohort of 152 stage-IV non-small-cell lung cancer patients (NSCLC)(73 discovery, 79 test, 903 CTs), who received nivolumab were retrospectively collected. We trained a neural network to identify morphological changes on chest CT acquired during patients' follow-ups. A classifier was employed to link imaging features learnt by the network with overall survival.

Results Our results showed significant performance in the independent test set to predict 1-year overall survival from the date of image acquisition, with an average area under the curve (AUC) of 0.69 ($p<0.01$), up to AUC 0.75 ($p<0.01$) in the first 3-5 months of treatment, and 0.67 AUC ($p=0.01$) for durable clinical benefit (6-months progression-free survival). We found the AI-derived survival score to be independent of clinical, radiological, PDL1, and histopathological factors. Visual analysis of AI-generated prognostic heatmaps revealed relative prognostic importance of morphological nodal changes in the mediastinum, supraclavicular and hilar regions, lung and bone metastases, as well as pleural effusions, atelectasis and consolidations.

Conclusions Our results demonstrate that deep learning can quantify tumor- and non-tumor related morphological changes important for prognostication on serial imaging. Further investigation should focus on the implementation of this technique beyond thoracic imaging.

4.1 Introduction

Recent advancements in the understanding of the tumor-immune cell interactions [LKA96; Ish+92] have enabled the development of novel drugs for the treatment of advanced-stage lung cancer. Immune checkpoint inhibitors, in particular, have been shown to provide sustained clinical benefit to patients, especially in the metastatic setting [Bor+15; Bra+15; Her+16].

Metastatic markers that can be used for patient selection (i.e. before the start of treatment), as well as for treatment monitoring (i.e. during treatment), are still under research [Ten+18; HCC19b; Ros+19]. In the context of oncological research, most predictive/prognostic markers are derived from tissue samples , routinely-extracted blood [Wan+19], or non-invasive radiological imaging (surrogate imaging markers). Tissue samples derived from biopsies (usually taken from anatomically accessible locations) often fail to account for inter- and intra-lesion heterogeneity, and response assessed during evaluation of tissue samples of only a few lesions does not necessarily mean that all lesions have responded in the same way. Furthermore, serial biopsies during longitudinal follow-up are cumbersome for the patient but also impractical. Regardless of biomarker source, monitoring of response to therapy remains challenging. As such, they are not part of the routine clinical workflow of patients.

Standard clinical imaging provides a non-invasive overview of the entire tumor burden and has the potential to more accurately evaluate the overall response of the patient to the treatment. Yet, imaging evaluation is currently limited to 2-dimensional "subjective" measurements of tumor size changes [Eis+09], time-consuming ROI delineation [Tre+19; Sun+18], and/or to values approximating metabolic activity (i.e. SUV values in PET) [Ten+18]. By limiting the use of imaging for response evaluation to only these approaches, many (potentially prognostic) imaging characteristics are ignored. For example, as the disease evolves in multiple distal sites, traditional imaging assessment methods would not account for

the microenvironment of each lesion, despite the fact that several potential prognostic factors (e.g. angiogenesis, inflammation, and lymphocytic infiltration) likely depend on that environment [Gar+19]. Since immunotherapy is a systematic treatment modality, changes indicating response are not limited to one location but can occur all over the body. This is particularly relevant in patients treated with anti PD-1 blockade where lymphadenopathy [NHH19; Tir+15], parenchymal inflammations, edema [Ale+19; Joh+16], and compression atelectasis (18), can be observed. Ideally, during image response evaluation these conditions, together with tumor growth, should be monitored and quantified as they might hold valuable prognostic information.

Using Artificial Intelligence (AI), treatment monitoring tools can be built, capable of rapidly assessing gross morphological changes between two (or more) follow-up images of the same patient [Bi+19a], in a fully-automatic manner, completely independent of human input. In this context, image registration can be used as the basis for such a method. At its core, image-to-image registration is the process of establishing a voxel-wise match between two radiological images. By establishing a match, we can measure voxel-level differences between corresponding objects represented in the images quantitatively. While conventional registration techniques are very limited for this application, deep learning-based methods have shown promise in image-to-image registration [HKY20]. There are three main advantages to using deep learning-based image registration as the core technique. The first advantage is that registration networks are trained to match a pair of images, voxel-wise. This creates a network that is explicitly trained to quantify differences between two images. By leveraging its internal features, we can effectively obtain feature vectors that represent these voxel-wise changes. These vectors can be used for classification purposes. The second advantage of using image registration is that it can be trained on large unlabeled datasets (i.e. lacking any kind of manual annotation, such as segmentations or RECIST-like measurements), while not compromising its ability to model voxel-wise details, that are

likely lost in a classical unsupervised autoencoder approach. The third advantage of using image-to-image registration is that, unlike standard RECIST, such a method could be fully automatic and not require any manual input (e.g. two-dimensional diameter measurements), and not be limited to changes in the tumour size, but it would also account for global morphological changes, whether tumour associated or not, throughout the body. Applying an image-registration-based AI algorithm in oncological follow-up imaging enables us to develop a novel method that can accurately measure gross morphological changes during treatment. Quantitative measurements of these changes can then be used for prognostication.

This study aims to investigate the potential prognostic value of AI-mediated monitoring on CT scans in non-small cell lung cancer (NSCLC) patients receiving anti-PD-1 immune checkpoint blockade. Relying on existing technical research on image-to-image registration, we hypothesize the existence of quantitative imaging features describing a set of gross morphological changes during treatment that hold prognostic value. To test this hypothesis, we developed a deep learning network for thoracic image-to-image registration and studied the prognostic value of features learnt by the network in NSCLC patients being treated with PD-1 blockade.

4.2 Materials and methods

4.2.1 Study cohort

For this study, we retrospectively included patients with stage IV NSCLC treated with anti-PD1 monotherapy within The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital (NKI-AVL; Amsterdam, The Netherlands) between 2014 and 2016. All patients underwent standardized, imaging-based tumor response assessment with contrast-enhanced computed tomography (CT), with follow-up (FU) intervals of 8-12 weeks. We retrieved all available FU scans within the first two years of treatment, together with a baseline

scan (BL) performed 8 weeks before and up to 1 week after start of treatment. To encode pre-treatment tumor spread, a pre-baseline scan (PBL), defined as the first available scan before BL, was also retrieved when available. The exact dates of each scan were recorded with respect to the start of treatment (in days). Patients with only one scan available throughout the entire treatment regimen, or whose scan would not fully cover the thorax, were excluded from the analysis. The cohort was divided into a discovery and independent test set based on the patient identifier: patients with even ID numbers were assigned to the discovery set, patients with odd ID numbers were assigned to the independent test set. The study was carried out at the NKI-AVL with the approval of the local Institutional Review Board (IRBd19-083). This cohort is a longitudinal expansion of a previously described NSCLC cohort [Tre+19].

4.2.2 Image acquisition

The CT scans were performed by either covering the chest or covering the chest and abdomen using multi-slice CT equipment (Toshiba Aquilion CX, Minato, Tokyo, Japan; Siemens Somatom Sensation Open, Erlangen, Germany) with a tube voltage of 120 kVp, slice thickness of 1 mm, and in-plane resolution of 0.75 x 0.75 mm. The bolus injection was performed at 3 ml/s (Omnipaque 300, GE Healthcare, Chicago, Illinois, US) not pre-warmed, with a total amount based on the patient weight + 40 cc (minimum of 90 cc and maximum of 130 cc) followed by a saline flush of 30 cc. The chest CT examinations were performed 40 seconds after contrast injection, whereas the chest and abdomen examinations were performed at 70 seconds.

4.2.3 Data curation

Radiological datasets are often heterogeneous. To mitigate differences in radiological image acquisition, all CT scans were cropped between

the liver and the lower neck region using the method proposed by Zhang et al. [ZWZ17], and linearly resampled to 2 mm isotropic voxel size. Hounsfield units were clipped between -120 (fat) and 300 (cancellous bone) and rescaled between 0 and 1. CT scans were further cropped to 192 x 192 x 160 voxels from the center point in order to provide the network with regular image shapes during training.

4.2.4 AI-mediated quantitative treatment monitoring

To harness AI for quantitative treatment monitoring, we developed a 3-dimensional convolutional neural network to perform image-to-image registration between subsequent follow-ups of the same patient (architecture shown in Figure 4.1), based on the research of Balakrishnan et al. [Bal+19] and Zaho et al. [Zha+19b]. The network comprised of two subsequent parts: the first performing affine registration aimed to provide alignment of the scans (i.e. to correct for different patient positions), the second section performing deformable registration and aimed to identify morphological changes during the course of the treatment (i.e. longitudinal tracking).

Architecture-wise, the first part of the network consisted of a VGG-like network comprised of a series of five convolutional blocks, and two fully-connected layers, regressing the 12 parameters of the affine transform. The output transform of the network was applied to the moving image, concatenated to the fixed image, and fed into the second part of the network. The second part of the network followed a U-Net architecture [RFB15], and it aimed to quantify non-linear anatomical differences between the input scans. This consisted of an encoding section, comprising 4 convolutional blocks downsampling the images by half the size via striding, a convolutional latent space with stride of 1, and 4 deconvolutional blocks each upsampling the inputs by double the size via striding. Skip connections were implemented between encoding and decoding layers following the implementation in the original paper. The network was trained to minimize the correlation coefficient loss [Zha+19b]. Unlike standard measurements of classical registration

procedures, this loss is easy to compute in the continuous case. Three penalties were also employed to mitigate for unlikely morphological deformations, each weighted 1/10 in the final loss. Adam optimizer was used during training, with an initial learning rate of 8 × 10-5. A curriculum learning scheme was implemented during training, such that the loss would be computed on a smoothed version of the images. The smoothing was implemented via average pooling, starting with a kernel size of 9, and reduced by 3 at epochs 100, 150, and 175. Batch size was set to 2. To mitigate negative effects resulting from the small batch size, group normalization was employed instead of batch normalization. Figure 4.1 shows a detailed overview of the model loss used. The network was trained on a publicly available dataset of 1010 patients of the lung image database consortium [McN+07; Arm+11; Cla+13] with 10% hold out during training to control for overfitting (i.e. patients whose ID were multipliers of 10 were held out). Our code can be found online[1].

4.2.5 Prognostication through quantitative monitoring

To explore the prognostic value of AI-mediated treatment monitoring, we trained a random forest classifier [B201], with wrapper feature selection, to predict survival based on network imaging features extracted from pairs of subsequent follow-up scans. More specifically, the RFC was trained longitudinally, on pairs of subsequent scans, to predict whether the patient would survive 1 year from the date of the latest of the two scans (see Figure 4.2). The input of the RFC consisted of 96 feature maps from the latent space of the decoder that represented the morphological changes between the prior and the subsequent scan. These are the deepest features found in the middle layer of the second section of the network — the one handling deformable registration. These features come in tensor shape, hence the name feature maps. For classification purposes, it is standard to transform the feature maps of

[1] code: github.com/nki-radiology/PAM.git

the network to a feature vector, to be fed into a classifier. Global average pooling is the technique commonly used to create a feature vector out of a set of feature maps: each entry of the feature vector is the average value of the corresponding feature map. Alongside the global average pooling, we also included standard deviation, skewness and kurtosis, as we deemed the feature maps too large to be represented just by the mean activation — 1000 values per feature map, compared to 49 of a classical ResNet architecture.

To correct for temporal discrepancies (e.g. differences in time between follow-ups), the amount of days elapsed between the two scans, and the days elapsed since the start of treatment were also fed to the RFC. Furthermore, morphological changes should be order invariant: the differences estimated between image A and B should be the same as the differences between image B and A. To provide order invariance, we applied element-wise multiplication of the feature maps generated by swapping the input scans. More specifically, we computed the feature maps for the scan pair prior-to-subsequent, and the feature maps for the pair subsequent-to-prior. Then we multiplied them together, element-wise. The multiplication preserved only those changes that were detected in both directions, therefore providing order invariance to our model. The discovery set was used for training, while testing was performed on the independent validation set. Both the registration network and the random forest classifier were trained on the partitioned data, at once, with their respective default parameters — no cross-validation or model selection was performed.

4.2.6 Prognostic heatmaps

Occlusion sensitivity was employed to visualize the parts of the image that were deemed prognostic of the outcome [ZF14]. The main idea of the occlusion algorithm is based on the assumption that removing a predictive section / region from the original image will change the algorithm prediction substantially. In contrast, removing a non-predictive section/region from the original image, the algorithm pre-

diction will stay unchanged. We occluded a section (or patch) of the input image presented to the RFC. The prognostic value of that patch is then computed as the difference of the RFC survival score produced by the occluded image vs the original unoccluded one. The resulting prognostic map is the result of the algorithm scrolling the ROI through the image, and repeating the procedure. This was filtered with the gross morphological changes map to produce a prognostic map of the gross morphological changes used for visual interpretation. Details of the algorithm reported in Algorithm 1. Visual assessment of the resulting prognostic maps was carried out by an expert reader (T.N.B., board certified radiologist, 2 years experience in thoracic imaging at a tertiary oncologic center), blinded to all clinical parameters, including survival. All scan pairs were assessed with the prognostic maps overlaid on top. The reader was tasked to identify the areas of activation (i.e. hot spots) in the scan pair, and report them categorized as tumor-related areas, secondary comorbidities, and general anatomical areas. Tumor-related areas and secondary comorbidities, which were not highlighted in the prognostic map, were recorded separately.

4.2.7 Independence from known prognostic factors

To test the independence of our AI model, we ran a multivariate analysis against known prognostic factors. Age and pathological cancer subtypes were extracted directly from the anonymized patient records. Changes in tumoral burden were computed based on the available manual segmentations of the total tumor — i.e. all visible and segmentable lesions in the body, except for bone and brain. To ensure comparability with 2D measurements from standard RECIST criteria, volumes where converted to pseudo-diameters via $d = \sqrt[3]{(6V/\pi)}$, where V is the total tumoral burden. This computes the diameter of the sphere equi-volumetric to the total tumour burden. Tumor PD-L1 expression scoring was performed according to the instruction manual of the qualitative immunohistochemical assay developed as a complementary diagnostic tool for nivolumab (PD-L1

IHC 22C3 pharmDx, Dako, Carpinteria, CA). PD-L1 expression levels were determined by observing complete circumferential or partial linear expression (at any intensity) of PD-L1 on the plasma cell membrane of viable tumor-cells. In parallel, the pattern of staining in CD4 stained slides, which also stain CD4+ lymphocytes and macrophages, was evaluated and compared to PD-L1 stained slides in order to avoid false positive assessment due to PD-L1 expressing macrophages in between tumor cells. Assessment of expression levels was performed in sections that included at least 100 tumor cells that could be evaluated.

Algorithm 1 Generation of Heatmaps for Model Explainability

procedure GENERATEHEATMAP($prior, subsq, \Delta_t, \Delta_{SoT}$)
    ref$_{score}$ $\leftarrow$ Survival score on the original images
    ROI $\leftarrow$ Cube of $64 \times 64 \times 64$ in the top left back corner
    occl $\leftarrow$ Set intensties within the ROI to zero in prior and subsq
    occl$_{score}$ $\leftarrow$ Compute the survival score on $occluded$
    ROI$_{importance}$ $\leftarrow$ |occluded$_{score}$ - ref$_{score}$|
    prog-map[ROI] $\leftarrow$ max(progmap[ROI], ROI$_{importance}$)
    if ROI has not scrolled through the whole image **then**
        Move the ROI 8 voxels along one of the axis
        Go to step 4
    def-map $\leftarrow$ Anatomical changes between prior and subsq
    return $smooth$ (def-map $\times$ prog-map)

4.2.8 Statistical analysis

To assess prognostic performance, the area under the receiver operating curve (ROC-AUC) was used. Confidence intervals were estimated via bootstrapping performed using repeated sampling with replacement (10000 times). Statistical significance was assessed via Mann-Whitney-U test. Kaplan Meier models were employed for survival analysis. Statistical significance of survival metrics

was assessed via log-rank test. Prognostic (treatment monitoring) performance was quantified in terms of overall survival from the date of the scan. Biomarker performance was quantified in terms of overall survival and durable clinical benefit (complete or partial response, or stable disease, for at least 6 months) from the start of treatment. Cox-Hazards models were used for comparison of known prognostic factors.

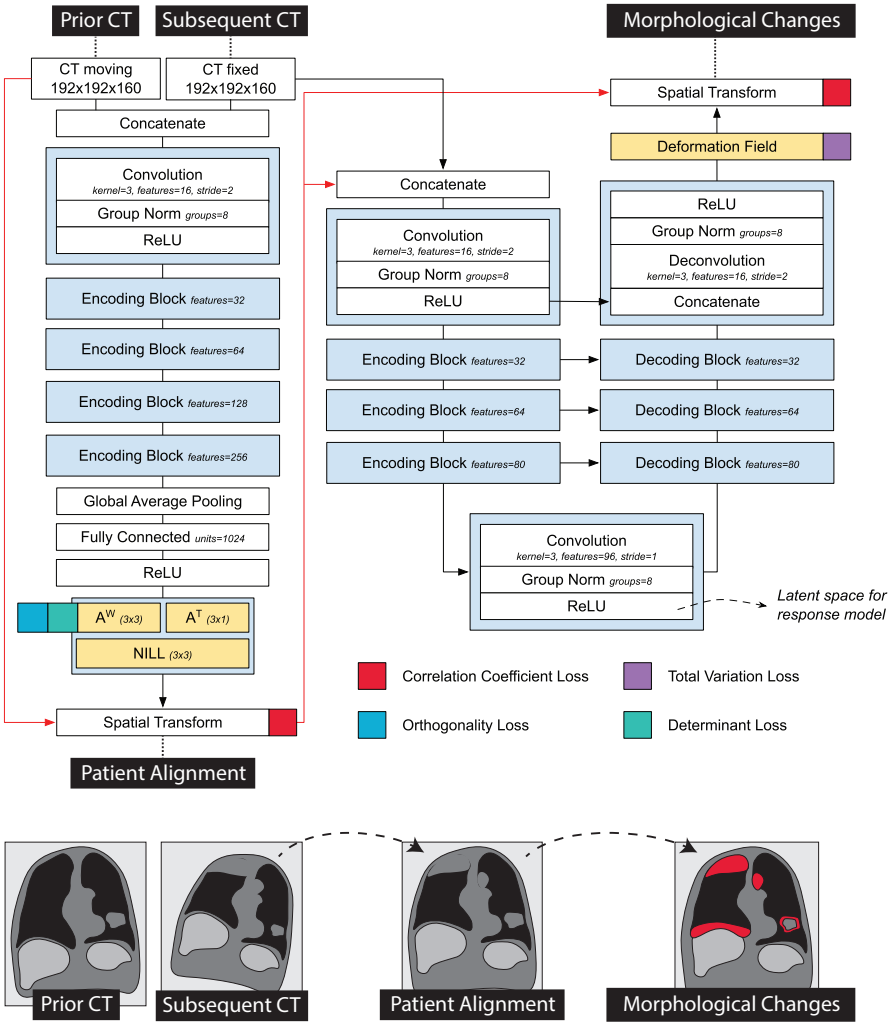

Figure 4.1: Detailed representation of the registration network used in the prognostic AI-monitoring framework.

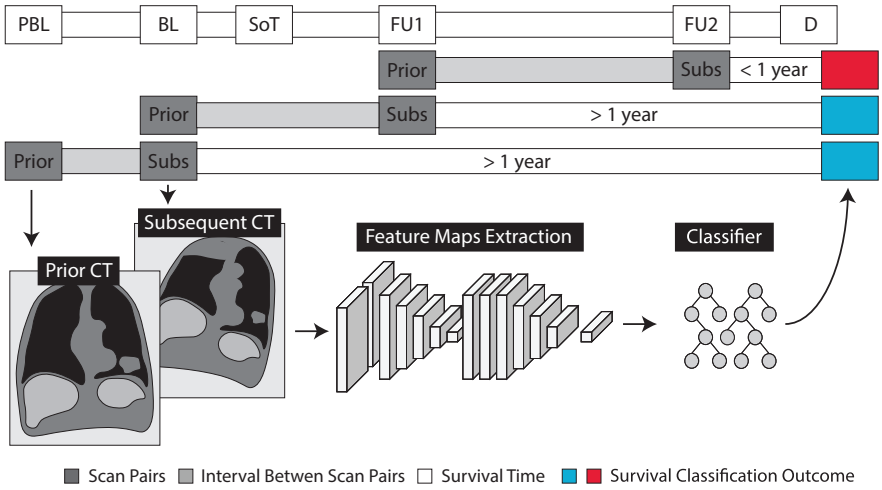

Figure 4.2: Schematic representation of the evaluation of prognostic values through quantitative monitoring. Radiological examinations are shown as pre-baseline (PBL), baseline (BL) and follow-up (FU), with respect to the start of treatment (SoT). Prediction of survival is made based on the time of death (D). For each pair of subsequent scans, we label the earlier one as prior and the subsequent as subsequent (Subs).

4.3 Results

4.3.1 Study cohort

A total of n=152 patients, n=903 CT scans, and n=611 scan matched pairs of subsequent CT scans were included in this study (see Figure 4.4). The discovery set consisted of n=73 patients (and n=276 scan pairs), while the independent validation set had n=79 patients (and n=335 scan pairs). Median age of the entire cohort was 64.4 (IQR 57.8 — 68.9), with a higher prevalence of males (57.9%). Adenocarcinoma was the most common subtype, reported in 61% of the cohort. No differences in clinical characteristics were encountered between discovery and validation set, except for survival. In comparison to the discovery set, the independent validation set had 180 days longer overall survival, and 101 days longer progression-free survival. Imaging-wise, we collected n=129 pre-baselines (PBL; 14.3%), n=149 baselines (BL; 16.5%), n=135 first follow-ups (FU1; 15.0%), and n=103 second follow-ups (FU2; 11.4%). Subsequent follow-ups (FU3+) constituted the remaining 42.9% of the dataset (n=387). Time-wise, BL scans were acquired on average 26 days before the start of treatment (IQR 37 — 14), while the first FU scan, 68 days after (IQR 46 — 77). Subsequent follow-ups were made on average every 77 days (IQR 55 — 95). Acquisition of non-contrast enhanced PET-CT instead of contrast enhanced CT was the main reason for lack of imaging during follow-up. Further patient characteristics in Table 4.1.

4.3.2 Image registration performance

We evaluated the performance of the registration algorithm merely to identify the cases where the registration algorithm failed. The evaluation of a registration algorithm is usually performed by evaluating the distance between two known corresponding landmarks in the registered image. This can be done automatically, in a circular fashion. Namely, by selecting N random points in an image, we can transform them to their new coordinates in the target image, and back, using the

registration functions T_{AB} to represent the transformation from source to target, and T_{BA} as the transformation from target to source. Ideally, these should be the inverse of one another. Practically however there is a registration error propagating from source to target and back. We estimate this error to be proportional to the euclidean distance between N and $T_{BA}(T_{AB}(N))$. It is not exactly the registration error, as this depends on two subsequent dependent registration steps. However, as registration is merely the auxiliary task in our model, a full evaluation of the registration procedure — also in terms of architecture and network components — is beyond the scope of this study. The purpose of this analysis is to analyze the worst cases, i.e. the failures of the algorithm.

We ran the evaluation for all scan pairs, with 100 randomly generated points that were transformed from prior to subsequent, and back to prior. The resulting error was 1.67 cm, on average (CI: 0.87 — 3.18). We selected for visual inspection the three three worst cases, with error 4.54, 3.76 and 3.75 cm, respectively (see Figure in 4.3). This can be considered the closest cases of failure of the algorithm. In each of these cases, we can notice the presence of unlikely deformation, like in the heart or the thoracic wall. Although a penalty was set to deter this behaviour, we would refrain from increasing it, as it might limit the ability of the network to model other deformations. The strength of the algorithm is represented by the classifier able to distinguish informative deformations from non-informative ones. Overall, in other locations of the image, the registration was still successful in matching anatomical structures properly.

4.3.3 Prognostic performance

We fed pairs of subsequent follow-up scans to our network trained for (CT chest) image-to-image registration, and trained a random forest classifier (RFC) on its feature maps to investigate the prognostic value of the imaging features learnt by the network. Overall results of the RFC survival score on the independent validation set show an AUC

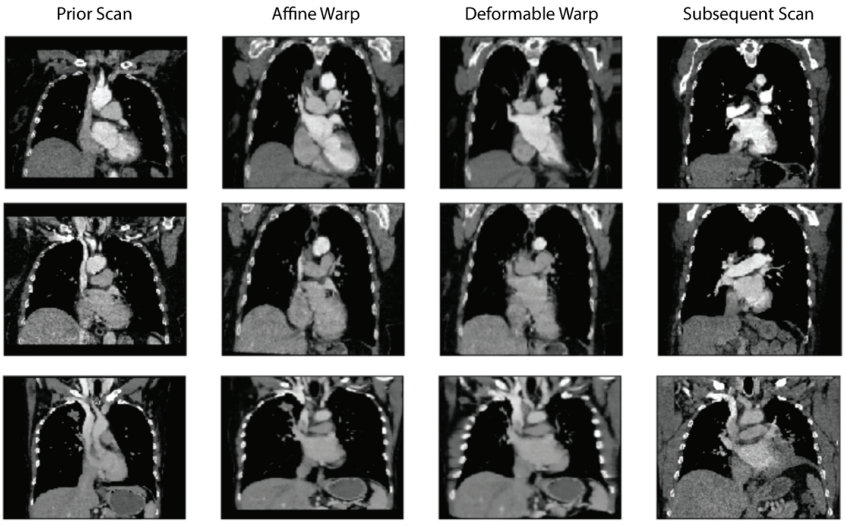

Figure 4.3: The three worst registration cases, and the result at each step of the deep learning registration pipeline.

of 0.68 (n=335, CI: 0.62 — 0.74, $p< 0.001$) to predict 1-year overall survival from the date of the later scan of the scan pair (see Figure 4.4b). The highest prognostic value can be found for the scan pair BL-FU1, reaching an AUC of 0.74 (n=61, CI: 0.61 — 0.86, $p< 0.001$), and for the scan pair FU1-FU2, reaching an AUC of 0.75 (n=42, CI: 0.58 — 0.89, p=0.002). A decrease in performance is observed during follow-ups, with a 0.71 AUC (n=42, CI: 0.50 — 0.89, p=0.02) for the pair FU2-FU3. None of these differences however reached statistical significance. Interestingly, RFC survival scores on the pair PBL-BL also showed prognostic value (0.69 AUC, n=51, CI: 0.54 — 0.83, p=0.01). After the fourth follow-up image, the prognostic performance of the model dropped (0.57 AUC, n=131, CI: 0.47 — 0.67, p=0.11). This trend becomes evident when looking at the performance with respect to the days between the

later scan in the scan pair, and the start of the treatment (see Figure 4.4c). In this respect, we divided the exam pairs in five groups, based on the time between the day of the later scan, and the day of start of treatment (i.e. before start of treatment, 0-90 days from start, 90-180 days and >365 days), and tested the performance in each group individually. Exam pairs performed before start of treatment showed an AUC of 0.72 (n=48, CI: 0.57 — 0.86, p=0.006), between start and 90 days after start of treatment showed an AUC of 0.73 (n=64, CI: 0.59 — 0.84, p<0.001), between 90 and 180 days showed an AUC of 0.68 (n=59, CI: 0.51 — 0.83, p=0.01), between 180 and 365 days an AUC of 0.66 (n=89, CI: 0.51 — 0.79, p=0.01). Exam pairs performed in the second year of treatment showed an AUC of 0.63 (n=75, CI: 0.50 — 0.75, p=0.04). Results summary in Table 4.2.

4.3.4 Biomarker performance

To investigate the prognostic value of AI-monitoring as a biomarker we ran a survival analysis on the scan pairs closest to the date of treatment start, i.e. PBL-BL and BL-FU1. High and low risk groups were defined for each scan pair by splitting the RFC survival scores on the median value. The scan pair BL-FU1 offered the highest prognostic performance (p= 0.02), with a median survival difference of 357 days (637 vs 280 days median survival respectively, p=0.02, see Figure 4.3d). A similar trend was observed for the PBL-BL pair, with a median survival difference of 239 days (467 vs 228 days median survival, respectively, see Figure 4.3e). This, however, did not reach statistical significance (p=0.16). For durable clinical benefit (6 months progression-free survival from start of treatment), we ran a classification analysis on the same scan pairs. This yielded a significant performance of 0.67 AUC (CI: 0.52 — 0.80, p=0.01) for the BL-FU1 pair, and a similar trend for the PBL-BL pair (0.61 AUC, CI: 0.44 — 0.77, p=0.10).

4.3.5 Combination of multiple time-points

To investigate the prognostic value of AI-monitoring across multiple time points, we combined the prognostic scores of PBL-BL monitoring, and BL-FU1 monitoring (see Figures 4.3f-g). For this particular analysis, we chose the start of treatment as reference, as differences in follow-up schemas might magnify when combining multiple time-points. Across the subset of patients analyzed (with PBL, BL and FU1 scans available, n=43), 53% survived 1 year after start of treatment (n=23). Patients with high expression of prognostic features during the monitoring of both PBL-BL and BL-FU1 (n=15) showed the highest increase in survival, with an enrichment from the baseline of 27% (80% survived 1 year after start of treatment). On the contrary, patients with low prognostic features on both PBL-BL and BL-FU1 (n=14) showed a diminution from baseline of 24% (29% survived 1-year after start of treatment). A point of interest is to be made for patients showing conflicting prognostic scores between PBL-BL and BL-FU1 (positive-negative and negative-positive, n=7, respectively). While these groups do not seem to show any deviation from the baseline (50% survived 1-year after start of treatment), further analysis on OS showed comparable results to the negative-negative group (p= 0.99) over a longer time span (2 year, see Figure 4.3h). The positive-positive group, on the other hand, kept showing significantly higher OS compared to both negative-negative (p= 0.01) and negative-positive (p= 0.003) groups.

4.3.6 Comparison with known prognostic factors

To compare the prognostic value of AI-monitoring against other known clinical prognostic factors, we ran a multivariate cox-hazards survival analysis. Specifically, we compared the RFC prognostic scores to age, cancer subtype, volumetric changes in total tumor burden between BL and FU1, and PDL1 expression at baseline. To mitigate collinearity, we reduced PBL-BL / BL-FU1 scores to a single score by principal component analysis. Complete data was available for 22 patients in the independent validation set. Results showed our

RFC survival score preserved statistical significance (0.35 HR, CI: 0.12 — 0.97, p= 0.04) against age (2.69 HR, CI: 1.20 — 6.05, p= 0.02), volumetric change of total tumor burden (2.36 HR, CI: 0.67 — 8.22, p = 0.18), >1% PDL-1 expression (0.26 HR, CI: 0.03 — 2.22, p= 0.22), adenocarcinoma (0.34 HR, CI: 0.03 — 4.43, p= 0.41) and squamous subtype (0.14 HR, CI: 0.01 — 3.01, p= 0.21).

4.3.7 Visual inspection of prognostic maps

The main idea behind predictive maps was to evaluate the predictive value of different regions of the image by removing those regions, one at the time, and estimating the difference in predicted survival. Figure 4.3i shows an example. The input scans are displayed in the first column. The second column shows the prognostic map generated by the occlusion algorithm (Algorithm 1). The patchy look of the overlay is the result of the cubic ROI, being scrolled around the image. Its intensity values are proportional to the change in predicted survival resulting from occluding that region. The third column is the deformation map, where hotspots correspond to regions of gross morphological changes (i.e. pleural effusion). The fourth column was the visualization presented to the reader. It is the result of the fusion between the prognostic map and the deformation map, and highlights the prognostic changes identified by the algorithm.

At visual inspection, lymph node metastases and lung lesions were common hotspots in the prognostic maps. Nodal metastases were present in 58% of scan pairs (n=57), and highlighted as prognostic in 81% of the cases (n=46). The mediastinum contained the most nodal hotspots, being highlighted in 80% of cases, followed by supraclavicular and hilar nodal metastases, highlighted in 67% and 57% of cases respectively. Axillary and pericardial nodal metastases were hotspots in 75% and 50% of cases, but found only in n=4 and 2 scan pairs respectively. Large lung masses were found in 45% of scan pairs (n=39), and highlighted as prognostic in 85% of cases. The same rate was observed for small lung nodules, while being less frequent, found in 30% of the

scan pairs (n=26). Bone metastases were found in 20% of scan pairs (n=17). Nonetheless, they were deemed prognostic by the algorithm in 82% of cases. Pleural masses, liver metastases and subcutaneous lesions, while being almost exclusively hotspots in the prognostic maps, accounted together for only 13 scan pairs. Among secondary comorbidities, pleural effusion, consolidations and atelectasis were the most common, accounting for 31%, 28% and 20% of scans pairs (n=27, 24, and 17, respectively). Hotspots were found in 94% cases of atelectasis (n=16), 93% cases of pleural effusions (n=25), and 83% cases of non-specific consolidation (n=20). Pericardial effusions were hotspots in 75% of the times, but found only in 8 cases. Only one case of ascites was reported, which the algorithm also highlighted as prognostic. Hotspots in anatomical regions included the spine in 56% of cases, the thoracic wall in 55% of cases, and various regions in the upper thorax, including periscapular (51%), shoulders (49%), neck (48%), and supraclavicular (45%), with the exception of the axilla, highlighted only in 13% of scan pairs. Normal lung parenchyma was highlighted in 28% of cases. Remaining hotspots include the great vessels (9%) and the breast (4%). Detailed summary reported in Table 4.3.

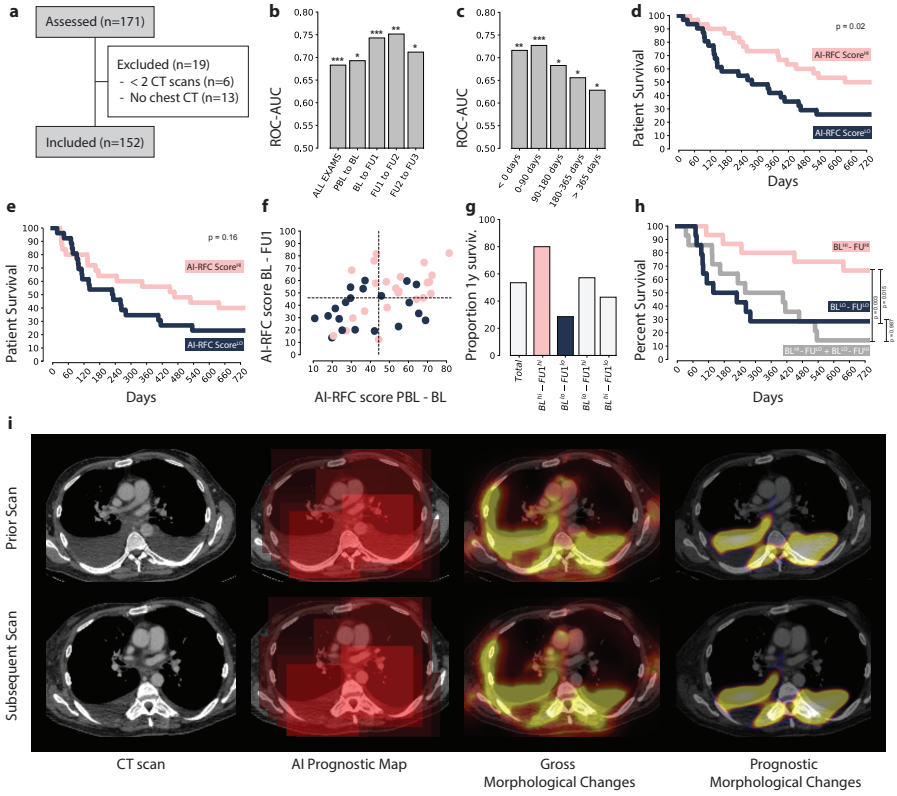

Figure 4.4: **(a)** CONSORT diagram **(b)** 1-year survival classification performance on the independent validation set, with respect to the clinical follow-up routine (highlighted in green the ROC-AUC of the scan pairs used for the 2-years survival analysis) and **(c)** corrected by time. **(d)** 2-years Kaplan-Meier curves of the RFC survival score of BL-FU1 and **(e)** PBL-BL. **(f)** Combination of the PBL-BL and BL-FU1 RFC survival scores with **(g)** enrichment of each of the four quadrants **(f)** and **(h)** survival of each of the four quadrants. **(i)** Example of the occlusion sensitivity method used for AI explainability and visualization.

4.4 Discussion

Advanced treatment monitoring through more detailed quantitative descriptors of the overall status of the patient, as visualized on routine imaging scans, could provide valuable prognostic information. Our aim was to investigate the potential prognostic value gained by AI-based treatment monitoring on imaging in NSCLC patients treated by PD-1 checkpoint inhibitors. To test this, we implemented a convolutional neural network for image-to-image registration, and trained it on a large public dataset of chest CT scans. The trained network was then used to longitudinally model gross morphological changes between subsequent scans of NSCLC patients receiving PD1 checkpoint inhibitors. Morphological changes identified by the network were then used to train a classifier to predict 1-year OS from the date of the latest scan.

Our results showed significant performance in the independent test set for the prediction of 1-year OS from the date of image acquisition, with an average AUC of 0.69, and up to 0.75 AUC for the first 3 to 5 months after start of treatment, and 0.67 AUC for durable clinical benefits, suggesting the presence of (AI-quantified) gross morphological changes encoding prognostic value. These results are comparable to state-of-the-art methods, which currently employs laborious and time-consuming segmentation procedures [Tre+19; Sun+18]. While the field of research has been focusing on single-lesion analysis — leveraging different known factors in cancer growth, including vascularity [Ali+19], oxygenation [Tun+], and metabolic activity [Mu+18] — our approach offers a novel fully automatic procedure which completely eradicates the need of time-consuming segmentations, and simultaneously offers a way to provide a full picture of the patient status as seen on chest imaging. While this does not preclude the usefulness of the single-lesion approach, it proposes a way for future multi-scale solutions that leverage both single lesion imaging biomarkers as well as whole image approaches that provide general quantitative information about the status of the patient receiving treatment. Research ef-

forts, however, have to be made in order to overcome the bottleneck of manual ROI delineation procedures, either in the form of automatic segmenters [Yan+18], or with implicit AI representations of the total tumor burden.

In addition to the statistical analysis of the performance, we investigated the choices the AI made by means of sensitivity occlusion [ZF14]. This resulted in a set of prognostic heatmaps, highlighting regions of morphological changes that the AI deemed prognostic relevance. Gross morphological changes in nodal and lung lesions held the highest prognostic value, especially nodal lesions in the mediastinum, hilum, and supraclavicular region. Further results suggested additional prognostic value for morphological changes affecting the lungs, either in the form of compression from the thoracic wall (due to pleural effusion or pleural masses), non-specific consolidations, or atelectasis. These results also seemed to extend to other regions, with ascites and pericardial effusions also being highlighted as prognostic, despite their rare occurrence. The AI seemed to pay particular attention to the skeleton, with the spine being the anatomical region most commonly highlighted by the AI in the prognostic maps, and bone metastases deemed prognostic in most cases where those were present. As common imaging follow-up schemas, such as RECIST [Sch+16; Sey+17a], do not account for tumor burden in the bones, our findings suggest that, on the contrary, such phenomena should not be ignored. Further investigations should lead to novel guidelines, which can provide valuable contribution from the imaging beyond diametrical measurements.

Particular attention should also be paid to nodal metastases and nodal growths during treatment. Imaging features of nodal metastases were found already to be correlated with disease progression for NSCLC, melanoma, and head and neck cancer [Tre+19; Yu+19], though no distinction was made between the location of the lymph nodes. However, both our findings and the current literature suggest that this information may be of value. This would be especially interesting in the light of regional (tumor draining) lymph nodes which play a critical role in

terms of anti-tumor immunity and priming (37), increased expression of cytokines and checkpoint markers [Ho+20a], and changes in the immune compartments resulting in a tumor favorable microenvironment [JPP18]. A major hurdle that remains in the analysis of lymph nodes is represented by the radiological assessment, often in contrast with the pathological one. Most radiomics studies so far focused on the detection of positive nodal metastases rather than their prognostic values [Ho+20b; Sha+20; Tan+19; Li+20; Zhe+20; Che+20].

The analysis of lung lesions is far more common. Imaging features from lung lesions have been reported to hold prognostic value for patients receiving immunotherapy in several studies [Tre+19; Mu+18; Tan+18; Tun+19; Sun+20b; Pat+19]. Indeed our findings confirm the association between lung lesions and treatment outcome, with about 85% percent of them being hotspots in the AI-generated prognostic maps, independent of size. Most of the studies published so far focus on the analysis of the tumor region and/or the peritumoral boundary, which may hold valuable information regarding tumor vascularization and inflammatory environment. In this study, the proposed AI model monitors the whole image including both the healthy tissue as well as the tumor(s). As the growth of a cancer lesion does not uniquely depend on the genetic makeup, but rather a complex interaction of microenvironmental features and favorable location for seeding, it would not be surprising to establish a link between a comprehensive modelling tool of cancer growth and its biological features. Even in this case however, further research is needed to establish any link between imaging features and tumor biology.

Following our results, we observed an increase in the prognostic performance of the AI resulting from the combination of multiple time points, namely pre-baseline, baseline and first follow-up. This analysis showed good OS for patients with higher AI-survival scores (AI-RFCˆhi) in both pre-baseline to baseline scan pair, and baseline to first follow-up — and worse OS for the opposite case (AI-RFCˆlo). Interestingly, patients with contradicting scores (AI-RFCˆhi for pre-baseline to baseline scan pair, and AI-RFCˆlo for the baseline

to follow-up, and vice versa) showed worse survival, similar to the double negative group. These results suggest the existence of a prognostic combination of pretreatment and early-treatment characteristics, both of which should be accounted for during patient stratification. Further insights could be achieved by more advanced AI methods that would account for larger time spans, or even the entirety of patients' treatment history.

The combined score was demonstrated to be an independent prognostic parameter even when corrected for other known prognostic parameters. This is of particular interest when we consider the possible role of such a tool, for example as an additional input to the tumor board during treatment decision making. Further research is required to study its implementation in the clinical settings.

4.4.1 Limitations and future outlook

Our study aimed to monitor AI-measured gross morphological changes between imaging follow-up for survival prediction in NSCLC patients receiving PD1 checkpoint antibodies. In this study, we pre-trained a neural network on a large dataset of chest CT scans, and fine-tuned it for survival on our smaller local immunotherapy dataset. Under the current settings, we limited the analysis to chest imaging which, in addition to the chest, frequently included the lower neck and the upper abdomen. While this limitation could hold for lung malignancy, extension to other cancer types would require this technique to be extended to include the whole body — i.e. the abdomen and, when available, the brain. Moreover, due to the limited amount of data, it was not possible to explore more complex machine learning algorithms for the prediction of survival, nor for more precise visualization of the prognostic maps. Expansion of the dataset, both in terms of patients and in terms of time points, would certainly allow for an increase in performance and better explainability of the AI algorithm. Specifically, an extension of the field of view of the algorithm to the whole body, as well as the usage of parameters

other than imaging, could potentially improve the performance of the algorithm to be usable in the clinics. Further clinical validation of the method is also needed. While this study presented a comparison of this method with response evaluation criteria (e.g. changes in total tumor burden) and biomarkers (e.g. PD-L1 expression), the primary objective for future studies should be a comparison with the clinical standard, namely the RECIST criteria. It remains to be investigated whether this method would be complementary to the current radiological response evaluation (i.e. RECIST). Furthermore, additional investigations are required to link biological features to tumor growth and gross morphological changes. Further analysis should also study the effects of different machine acquisition parameters, and the sensitivity of the method to imaging acquisition parameter variability. Looking into the future, we envision that an AI solution could be set up as a clinical decision support system capable of providing information to the treating physician complementary to traditional clinical and pathological input data.

4.5 Conclusion

In this study, we aimed to investigate the potential prognostic value of AI-mediated monitoring in NSCLC patients receiving PD-1 blockade. We hypothesized the existence of quantitative imaging features describing a set of gross morphological changes happening during treatment that hold prognostic information. Our results demonstrate the existence of such factors (as described by the AI on imaging), that are tumor-related, such as nodal, lung and bone lesions, as well as non-tumor related, such as pleural effusions, atelectasis and non-specific consolidations. Further investigation should focus on the development of more flexible models that can extend beyond thoracic imaging, as well as on external validations.

	Entire Dataset	Discovery Set	Validation Set
Patient Characteristics			
N	152	73	79
Age [median, IQR]	64.4 (57.8-68.9)	64.5 (58.3-69.2)	64.2 (56.2-68.2)
Gender [N, %]	88 Males (57.9%)	44 Males (60.3%)	44 Males (55.7%)
Survival [median days]	341	269	449
Adenocarcinoma [N, %]	92 (60.5%)	46 (63.0%)	19 (26.0%)
Squamous [N, %]	35 (23.0%)	46 (58.2%)	16 (20.3%)
Radiological Follow-up			
All scan pairs	611	276	335
— PB-BL to BL [N, %]	93 (15.2%)	42 (15.2%)	51 (15.2%)
— BL to FU1 [N, %]	116 (19.0%)	55 (19.9%)	61 (18.2%)
— FU1 to FU2 [N, %]	100 (16.4%)	50 (18.1%)	50 (14.9%)
Days b/w scans in any scan pairs [median, IQR]	77.0 (55.0-95.0)	77.0 (50.0-97.2)	77.0 (56.0-94.0)
— Pre-baseline to baseline [median, IQR]	76.0 (55.0-113.0)	75.0 (47.0-114.8)	76.0 (61.0-98.0)
— Baseline to follow-up 1 [median, IQR]	85.5 (69.0-105.0)	86.0 (68.5-107.0)	85.0 (70.0-104.0)
— Follow-up 1 to follow-up 2 [median, IQR]	57.0 (44.0-78.2)	53.5 (43.0-75.0)	72.0 (47.5-83.5)
Days b/w treatment start and BL [median, IQR]	-26.0 (-37.0–14.0)	-25.0 (-34.8–12.2)	-27.0 (-37.0–14.0)
Days b/w treatment start and FU1 [median, IQR]	68.0 (46.0-77.2)	67.0 (46.5-73.0)	68.0 (46.0-78.0)

Table 4.1: Patient and imaging data characteristics.

	N -	N +	p-value	Area under the ROC curve
With respect to the follow-up sequence				
All	128	207	< 0.001	0.68 (CI: 0.62 — 0.74)
PBL — BL	27	24	0.010	0.69 (CI: 0.54 — 0.83)
BL — FU1	30	31	< 0.001	0.74 (CI: 0.61 — 0.86)
FU1 — FU2	18	32	0.002	0.75 (CI: 0.58 — 0.89)
FU2 — FU3	14	28	0.015	0.71 (CI: 0.50 — 0.89)
FU3 +	39	92	0.112	0.57 (CI: 0.47 — 0.67)
With respect to days from start of treatment				
< 0	25	23	0.0057	0.72 (CI: 0.56 — 0.86)
0 — 90	33	31	< 0.001	0.73 (CI: 0.60 — 0.84)
90 — 180	19	40	0.013	0.68 (CI: 0.51 — 0.83)
180 — 365	26	63	0.011	0.66 (CI: 0.51 — 0.79)
365 +	25	50	0.037	0.63 (CI: 0.50 — 0.75)

Table 4.2: Prognostic and predictive performance.

	ALL	PBL — BL	BL — FU1
Tumor Related			
Lymph Nodes	46/57 (80.70%)	21/27 (77.78%)	25/30 (83.33%)
— Pericardial	1/2 (50.00%)	1/1 (100.00%)	0/1 (0.00%)
— Mediastinal	42/53 (79.25%)	18/25 (72.00%)	24/28 (85.71%)
— Hilar	16/28 (57.14%)	7/12 (58.33%)	9/16 (56.25%)
— Supraclavicular	16/24 (66.67%)	5/10 (50.00%)	11/14 (78.57%)
— Axillary	3/4 (75.00%)	1/2 (50.00%)	2/2 (100.00%)
Large Lung Nod.	33/39 (84.62%)	16/20 (80.00%)	17/19 (89.47%)
Small Lung Nod.	22/26 (84.62%)	8/11 (72.73%)	14/15 (93.33%)
Bone Metastases	14/17 (82.35%)	7/7 (100.00%)	7/10 (70.00%)
Pleural Masses	6/6 (100.00%)	3/3 (100.00%)	3/3 (100.00%)
Liver Metastases	5/6 (83.33%)	2/3 (66.67%)	3/3 (100.00%)
Subq. Lesions	1/1 (100.00%)	—	1/1 (100.00%)
Secondary Comorbidities			
Pleural Effusion	25/27 (92.59%)	12/12 (100.00%)	13/15 (86.67%)
Consolidation	20/24 (83.33%)	10/12 (83.33%)	10/12 (83.33%)
— Post-radiation	3/3 (100.00%)	2/2 (100.00%)	1/1 (100.00%)
Atelectasis	16/17 (94.12%)	9/9 (100.00%)	7/8 (87.50%)
— Post-obstructive	7/8 (87.50%)	4/4 (100.00%)	3/4 (75.00%)
Pericardial Effusion	6/8 (75.00%)	2/3 (66.67%)	4/5 (80.00%)
Ascites	1/1 (100.00%)	—	1/1 (100.00%)
General Anatomical Areas			
Spine	48/86 (55.81%)	26/43 (60.47%)	22/43 (51.16%)
Thoracic Wall	47/86 (54.65%)	25/43 (58.14%)	22/43 (51.16%)
Periscapular	44/86 (51.16%)	20/43 (46.51%)	24/43 (55.81%)
Shoulder	42/86 (48.84%)	23/43 (53.49%)	19/43 (44.19%)
Neck	41/86 (47.67%)	20/43 (46.51%)	21/43 (48.84%)
Periclavicular	39/86 (45.35%)	19/43 (44.19%)	20/43 (46.51%)
Lung Parenchyma	24/86 (27.91%)	13/43 (30.23%)	11/43 (25.58%)
Axilla	11/86 (12.79%)	6/43 (13.95%)	5/43 (11.63%)
Great Vessels	8/86 (9.30%)	5/43 (11.63%)	3/43 (6.98%)
Breast	3/86 (3.49%)	1/43 (2.33%)	2/43 (4.65%)

Table 4.3: Results from the visual inspection of the AI-generated prognostic maps. **Subq.** = subcutaneous, **Nod.** = Nodule.

5

Whole-body imaging-based prognostic monitoring

Stefano Trebeschi et al. "Development of a prognostic AI-monitor for metastatic urothelial cancer patients receiving immunotherapy". In: *Submitted for publication.* (2021).

Abstract

Background Immune checkpoint inhibitor efficacy in advanced cancer patients remains difficult to predict. Imaging is the only technique available that can non-invasively provide whole body information of a patient's response to treatment. We hypothesize that quantitative whole-body prognostic information can be extracted by leveraging artificial intelligence (AI) for treatment monitoring, superior and complementary to the current response evaluation methods.

Methods To test this, a cohort of 74 stage-IV urothelial cancer patients (37 in the discovery set, 37 in the independent test, 1087 CTs), who received anti-PD1 or anti-PDL1 were retrospectively collected. We designed an AI system able to identify morphological changes in chest and abdominal CT scans acquired during follow-up, and link them to survival. We termed this system the prognostic AI-monitor (PAM).

Results Our findings showed significant performance of PAM in the independent test set to predict 1-year overall survival from the date of image acquisition, with an average area under the curve (AUC) of 0.73 ($p<0.001$) for abdominal imaging, and 0.67 AUC ($p<0.001$) for chest imaging. Subanalysis revealed higher accuracy of abdominal imaging around and in the first 6 months of treatment, reaching an AUC of 0.82 ($p<0.001$). Similar accuracy was found by chest imaging, 5 to 11 months after start of treatment. At univariate and multivariate comparison with current monitoring methods (laboratory results and radiological assessments) PAM remained significant with higher or similar, suggesting its complementary value.

Conclusions Our study demonstrates that a comprehensive AI-based method such as PAM, can provide prognostic information in advanced urothelial cancer patients receiving immunotherapy, leveraging morphological changes not only in tumour lesions, but also tumour spread, and therapy induced non-tumoral lesions. Further investigations should focus beyond anatomical imaging. Prospective studies are warranted to test and validate our findings.

5.1 Introduction

Durable clinical benefit to immune checkpoint inhibitors in metastatic setting led to approval in several malignancies [Pla+18; 19; Bal+17]. Unlike traditional cancer treatments, such as chemotherapy and radiotherapy, which are administered for a predefined amount of time, immunotherapy is generally administered until there are tangible clinical benefits or until progressive disease/adverse events deem it unsuitable — for a maximum of 2 years. To achieve this, an accurate treatment evaluation method is required.

Whole-body Computed Tomography (CT) provides information on the full-picture of the patient. Beyond tumor size dynamics, CT imaging allows assessment of immune-related side-effects and/or disease related complications.

Therapy response evaluation following CT is measured according to the response evaluation criteria in solid tumour (RECIST) [Eis+09], or iRECIST, adapted for immunotherapy [Sey+17b]. This involves prospective tracking of preselected lesions by measuring 2-dimensional diameters. Various immune-related toxicities and cancer related complications that inform clinical practice may also be identified on CT scans, but are not accounted for in current RECIST criteria. So far, a comprehensive quantitative approach that involves quantitative response evaluation and clinically relevant conditions is lacking.

Quantitative approaches, such as radiomics, have been explored in the past [Tre+19; Sun+18]. While these led to satisfactory results in the field of prognostication, these rely mostly on manual segmentations, which are time-consuming and prone to human operator error. A comprehensive non-invasive method that comprises the assessment of tumour size dynamics and side-effects or other cancer-induced conditions, in an automatic and precise quantitative manner, would be preferable.

Novel techniques of computational imaging and artificial intelligence (AI) can be the basis for quantitative methods for treatment monitoring [Tre+21b]. Specifically, AI algorithms can be seen as methods to capture, measure, and quantify complex highly-variable anatomical phenomena for prognostic purposes, in a robust and time-efficient manner. To this end, we have developed an AI algorithm that performs automated tracking and quantification of morphological changes based on longitudinal CT imaging in immunotherapy treated patients, allowing correlations with overall survival. We term our AI system the Prognostic AI-monitor (PAM). Recently, a similar pilot approach was tested in a study on chest imaging of a NSCLC cohort [Tre+21b], demonstrating accurate response prediction and a correlation with overall survival. In this study, we aim to extend the model to thoracoabdominal imaging, and validate it on a cohort of metastatic urothelial cancer patients treated with anti-PD1/PDL1. The model accuracy will be assessed at various time points within the treatment timeline, and the explicability through qualitative investigation of AI-generated prognostic heatmaps.

5.2 Materials and methods

5.2.1 Study cohort

We retrospectively included stage-IV urothelial cancer patients treated with anti-PDL1 or anti-PD1 monotherapy that had started follow-up imaging at the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital (NKI-AVL, Amsterdam, The Netherlands) between 07-2014 and 06-2018. Response evaluation was done using regular contrast-enhanced CT scans of the abdomen, chest, or both. For all patients, CT imaging data from 6 months prior to immunotherapy, up to 2 years follow-up was retrieved. Inclusion criteria were high resolution images (slice thickness < 5mm), and the presence of at least thorax or abdomen in the scan field. These criteria were verified automatically via the DICOM tag, or via the

automatic localization algorithm proposed by Zhang et al. [ZWZ17], respectively. For each patient, we recorded age at start of treatment, date of start of treatment, and date of death. Additionally, to compare PAM with current treatment monitoring standards, we collected parameters of radiological assessments (progression and response), as well as routine clinical blood analyses (hemoglobin, leukocyte count, thrombocyte count, and erythrocyte count). The entire dataset was divided into a discovery and an independent test set based on the patients' ID (even IDs were assigned to the discovery set, odd IDs were assigned to the independent test set, creating a 50/50 split). The study was conducted at the NKI-AVL after approval of the local Institutional Review Board (IRBd19-083).

5.2.2 Data harmonization

A data harmonization protocol was applied to mitigate heterogeneity from typical real-world imaging datasets. This consisted of isotropic linear resampling of the scans at 2 mm, clipping of the Hounsfield units between -120 (fat) and 300 (cancellous bone), and rescaling of the intensities between 0 and 1. All images were cropped and padded to $192 \times 192 \times 192$ voxels (160 axial coordinate for chest imaging).

5.2.3 Prognostic AI-monitor

PAM is composed of three AI-modules. The first module, termed *localizer*, consisted of a VGG-like convolutional network, tasked to crop out the chest and the abdomen in two separate images, each according to standardized anatomical locations. These were defined as the space between the lower neck and the lower diaphragm, and the space between the upper diaphragm and the lower pelvis, respectively. The second and third modules, termed *trackers*, consisted of two instances of the same convolutional network, one trained for chest imaging, and one for abdominal imaging, tasked to quantify morphological changes between pairs of images. We

termed these modules the chest and abdominal tracker, respectively. Their architecture was based on radiological deep learning-based image-to-image registration. At its core, each tracker is tasked to match anatomical landmarks and shapes of two 3D radiological images. In doing so, the network learns to quantify anatomical differences between pairs of scans. We leveraged the tracker network knowledge (i.e. its latent representation) to extract quantitative imaging feature vectors, representing morphological changes between follow-up scans of the same patient, and fed them into a classifier trained to predict survival. Time from start of treatment, and time between scans were also fed into the classifier, for temporal reference.

5.2.4 Localizer module

The localizer module was designed following the research of Zhang et al. [ZWZ17]. The authors showed how a convnet trained to sort slices in a specific order (e.g. from head to toe) can be used for anatomical localization. The network followed a siamese learning scheme. It received a pair of CT slices from a single scan, and had to learn which of the two slices would be on top of the other in the original CT scan. The only way for the network to learn to perform this task would be for the network to assign to each anatomical location a number i that would increase from head to toe. Once the training was complete, we used the network to retrieve a specific location by searching for their assigned number (for example, in our case, the upper-most point of the diaphragm was always assigned to be around $i = 25$). This algorithm idea is particularly powerful, as the ground truth (i.e. the order of the slices) can be automatically extracted from the CT scan, and therefore it does not require any manual labelling.

Our localizer module was built largely based on Zhang's architecture design — the exact architecture we used is shown in Figure 5.1a. The network was trained following the same siamese learning scheme of

the original research [ZWZ17]. Binary cross-entropy was the loss function chosen, the optimizer was Adam with an initial learning rate of 0.001, and the batch size was set to 8 (i.e. 8 random scans, one random pair of slices per scan). As it was difficult to set a number of epochs (considering it could be based either on the number of scans or on the number of slices), we chose to set a general number of iterations, namely 50,000. RANSAC regression [FB81] was used to model the relation between the network score and the actual slice number for each scan. We chose RANSAC for its robustness to irregularities provided by the localizer algorithm. Figure 5.1b shows the localizer network applied to a scan.

5.2.5 Tracker module

The tracker module was designed following the research of Balakrishnan et al. [Bal+19] and Zhao et al. [Zha+20], as well as our previous work on chest imaging in NSCLC [Tre+21b]. The network receives two images as input (i.e. a moving and fixed one) concatenated along the channel axis. The architecture of the network processes the input in two subsequent parts. The first part, consisting of VGG-like convnet, parses the images through a series of five subsequent convolutional blocks and two fully connected layers, to regress the 12 parameters of the affine transform. This is used to give a linear pre-alignment between the input images, correcting for different patient positions. The second part of the network follows a U-Net architecture [RFB15], where the inputs (i.e. the affine warped moving image and the fixed image) are processed together to regress a displacement field. The displacement field specifies for each voxel a 3D vector. The vector indicates where the voxel in that location of the moving image would be displaced to, in order to match the corresponding anatomical structure in the fixed image. This part of the network consisted of an encoder with four convolutional blocks downsampling the images by half the size via striding, a convolutional latent space with stride of one, and four deconvolutional blocks each upsampling the inputs by double the

size via striding. Skip connections were implemented between encoding and decoding layers following the implementation in the original paper. Both affine and deformable parameters are applied to the moving image through a spatial transformation layer.

The network was trained to minimize the correlation coefficient loss [Bal+19]. Three penalties were also employed to mitigate for unlikely morphological deformations: two on the affine loss (weighted 1/10), and one on the deformable loss (weighted 1/100). We decided to decrease the weight on the deformable loss to give to the model more freedom in modeling abdominal changes. Adam optimizer was used during training, with an initial learning rate of 3×10^{-4}. A curriculum learning scheme was implemented during training, such that the loss would be computed on a decreasingly smoother version of the images. The smoothing was implemented via average pooling, starting with a kernel size of 9, and reduced by 3 at epochs 100, 150, and 175. Batch size was set to 2. To mitigate negative effects resulting from the small batch size, group normalization was employed instead of batch normalization [WH18]. Figure 5.2a shows a detailed overview of the model and the loss used.

Both the localizer and trackers were unbiased towards both cancer and treatment, and could be trained on unlabeled data. Using The Cancer Imaging Archive (TCIA) [Cla+13], we collected[1] all available radiological images, and excluded scans with non-axial acquisition, low resolution ($>$ 5mm), animals (e.g. mice, suine) and phantoms. Based on thorax-abdominal CT scans, we then trained the localizer module on a lymphadenopathy dataset and extracted abdomen slices from all archived CT scans. Next, the isolated set of abdominal CT scans [Smi15; Bei+15; Kin+19; Gro+20; Hel+19; Kin+17; Bak+17; Rot+15; Bos+15; PPT19b; Kur+15; MWK15; Val+15; Aer+15; Wee+19; BJJ15; Yor+19; Val+17; Aer+19] were employed to train the abdominal tracker PAM module. We kept a 10% hold out during training to control for overfitting (i.e. patients whose ID were multiples of 10 were held out).

[1] Accessed on the 21st of April 2020

For the chest AI tracker module, we leveraged the trained weights from the NSCLC study [Tre+21b]. The code of both tracker and localizer have been added to the department AI repository[2].

5.2.6 Association with survival

In order to predict survival, we trained a logistic regression classifier based on the quantitative features extracted from the tracker. More specifically, we leveraged the feature maps in the deepest layer of the U-Net (this is shown in Figure 5.2a). To obtain a feature vector that can be used for the standard logistic classifier, we applied global average pooling. The resulting feature vector (96 entries or features) was fed into the logistic regression model to predict whether the patient would die within one year after the date of the latter scan, see Figure 5.2b. Time from start of treatment, and time between scans were also fed into the classifier, for temporal reference.

5.2.7 Comparison to clinical sstandards for monitoring

We compared PAM against radiological assessments and blood values. For simplicity, we limited the analysis to PAM-scores of abdominal imaging. We employed both univariate and multivariate comparison. The large majority of scans included in the analysis of PAM did not have a corresponding radiological assessment, or blood exam done on the same day — in other words, there was no one-to-one matching for the majority of the cases. To overcome this limitation, we averaged the values of both radiological assessments and blood work over a window of 6 weeks, centered on the date of the CT scan analyzed by PAM. Since PAM leverages tracking of morphological changes, we applied the same principle to the blood values. Namely, we estimated the rate of change of each blood value over time, i.e. $(v_s - v_p)/dt$, where v_p and v_s is the blood values at prior and subsequent scan, respectively, and dt is the time in between.

[2] code: github.com/nki-radiology/PAM.git

Radiological progression and response were assessed based on an increase in diameter of 20% or decrease of 30% in diameter, respectively, according to RECIST standards. Diameters were derived using $d = \sqrt[3]{(6V/\pi)}$, where V is the tumour volume delineated by a radiologist (PA). As these assessments already represented longitudinal change, they were left untouched, allowing for the creation of two classes (i.e. "response" and "progression").

5.2.8 Prognostic heatmaps

In order to interpret the results from PAM, we employed an occlusion sensitivity method [ZF14]. With this method, we occluded a section (or patch) of the image to the AI, by setting its voxel intensities to zero. We collected the prediction made by the AI on the occluded image, and compared it with the prediction on the original image. The importance of that patch was defined as the absolute difference between the predictions made on the occluded and the prediction made on the original image. A heatmap was generated by scrolling the occluded patch through the image, and collecting the relative importances of each patch. We termed the resulting visualization the prognostic heatmap. A board certified radiologist (TNB, specialized in thoracic and abdominal oncologic imaging, blinded to the outcome) was tasked to visually analyse the prognostic maps for a subcohort of the validation set. These were patients that had both thoracic and abdominal imaging. We chose the first available scan pair closest to the start of treatment — namely baseline and first follow-up. The radiologist was tasked to identify the location of highlights on the heatmaps, as well as pathologies/anomalies that were not highlighted, i.e. "hotspots" and "coldspots", respectively. Expert assessments were categorized based on whether they were hotspots or coldspots. This resulted in three classes of interest: hotspots on tumour or therapy related lesions, hotspots on seemingly healthy parenchyma, and coldspots on tumour or therapy related lesions. Coldspots on healthy tissues are trivial, and therefore not accounted for.

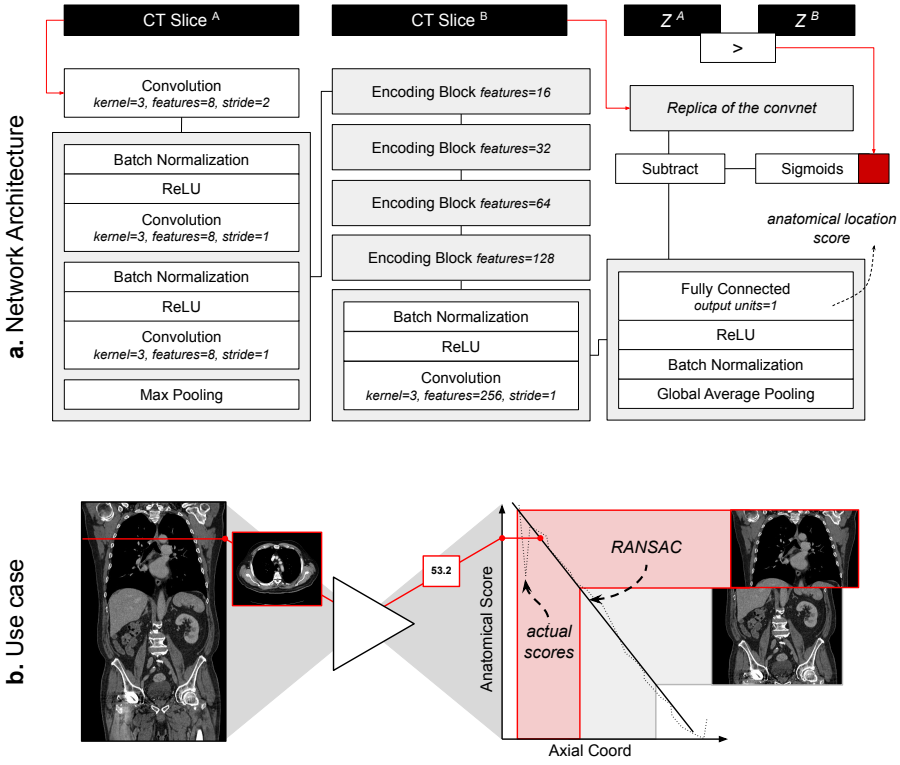

Figure 5.1: **(a)** Schematic representation of the localizer architecture and its training. The CT slices in input are axial slices taken from the same CT scan. Z_1 and Z_2 are the axial coordinates of each input slice, respectively. The red square symbolizes the binary cross entropy loss function used during training. **(b)** A use-case of the localizer. Each axial slice is processed through the network to generate a score. A linear relation between the scores and the axial coordinate is estimated. Cropping of the thorax and abdomen is done based on the anatomical scores, and corresponding axial slice.

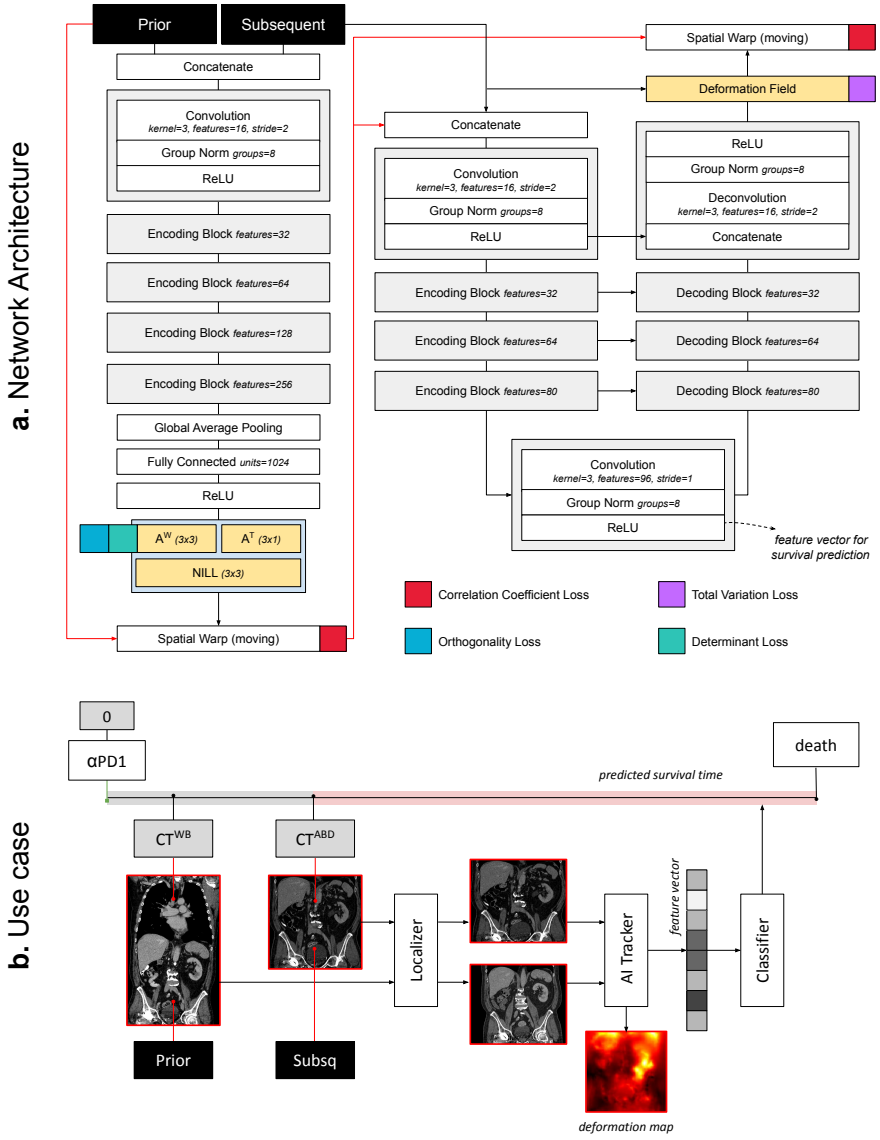

Figure 5.2: **(a)** Schematic representation of the tracker architecture and its training. **(b)** Use case of the tracker within PAM.

5.3 Statistical analysis

PAM aims to predict whether the patient will die within one year after the date of the latter scan. As this is done through a classification system, we evaluated the performance of the model using classical classification statistics. Namely, we assessed specificity, sensitivity, and area under the receiver operating curve (ROC-AUC). Statistical significance was assessed using the Mann-Whitney-U test. Confidence intervals were estimated via bootstrapping performed using sampling with replacement (1000 times). Statistical comparison between ROC-AUC was performed via McNeils' test. Multiple hypothesis testing was corrected with the false discovery rate (FDR) method with alpha set at 10%. A generalized multivariate linear regression was employed to evaluate the significance of PAM against current clinical standards (radiology and blood work).

5.4 Results

5.4.1 Study cohort

A total of n=103 patients were included in this study. Ten patients had only one scan available, making it impossible to model longitudinal changes, and therefore had to be excluded from the analysis. Nineteen patients did not have enough time between imaging date and censor date, and were excluded (see Figure 5.3a). The median age in this cohort was 65 years (IQR: 55 — 72). Upon stratification, n=37 patients were assigned to the training set, and n=37 in the validation set. In terms of overall survival, the median was reached in about 1 year (345 days).

Imaging-wise, we included a total of n=1087 CT scans between 6 months before start of treatment and up to two years after. These were used to create the scan pairs needed for PAM to model morphological changes. In total, we found n=2339 abdominal, and n=7431 chest scan pairs. We further excluded all scan pairs of living patients whose

time between the latest scan and censor was less than 1 year, and whose time between scans in the scan pair was more than 1 year. This resulted in n=1209 abdominal scan pairs, and n=3701 chest scan pairs in the discovery set and n=614 and n=1937, in the validation set, respectively. We chose not to limit the analysis to only subsequent scans, as the time points of when they were taken, and the time interval between them might vary. We rather chose to include all feasible pairs, within a given time-interval.

With respect to the unlabelled data used for training, we retrieved a total of n=37,573 CT scans from TCIA. The localizer was trained first, on n=176 thoracoabdominal CT scans from the lymphadenopathy dataset. The abdominal tracker was trained on n=3137 abdominal CT scans, resulting from the automatic inclusion procedure.

5.4.2 Prognostic performance

We assessed the ability of the classifier (trained on the imaging features of the tracker module) to predict 1 year survival after the latter scan of the scan pair. Across all scan pairs, the overall performance on the independent validation set was 0.73 AUC (CI: 0.69 — 0.76, $p<0.001$) for abdominal images, and 0.67 AUC (CI: 0.64 — 0.69, $p<0.001$) for chest images. Specificity and sensitivity were 0.74 (CI: 0.69 — 0.80) and 0.60 (CI: 0.56 — 0.64) for abdominal images; and 0.71 (CI: 0.68 — 0.74) and 0.58 (0.56 — 0.60) for chest images, respectively.

A performance analysis on a 6 months moving window was employed to assess the predictive value of PAM longitudinally during follow-up. The analysis was run on temporal windows with at least 10 positive and 10 negative samples to limit statistical noise. For abdominal scans, the highest prognostic performance was reached in the first 6 months of treatment (7 to 189 days), with an ROC-AUC of 0.82 (CI: 0.72 — 0.89, $P<0.001$). In general, the temporal windows around and up to the first 8 months of treatment seem to be the ones carrying the highest predictive value, staying significant after correction for multiple hypoth-

esis testing. Similar results were obtained for chest scans. The highest prognostic performance was reached later than the abdominal model, around 5 to 11 months after start of treatment, with a ROC-AUC of 0.83 (CI: 0.71 — 0.92, $P<0.001$). Unlike abdominal scans, which were observed to have a prognostic value both around and during treatment, chest scans carried much higher prognostic value during treatment rather than around the start date. Detailed results of the prognostic performance over time are shown in Figures 5.3b and 5.3c.

To investigate PAM as a biomarker, we analyzed the scans taken before the start of treatment. Namely, we investigated the scan pairs whose scans were taken between 12 weeks prior and start of treatment. This resulted in 31 scan pairs of 26 patients in the external validation set. Four patients had multiple scan pairs. We aggregated multiple scan pairs per patient by taking the average PAM prediction. This resulted in an AUC of 0.70 (CI: 0.50 — 0.88, p=0.054) for the prediction of 1 year survival from the moment of start of treatment. Specificity and sensitivity were 0.69 (CI: 0.50 — 0.87) and 0.82 (CI: 0.58 — 1.00), respectively. Further analysis on PAM predicted survival at baseline showed a significant difference of $>$464 days between low and high risk patients (p = 0.012, log rank test), with the high risk group had a median survival of 266 days, and the low risk group did not reach median survival within the first 2 years of treatment.

5.4.3 Comparison with current monitoring standards

Univariate analysis for current monitoring standards showed significant performance for both radiological assessments, as well as laboratory (hemoglobine, erythrocytes, leukocytes, and thrombocytes) results. Radiological progression and response reached an AUC of 0.64 (CI: 0.58 — 0.70, $p<0.001$) and 0.66 (CI: 0.62 — 0.69, $p<0.001$), respectively. In terms of blood markers, increases in erythrocyte counts (0.57 AUC, CI: 0.51 — 0.62, p=0.019), hemoglobin (0.62 AUC, CI: 0.57 — 0.66, $p<0.001$), and leukocyte counts (0.55 AUC, CI: 0.49 — 0.61, p=0.039) were all significant. None of these markers

performed better than PAM. PAM performance remained statistically significant against these other biomarkers using multivariate analysis. Other factors that retained significance were radiological progression (p<0.001), leukocyte count (p=0.045), hemoglobin (p=0.016), erythrocyte count (p=0.011) and age (p=0.001). Radiological response and thrombocyte count were not significant. Results of both univariate and multivariate are presented in Table 5.1 and Figure 5.3b.

5.4.4 Visual analysis of adbominal heatmaps

Results from visual analysis were classified based on highlighted areas (hotspots), and whether they were cancer lesions, cancer-spread complications, therapy-induced complications, or seemingly healthy tissue. If cancer lesions and cancer-spread or therapy-induced complications were not covered by a hotspot, these were flagged as coldspots. In total, N=31 cases were analyzed. Table 5.2 shows a summary of the results. A heatmap example is shown in Figure 5.3d.

In the abdomen, primary bladder tumours (n=13), involved lymph nodes (n=18) and liver metastases (n=10) were flagged as prognostic by the AI algorithm in most cases where they were present — namely, hotspots in 85%, 83% and 80% of cases, respectively. Similar frequencies were observed for bone (n=7) and peritoneal metastases (n=7), having been flagged in 86% and 71% of cases. Rare occurrences of an adrenal metastasis, as well as abdominal wall metastasis and a ureter mass were also found, both as hotspots and coldspots. Low occurrence was also observed for cancer spread-related complications. These were hydronephrosis (n=5), ascites (n=3) and pleural effusion (n=1). Hydronephrosis and ascites were highlighted in n=4 and 2 cases, respectively. Far more common were hotspots observed on seemingly healthy tissue, including the hip region (n=27), pelvic bone (n=26), spine (n=25), liver and bowels (n=20), kidneys (n=17), and spleen (n=16). It was further observed that, in the large majority of

cases, only part of the tissue would be highlighted, but never the full organ.

5.4.5 Visual analysis of chest heatmaps

In the thorax, n=7 out of 11 lung lesions were highlighted (64%). The mediastinum, chest wall, and upper spine were the most common hotspots in seemingly healthy areas. Other lesion types, such as lymph node metastases and bone metastases, were also present but low in numbers. Observed cancer spread-related complications include pleural effusion (hotspot in n=1 out of 2 cases), and ascites (coldspot). Pneumonitis and sarcoid-like disease were also present as therapy related complication hotspots, but both as single cases. A summary of the results is shown in Table 5.2.

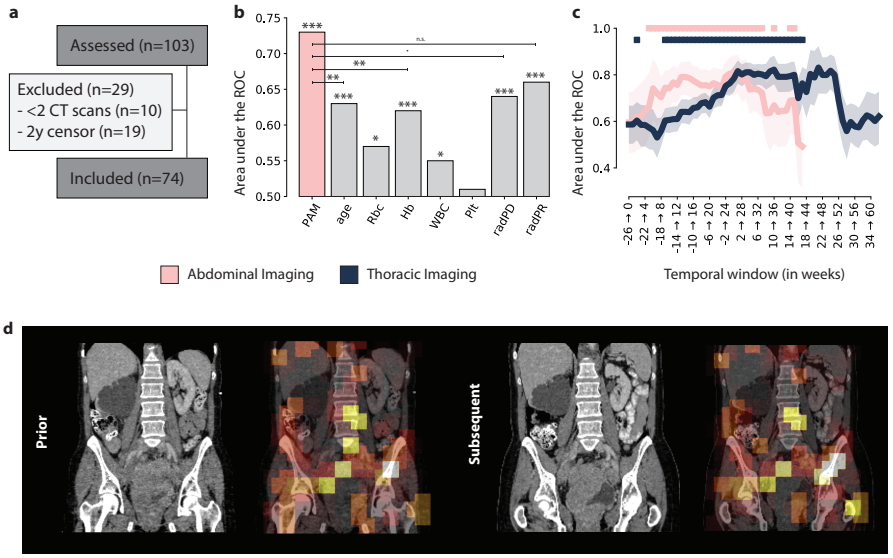

Figure 5.3: **(a)** Consensus diagram, **(b)** PAM Abdominal tracker performance compared to other standard factors used for treatment monitoring. Significance levels are reported for p<0.001 (***), 0.01 (**), 0.05 (*), 0.1 () and n.s. for non-significant **(c)** PAM Abdominal and Thoracic monitoring performance, over time, with respect to the start of treatment, in weeks. The ■ indicates statistical significance after FDR correction **(d)** Example of the prognostic heat map overlaid on top of the original abdominal scan.

5.5 Discussion

Advanced and non-invasive imaging methods for evaluation of treatment response, which would provide comprehensive and reliable information on how the patient responds to treatment, could improve accurate clinical decision making. Our aim was to assess the prognostic value of AI-enriched thoraco-abdominal CT response assessment in stage-IV urothelial cancer patients undergoing immune checkpoint inhibitors. We set up a fully-automatic AI-system that would track

changes between follow-up thoraco-abdominal CT scans, and linked their quantitative descriptors to overall survival. We term this method prognostic AI-monitor (PAM).

Our findings showed that PAM reached significant predictive performance for both thoracic and abdominal CT, with AUCs of 0.67 and 0.73, respectively, for the prediction of 1-year overall survival from the moment of the scan. In-depth analysis revealed stark differences in the prognostic value of morphological changes depending on the time point of treatment, with the first 9 months of treatment being the most predictive and significant AUCs $>$0.70, peaking to over 0.80 for both abdominal imaging, and thoracic imaging. Similar findings were observed in our previous study on NSCLC [Tre+21b], where the changes recorded by the algorithm in the first 3 to 5 months of treatment were observed to have a higher prognostic value. In the present study, we extended the system to include both the thorax and abdomen, and trained with far larger datasets both in terms of pre-training as well as survival association. The AI algorithm designed in this study was significantly extended to a comprehensive AI-system (i.e. PAM), able to scan imaging data, identify the regions of interests, and analyze them for the purpose of monitoring and prognostication. By including abdominal images, we also showed that the previous system [Tre+21b] can be extended to multiple parts of the human body.

To the best of our knowledge, this is one of the first studies employing artificial intelligence for prognostication in immunotherapy-treated urothelial cancer patients. In the study by Park et al. [Par+20b], the authors developed a radiomics model for the prediction of objective response and overall survival in a similar population. Machine learning was also employed on imaging (radiomics) features, however these were extracted via manually delineated lesions. The authors reported an AUC of 0.88 (CI: 0.65 — 0.97) for objective response prediction of bladder tumors in a cohort of N=21 patients, with a significant difference in overall survival between (radiomics-identified) higher and lower risk groups. Our findings also showed significant differences in survival, but in contrast to the

above study, we looked at the whole body CT changes, not only those of the tumoral lesions but also the non-tumoral treatment or cancer related CT changes. Our results are comparable to state-of-the-art methods based on time-consuming, error-prone, manual delineations [Tre+19; Sun+18]. Till now, single lesion analysis has allowed the field to develop, however, it has restricted the usage of the image only to selected areas-of-interest, accounting for $<5\%$ of the total data in the scan. While these methods have been refined to leverage known factors in cancer growth, including vascularity [Ali+19], oxygenation [Tun+], and metabolic activity [Mu+18] — our approach is different. Not only do we offer a novel fully automatic procedure which completely eradicates the need of time-consuming segmentations, but it also makes use of the whole body image of the patient, to evaluate the patient's status and estimate survival.

We analyzed the PAM further, by means of visualization. More specifically, we employed a visualization method [ZF14] to generate heatmaps, which highlighted regions of the image that carried higher predictive value, according to PAM. In our case, hotspots would correspond to gross morphological changes that the AI algorithm deemed of prognostic relevance. An expert radiologist was tasked to visually confirm these findings. Our findings show that changes in the primary tumour of the bladder, as well as metastases in lymph nodes, liver, peritoneum, and skeleton were among the most predictive for the algorithm.

Interestingly, there are similarities between our results, and the results from the NSCLC study [Tre+21b]. In both cases, the region of the primary tumour, as well as lymph nodes and bone lesions were closely inspected by the algorithm. Additionally, in the present study, the algorithm is also tracking changes in liver and peritoneal metastases. Unlike the present study however, the AI in the NSCLC cohort was working only on chest imaging, therefore unable to access the abdominal cavity.

There is evidence, in both studies, that bone lesions should be ac-

counted for in imaging evaluation schedules. These are considered non-target lesions in the current response criteria and are notoriously difficult to assess [Eis+09; Sey+17b; Sch+16]. Both bladder and lung cancer generated evidence to support the further investigation for the inclusion of CT changes in the bone among the target lesions.

Generally speaking, these findings suggest an unequal effect of cancer lesions on survival. While this might seem trivial at first (e.g. brain metastases are known to have worse prognosis), all current imaging methods for response evaluation and prognostication (like the RECISTs [Eis+09; Sey+17b; Sch+16]) do not distinguish between lesion types. RECIST methods are based on the change in the sum of diameters of a (limited) set of lesions. In other words, the growth of lesions in one organ is measured and weighed in the same way as the growth of another lesion in a different organ — no distinction is made. Our results however suggest that these factors should be accounted for, which would therefore require a more comprehensive evaluation scheme.

In this study, we proposed a method that is based on image-to-image registration, leveraging the properties of this technique in finding corresponding anatomical landmarks in pairs of images, and therefore constructing a model able to track not only tumours but also tumour- and therapy induced changes, as well as seemingly healthy parenchyma. Our method does not preclude the usage of other techniques and methods. As we have observed, commonly used clinical response evaluation tools also retained significance when compared against PAM, suggesting PAM as a complementary value to the current clinical standards. An optimal approach to the utilization of PAM would be integration of this method with other diagnostic tools currently available [BTB18].

In this study, we proposed and presented an AI-system termed PAM, for the analysis of bladder cancer patients receiving immunotherapy. The study was limited to whole-body CT, which is the workhorse in

standard clinical practice. As brain imaging is not part of the standard whole-body CT protocol, anatomical and functional changes in the brain as captured on MRI and PET/CT are yet to be explored. We envision a more comprehensive usage of this technique, where all available imaging during follow-up is leveraged for prognostication purposes. It is also yet to be confirmed whether the PAM approach would extend to other treatments and cancer types, and to which extent survival associations would be interchangeable.

Another limitation of the study is the monocentric nature of the analysis. While CT data is generally acknowledged to have higher level of reproducibility across vendors than MRI, it is yet to be seen whether this would hamper the association with survival, and to what extent. Nonetheless, we made sure to train the tracker and localizer modules on large publicly-available datasets that would, in theory, provide a larger pool of variations in image acquisition protocols.

In conclusion, as a future outlook, we envision an extended PAM-like algorithm to be set up as a clinical decision support system in tumour boards, providing continuous monitoring and prognostication information, in order to assist physicians in the treatment decision process.

5.6 Conclusions

In this study, we investigated the prognostic information of AI-derived whole-body imaging monitoring markers in advanced urothelial cancer receiving checkpoint inhibitors. We hypothesised that quantitative AI-derived features describing morphological changes happening during the course of treatment could hold prognostic information. To this end, we designed and implemented a prognostic AI-monitor (PAM). Our findings demonstrate that PAM is complementary to existing monitoring methods, while reaching comparable or superior accuracy. We argue that this could be the result of PAM's ability to analyze the whole body, including non-target cancer lesions and non-cancer

lesions. Further investigation should focus on the development of a comprehensive pipeline beyond anatomical imaging, as well as on external validations.

	Univariate Analysis				
	N- / N+	***p-val***	**ROC-AUC (95CI)**	**Sensitivity**	**Specificity**
Erythrocyte count (/dt)	358 / 110	0.019	0.57 (0.51 - 0.62) *	0.47 (0.43 - 0.52)	0.42 (0.34 - 0.49)
Hemoglobin (/dt)	372 / 122	<0.001	0.62 (0.57 - 0.66) *	0.47 (0.43 - 0.51)	0.38 (0.31 - 0.45)
Leukocyte count (/dt)	366 / 116	0.039	0.55 (0.49 - 0.61)	0.52 (0.48 - 0.56)	0.56 (0.49 - 0.64)
Thrombocyte count (/dt)	366 / 116	0.421	0.51 (0.45 - 0.56)	0.51 (0.47 - 0.56)	0.54 (0.46 - 0.61)
Radiological Progression	145 / 65	<0.001	0.64 (0.58 - 0.70)	0.87 (0.82 - 0.91)	0.42 (0.31 - 0.52)
Radiological Response	145 / 65	<0.001	0.66 (0.62 - 0.69) *	0.69 (0.62 - 0.75)	0.00 (0.00 - 0.00)
AI-score (Abdomen)	437 / 117	<0.001	0.73 (0.69 - 0.76)	0.60 (0.56 - 0.64)	0.74 (0.69 - 0.80)
AI-score (Thorax)	1421/ 516	<0.001	0.67 (0.64 - 0.69)	0.58 (0.56 - 0.60)	0.71 (0.68 - 0.74)
	Multivariate Analysis				
	Coef	***Std***	***95 CI***		***p-val***
Intercept	0.33	3.47	- 6.5	7.1	0.924
Age	-9.80	3.07	- 15.8	- 3.8	0.001
Erythrocyte count (/dt)	-14.90	5.84	- 26.4	- 3.4	0.011
Hemoglobin (/dt)	13.30	5.51	2.5	24.1	0.016
Leukocyte count (/dt)	9.90	4.93	0.2	19.6	0.045
Thrombocyte count (/dt)	1.52	2.20	- 2.8	5.8	0.488
Radiological Progression	-2.69	0.71	- 4.1	1.3	<0.001
Radiological Response	23.31	>100	>100	>100	0.999
AI-score (abdomen)	-8.20	1.76	- 11.6	- 4.7	<0.001

Table 5.1: Prognostic performance of PAM against current monitoring tools. AUCs < 0.5 were inverted for readability, indicated by (*).

	10% - 50%	Frequent (>50%)
	Abdominal Imaging	
H.T	Bone mets (6), Peritoneal (5), Bladder Ca (11), Lymph Nodes Mets (15), Liver Mets (8)	
H.Tr	Hydronephrosis (4)	
H.H	Chest wall (7), Pancreas (6), Abdominal Wall (13), Stomach (12)	Bowel (20), Liver (20), Spleen (16), Kidneys (17), Spine (25), Pelvic Bone (26), Hip Region (27)
C.T		
	Chest Imaging	
H.T	Lung Mets (7), Liver Mets (4)	
H.Th		
H.H		Mediastinum (24), Chest Wall (26), Upper Abdomen (21), Spine (25)
C.T		

Table 5.2: Visual analysis of PAM generated prognostic maps. Number of cases between parenthesis (N). The column corresponding to rare occurences (<10%) has been omitted for readability, available in the original publication. **H.T** = hotspot tumour, **H.Tr** = hotspot tumour-related, **H.Th** = hotspot therapy related, **C.T** = coldspot tumour. **Ca** = cancer, **Mets** = metastases.

6

Prognostic response patterns in brain imaging

Stefano Trebeschi, Thi Dan Linh Nguyen-Kim et al. "AI-driven identification of prognostic response patterns to immunotherapy of melanoma brain metastases". In: *Submitted for publication.* (2021).

Abstract

Background. The response of brain metastases (BM) to systemic cancer immunotherapy is still an open research question. Normally, the brain blood barrier would separate the brain from the rest of the body. In the presence of brain metastases however the barrier is shown to be compromised, but not disrupted, which results in brain-specific tumour microenvironments, ultimately leading to a different response than the one seen in the body. We hypothesize that AI-analytic pipelines (specifically prognostic monitoring) can be used on routine brain imaging to identify brain-specific response patterns different from the one detected from standard clinical methods, and that can be used for patient treatment monitoring.

Methods. We collected brain MR of melanoma patients with BM receiving anti-PD1 in two different centers. Alongside imaging, we collected brain-derived markers, for the purpose of comparison between our AI-system and the current clinical standards. A brain-specific version of the prognostic AI-monitor (PAM) is proposed.

Results. We collected a total of 49 patients in the discovery set and 26 in the external validation set. Significant performance were observed in the prediction of overall survival (0.64 AUC, p=0.04) and progression-free survival (0.66, p=0.005). Higher performance, up to 0.75 and 0.81 AUC, were reported for imaging acquired outside of local treatment (radiotherapy or surgery). PAM predictions remained significant against, and independent from, other clinical, radiological and molecular factors.

Conclusions. Our findings show that AI-systems (like PAM) can be used to model complex response patterns in patients receiving immunotherapy for the purpose of whole-body monitoring and prognostication.

6.1 Introduction

The management of intracranial metastatic spread has gained interest in recent years thanks to novel systemic therapies, which improved the overall survival of patients, but poorly managed to cross the blood-brain barrier [Arv+16]. The role of the blood-brain barrier is to separate and protect the brain from the rest of the body. In the presence of brain metastases (BM), the blood-brain barrier is shown to be compromised, but not fully disrupted [Lyl+16]. Under these conditions, brain metastases can result in tumor microenvironments with complex, brain-specific, cancer-promoting features [Jua+18] that can lead to poor intracranial response to systemic treatment. In immunotherapy clinical trials, these factors have led clinicians to exclude any potential for disease control, despite the success of the agent even in advanced stages of the disease [Ber+16]. Stereotactic brain radiotherapy and surgery remain the most common treatment options for BM.

There is evidence, however, to suggest that the tumor microenvironment of BM can also present an active inflammatory component [Ber+16], allowing for immunotherapeutic treatment options. Recently-published meta-analyses provided evidence of improved overall survival and long-term disease control in patients with BM receiving immune checkpoint-inhibitors in addition to standard stereotactic radiotherapy [Pet+19]. Moreover, the combination of different immunotherapy agents helped achieve better intracranial objective response [Rul+19] — though not without neurotoxic side effects [Opi+20]. Currently, the focus of clinicians is to identify the optimal time window when immunotherapy can be added to the standard treatment of BM [Lu+19].

Due to the location of the lesions, collection of tissue samples for biomarker studies is challenging. Imaging has been investigated as a potential, non-invasive alternative [Gal+20]. These are still limited in number, and constitute only pilot evidence. Some studies, for example, have uncovered links between blood perfusion and vascularisation to pseudo-progression [Coh+16; Ume+20], while

hyper-progression remains not addressed — despite the higher incidence [Cha+18b]. The main challenge in these types of studies remains the high dimensionality of MR imaging and the limited availability of data samples.

In this sense, artificial intelligence (AI) can be used to explore complex patterns in high dimensional, longitudinal imaging data. Specifically, we aim to employ AI to track and quantify, longitudinally, intracranial treatment response patterns. We termed this process Prognostic AI-monitoring (PAM) [Tre+21b]. Through PAM, we hypothesize that we can isolate prognostic morphological changes of brain metastatic melanoma patients receiving anti-PD1 monotherapy, as depicted on standard, clinically-available MRI imaging. We test our hypothesis on an external validation set, and compare the resulting AI prognosis to clinical, radiological, molecular markers.

6.2 Materials and methods

6.2.1 Study cohorts

Discovery set. We retrospectively included stage-IV melanoma patients treated with anti-PD1 monotherapy between 2014 and 2016 at the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital (NKI-AVL, Amsterdam, The Netherlands). Patients with clinical suspicion for BM underwent standard brain magnetic resonance imaging (MRI) with follow-ups in case of confirmed disease. Patients without any brain MRI within the first year of treatment, were excluded. Additionally, since PAM works by modelling changes between follow-up exams, patients with only one MR scan available were also excluded. The collection of this dataset was approved by the local Institutional Review Board (IRBd19-083) as imaging and longitudinal expansion of a previously described melanoma cohort [Tre+19].

Extenal validation set. We retrospectively included stage-IV melanoma patients, with brain cancer specimens, treated with

anti-PD1/PDL1 monotherapy between 2013 and 2019 at the University Hospital of Zurich (USZ, Zurich, Switzerland). Patients without any brain magnetic resonance imaging (MRI) within the first year of treatment, were excluded. The collection of this dataset was approved by the local Institutional Review Board (EK-Nr.647, KEK-ZH-NR:2010-0117/0).

For both datasets, we recorded the date of start of immunotherapy, and death. Additionally, for patients in the validation set, we also included dates of brain radiotherapy, brain surgery, and intracranial progression — assessed by a board certified radiologist (TDLNK).

6.2.2 Data curation

T2-weighted imaging was selected for this analysis, as this provides high soft-tissue contrast, as well as clinical availability and standardization. For all scans, the date of acquisition was recorded. To mitigate differences in image acquisition, all scans were linearly resampled at 1.5 mm isotropic voxel size, and resized to 192x192x192 voxels using padding and cropping. Intensities were standardized, clipped (3σ), and rescaled on the interval [0, 1].

6.2.3 Prognostic AI-monitoring

Leveraging AI for prognostic monitoring comprised two aspects. The first aspect was the ability of the AI to quantify morphological changes between pairs of subsequent follow-up scans. In PAM, this is implemented leveraging image-to-image registration. Image-to-image registration is a computational technique where a moving image is deformed (according to a set of estimated parameters) to match the corresponding landmarks of a fixed image. In doing so, the AI learns to estimate differences between the two images. There are two main advantages in using registration as a base technique for parameterizing differences between scans. First, as the objective of the network (i.e.

how well two images match) can be computed from the images itself, with no need for manual annotations. Second, given that the network can be trained with no manual annotations, we could leverage this large publicly available datasets of brain imaging — we downloaded all MRI data available on The Cancer Imaging Archive[1] [Cla+13]. The specific implementation we used in this study was the one presented in the original pilot study on NSCLC chest CTs [Tre+21b], and it was based on the research of Balakrishnan et al. [Bal+19] and Zaho et al. [Zha+19b]. The code of our network can be found on the AI repository of our department [2].

The second aspect that is needed to leverage AI and imaging monitoring for prognostication is the ability of the AI to link morphological changes to survival. In PAM, this was implemented using a classifier on quantitative features estimated from the registration network. These features are quantitative parameters describing morphological deformations between pairs of follow-up images of the same patient (we term these as scan pairs, composed by a prior scan and a subsequent one). The classifier is trained to use the parameters describing the change between prior and subsequent scan to identify situations of risk, which we define as whether the patient will survive 6 months after the date of the subsequent scan. We set a maximum time delta between prior and subsequent scan of 6 months. A schematic representation of the process is shown in Figure 6.1.

6.2.4 Training of PAM

The registration network was trained to minimize the correlation coefficient loss [Bal+19], alongside with three penalties to mitigate for unlikely morphological deformations, namely two affine penalties, weighted 1/10, and one deformable penalty, weighted 1/100. Moreover, we implemented a curriculum scheme on the loss during

[1] Accessed on the 21st of April 2020
[2] code: github.com/nki-radiology/PAM.git

training, such that the loss would be computed on a decreasingly smoother version of the images. This was done via average pooling, with a kernel of 9, reduced by 3 at epochs 100, 150, and 175. Batch size was set to 2. Due to the small batch size, batch normalization was replaced by group normalization [WH18]. The network was trained on all available T2-weighted brain MRI found on the cancer imaging archive [Cla+13]. We kept 10% hold out during training to control for overfitting.

The classifier was trained to identify situations of risks, namely to predict whether a patient will die within 6 months from the date of the subsequent scan. More specifically, for each scan pair, we extracted the feature maps from the lowest layer of the deformation network. These were 96 feature maps, reduced to a feature vector by average pooling. The feature vector was passed to a logistic regression classifier, together with the number of days passed since start of treatment, and the time between prior and subsequent scan. These two additional inputs were included to correct for instabilities derived from temporal discrepancies.

6.2.5 Factor analysis

To assess the independence and novelty of our method, we compared the AI-score against other (mostly brain-specific) clinical, radiological and molecular factors. Clinical factors were age at start of treatment, immune-related side-effects developed during the course of the treatment, autoimmune diseases, and presence of diabetes type-II. We also included time information on the treatments — immunotherapy, radiotherapy, and surgery. All these factors were collected from the clinical records of the hospital. Brain tissue-derived molecular factors consisted of the presence of the following mutations: BRAFv600 (or wild type), NRAS, CKIT, CD117, Ki67. Brain-specific radiological parameters were derived from the volumetric segmentations. This was performed by a board certified radiologist (TDLNK) with

3D Slicer[3]. The segmentations were performed using T2, FLAIR, post-contrast and pre-contrast imaging, all resampled at 1mm voxel spacing and re-aligned to the Talairach template using elastix [Kle+10; Sha+13]. From the manual labels, we derived the following parameters: increase in viable, increase in necrosis, increase in edema, relapse, complete response, and whether imaging was acquired after intracranial progression.

6.2.6 Statistical analysis

To assess the performance of the model, we employed the area under the receiver operating curve (ROC-AUC) and the Mann-Whitney-U test. Confidence intervals were estimated via bootstrapping performed using repeated (1000 times) sampling with replacement. Multivariate analysis was performed to assess the independence of the classifier survival score, we employed a multivariate binomial generalized linear model. To ensure the validity of the multivariate analysis, we removed all parameters with absolute correlation coefficient $\rho > 0.90$ (collinearity), and removed variables with less than 10% positive or negative events (imbalance).

[3] url: slicer.org

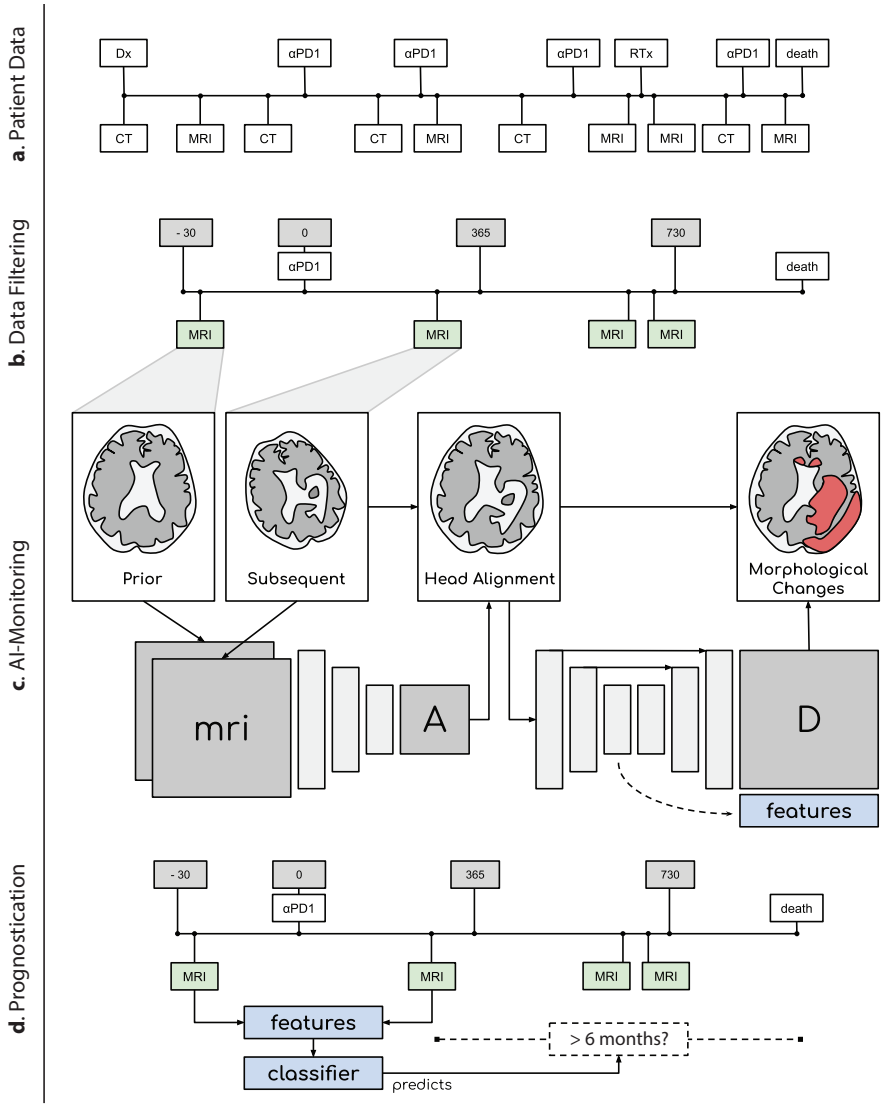

Figure 6.1: **(a)** All imaging and therapy related data generated during treatment **(b)** Filtered data used in this project **(c)** The base deep-learning framework based on image registration **(d)** Discovery of prognostic morphological changes.

6.3 Results

A total of n=49 patients were included in the discovery set, and n=26 in the external validation set (CONSORT diagrams in Figure 6.2a-b). The discovery set had a median age of 58 years (IQR 48 — 68), equally distributed between males and females (n=25 vs 24, respectively). The validation set had a median age of 56 years (IQR 49 — 72), and were mostly males (n=21 vs 5). Both sets did not reach median overall survival within the first year of treatment. Progression free survival was available only for the external validation set, and was of 523 days from start of treatment.

Imaging-wise, we collected a total of n=260 brain MRIs, namely n=165 in the discovery set and n=95 in the external validations set, with a median of 3 (IQR 2 — 4) and 3 (IQR 3 — 4) MR scans per patient, respectively (distributions shown in Figure 6.2c). From here, scans were rearranged to form the input to the AI. Namely, as the AI algorithm (PAM) aims to identify prognostic morphological changes between follow-up scans of the same patient, the dataset had to be arranged with scan pairs. We collected a total of n=236 scan pairs in the discovery set, and n=140 pairs in the external validation set. Each patient had a median of 3 scan pairs (IQR 1 — 6) in the discovery set, and 4.5 scan pairs (IQR 3 — 6) in the external validation set.

Additional clinical parameters were collected in the external validation set. This included information about brain radiotherapy, surgery, and immune related adverse effects (irAE). A total of n=20 patients received brain radiotherapy (77%), n=12 of which were commenced after the start of immunotherapy. The overall median time for the first radiotherapy treatment was 5.5 days after the start of immunotherapy, with a stark difference between patients that started before vs after the start of immunotherapy (-20 vs 149 days, respectively). A total of n=13 patients received brain surgery (45%), four of which happened after the start of immunotherapy. Patients received surgery on average 21 days before the start of immunotherapy, while the remaining four had it on 8, 44, 110 and 416 days after, respectively. Seventeen patients

experienced immune-related side-effects (irAE). Five patients had autoimmune disorders, and another n=5 had diabetes type-II (only one patient had both).

From the brain cancer samples (surgery specimens or biopsies) we collected molecular markers. Twenty-eight patients were positive for Ki67, n=15 for BRAFv600, n=14 had wild type, n=11 were positive for NRAS, and n=3 for CKIT. None was positive for CD117.

6.3.1 AI training

The AI model was trained in two steps. The first step was the training of the registration module, used to quantify differences between pairs of brain scans. This was trained on n=1832 T2-weighted MRI of the brain retrieved from the cancer imaging archive. We termed this pre-training dataset, and should not be confused with the study cohort. This was created automatically, through heuristics selections (e.g. remove datasets named "prostate") and internal MR sequence detection algorithms. The final dataset consisted of a heterogeneous set of primary and metastatic brain cancer from 9 different studies [Kin+18; Sha+16; PPT19a; MK16; Eri+17; Sca+19; SP19; Kin+19; Gro+20]. During training, we randomly selected two images in the set, and fed them to the model to register them. We shared the code of our network. Once trained, the resulting network was used to extract radiomic features representing gross morphological changes between follow-up scans of the same patient (scan pairs). We collected a total of n=236 PAM signatures in the discovery set, and n=140 in the external validation set — one signature per scan pair. Survival was the outcome variable. We encoded as failure those scan pairs where the patient would not be alive within 6 months after the date of the latter scan in the scan pair. There were n=36 positive scan pairs in the discovery set (15%), n=17 in the external validation set (12%).

6.3.2 Prognostic and predictive performance

The overall performance in the external validation set was 0.64 AUC (CI: 0.54 — 0.73, p=0.04) for the prediction of 6 months survival from the date of the subsequent scan. Sensitivity and specificity were 0.53 (CI: 0.44 — 0.62) and 0.65 (0.44 — 0.83), respectively. As these patients received combinations of immunotherapy with radiotherapy and/or surgery, we performed a sub-analysis of the results, stratified according to the different treatments. Here, we consider a scan pair to be affected by radiotherapy or surgery if either of them happened within 90 days prior to imaging date (we assume that the effects of these treatments on the anatomy wear off after 90 days), and stratify the dataset in two groups: around radiotherapy or surgery, and outside radiotherapy and surgery. The highest prognostic performance of the model was reached in scan pairs acquired outside the time frame of radiotherapy and surgery, with an AUC of 0.75 (CI: 0.63 — 0.85, p=0.02). This was observed to decrease in scan pairs acquired within the time frame of either radiotherapy or surgery (0.53 AUC, CI: 0.36 — 0.70, p=0.40). For the prediction of intracranial progression within 6 months from the date of imaging, the AI reached similar performance as for the overall survival, with an AUC of 0.66 (CI: 0.57 — 0.74, p=0.005). As for the previous case, scan pairs acquired outside of radiotherapy or surgery yield the highest performance, with an AUC of 0.81 (CI: 0.67 — 0.92, p=0.002), whereas scan pairs acquired around radiotherapy or surgery yielded the lowest performance (0.49 AUC, CI: 0.35 — 0.65, p=0.50). A full overview of the results is shown in Table 6.5 and Figures 6.2d-e.

6.3.3 Factor analysis

To assess the independence/novelty of the AI score, we ran a multivariate generalized linear model binomial regression analysis against known clinical, radiological and molecular prognostic factors (20 in total, see Table 6.5) to predict 6 months survival from the moment of the scan. To control for collinearity, we removed factors with a Pearson correlation coefficient $\rho > 0.90$. This removed BRAF wild type,

which correlated with BRAFv600. Rare events, defined as occurring $<$10%, were also removed. These were: complete response, relapse, Ki67 and diabetes type-II. The results of the multivariate analysis show that the AI score retained its statistical significance (p=0.04) against all other factors. Other factors in the analysis were also significant, including presence of autoimmune disease (p=0.045), intracranial radiological progression (p=0.04), imaging acquired around brain radiotherapy (p=0.006) and increase of viable tumour (p=0.018). No correlation is reported among these factors (see Figure 6.2f). Individual analysis of each factor shows comparable significant performance, most of them between 0.63-0.64 AUC, but with remarkable differences between sensitivity and specificity (see Table 6.5 and Figure 6.2g), with the highest sensitivity being from presence of autoimmune disorders (0.91, CI: 0.86 — 0.96) and intracranial progression (0.88, CI: 0.82 — 0.93), and the highest specificity from our AI algorithm (0.65, CI: 0.44 — 0.83) and radiotherapy (0.59, CI: 0.38 — 0.79).

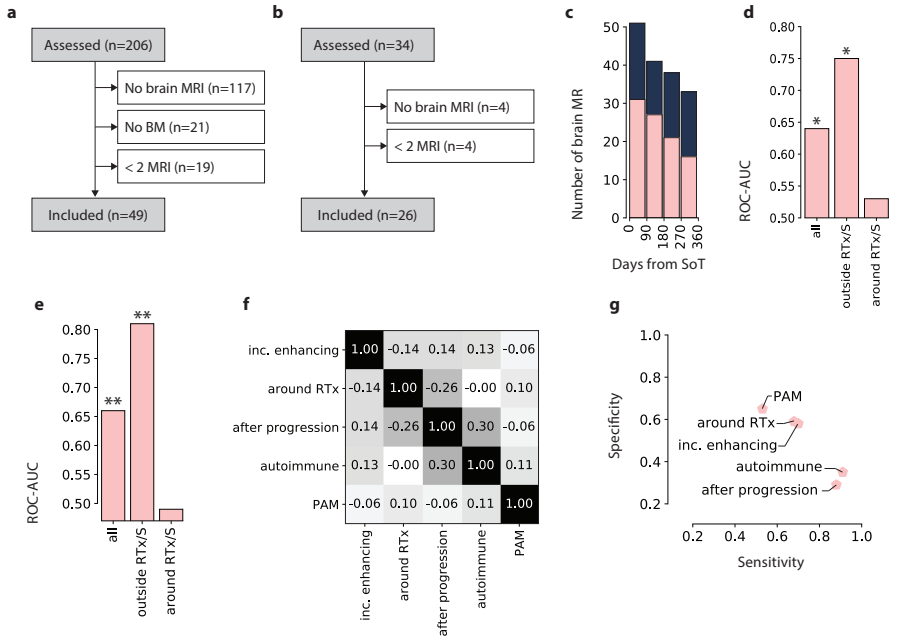

Figure 6.2: **(a)** All imaging and therapy related data generated during treatment **(b)** Filtered data used in this project **(c)** The base deep-learning framework based on image registration **(d)** Discovery of prognostic morphological changes.

6.4 Discussion

Advanced AI treatment monitoring could be used to study complex response patterns in imaging data. Our aim was to investigate prognostic values and novelty of response patterns seen by prognostic AI-monitoring (PAM) in melanoma brain metastases patients receiving PD-1 checkpoint inhibitors. To this end, we implemented a deep neural network for image-to-image registration, which we trained on a large public MR dataset of brain cancer patients. The tranied AI was used to model longitudinal changes occurring during

anti-PD1 therapy in the brain of metastatic melanoma patients. Morphological changes in the brain, identified by the network, were then used in a logistic classifier to predict situations of poor outcome, defined as death within 6 months from the date of image acquisition.

Our results showed significant performance of PAM in the identification of cases at risk. We show that the model predictions are linked to survival and intracranial response, as well as that they are an independent marker, novel and complementary to existing clinical standards, and have the potential to be used in conjunction with them. Moreover, following the results of the sub-analysis, we observed differences in the AI-response to patients that received local treatment for brain metastases, where the performance was reported to be higher in patients that did not receive either radiotherapy nor surgery around the time of the imaging. Furthermore, following the result of the factor analysis, we observed that our AI-model offers a novel view of brain response, different from current clinical standards like radiological progression [Oka+15], or more general increase in enhancing tumour volume [Del+19]. To the best of our knowledge, this is the first study which leverages an AI to investigate morphological changes in patients with brain metastases receiving immunotherapy. Bhatia et al. [Bha+19] found a radiomic signature in a cohort of 88 patients, associated with overall survival in a cohort of patients receiving immune checkpoint inhibitors. Their results were confirmed in an independent cohort. A similar approach was pioneered by Cha et al. [Cha+18a], where the authors also used convolutional neural networks, but for the prediction of response to stereotactic radiosurgery. Other studies in the literature focused more on the diagnostic part, such as the differentiation of brain metastases from primary brain cancers [ABB19; Qia+19], prediction of the BM origin [Kni+19; Ort+17; Ort+18], or differentiation of BM recurrence from radionecrosis [HMM19; Loh+18; Pen+18]. Only one study was found that, similarly to the present study, looks at serial imaging and changes over time, rather than a single time point [Zha+18]. Differently from the present study, the authors used

this type of analysis to distinguish radionecrosis from tumour progression.

PAM was first introduced in a pilot study on thoracic imaging of NSCLC patients [Tre+21b]. As in the original study, this study demonstrates that PAM can be used for fully-automatic response evaluation, namely tool to high risk patient, and that PAM would work as an independent factor from current clinical standards for response evaluation and monitoring. In this study, we expand PAM to include MRI imaging, known to lack standardization and reproducibility, and test its performance on an independent, external validation set. Moreover, we selected that set to include brain molecular markers to exclude possible pathological confounders. Furthermore, unlike our pilot study, where we analyzed the AI-response at different timepoints during and before treatment, in this study we chose to turn our attention to concurrent local treatments for the sub-analyses, as this aspect is more relevant in patients with BM. In specific, we analysed the AI-performance with respect to the start of immunotherapy, brain radiotherapy and brain surgery. Results confirmed that these local treatments have an impact on the response pattern, as seen in imaging, as they set themselves apart. The low number of cases in the validation set did not allow us to further investigate the exact nature of this difference.

PAM predictions remained significant in multivariate analysis against a set of clinical, radiological and molecular factors. Interestingly, other radiological factors, such as radiological progression and the increase in enhancing tumour volume; and clinical factors, such as radiotherapy and presence of autoimmune disorders, were also significant at multivariate analysis. This suggests that PAM is an independent and complementary tool to current clinical standard examination, and could potentially be envisioned to have a synergistic role in treatment planning and monitoring. This was confirmed with the correlation analysis, and the subsequent analysis of the predictive performance of each individual factor.

It shall be noted however, how some of these factors, such as mutational status, were patient-wise factors. In the current clinical practice, due to the invasive nature of biopsy, it is not common to develop a comparative response evaluation or monitoring marker on pathological and genetic markers. Future development in biopsy, such as liquid biopsy-derived markers [Tie20; SS16], might pave the way to molecular — possibly AI-driven — monitoring. In such a scenario, PAM should be extended to integrate not only imaging data, but also laboratory work, fluid biopsies, and clinical reports generated during the course of treatment [BTB18].

6.4.1 Limitations and future outlook

In this work, we extend previous research in prognostic AI-monitoring [Tre+21b] to brain imaging, we explore the reproducibility of results by means of an external validation set, and we investigate further the relation of PAM-markers to current clinical, radiological, and molecular markers. We only collected patients with brain biopsy. This allowed us to compare PAM with a wider range of brain specific markers, but limited the size of the dataset. Moreover, the lack of standardization for the treatment and imaging of BM in patients receiving immunotherapy further limited not only the dataset, but also the analysis. Future studies should address this limitation, to include more patients from different centers, even without the restriction of biological samples. Such study would allow the in-depth analysis of the relation between different localized treatments, and the systemic one, in the context of response evaluation and prognostication.

6.5 Conclusions

In this study, we present an expansion of the PAM analytic pipeline to brain imaging of BM patients receiving immunotherapy. Our results demonstrate that PAM can be extended to imaging modalities beyond CT, and be used to capture prognostic response patterns that are

unique and complementary to a wide range of different brain-specific markers, currently used in the clinics. Further investigation should focus on larger, multi-center imaging (and non-imaging) datasets.

	Pts	N-	N+	ROC-AUC	Sensitivity	Specificity	*p-value*
Death (with 6 months from imaging)							
All	26	91	17	0.64 (0.54 - 0.73)	0.53 (0.44 - 0.62)	0.65 (0.44 - 0.83)	**0.036**
Outside RTx and surgery	21	48	7	0.75 (0.63 - 0.85)	0.58 (0.47 - 0.70)	1.00 (1.00 - 1.00)	**0.019**
Around RTx and surgery	18	43	10	0.53 (0.36 - 0.70)	0.49 (0.36 - 0.61)	0.40 (0.14 - 0.67)	0.406
Radiological progression (within 6 months from imaging)							
All	25	54	38	0.66 (0.57 - 0.75)	0.63 (0.53 - 0.74)	0.69 (0.56 - 0.81)	**0.005**
Outside RTx and surgery	19	33	10	0.81 (0.67 - 0.92)	0.52 (0.35 - 0.71)	0.91 (0.71 - 1.00)	**0.002**
Around RTx and surgery	13	16	15	0.49 (0.35 - 0.65)	0.44 (0.22 - 0.67)	0.50 (0.35 - 0.66)	0.496

Table 6.1: Performance of PAM on the external validation set, stratified according to local treatment.

	Coefficient	Std Error	Z	p-value	0.025% CI	0.975% CI
Intercept	8.9045	3.502	2.543	0.011	2.04	15.769
			Radiological Parameters			
Increase of enhancing	**-2.2404**	**0.945**	**-2.371**	**0.018**	**-4.092**	**-0.389**
Increase of non-enhancing	-0.6158	1.023	-0.602	0.547	-2.621	1.39
Increase of edema	-0.4003	0.838	-0.478	0.633	-2.043	1.243
After intracranial progression	**-2.362**	**1.151**	**-2.052**	**0.04**	**-4.618**	**-0.106**
			Clinical Parameters			
Around brain radiotherapy	**-2.962**	**1.074**	**-2.759**	**0.006**	**-5.066**	**-0.858**
Around brain surgery	21.6885	¿100	0.001	0.999	< -100	>100
Number of days after start of treatment	-1.0231	2.197	-0.466	0.641	-5.329	3.282
Age at start of treatment	2.2016	1.912	1.151	0.25	-1.546	5.949
Immune-related adverse effects (irAEs)	-0.6337	1.348	-0.47	0.638	-3.276	2.008
Autoimmune disorders	**-1.9755**	**0.988**	**-2**	**0.045**	**-3.911**	**-0.04**
			Molecular Parameters			
BRAFv600	-2.3682	1.813	-1.306	0.191	-5.921	1.185
NRAS	-0.0735	1.855	-0.04	0.968	-3.709	3.562
			Prognostic AI-Monitoring			
PAM Prediction	**-5.3581**	**2.616**	**-2.048**	**0.041**	**-10.486**	**-0.23**

Table 6.2: Multivariate generalized linear model. Complete response, relapse, Ki67, CD117, cKIT and diabetes type-II occured <10% and were removed. BRAF wild type correlated with BRAFv600 and was removed.

	ROC-AUC	Sensitivity	Specificity	p-value
Increase in enhancing tumor volume	0.64 (0.54 - 0.75)	0.70 (0.62 - 0.78)	0.58 (0.38 - 0.79)	**0.010**
Around brain radiotherapy	0.63 (0.52 - 0.74)	0.68 (0.60 - 0.76)	0.59 (0.38 - 0.79)	**0.017**
After intracranial progression	0.59 (0.49 - 0.69)	0.88 (0.82 - 0.93)	0.29 (0.11 - 0.50)	**0.034**
Autoimmune disorders	0.63 (0.52 - 0.74)	0.91 (0.86 - 0.96)	0.35 (0.16 - 0.57)	**0.0015**
PAM prediction	0.64 (0.53 - 0.73)	0.53 (0.44 - 0.62)	0.65 (0.44 - 0.83)	**0.036**

Table 6.3: Univariate analysis of the significant cofactors.

7

The future of artificial intelligence immunotherapy trials

Zuhir Bodalal, Stefano Trebeschi et al. "The future of artificial intelligence applied to immunotherapy trials". In: *Neoadjuvant Immunotherapy Treatment of Localized Genitourinary Cancers: Multidisciplinary Management*. Book chapter accepted for publication. Springer, 2021.

Abstract

Clinical trials serve as a barrier of entry for new interventions and treatments prior to implementation in routine clinical practice. At its essence, the primary role of a clinical trial is to monitor a patient longitudinally using the diagnostic disciplines (radiology, pathology, and laboratory medicine) to assess clinical outcomes. As the diagnostic fields have begun to fully digitalise, large volumes of data are being generated per patient - creating a ripe environment for the implementation of artificial intelligence (AI). In recent years, AI has found multiple applications in the medical field, most notably in radiology. In this book chapter, we will explore how artificial intelligence has been applied in each of these diagnostic disciplines and discuss how this may influence clinical trials in the future.

7.1 Introduction

Clinical trials are a cornerstone of medical research, especially in the context of oncology. Rigorous trials act as a barrier of entry for novel interventions, treatments, or treatment combinations. Traditional clinical trial approaches have been mostly unchanged for decades; a patient is assigned to either an experimental or control group, and their clinical outcome is recorded. However, the emergence of novel data analysis techniques provides a unique opportunity to augment decisions being made during a trial (e.g. prediction of resistance or adverse events) or even change to the way clinicians conduct trials (e.g. patient enrolment using AI or dynamic re-assignment of patients between various treatment arms). Artificial intelligence (AI), in particular, has become the focus of a significant amount of research where groups explore the possible integration of these computational methods into a potential clinical decision support system.

In order to bring novel AI methods into clinical trials, it would be necessary to break down the components of a clinical trial and see where AI can be of added value. In its simplest form, we can think of a clinical trial as a process where an intervention/medication is first administered to a patient and then subsequently monitored continuously by diagnostic disciplines.

The principal diagnostic disciplines in the context of a clinical trial would be imaging (radiology), pathology, and laboratory medicine. Each of these disciplines would generate large volumes of data per patient from the beginning of the trial to its conclusion. AI, in general, and deep learning, in particular, are notorious for their demand for data. As such, the implementation of AI has been most prevalent in these fields, particularly radiology.

In each of the principal diagnostic disciplines, several use-cases for artificial intelligence methods have emerged. In the following sections, we will highlight the major applications of AI in these fields and discuss what a future "AI-powered clinical trial" could look like.

7.2 AI in medical imaging

Artificial intelligence can be applied to many fields within medicine; however, generally speaking, a prerequisite is that AI requires access to vast amounts of data [SM19; Ahu19]. The development of three-dimensional magnetic resonance imaging (MRI), computed tomography (CT), and other multiparametric imaging modalities significantly increased the availability of anatomical, functional and molecular information within routinely generated clinical images. AI algorithms have emerged to extract meaningful imaging properties, or features, from these imaging modalities that can be linked back to clinical endpoints [Gri+17b; Aer+14]. Research into medical image analysis has since blossomed with increasing interest from both radiologists and other physicians in the clinical implementation of these algorithms [Ros+15].

The novel domain of imaging features can be chiefly divided into qualitative, or semantic, imaging features and quantitative imaging features (Figure 7.1). Semantic features are obtained from experienced readers (e.g. radiologists) who assess medical images and score specific parameters (e.g. presence of necrosis, lesion size and shape), thereby generating a feature vector (or a collection of features) that can be used for statistics or AI model construction. The scoring of semantic features is generally dependent on expertise. Quantitative imaging features, however, are extracted by applying mathematical algorithms on the images. Examples are attenuation, tumour diameter, anatomically relevant angles, or radiomics, among others [Yip+17a; Bod+19; Rah+19; GKH16]. Both types of features can be used in artificial intelligence models to link imaging data to clinical endpoints.

7.2.1 Semantic features

Semantic (qualitative) features reflect intuitive tumour properties such as lesion size, shape, number of lesions, location, intensity and others

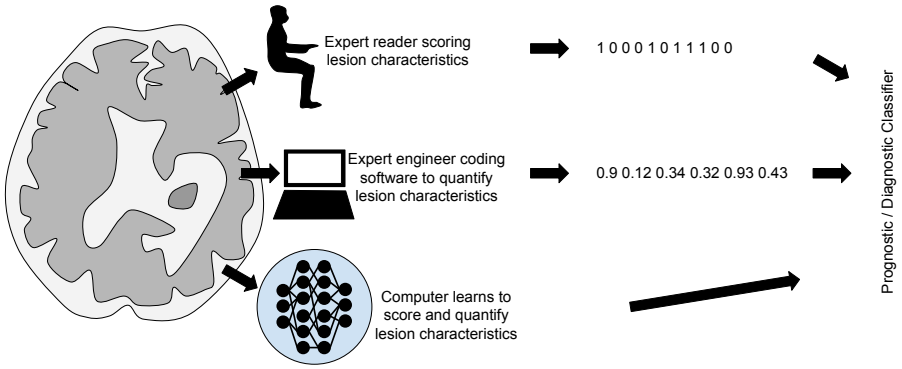

Figure 7.1: Schematic representation of the methods of imaging feature generation and application. In the top path, an expert reader scores specific (usually binary) parameters in the image. In the second path, handcrafted radiomic features are extracted automatically using predefined algorithms/formulae. In the final path, a deep learning neural network acts as an end-to-end solution, where the image is input, features are automatically learned, and a classification is made based on the clinical endpoint.

[Bod+19; GKH16; Kis+16; Riz+18]. Several of these features have been added to the routine clinical workflow as they provide the radiologist extra information during the diagnosis of diseases and treatment response monitoring [Rah+19].

A number of semantic imaging features were significantly associated with progression-free survival (PFS) and overall survival (OS) in glioma and glioblastoma multiforme patients [Pee+18; Pop+05]. In non-small cell lung cancer (NSCLC), semantic features were able to distinguish tumours with different genetic mutational statuses. ALK-positive nodules tended to show larger volume multi-focal thoracic lymphadenopathies on CT imaging [Hal+14] while pleural retraction [Riz+16], smaller nodules [Riz+16; Lv+18], or spiculation [Lv+18] were indicative for an EGFR mutation. Tumour characteristics such as round shapes [Riz+16], the presence of multiple small nodules

[Lv+18], or nodules in non-tumour lobes were associated with KRAS mutation [Riz+16].

Despite the implementation of different semantic features in the daily clinical workflow of radiologists, semantic features suffer from specific shortcomings, most notably standardisation [Lv+18]. Because semantic features are subject to human bias, two radiologists can score tumour properties very differently. Differences in experience between readers could lead to different results in relation to diagnosis or treatment response [Kis+16]. Semantic features suffer from inter- and intra-observer variability [PT17; Van+15; Wet+02; Rid+16], and a learning curve exists for readers to generate appropriate and accurate features [Li+18].

Another disadvantage of semantic features is that they are bound to what is discernible to the human eye. This limitation could lead to the missing of high-dimensional and potentially important imaging traits which, as a result, will not be taken into account [Aer+14; Sun+20a; Wu+19]. The last limitation can generally be overcome with quantitative analysis [Tim+20; WN11].

7.2.2 Radiomics

In the field of radiomics, advanced mathematical algorithms are applied to medical images to convert them into quantitative minable data [Riz+18; PMK20; TM20; May+20]. The field of radiomics is built on the principle that medical images contain valuable information beyond what is discernible to the human eye. Radiomics is capable of extracting predictive high-dimensional information from the images by means of quantitative image analysis [Rog+20; Hec+20].

Radiomic data could improve our understanding of medical domains such as treatment-related adverse events, therapeutic, and post-therapeutic changes, and underlying biology, among others [Aer+14]. Depending on the way that radiomic features are extracted, the field of radiomics can be divided into two main approaches:

handcrafted, or classical, radiomics and deep learning radiomics [Bod+19; Rog+20].

In handcrafted/classical radiomics, visual aspects in the medical images are converted into features by means of predefined mathematical formulae [Bod+19; Rog+20; Afs+19]. These types of features are generally based on morphological, phenotypic characteristics, such as image intensities, shape, or textural attributes [Aer+14; Hag+19]. A prerequisite for handcrafted radiomic features, and generally considered a limitation, is the requirement of manual delineation of regions of interest. An experienced radiologist generally performs the region of interest delineation. The workflow of manual delineation followed by handcrafted feature extraction is characteristic of classical radiomics. Classical radiomics can then be analysed by conventional statistical methods or by machine learning artificial intelligence models.

The second and upcoming approach that makes use of radiomic features is deep learning. Deep learning, in itself, is a term that can be explained in several books. The concept of deep learning describes a number of neural networks that can be taught, a process commonly known as training, to generate features by itself. While creating these features, deep neural networks can perform classification without human involvement [Maz+19; Kim19]. Within the medical community, the convolutional neural network (CNN) is generally accepted as the go-to class of deep neural networks as it is free from human interference and is capable of extracting many more features than classical radiomics or semantic features [Bod+19]. A major advantage of deep learning is that feature extraction, selection, and classification all happen within the same network.

Radiomics, with either machine learning or deep learning, have been employed in a number of applications in medical imaging [Nie+19; Che+17; Liu+19]; two prominent ones can be put forward in this book chapter:

7.2.3 Prediction of response to immunotherapy

AI has been used for different objectives within the field of medicine. One of the proposed applications is response prediction by distinguishing probable responders from non-responders, with the predictive performance being measured by the area under the receiver operating characteristic curve (AUC). Ultimately, such predictive AI algorithms may exclude patients from exposure to unnecessary treatment, mitigating both potential adverse effects for the patients [Man+13] and loss of precious funds for the healthcare facility [Gha+18].

Publications about radiomics/deep learning for prediction of response to immunotherapy are relatively scarce but are increasingly being published. Trebeschi et al. used CT-based radiomic biomarkers to predict treatment response to immune checkpoint blockade in melanoma and non-small-cell lung cancer (NSCLC) [Tre+19]. Similarly, non-invasive CT biomarkers were able to distinguish between high tumour mutational burden and low tumour mutational burden in patients with NSCLC (AUC = 0.81). These biomarkers were also able to predict clinical outcomes of NSCLC patients receiving anti-PD-1/PD-L1 treatment [He+20]. Radiomic features derived from PET/CT showed promising results in determining which NSCLC patients would likely benefit from anti-PD-1/PD-L1 immunotherapy [Mu+20; Pol+20]. A deep learning model trained on FDG-PET images also appeared to be predictive of immunotherapy in patients with lung adenocarcinoma [Par+20a]. Another study used time-to-progression and pretreatment CT-based features to identify NSCLC patients unlikely to benefit from immunotherapy [Tun+19]. Similarly, delta radiomic features (resulting from the subtraction of pre- and post-treatment features) could recognise early immunotherapy response in NSCLC patients [Kho+20].

Sun et al. found that a combination of CT-based features and CD8 gene expression signature showed promising results when predicting clinical outcome in four independent cohorts with advanced solid tumours

treated with immunotherapy [Sun+18].

Response prediction has also been studied for urothelial cell cancer patients treated with immunotherapy. In metastatic urothelial carcinoma, a radiomics model involving CT-based features showed promising results when predicting immunotherapy response and survival outcome (AUC = 0.88) [Par+20b]. The use of a deep learning network radiomics pipeline in bladder cancer achieved 86% accuracy when distinguishing between potential responders and non-responders to immunotherapy [Run+19].

7.2.4 Radiogenomics for prognostication and precision medicine

Radiogenomics is a novel research field that links imaging phenotypes to genetic characteristics (such as gene profiles, gene expression, and gene mutation status) [KJ14; RK09]. The term has expanded its meaning beyond 'just genomics' and now also encompasses the linking of imaging markers with other biological parameters in the tumour microenvironment, such as proteomics and metabolomics [Bod+19].

Radiogenomics addresses a number of the shortcomings of conventional biopsy-based approaches. Biopsies suffer from sampling bias, fail to take interlesional and intralesional heterogeneity into account, and have increased morbidity and invasiveness for patients. Additionally, biopsies are limited to sites of the body that are more accessible. Using AI radiomics, we can gain insight into the tumour biology of the full tumour burden (i.e. the primary lesion and metastases) across time. A key advantage of radiogenomics is the possibility of using serial and longitudinal imaging, whereas serial or multi-lesion biopsies are often unavailable.

By linking different biopsy results to radiomic features, radiogenomics could potentially map the genomic landscape of a patient's entire tumoural burden, completely non-invasively. In the context of a clinical trial, radiogenomics can be used to identify early biological markers

of resistance or the emergence of a targetable mutation for precision medicine. The value of visualizing the genomic profile of the full tumour burden longitudinally cannot be overstated.

We envision that, in the future, a generalisable radiogenomics model may ultimately be used as a non-invasive approach that mimics and expands biopsy function.

7.3 AI in pathology

The application of artificial intelligence in medical imaging has started with considerable focus on 2D imaging, such as histopathological slides. In an effort to promote digitalisation and long-term preservation of the biological data collected, pathology departments have long started the digitalisation of image acquisition, processing, and storage. This, in turn, allowed AI researchers to make use of such data for purposes of diagnosis and prognostication (Figure 7.2).

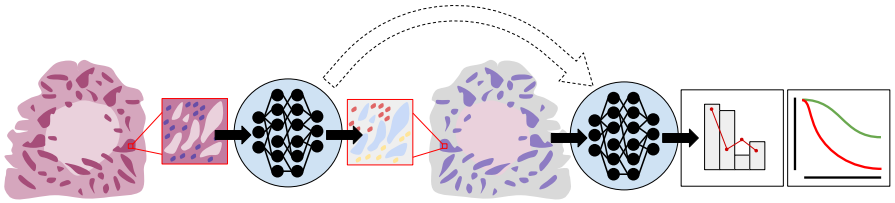

Figure 7.2: An illustration of possible applications of AI in pathology. Using AI, we can apply computational stainings on pathological slides (as shown by the transition from an H&E image to an IHC image). Additionally, a digital image of a slide can be used for biomarker identification (or for classification purposes).

Several aspects are unique to pathology, including the multi-dimensional information of the staining encoded in the colours, the relatively high dimension of the image derived from the microscopic

scale resolution, and the level of heterogeneity resulting from different intra- and inter-patient variability, as well as biopsy parameters. These aspects, among others, have been reported to require specific, tailored-made solutions [Sal+18; Zha+19a; Eht+17; Cou+18].

7.3.1 AI-assisted pathological assessment

In the realm of AI-driven pathological diagnosis, one of the first applications is the automatic extraction of biological parameters of interest. In this sense, AI was developed to be used as a tool to convert complex, high dimensional data to cellular and tissue phenotypes at a large scale, to be used for scientific research and prognostication. The simplest example is the identification and segmentation of single cells [Al-+18; Fal+19; HB17]. In this case, the algorithm would automatically discern pixels that belong to cells from background ones, separating different cells from each other. This enables high-throughput processing of information regarding cellularity and cellular distribution from massive datasets, which would have otherwise been impossible by manual labelling. Further research redefined and proposed more precise methods for the segmentation of finer structures, like cell nuclei and cytoplasm [Nay+17; Mah+19; Sir+16], classification of tumour and immune cells [Sal+18; Tur+16], tumour epithelium and stroma [Du+18; Al-+19; Bej+17]. AI promises to unlock further information encoded in these pathological slides with the aim of supporting scientists and clinicians.

7.3.2 AI-driven pathological biomarker discovery

AI has the potential to unlock automatic quantification of biological parameters of interest through the identification and quantification of known structures and patterns. This raises the question of whether AI-based features can be used for biomarker discovery and the extent of their application. A study from Stein et al. revealed how common pathological scoring, featuring immune activation, cell death, tissue

repair, and regression grade, had the potential for pan-tumour scoring of response to anti-PD1 therapies [Ste+20]. If these findings hold, it would be a logical step to explore AI applications able to quantify these aspects from pathological slides and potential associations with pan-cancer therapy-specific response. Steps in this direction have been taken already with deep learning methods developed for predicting known biological immune-biomarkers. These include PD-L1 status [Sha+19], TMB [JM; Wan+20], and microsatellite instability [Kat+19], among others. It is yet to be seen whether these AI algorithms can generalise beyond the tumour type in their training set. Most statistical methods used for these applications (of which deep learning networks are part of) are based on the creation and synthesis of domain-specific knowledge. Whether the AI would manage to extrapolate its knowledge to cancer types, it was not trained upon has yet to be seen.

7.3.3 Unique aspects of AI pathology imaging and immunotherapy

Most of the approaches reported so far encompass the application of AI for the quantification of specific biological quantities of interest. However, immunotherapy depends on more complex mechanisms which frequently rely on the relation and distribution of biological entities (and their respective properties) in the microenvironment. In this case, it would be beneficial to harness the full potential of these large computational models (often $>$1M parameters) to track and quantify these complex patterns. In a study from Saltz et al. [Sal+18], researchers found an association between tumour infiltrating lymphocyte (TIL) patterns (exposed by AI from haematoxylin and eosin (H&E) staining) and tumour and immune molecular features, and ultimately treatment outcome. These were automatically extracted and analysed from a public cohort of 5202 H&E high-resolution pathological slices, a study otherwise unfeasible if it were to be completed by human manual labour. These aspects are particularly researched in the immune-oncological world, as

they enable scientists to gain additional insights into the complex immunotherapy functioning mechanisms. The application of AI in pathology imaging of immunotherapy patients is becoming more relevant with the enlargement of immunotherapy to the neoadjuvant settings. As these patients often present tumour in-situ, AI would allow to perform an initial assessment of the sample, classifying the microenvironment, quantifying immune-infiltration, and estimating the likelihood of micrometastases in the surrounding tissue.

7.3.4 Data rediscovery

The last aspect for which AI is set to transform the field of pathology is through data rediscovery or, in other words, rediscovering older datasets for novel purposes. Currently, clinical pathology is moving away from standard H&E staining to more complex and customised stains for the purpose of personalised medicine. This is the case for immunohistochemistry, that is now becoming the de-facto standard in immunotherapy clinical settings. While normally this would mean that retrospective observational studies would not be able to harness the data collected in the time when H&E was the only standard, AI allows us to do just that. This process is commonly termed computational staining and has been used to generate H&E imaging from unstained tissue samples [Ran+18], staining of TILs from H&E [Sal+18], and even commercial solutions for computational IHC staining from H&E. Once deployed, these technologies will enable us to harness the full potential of the datasets collected by hospitals during the last decades, and gain novel insights that are still hidden in old data.

7.4 AI in laboratory medicine

Laboratory medicine represents a vast source of healthcare data [Shi+15], paving the way for numerous potential AI applications, including laboratory operations optimisation, laboratory tests

analysis, early diagnosis, personalised patient care, among others [Gru+19]. In addition, laboratory data can be particularly appealing in machine learning because of its tabular and codified nature. Traditional machine learning algorithms depend on structured data and are typically organised in a tabular format on which to train. Despite this immense potential, the application of AI to traditional laboratory results appears to be relatively unexplored [CB18].

In clinical oncology, the treatment technology is rapidly improving with advanced techniques and new types of therapies, including immunotherapy, chemotherapy, and radiotherapy. Developing precision medicine with the aid of AI techniques is becoming a major trend. AI research in laboratory medicine also has been growing, although the total number of publications remain relatively low, especially in immunotherapy. Recently, machine learning was used to analyse rhabdomyosarcoma patients treated with vincristine (IVA) chemotherapy to predict their blood cell count dynamics and reproduce the dynamic profiles of the haematologic toxicities. In the study, twenty-four patients with rhabdomyosarcoma treated with IVA chemotherapy courses were included and during each cycle, routine multiple blood cell counts were performed [CA20]. Such kinds of AI-based studies could also be extended and applicable in immunotherapy treatments.

In immunotherapy trials, laboratory parameters have been investigated as predictive and prognostic biomarkers and as a monitoring tool for treatment response. Particularly, biomarkers exploring the immunologic aspects of the tumour and its microenvironment (e.g. PD-L1 expression, tumour infiltrating lymphocytes, TMB) have been broadly studied [Hen+17; Hel+18b; CT19].

However, these require a tissue biopsy that can be challenging-to-obtain considering the invasiveness of the procedure, high tumour heterogeneity, high turnaround time or tissue insufficiency [Gan+18a; PU19]. Thus, routinely collected laboratory values,

which are easily obtained at baseline and follow-ups, have also gained interest as biomarkers. Serum based markers such as the neutrophil-to-lymphocyte ratio [Pet+; Ros+18; Soy+18; Möl+20; Pen20], and lactate dehydrogenase [Ros+18; Pen20; Gam+18], for instance, have been found to have prognostic or predictive power in NSCLC and melanoma patients receiving immunotherapy. Blood TMB was shown to be correlated with tissue TMB and associated with longer progression-free survival in NSCLC patients [Gan+18a], which could, therefore, obviate the need for resorting to tumour tissue. Biomarkers that can be derived directly from routinely collected laboratory values do not require extra resources [Vou20] and allow to quickly evaluate dynamic changes during treatment [Soy+18], assisting, for example, in restaging decisions when imaging assessment is uncertain [Bil+19].

7.5 Integrated artificial intelligence

As seen in the previous sections, artificial intelligence paves its way into many branches of medicine, for a wide variety of use cases. Quite often, the key advantage offered by AI methods is that of automation and labour reduction. This is mainly due to the historical advantage that computational methods have when it comes to monotonous, repetitive tasks. AI models tend to perform better when there is less variability with a single class in the input. However, real-world data is heterogeneous, particularly in the medical setting. In part, due to this high variability, the predictive performance of AI models in each of the disciplines has often been below what is expected from trusted clinically implemented models (AUC$>$0.9).

This limitation in the predictive performance of medical AI models has triggered the call for the development of an "Integration AI System" (also known as Integrated Diagnostics) [BTB18]. The fundamental hypothesis of this concept is that data from different medical disciplines

would contain complementary information that could be harnessed by a neural network to boost its predictive performance (Figure 7.3).

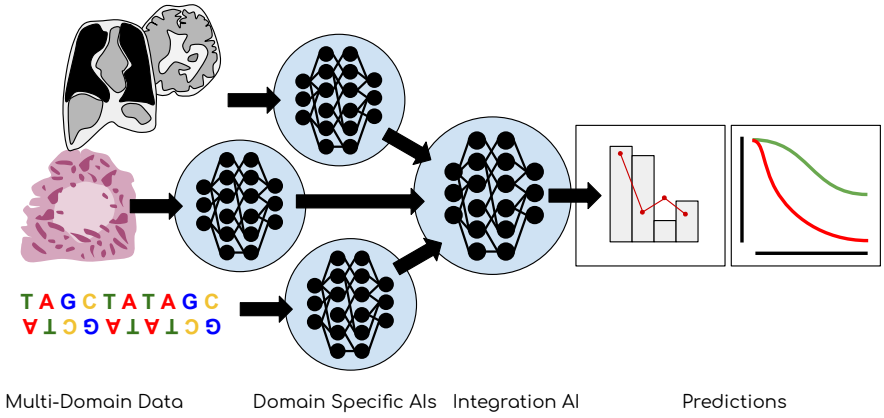

Figure 7.3: Representation of an integrated AI system, where individual predictive models receive a separate input and then the output of those models are integrated into a new neural network.

In modern healthcare facilities, it is very often the case that a patient generates data from the moment that they enter the front door. A clinician will take their medical history and perform a physical examination, radiological images and pathological slides are obtained, and even genomic data/fluid biopsy data is acquired. Each of these disciplines has identified prognostic and predictive markers within their respective fields. However, it is likely that individual biomarkers alone might not have sufficient predictive power. We believe that future immunotherapy biomarkers with sufficient discriminatory power to predict response/prognosis will involve a comprehensive multi-parametric approach, including multi-dimensional biomarkers obtained from whole-body imaging, pathology, peripheral blood markers and omics-based biomarkers rather than single-analyte biomarkers alone. By simultaneously integrating composite biomarkers and their dynamic interactions,

machine learning allows superior response prediction and prognostic performance when compared to manual biomarker selection.

7.6 The clinical trial of the future

Newly discovered treatments for patients have to endure a clinical trial before they can enter the market. However, problems that arise with developing a clinical trial are numerous, driven by the ever-increasing complexity of these trials [Siu+17; Ana+17; HGM17; Hwa+16; Fog18]. AI has the potential to counter these problems and can improve parts of the clinical trial workflow, such as trial design, patient selection, and patient monitoring, which could have a massive impact on the speed of implementing cures for cancer [Fog18; Woo19]. AI can explore massive datasets and find relations humans cannot comprehend. However, it holds new challenges for implementation [Kel+19]. The objective of specialised AI algorithms that we have at the moment is not to supplant clinical trials but rather augment them - with the ultimate aim of optimising clinical benefit.

In this section, we discuss the key challenges in clinical trials for immunotherapy, and propose AI-based solutions, hypothesising the 'clinical trial of the future' (Figure 7.4).

7.6.1 Challenges currently faced by clinical trials

Testing a newly discovered immunotherapeutic agent in a clinical trial for FDA approval is laborious, costly, and logistically challenging. Here, the design of the trial is of the utmost importance. The inclusion and exclusion criteria should generate a statistically similar population to the targeted population [Ana+17]. The inclusion of unsuitable patients in the cohort is detrimental, and the response criteria should be accurately defined since an offset in thresholds for inclusion criteria can doom the entire trial [Siu+17; HGM17]. The lack of a standardised method to determine tumour immune response

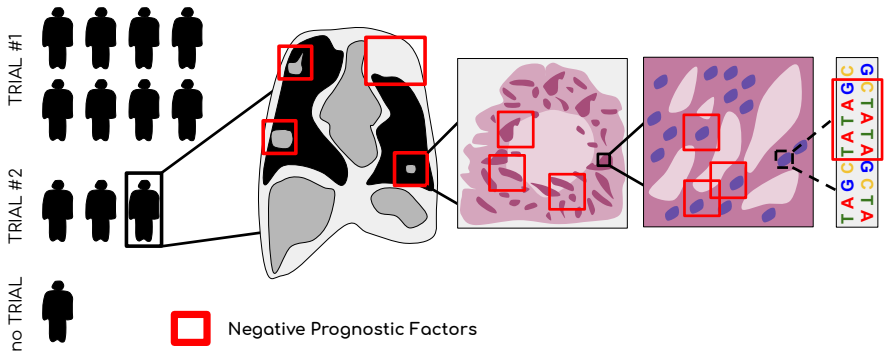

Figure 7.4: Visual representation of how AI could be implemented in clinical trials. In this figure, patients are classified into different trials (or recommended against a trial altogether) based on the presence of predictive and prognostic factors in the input data (e.g. radiological image, pathological slides, genetic data).

that is universally accepted troubles this process further [Siu+17], although the irRC, irRECIST, and iRECIST criteria offer guidelines [Wol+09; Nis+13; Sey+17b]. However, these guidelines track tumour progression by measuring the diameters of the tumour, resulting in possible inadequate estimations of the tumour growth due to the inability to calculate tumour volume. Biomarkers reflect dynamics of deeper mechanisms on the molecular level. Generally, biomarkers are obtained upon assessment of biopsies from single-tumour lesions. Here, biomarker dynamics, tissue heterogeneity and spatio-temporal dynamics in biomarker expression limit the reliability and representability of potential biomarkers obtained from limited tumour tissue [Hof19].

Furthermore, preventing underpowered clinical trials due to troublesome patient recruitment is another challenge [Fog18]. A study found that 25% of cancer trials did not meet the required enrolment [Fel15]. Other studies found that for a given cohort of patients, only 5% end up

enrolled in a cancer trial [Fou+13], while only 18% of the cohort would be ineligible for a trial [Ung+16]. Moreover, clinical trials compete for patients to enrol, causing a higher risk of underpowered results for competing trials.

7.6.2 Solutions that can be offered by AI

AI's ability to navigate through the massive maze of clinical trial criteria exceeds the capability of humans. It can learn features from thousands of trials and is able to return the best matches based on the results of previous trials. This can yield problems for normal computer programs as they are unable to extract contextual information. Natural Language Processing (NLP) is a specific AI method to retrieve and process the textual context. Currently, GPT-3 [Bro+20] is an incredibly powerful NLP model that can accurately extract context out of the text and generate text itself if the user describes the concept.

- It can compare the proposed specifics of the study to the ongoing studies with a similarity measure when presenting a new study. To prevent competing trials, the authors can decide whether they want to cancel a trial when the proposed study is almost identical to an ongoing one. Moreover, one could think of an application database where studies are preliminary compared to detect potential collaboration, increasing trial study capacity.
- To improve patient enrolment and decrease dropout, the NLP model can analyse information about the patient to retrieve the best fit for a study, yielding higher power for statistical analysis.
- To predict the probability of success, the AI can analyse previous studies, indicating the risks stakeholders are taking. As long as the studies are documented in an accessible database, such an NLP model can continuously learn. These applications have the potential to improve the pipeline of clinical trials but still rely on human designs for the trials. Here, AI can boost the design to enhance cohort composition and patient monitoring [Har+19].

> Based on the specifics, like cancer type and molecular structure of the proposed therapeutic agent, AI can predict optimal inclusion and exclusion criteria to involve. Furthermore, AI can learn relations in biomarker expression and discover complex characteristics beyond human ability. By clustering the AI's predictions based on the biomarkers, we can gain insight into the discovered relations.

AI can improve the tracking of tumour growth by automatically segmenting the tumour within seconds after acquiring the imaging scan. By tracking the segmented tumour volume over time, the AI model gives a more accurate indication of the response to the treatment than the current state of the art of measuring the diameters in three axes, which will improve the robustness of the clinical trials. Naturally, clinical trials, as we currently conduct them, will benefit greatly from the implementation of artificial intelligence in each of the diagnostic disciplines that form the backbone of the clinical trial.

7.6.3 An AI-powered clinical trial

Traversing even further into the (far) future, the most exciting concept would be to transcend traditional clinical trials completely. Imagine a powerful AI model that enables feature matching of genomes to generate patient-tailored therapeutic agents that activate the right immune response. Here, AI has the potential to not only accurately predict the success rate of the discovered agent, but also predict the side effects for the patient-agent combination. The AI model can find the optimal agent for each specific patient by processing the vast amount of data acquired over the years. This, in turn, could yield therapeutic agents' characteristics that are effective for particular cancerous cells without damaging healthy cells. The aim is not to test the therapeutic agent that targets specific cells in a clinical trial, but rather the AI that generates these agents. Once the trained AI can generate reliable, safe therapies for each individual patient, it only needs to be validated once versus the current state-of-the-art in a massive trial. The CE and FDA can

approve the method of generating the therapies, instead of approving each patient-tailored therapy after a separate trial. The moment such an immensely powerful method is available and validated, clinical immunotherapy trials will become redundant.

Regulations in clinical trials are of utmost importance to guarantee patient safety and standardised comparison methods. When AI is available to improve trials, regulations specifically for AI should already have been constructed. The SPIRIT-AI and CONSORT-AI guidelines are defined for developing and reporting clinical trials involving AI [Riv+20; Liu+20]. Researchers should always be up-to-date with the current guidelines since the rapidly evolving nature of AI will constantly result in new challenges and concepts to regulate.

To conclude, clinical trials in immunotherapy suffer from challenges that prevent the treatment from reaching its full potential. AI can improve the pipeline of clinical trials and has the potential to resolve multiple challenges. When regulated correctly, it holds tremendous capabilities, and we hypothesise that AI has to be part of the clinical trial of the future to develop a patient-tailored immunotherapy treatment.

7.7 Conclusions

Artificial intelligence has only just begun to have an impact on the medical field. Research is ongoing in the diagnostic disciplines, particularly radiology, to test the limits of existing predictive algorithms. The future of radiomics is especially promising considering the trend towards increasing resolution of radiological images. This trend could unlock even more information encoded in the image and yield better imaging markers/phenotypes. Despite the impressive achievements of predictive models in research, many of these networks are only tasked with detecting small abnormalities, neglecting a number of biological/clinical characteristics of the tumour for the sake of simplicity

[OGB20]. This serves as a reminder that AI is not intended as a replacement for the healthcare team but rather as a support tool to help guide decisions.

One long-standing challenge for medical AI has been generalizability of the predictive model to real-world data. A classical technical solution would be to expand the training data, but in the medical setting, this is often infeasible due to limited patient cohorts. This challenge may yet be overcome with the integration of different data types within an integrated AI system.

Finally, and possibly the most profound question that needs to be solved before AI can be implemented in the clinics is: 'Who is responsible for the predictions an AI algorithm makes?' [Rec+20]. This socio-philosophical/medico-legal enigma remains unsolved and may prove to be one of the largest hurdles for mainstream adoption of AI in clinical trials and daily practice. While many open questions have yet to be answered, the impact that artificial intelligence will have on the domain of healthcare is undeniable. Man and machine need to work together for the betterment of patient care.

8

Towards integrated healthcare

Zuhir Bodalal, Stefano Trebeschi, and Regina Beets-Tan. "Radiomics: a critical step towards integrated healthcare". In: *Insights into imaging* 9.6 (2018), pp. 911–914.

Abstract

Medical imaging is a vital part of the clinical decision making process, especially in an oncological setting. Radiology has experienced a great wave of change and the advent of quantitative imaging has provided a unique opportunity to analyze patient images objectively. Leveraging radiomics and deep learning, there is increased potential for synergy between physicians and computer networks – via computer aided diagnosis (CAD), computer aided prediction of response (CARP), and computer aided biological profiling (CABP). The ongoing digitalization of other specialties further opens the door for even greater multidisciplinary integration. We envision the development of an integrated system composed of an aggregation of sub-systems interoperating with the aim of achieving an overarching functionality (in this case better CAD, CARP, and CABP). This will require close multidisciplinary cooperation between the clinicians, biomedical scientists, and (bio)engineers as well as an administrative framework where the departments will operate not in isolation but in successful harmony.

8.1 Introduction

Medical imaging has historically played a key role in cancer screening, diagnosis, staging, and therapeutic response monitoring. On a daily basis, treating physicians rely on input from imaging to help formulate patient management plans [Sev+17]. This is especially true within the context of modern oncological guidelines, where patients are stratified into increasingly complex subgroups based on biological, clinical, and radiological parameters.

Historically, qualitative semantic features were used to describe tumour morphology – as observed in the patient image. These descriptions were a reflection of a scoring system based on visual assessment. Semantic features were shown in literature to have correlations with stage, prognosis, and even response prediction [Yip+17b]. However, as one could imagine, this method suffered from shortcomings rooted in its dependence on subjective scoring and the limited sensitivity of the human eye.

Modern, ubiquitous imaging modalities, such as CT, MRI, and PET in radiology (and digital images in pathology) are primarily quantitative in nature. This characteristic is harnessed, using computational algorithms, to extract quantitative features and generate mineable data. Rather than relying solely on subjective interpretation of images, these quantitative features can be used to objectively characterize tumour morphology.

In radiomics, medical images are processed to generate quantitative features and this mineable data can then be used for clinical purposes. Radiomic features serve the purpose of describing morphological characteristics (e.g. density distribution, recurrent patterns and textures, shape and outline…) in an objective, quantitative manner. The ambition is to find completely non-invasive radiomic features that could be used as predictive and prognostic biomarkers.

8.2 The promise of radiomics

The advent of radiomics has opened a brand new avenue in cancer research and presents a unique opportunity to data scientists and radiologists alike. Broadly speaking, two prominent potential have emerged for radiomics – tumour characterization and therapeutic response prediction (Figure 8.1).

The search has begun to identify imaging markers that could be used to assess biological parameters (i.e. genetic mutations or surface expression of particular molecules) in the tumour. Normally, such biological assessment of a tumor is achieved by biopsy – a process that is highly invasive, carries potential risk for patient morbidity, and can only elucidate information for lesions in sites easily accessible to surgeons. Radiomics provides the opportunity to non-invasively assess the biological profile (i.e. surface marker expression, genetic mutational status, blood markers etc...) of all the lesions simultaneously and instantaneously. With the increased use of computer models to diagnose conditions and predict response to therapy, this new field where biological parameters can non-invasively be assessed using quantitative features and computer models can be termed Computer Aided Biological Profiling (CABP).

One of the earlier studies to leverage radiomic features in the assessment of genetic mutational status (i.e. radiogenomics) was the work of Segal et al. in human liver cancer where combinations of twenty-eight imaging traits were shown to be capable of reconstructing 78% of the global gene-expression profile (i.e. mRNA levels) of these tumours [Seg+07]. Further research ensued on a number of tumour types – with varying degrees of success.

With rise of deep-learning based image analysis, computer algorithms could be used to extract radiomic features on a large scale which could then be linked to predictive and prognostic biomarkers in cancer (that would otherwise be obtained surgically).

Unlike more traditional radiomics approaches where feature extraction and data analysis consisted of two separate steps, deep learning fuses these processes together and iteratively optimizes one with respect to the other. In other words, deep learning provides radiomics models with optimal features and optimal data analysis for a specific clinical problem. This advanced form of computer-aided biological profiling (where a neural network can extract features and link them together on a massive scale) can be termed as Deep Learning Mediated Tumour Profiling (DL-TP).

The next application of radiomic features in cancer research was prediction of response to different forms of treatment (i.e. Computer Aided Response Prediction (CARP)). In non-small cell lung cancer (NSCLC), Coroller et al. identified seven features that were predictive for pathological gross residual disease and one feature for pathological complete response [Cor+16]. Further studies later identified other radiomic features that would predict response to conventional treatment (i.e. chemo/radiotherapy) in bladder cancer [Cha+17] and locally advanced rectal cancer [Lov+18].

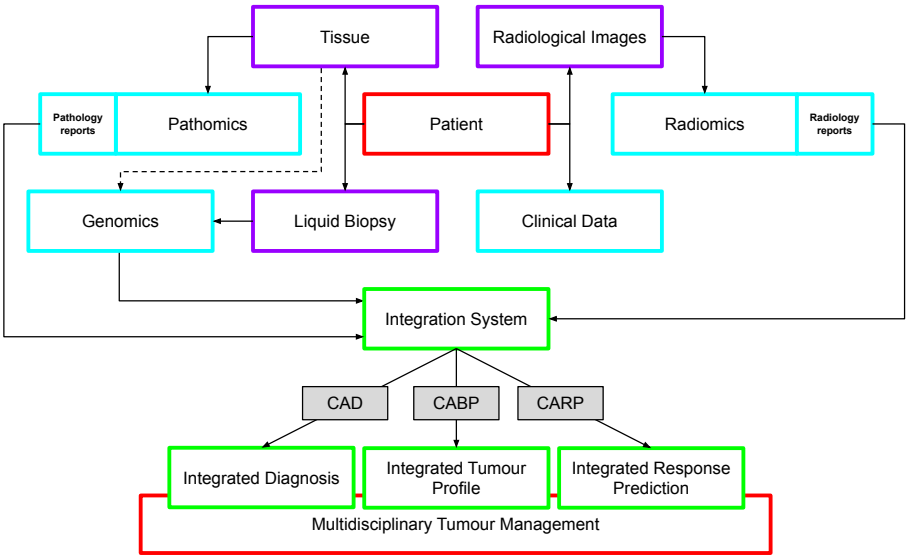

Figure 8.1: A schematic of a future radiomics pipeline highlighting a simplified workflow for CABP and CARP wherein patient images are input into a specialized (series of) AI algorithm(s) and based on the outcome, can be classified. CABP algorithms assess the profile of the tumour (for stratification) while CARP algorithms focus purely on the prediction of response to (and ultimately selection of) therapy. **CAD** = Computer Aided Diagnosis, **CARP** = Computer Aided Response Prediction, **CABP** = Computer Aided Biological Profiling.

8.3 Integrated systems in healthcare

While the suffix of "–omics" has come to denote the idea of extracting valuable information from datasets, radiomics is only the latest addition to the ever-growing list of new fields of study within the fusion of advanced technology and modern medicine. Images derived from tissue (e.g. general microscopy, immunohistochemistry etc…) have also been subject to quantitative analysis and new information is being generated beyond what would be observed by a pathologist i.e. path-

omics. Genomics, the branch of molecular biology concerned with the mapping of the human genome, has helped to identify many genetic mutations and pathways that have been used for prognostication or as novel targets for modern therapeutics and has heavily relied on computer models developed by skilled bioinformaticians.

As it currently stands, different medical disciplines have developed different stratification methods, primarily based on their own field (i.e. radiological classifications, pathological and chemical laboratory classifications, clinical checklists used for prognostication etc…) – quite often to the exclusion of other departments. As these traditional scoring systems were often based on subjective interpretations of analogue readouts, combining these disparate outputs is quite challenging. The rise of the quantitative aspects of various medical disciples (i.e. the "–omics") presents a remarkably unique opportunity wherein information from different diagnostic modalities can be objectively integrated.

8.4 Conclusion

The long term vision for precision medicine should focus on the development of integration strategies, wherein data derived from the patient themselves could be used to guide the treating physician. Through intricate analyses that integrate clinical data, blood markers, pathomics, radiomics, and genomics, we envision that a patient can be provisionally diagnosed (via computer aided diagnosis), stratified into a molecular subtype of their tumour (via computer aided biological profiling), and have a recommended treatment formulated (via computer aided response prediction). This aggregation of sub-systems cooperating with the aim of achieving an overarching functionality (in this case better CAD, CARP, and CABP) is termed as the integration system (see Figure 8.2). This will require hand-in-hand multidisciplinary collaboration between the biomedical field (i.e. clinicians, geneticists, radiologists, pathologists, clinical chemists),

and the technical field (i.e. computer scientists, physicists, engineers, statisticians, and mathematicians) as well as an organizational structure wherein the departments will operate not in isolation but in successful integration.

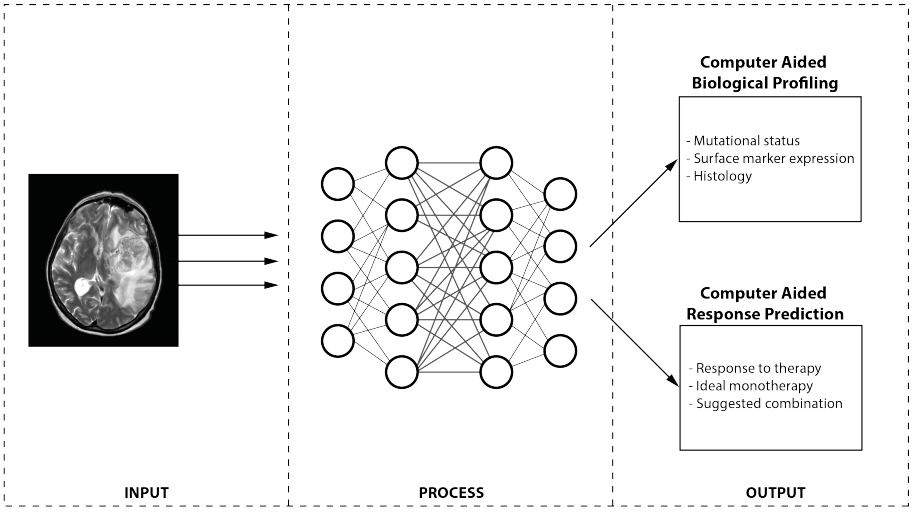

Figure 8.2: A schematic flow chart envisioning the usage of patient-derived data (in light blue) from raw materials (in purple) as a means to improve CAD, CARP, and CABP and ultimately help guide decisions by the multidisciplinary management team. **CAD** = Computer Aided Diagnosis, **CARP** = Computer Aided Response Prediction, **CABP** = Computer Aided Biological Profiling.

9
Discussion

Cancer immunotherapy drugs have been shown to improve outcomes in advanced stage cancer patients [Lar+18; Vok+18; Alb+20; Pai+19]. As these treatments will expand in the future (in terms of the number of monotherapy drugs and their combinations [Bla+18; Mot+19; Hel+19], or by exploiting synergistic effects of non-systemic treatments [PCW19; Twy+15]) the need for personalized medicine, and therefore markers able to select the right treatment for each patient, will increase accordingly [Lüt+20; LN19; MSG17]. Currently, biopsy-derived markers remain the most common type of marker under research [Bai+20; DC19; McK+20]. However, given the invasiveness of the procedure, the risk for sampling error and the fact that it provides information only of few biopsied lesions which does not always reflect the information of the entire tumour load, there is a need for non-invasive biomarkers that could reliable provide us the whole body information on how the patient will respond to treatment [SMH19; Du+19].

We hypothesized that advanced analytical methods of radiomics and artificial intelligence on clinical-routine whole-body imaging data could be used for a non-invasive prediction of treatment response and prognostication of outcome in patients receiving immunotherapy.

To test our hypothesis, we first investigated whether there are lesion-wise radiomic computer tomography (CT) features that could be associated with lesion specific biological profiles and tumor growth (chapters 2 and 3). The results would allow us to build an AI model for lesion response prediction in lung cancer and melanoma patients receiving immunotherapy. From these data, we observed that patients with mixed responses (i.e. some lesions growing while others shrinking) were associated with worse survival (chapter 2).

To investigate this further, we set up a different image analysis pipeline, that would enable us to gain insights into different patterns of response, and their relation to survival. We therefore developed an image analytic AI-system, able to quantify whole-body morphological changes in the interval between follow-up scans of the same patient, and investigated whether these changes would be correlated with

overall survival (chapters 4 and 5). Our results confirmed that the association between response patterns and survival outcomes cannot be explained solely by growth or shrinkage of single-lesions (or a sparse group thereof), nor by a group of tissue-based markers. Instead, our results suggest that specific subsets of lesions, which differ depending on the cancer type, as well as tumour and treatment-induced changes, were equally or better suited to explain overall survival than standard clinical, patho-immunological, and radiological factors (chapters 4, 5 and 6).

In this work, we applied different AI techniques for the study and development of imaging markers and imaging analytic pipelines. Chapter 2 was developed using engineered, handcrafted features, which allowed to focus the analysis to the tumor region only, and prove the existence of predictive imaging features within the tumour — albeit with a significant demand for manual labelling. The region of interest was extended to include the surrounding tissue in chapter 3. Here, we made use of a more advanced AI method for building the analytic pipeline, namely deep transfer learning. Through visualization techniques, we could observe that the AI algorithm was making use of inner-tumour features, as well as features from the surrounding tissue, highlighting the potential of these advanced techniques to handle more complex data. We took this idea further, by developing an AI-algorithm (chapters 4, 5 and 6) able to analyse the whole body — which of course includes tumors and surrounding tissues. We made use of novel deep self-supervised learning to remove the need of manual labelling, and deep learning based image registration to tailor the algorithm to the specific clinical need of response pattern modelling. This led to the development of the prognostic AI-monitor (PAM), which we envision to adapt and extend further into the clinics for imaging and non-imaging data.

The research work in this thesis provided evidence of the existence of morphological patterns in imaging, as processed by AI algorithms, that can function as predictive and prognostic markers in cancer patients receiving immunotherapy. Moreover, our results highlight how

imaging is the only routinely available tool which allows us to search, localize and connect data patterns across the entire body. Other, more invasive methods, might provide good predictive value as well (despite the fact that this was not observed for both PD-L1 expression, as well as laboratory results), but none of them is able to collect, process and prognosticate based on information that span across the whole body. This is a fundamental aspect in immunotherapy treated patients, as these often present with metastatic disease.

Our work culminates in the development of PAM: an AI-system that allows for comparative analysis between whole-body scans. Conceptually, the closest tool used in the clinics is the *Response Evaluation Criteria in Solid Tumour* (RECIST) [Sch+16; Eis+09; Sey+17a]. Our findings suggest that aspects that are not included in the current RECISTs can be easily, quickly and fully automatically assessed via AI-radiomics systems, such as PAM. We've proved that in patients receiving immunotherapy, response patterns evaluated from AI methods are equally or better suited to describe survival than approximate tumour growth (i.e. growth corresponds to decreased survival). In this cohort, it is evident that different lesions contribute differently to the prognosis (chapters 4, 5 and 6), suggesting that a different (possibly cancer-specific) selection of target lesions might be needed. It is also evident that non cancer lesions, such as treatment- and cancer-induced complications (chapters 4 and 5) and their diagnostic profiles (chapter 3), should be quantitatively accounted for in response evaluation, as they influence the survival of patients receiving immunotherapy.

A comprehensive overview is also of the current status of immunotherapy imaging marker research is also given in chapter 7. When compared to the work presented in this thesis, we see that single lesions analysis (chapters 2 and 3) are the most common in the literature. Response prediction studies have been proposed, with various degrees of novelty, by other studies [Nar+20; He+20; Maz+20; Bas+20; Mu+20; Tun+19; Vai+20; Par+20c], and led to similar results. Most of these studies [Nar+20; Bas+20; Mu+20; Tun+19; Vai+20;

Par+20c] made use of the original experimental design presented in chapters 2 and 3. Different designs included AI-radiomics models built to predict biological factors relevant to immunotherapy, like tutal mutational burden [He+20; Vee+20], microsatellite instability [Del+20], tumour micro-environment [Maz+20]. Only one study, by Del Re et al. [Del+20] proposed the integration of radiomics to biopsy-derived markers, which led to higher accuracies.

In chapters 4, 5 and 6 we proposed a completely different approach than the one applied in these studies. Conceptually, it could be associated with delta radiomics [Kho+20]. In delta radiomics, the difference between radiomics signature on serial imaging is used for response prediction and prognostication. This however is mostly based on radiomics signatures extracted from time-consuming segmentations, which also limits the analysis to the regions of the tumour, and does not account for any other prognostic or predictive factor that is not a cancer lesion. Instead, our approach is fast, comprehensive, fully automatic, and easily extendable to other imaging types, cancer types and possibly therapies.

The role of AI in clinical decision making

None of the aforementioned studies, including ours, address exhaustively how these methods would fit into a possible clinical scenario. The treatment of cancer patients is currently decided in a multidisciplinary meeting of medical specialists, i.e. the tumour board [El +14]. Here, the case of each cancer patient is presented, and discussed. The outcome, a patient-specific treatment and care plan, is the result of diagnostic information, oncologic guidelines [Bes+14; Mic+19; Bel+14] and expert knowledge, carefully knitted together, to achieve an optimal balance of different objectives, namely: prolonged survival, sufficient quality of life, and accomodation of patient's own requests [Sle+90; BTO01; Hir+05; McQ+95; Wee+98; Mar+14]. In this complex interplay of experts and specialties, it might be worth to

figure out where exactly an AI system like PAM (chapters 4, 5 and 6), or the lesion-wise correspondent (chapters 2 and 3), would fit.

The easiest scenario would see the AI being actively interrogated by clinicians in order to estimate the treatment with the highest prognostic likelihood. In this scenario however, it would be unclear to what extent such AI should be trained: should this include only the main treatment, or should this also account for side and palliative ones? The infeasibility of this solution becomes clearer when considering the plethora of treatment combinations that will be available in the near future [JD19; Roc+19; BBB18].

In the short term, a more likely scenario might see the training of main AI systems for the main available drugs (including immunotherapeutic ones). Lesions or side-effects likely to impair the success of the systematic treatment would be treated locally instead, via e.g. radiotherapy or radiofrequency ablation. The usefulness of such an approach is particularly evident in immunotherapy, where patients can remain under treatment up to 1 year, or until there are visible clinical benefits in doing so — namely the disease remains stable and under control [BC18; BKG20; Smi+20]. In this scenario, the AI would therefore accompany the tumour board, quickly analysing baseline and follow-up data to identify situations of risk to be discussed within the board. Such a scenario would see the clinicians in the board empowered with an "AI-eye", while still retaining control of the treatment planning.

It is important to note that the role of the AI to be played in this scenario does not replace the therapeutic specialities of the tumor board, but rather acts as complementary to them. However, the role of diagnostic specialties in relation to the AI would have to be defined. Our results from the multivariate analyses of chapters IV, V and VI show that currently used clinical methods can be complementary to AI-based methods. While we cannot preclude future development of other AI systems to replace current diagnostic methods and roles, current evidence is insufficient to support the replacement of any clinical role by AI.

The position of AI in integrated diagnosis

The large amount of data generated by current diagnostic departments (chapter 7), as well as the call for integrative diagnostic (chapter 8), requires specialized integrative AI-systems able to process high-dimensional, heterogeneous data, and scale it down, for expert interpretability. In this work, specialists in diagnostic disciplines (i.e. radiologists) were interpreters of AI-machine readings. In a future scenario where an "orchestra" of AIs will be running simultaneously within the department, the diagnostic specialist will have to assume the role of "conductor", collecting the results from different machines, interpreting and filtering them, and sharing them with the tumour board. Only the information relevant to the final objective should be provided to the board.

The objective itself remains the main hurdle. While improvement of survival is an objective, measurable function that can be analyzed mathematically, quality of life and accommodation of patient-specific requests require human comprehension and understanding. To the best of our knowledge, it is still unclear whether, in the near future, there will be an AI-machine able to create treatment plans which, for example, could accommodate even simple requests, such as "regardless of the prognosis, avoid amputation" or "spare bowel control, as much as possible". One could argue that an AI system could be trained to replicate the choices made in past cases. Even assuming that such an AI could be trained, and ignoring all practical issues related to it, such as the lack of necessary data to train such a system (e.g. conversations between physicians and patients), doubts would persist of whether this scenario were the wisest. Current accomplishments in cancer treatment are the results of years of multidisciplinary dialogue between different fields of science, which has the physicians as central coordinating nodes, thanks to their practical experience with patients. Were this node to be replaced by an AI trained to replicate treatment plans from the past (or variation of them), the entire innovation process might collapse, with negative

long term outcomes far overreaching the money saved from a reduction of personnel. We therefore argue that the optimal, and safest scenario is the one proposed, where the AI and the physician in the tumour board interact, interrogate and improve each other.

In this work, we aimed to assess the predictive and prognostic value of AI-radiomics on routinely available whole-body imaging in cancer patients receiving immunotherapy. Thanks to the usage of these systems, we gained deeper insights into the relation between tumour morphology and tumour response to immunotherapy, and immunotherapy response patterns to overall survival. This culminated in the development of an AI-system, PAM, tailored for follow-up analysis of cancer patients undergoing immunotherapy, and publicly shared with the community.

We envision the shared model to spark additional development from the clinical, as well as technological side. An extension of the current models, for example, to include more cancer types or stages, and different treatments. This process would be straightforward, and require relatively small fine-tuning procedures. It is to be seen if the same procedures and principles can be applied to different, non-imaging data that are routinely available during follow-up, such as laboratory, genetic and tissue data. Namely, if the same principle of tracking changes, evaluation response and relating it to prognostic outcomes using AI-systems still holds for non-imaging data. Even more, the question remains of how these systems would play in an integrative diagnostic scheme, where different AIs, working with different data types and different objectives are orchestrated together to provide clinically relevant information. We encourage further research in these questions prior to a clinical prospective validation.

In conclusion, this work presents a prospective of how AI-radiomics systems could benefit oncological care, with practical application in immunotherapy: one of the most promising, yet challenging treatment modality currently available. We showed not only how these systems can be used for prediction and prognostications, but also how they

can be used to gain additional medical and oncological insights, which relevance outreaches the original fields of radiology or computer science.

Bibliography

[Aer16] Hugo J W L Aerts. "The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review". en. In: *JAMA Oncol* 2.12 (Dec. 2016), pp. 1636–1642.

[AH16] Hugo J W L Aerts and Hugo J W. "The Potential of Radiomic-Based Phenotyping in Precision Medicine". In: *JAMA Oncology* 2.12 (2016), p. 1636.

[Aer+19] Hugo J W L Aerts et al. *Data From NSCLC-Radiomics*. 2019.

[Aer+15] Hugo J W L Aerts et al. *Data From NSCLC-Radiomics-Genomics*. 2015.

[Aer+14] Hugo J W L Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nat. Commun.* 5 (2014).

[Afs+19] P Afshar et al. "From Handcrafted to Deep-Learning-Based Cancer Radiomics: Challenges and Opportunities". In: *IEEE Signal Process. Mag.* 36.4 (July 2019), pp. 132–160.

[Ahu19] Abhimanyu S Ahuja. "The impact of artificial intelligence in medicine on the future role of the physician". en. In: *PeerJ* 7 (Oct. 2019), e7702.

[Alb+20] Laurence Albiges et al. "Nivolumab plus ipilimumab versus sunitinib for first-line treatment of advanced renal cell carcinoma: extended 4-year follow-up of the phase III CheckMate 214 trial". en. In: *ESMO Open* 5.6 (Nov. 2020).

[Ale+19] Francesco Alessandrino et al. "Frequency and imaging features of abdominal immune-related adverse events in metastatic lung cancer patients treated with PD-1 inhibitor". en. In: *Abdom Radiol (NY)* 44.5 (May 2019), pp. 1917–1927.

[Ali+19] Mehdi Alilou et al. *Quantitative vessel tortuosity radiomics on baseline non-contrast lung CT predict response to immunotherapy and are prognostic of overall survival*. 2019.

[Ana+17] V Anagnostou et al. "Immuno-oncology trial endpoints: capturing clinically meaningful activity". In: (2017).

[Arm+11] Samuel G Armato et al. *The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans*. 2011.

[ABB19] Moran Artzi, Idan Bressler, and Dafna Ben Bashat. "Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis". en. In: *J. Magn. Reson. Imaging* 50.2 (Aug. 2019), pp. 519–528.

[Arv+16] Nils D Arvold et al. "Updates in the management of brain metastases". en. In: *Neuro. Oncol.* 18.8 (Aug. 2016), pp. 1043–1065.

[Ask+99] Johan Askling et al. "Increased risk for cancer following sarcoidosis". In: *American Journal of Respiratory and Critical Care Medicine* 160.5 (1999), pp. 1668–1672.

[Aye+17] Mark Ayers et al. "IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade". In: *J. Clin. Invest.* 127.8 (Aug. 2017), pp. 2930–2940.

[Bai+20] Rilan Bai et al. "Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors". en. In: *Biomark Res* 8 (Aug. 2020), p. 34.

[Bak+17] Shaimaa Bakr et al. *Data for NSCLC Radiogenomics Collection*. 2017.

[Bal+19] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". en. In: *IEEE Trans. Med. Imaging* (Feb. 2019).

[Bal+17] Arjun V Balar et al. “Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial”. In: *Lancet* 389.10064 (Jan. 2017), pp. 67–76.

[BTO01] C E Balmer, P Thomas, and R J Osborne. “Who wants second-line, palliative chemotherapy?” en. In: *Psychooncology* 10.5 (Sept. 2001), pp. 410–418.

[BC18] Shanta Bantia and Nirmal Choradia. “Treatment duration with immune-based therapies in Cancer: an enigma”. en. In: *J Immunother Cancer* 6.1 (Dec. 2018), p. 143.

[Bas+20] Lucas Basler et al. “Radiomics, Tumor Volume, and Blood Biomarkers for Early Prediction of Pseudoprogression in Patients with Metastatic Melanoma Treated with Immune Checkpoint Inhibition”. en. In: *Clin. Cancer Res.* 26.16 (Aug. 2020), pp. 4414–4425.

[Bei+15] Reinhard R Beichel et al. *Data From QIN-HEADNECK*. 2015.

[Bej+17] Babak Ehteshami Bejnordi et al. “DEEP LEARNING-BASED ASSESSMENT OF TUMOR-ASSOCIATED STROMA FOR DIAGNOSING BREAST CANCER IN HISTOPATHOLOGY IMAGES”. en. In: *Proc. IEEE Int. Symp. Biomed. Imaging* 2017 (Apr. 2017), pp. 929–932.

[Bel+14] J Bellmunt et al. “Bladder cancer: ESMO Practice Guidelines for diagnosis, treatment and follow-up”. en. In: *Ann. Oncol.* 25 Suppl 3 (Sept. 2014), pp. iii40–8.

[Ber+16] Anna S Berghoff et al. “Immune Checkpoint Inhibitors in Brain Metastases: From Biology to Treatment”. en. In: *Am Soc Clin Oncol Educ Book* 35 (2016), e116–22.

[Bes+14] B Besse et al. "2nd ESMO Consensus Conference on Lung Cancer: non-small-cell lung cancer first-line/second and further lines of treatment in advanced disease". en. In: *Ann. Oncol.* 25.8 (Aug. 2014), pp. 1475–1484.

[BBB18] Neeraj Bhalla, Rachel Brooker, and Michael Brada. "Combining immunotherapy and radiotherapy in lung cancer". en. In: *J. Thorac. Dis.* 10.Suppl 13 (May 2018), S1447–S1460.

[Bha+19] Ankush Bhatia et al. "MRI radiomic features are associated with survival in melanoma brain metastases treated with immune checkpoint inhibitors". en. In: *Neuro. Oncol.* 21.12 (Dec. 2019), pp. 1578–1586.

[Bi+19a] Wenya Linda Bi et al. "Artificial intelligence in cancer imaging: Clinical challenges and applications". en. In: *CA Cancer J. Clin.* 69.2 (Mar. 2019), pp. 127–157.

[Bi+19b] Wenya Linda Bi et al. "Artificial intelligence in cancer imaging: clinical challenges and applications". In: *CA: a cancer journal for clinicians* 69.2 (2019), pp. 127–157.

[Bil+19] Mehmet A Bilen et al. "The prognostic and predictive impact of inflammatory biomarkers in patients who have advanced-stage cancer treated with immunotherapy: Inflammatory Biomarkers in Immunotherapy". In: *Cancer* 125.1 (2019), pp. 127–134.

[BKG20] Salem Billan, Orit Kaidar-Person, and Ziv Gil. "Treatment after progression in the era of immunotherapy". en. In: *Lancet Oncol.* 21.10 (Oct. 2020), e463–e476.

[Bir+17] Mathew R Birnbaum et al. "Nivolumab-related cutaneous sarcoidosis in a patient with lung adenocarcinoma". In: *JAAD Case Reports* 3.3 (2017), pp. 208–211.

[Bla+18] Christian U Blank et al. "Neoadjuvant versus adjuvant ipilimumab plus nivolumab in macroscopic stage III melanoma". en. In: *Nat. Med.* 24.11 (Nov. 2018), pp. 1655–1661.

[BJJ15] B Nicolas Bloch, Ashali Jain, and Conrade Carl Jaffe. *Data From BREAST-DIAGNOSIS*. 2015.

[BTB18] Zuhir Bodalal, Stefano Trebeschi, and Regina Beets-Tan. "Radiomics: a critical step towards integrated healthcare". In: *Insights into imaging* 9.6 (2018), pp. 911–914.

[Bod+19] Zuhir Bodalal et al. "Radiogenomics: bridging imaging and genomics". In: *Abdominal Radiology* 44.6 (2019), pp. 1960–1984.

[Bon+15] Martina Bonifazi et al. "Sarcoidosis and cancer risk: systematic review and meta-analysis of observational studies". In: *Chest* 147.3 (2015), pp. 778–791.

[Bor+15] Hossein Borghaei et al. "Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer". en. In: *N. Engl. J. Med.* 373.17 (Oct. 2015), pp. 1627–1639.

[Bos+15] Walter R Bosch et al. *Data From Head-Neck_Cetuximab*. 2015.

[Bra+15] Julie Brahmer et al. "Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer". en. In: *N. Engl. J. Med.* 373.2 (July 2015), pp. 123–135.

[B201] L Breiman - Machine learning and 2001. "Random forests". In: *Springer* (2001).

[Bro+20] Tom B Brown et al. "Language Models are Few-Shot Learners". In: (May 2020).

[CB18] Federico Cabitza and Giuseppe Banfi. "Machine learning in laboratory medicine: waiting for the flood?" In: *Clin. Chem. Lab. Med.* 56.4 (2018), pp. 516–524.

[Cap+16] Selene Capitanio et al. "PET/CT in nononcological lung diseases: current applications and future perspectives". In: *European Respiratory Review* 25.141 (2016), pp. 247–258.

[Car+17] David P Carbone et al. “First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer”. en. In: *N. Engl. J. Med.* 376.25 (June 2017), pp. 2415–2426.

[Cha+17] Kenny H Cha et al. “Bladder cancer treatment response assessment in CT using radiomics with deep-learning”. In: *Scientific reports* 7.1 (2017), pp. 1–12.

[Cha+18a] Yu Jin Cha et al. “Prediction of Response to Stereotactic Radiosurgery for Brain Metastases Using Convolutional Neural Networks”. en. In: *Anticancer Res.* 38.9 (Sept. 2018), pp. 5437–5445.

[Cha+18b] Stéphane Champiat et al. “Hyperprogressive disease: recognizing a novel pattern to improve patient management”. en. In: *Nat. Rev. Clin. Oncol.* 15.12 (Dec. 2018), pp. 748–762.

[Che+17] Bojiang Chen et al. “Development and clinical application of radiomics in lung cancer”. en. In: *Radiat. Oncol.* 12.1 (Sept. 2017), p. 154.

[Che+20] Jiaming Chen et al. “Noninvasive CT radiomic model for preoperative prediction of lymph node metastasis in early cervical carcinoma”. en. In: *Br. J. Radiol.* 93.1108 (Apr. 2020), p. 20190558.

[Cla+13] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. en. In: *J. Digit. Imaging* 26.6 (Dec. 2013), pp. 1045–1057.

[Coc+19] Kim Cocks et al. “A Q-TWiST Analysis Comparing Nivolumab and Therapy of Investigator’s Choice in Patients with Recurrent/Metastatic Platinum-Refractory Squamous Cell Carcinoma of the Head and Neck”. en. In: *Pharmacoeconomics* 37.8 (Aug. 2019), pp. 1041–1047.

[Coh+20] I Glenn Cohen et al. “The European artificial intelligence strategy: implications and challenges for digital health”. en. In: *Lancet Digit Health* 2.7 (July 2020), e376–e379.

[Coh+16] J V Cohen et al. *Melanoma Brain Metastasis Pseudoprogression after Pembrolizumab Treatment*. 2016.

[CK07] Philip R Cohen and Razelle Kurzrock. “Sarcoidosis and malignancy”. In: *Clinics in dermatology* 25.3 (2007), pp. 326–333.

[Cor+16] Thibaud P Coroller et al. “Radiomic phenotype features predict pathological response in non-small cell lung cancer”. In: *Radiotherapy and oncology* 119.3 (2016), pp. 480–486.

[CT19] Tricia Cottrell and Janis M Taube. “PD-L1 and Emerging Biomarkers in PD-1/PD-L1 Blockade Therapy”. In: (2019), p. 14.

[Cou+18] Nicolas Coudray et al. “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning”. en. In: *Nat. Med.* 24.10 (Oct. 2018), pp. 1559–1567.

[Cox58] D R Cox. “The Regression Analysis of Binary Sequences”. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 20.2 (1958), pp. 215–242.

[CSA19] Ashleigh Cruickshank, Geoff Stieler, and Faisal Ameer. “Evaluation of the solitary pulmonary nodule”. In: *Internal Medicine Journal* 49.3 (2019), pp. 306–315.

[Cun+15] J J Cunningham et al. *Divergent and convergent evolution in metastases suggest treatment strategies based on specific metastatic sites*. 2015.

[CA20] Vesna Cuplov and Nicolas André. “Machine Learning Approach to Forecast Chemotherapy-Induced Haematological Toxicities in Patients with Rhabdomyosarcoma”. In: *Cancers* 12.7 (2020), p. 1944.

[19] "Cutaneous melanoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". In: *Ann. Oncol.* 30.12 (Dec. 2019), pp. 1884–1901.

[Dan+16] François-Xavier Danlos et al. "Nivolumab-induced sarcoid-like granulomatous reaction in a patient with advanced melanoma". In: *Chest* 149.5 (2016), e133–e136.

[Del+20] Marzia Del Re et al. "A multiparametric approach to improve the prediction of response to immunotherapy in patients with metastatic NSCLC". en. In: *Cancer Immunol. Immunother.* (Dec. 2020).

[Del+19] Marta Della Seta et al. "A 3D quantitative imaging biomarker in pre-treatment MRI predicts overall survival after stereotactic radiation therapy of patients with a singular brain metastasis". en. In: *Acta radiol.* 60.11 (Nov. 2019), pp. 1496–1503.

[Dim+18] Florentia Dimitriou et al. "Sarcoid-like reactions in patients receiving modern melanoma treatment". In: *Melanoma research* 28.3 (2018), p. 230.

[Du+19] Yang Du et al. "Noninvasive imaging in cancer immunotherapy: The way to precision medicine". en. In: *Cancer Lett.* 466 (Dec. 2019), pp. 13–22.

[Du+18] Yue Du et al. "Classification of Tumor Epithelium and Stroma by Exploiting Image Features Learned by Deep Convolutional Neural Networks". en. In: *Ann. Biomed. Eng.* 46.12 (Dec. 2018), pp. 1988–1999.

[DC19] Michael J Duffy and John Crown. "Biomarkers for Predicting Response to Immunotherapy with Immune Checkpoint Inhibitors in Cancer Patients". en. In: *Clin. Chem.* 65.10 (Oct. 2019), pp. 1228–1238.

[Egg+19] Hendrik Eggers et al. “Sarcoid-Like Lesions Mimicking Pulmonary Metastasis: A Case Series and Review of the Literature”. In: *Oncology research and treatment* 42.7/8 (2019), pp. 382–386.

[Eht+17] Babak Ehteshami Bejnordi et al. “Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer”. en. In: *JAMA* 318.22 (Dec. 2017), pp. 2199–2210.

[Eis+09] E A Eisenhauer et al. “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)”. en. In: *Eur. J. Cancer* 45.2 (Jan. 2009), pp. 228–247.

[El +14] Nagi S El Saghir et al. “Tumor boards: optimizing the structure and improving efficiency of multidisciplinary management of patients with cancer worldwide”. en. In: *Am Soc Clin Oncol Educ Book* (2014), e461–6.

[Eri+17] Bradley Erickson et al. *Data from LGG-1p19qDeletion*. 2017.

[Eura] European Medicines Agency. *Keytruda: EPAR - Product Information*. `http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/003820/WC500190990.pdf`. Accessed: 2018-9-5.

[Eurb] European Medicines Agency. *Opdivo: EPAR - Product Information*. `http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/003985/WC500189765.pdf`. Accessed: 2018-9-5.

[Eurc] European Medicines Agency. *Yervoy*. `https://www.ema.europa.eu/en/medicines/human/EPAR/yervoy`. Accessed: 2020-12-17.

[Fal+19] Thorsten Falk et al. “U-Net: deep learning for cell counting, detection, and morphometry”. en. In: *Nat. Methods* 16.1 (Jan. 2019), pp. 67–70.

[Fav+17] X Fave et al. “Using Pretreatment Radiomics and Delta-Radiomics Features to Predict Non–Small Cell Lung Cancer Patient Outcomes”. In: *Int. J. Radiat. Oncol. Biol. Phys.* 98.1 (May 2017), p. 249.

[Fel15] S Feller. “One in four cancer trials fails to enroll enough participants”. In: *UPI.* (2015).

[Fer+18] Robert L Ferris et al. “Nivolumab vs investigator’s choice in recurrent or metastatic squamous cell carcinoma of the head and neck: 2-year long-term survival update of CheckMate 141 with analyses by tumor PD-L1 expression”. en. In: *Oral Oncol.* 81 (June 2018), pp. 45–51.

[FB81] Martin A Fischler and Robert C Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Commun. ACM* 24.6 (June 1981), pp. 381–395.

[Fog18] David B Fogel. “Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review”. en. In: *Contemp Clin Trials Commun* 11 (Sept. 2018), pp. 156–164.

[Fou+13] Mona N Fouad et al. “Enrollment of patients with lung and colorectal cancers onto clinical trials”. en. In: *J. Oncol. Pract.* 9.2 (Mar. 2013), e40–7.

[Fra+18] Daniel Franzen et al. “Ipilimumab and early signs of pulmonary toxicity in patients with metastastic melanoma: a prospective observational study”. In: *Cancer Immunology, Immunotherapy* 67.1 (2018), pp. 127–134.

[FPP16] Claire F Friedman, Tracy A Proverbs-Singh, and Michael A Postow. “Treatment of the immune-related adverse effects of immune checkpoint inhibitors: a review”. In: *JAMA oncology* 2.10 (2016), pp. 1346–1353.

[Gal+20] Norbert Galldiks et al. “Imaging challenges of immunotherapy and targeted therapy in patients with brain metastases: response, progression, and pseudoprogression”. en. In: *Neuro. Oncol.* 22.1 (Jan. 2020), pp. 17–30.

[Gam+18] T Gambichler et al. “Baseline laboratory parameters predicting clinical outcome in melanoma patients treated with ipilimumab: a single-centre analysis”. In: *J. Eur. Acad. Dermatol. Venereol.* 32.6 (2018), pp. 972–977.

[Gan+18a] David R Gandara et al. “Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab”. In: *Nat. Med.* 24.9 (2018), pp. 1441–1448.

[Gan+18b] Leena Gandhi et al. “Pembrolizumab plus Chemotherapy in Metastatic Non–Small-Cell Lung Cancer”. In: *N. Engl. J. Med.* 378.22 (Apr. 2018), pp. 2078–2092.

[Gar+19] Roberto Garcıa-Figueiras et al. *How clinical imaging can assess cancer biology*. 2019.

[Gha+18] Sameer R Ghate et al. “Economic Burden of Adverse Events Associated with Immunotherapy and Targeted Therapy for Metastatic Melanoma in the Elderly”. en. In: *Am Health Drug Benefits* 11.7 (Oct. 2018), pp. 334–343.

[GWA16] Geoffrey T Gibney, Louis M Weiner, and Michael B Atkins. “Predictive biomarkers for checkpoint inhibitor-based immunotherapy”. en. In: *Lancet Oncol.* 17.12 (Dec. 2016), e542–e551.

[GKH16] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. “Radiomics: Images Are More than Pictures, They Are Data”. en. In: *Radiology* 278.2 (Feb. 2016), pp. 563–577.

[Gol+16] Simone M Goldinger et al. “Cytotoxic cutaneous adverse drug reactions during anti-PD-1 therapy”. In: *Clinical Cancer Research* 22.16 (2016), pp. 4023–4029.

[Gra+16] Richard NJ Graham et al. “Return of the pulmonary nodule: the radiologist’s key role in implementing the 2015 BTS guidelines on the investigation and management of pulmonary nodules”. In: *The British Journal of Radiology* 89.1059 (2016), p. 20150776.

[GM12] Mel Greaves and Carlo C Maley. “Clonal evolution in cancer”. en. In: *Nature* 481.7381 (Jan. 2012), pp. 306–313.

[Gre+17] Richard Green et al. “Management of pulmonary nodules in head and neck cancer patients–Our experience and interpretation of the British Thoracic Society Guidelines”. In: *The Surgeon* 15.4 (2017), pp. 227–230.

[Gri+17a] Joost J M van Griethuysen et al. “Computational Radiomics System to Decode the Radiographic Phenotype”. en. In: *Cancer Res.* 77.21 (Nov. 2017), e104–e107.

[Gri+17b] Joost J M van Griethuysen et al. “Computational Radiomics System to Decode the Radiographic Phenotype”. en. In: *Cancer Res.* 77.21 (Nov. 2017), e104–e107.

[G217] J J Griethuysen - Cancer Research and 2017. “Computational Radiomics System to Decode the Radiographic Phenotype”. In: (2017).

[Gro+20] Aaron Grossberg et al. *HNSCC*. 2020.

[Gro+17] Patrick Grossmann et al. “Defining the biological basis of radiomic phenotypes in lung cancer”. en. In: *Elife* 6 (July 2017).

[Gru+19] Damien Gruson et al. “Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation”. In: *Clin. Biochem.* 69 (2019), pp. 1–7.

[Hag+19] Akihiro Haga et al. “Standardization of imaging features for radiomics analysis”. In: *J. Med. Invest.* 66.1.2 (2019), pp. 35–37.

[Hal+14] Darragh F Halpenny et al. “Are there imaging characteristics associated with lung adenocarcinomas harboring ALK rearrangements?” en. In: *Lung Cancer* 86.2 (Nov. 2014), pp. 190–194.

[Har+19] Stefan Harrer et al. “Artificial Intelligence for Clinical Trial Design”. en. In: *Trends Pharmacol. Sci.* 40.8 (Aug. 2019), pp. 577–591.

[HKY20] Grant Haskins, Uwe Kruger, and Pingkun Yan. *Deep learning in medical image registration: a survey*. 2020.

[HB17] Nuh Hatipoglu and Gokhan Bilgin. “Cell segmentation in histopathological images with deep learning algorithms by utilizing spatial relationships”. en. In: *Med. Biol. Eng. Comput.* 55.10 (Oct. 2017), pp. 1829–1848.

[HCC19a] Jonathan J Havel, Diego Chowell, and Timothy A Chan. “The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy”. en. In: *Nat. Rev. Cancer* 19.3 (Mar. 2019), pp. 133–150.

[HCC19b] Jonathan J Havel, Diego Chowell, and Timothy A Chan. “The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy”. en. In: *Nat. Rev. Cancer* 19.3 (Mar. 2019), pp. 133–150.

[Hay+77] J L Hayward et al. “Assessment of response to therapy in advanced breast cancer”. en. In: *Br. J. Cancer* 35.3 (Mar. 1977), pp. 292–298.

[He+20] Bingxi He et al. “Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker”. In: *Journal for ImmunoTherapy of Cancer* 8.2 (2020), pp. 1–10.

[He+17] Yayi He et al. “PD-1, PD-L1 Protein Expression in Non-Small Cell Lung Cancer and Their Relationship with Tumor-Infiltrating Lymphocytes”. en. In: *Med. Sci. Monit.* 23 (Mar. 2017), pp. 1208–1216.

[Hec+20] Stefanie J Hectors et al. “MRI radiomics features predict immuno-oncological characteristics of hepatocellular carcinoma”. en. In: *Eur. Radiol.* 30.7 (July 2020), pp. 3759–3769.

[Hel+19] Nicholas Heller et al. *C4KC KiTS Challenge Kidney Tumor Segmentation Dataset*. 2019.

[Hel+18a] Matthew D Hellmann et al. “Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden”. en. In: *N. Engl. J. Med.* 378.22 (May 2018), pp. 2093–2104.

[Hel+18b] Matthew D Hellmann et al. “Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden”. In: *N. Engl. J. Med.* 378.22 (2018), pp. 2093–2104.

[Hen+17] Shona Hendry et al. “Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group”. In: *Adv. Anat. Pathol.* 24.5 (2017), pp. 235–251.

[HGM17] Carl Heneghan, Ben Goldacre, and Kamal R Mahtani. “Why clinical trial outcomes fail to translate into benefits for patients”. en. In: *Trials* 18.1 (Mar. 2017), p. 122.

[Hen+02] Claudia I Henschke et al. “CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules”. In: *American Journal of Roentgenology* 178.5 (2002), pp. 1053–1057.

[Her+16] Roy S Herbst et al. “Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial”. en. In: *Lancet* 387.10027 (Apr. 2016), pp. 1540–1550.

[Hir+05] Takashi Hirose et al. “Patients preferences in chemotherapy for advanced non-small-cell lung cancer”. en. In: *Intern. Med.* 44.2 (Feb. 2005), pp. 107–113.

[Ho+20a] Tsung-Ying Ho et al. “Classifying Neck Lymph Nodes of Head and Neck Squamous Cell Carcinoma in MRI Images with Radiomic Features”. en. In: *J. Digit. Imaging* (Jan. 2020).

[Ho+20b] Won Jin Ho et al. *Multipanel mass cytometry reveals anti–PD-1 therapy–mediated B and T cell compartment remodeling in tumor-draining lymph nodes*. 2020.

[Hod+10] F Stephen Hodi et al. “Improved survival with ipilimumab in patients with metastatic melanoma”. In: *New England Journal of Medicine* 363.8 (2010), pp. 711–723.

[Hof19] Paul Hofman. “The challenges of evaluating predictive biomarkers using small biopsy tissue samples and liquid biopsies from non-small cell lung cancer patients”. en. In: *J. Thorac. Dis.* 11.Suppl 1 (Jan. 2019), S57–S64.

[Hos+18] Ahmed Hosny et al. “Artificial intelligence in radiology”. en. In: *Nat. Rev. Cancer* (May 2018).

[HMM19] Masatoshi Hotta, Ryogo Minamimoto, and Kenta Miwa. “11C-methionine-PET for differentiating recurrent brain tumor from radiation necrosis: radiomics approach with random forest classifier”. en. In: *Sci. Rep.* 9.1 (Oct. 2019), p. 15666.

[Hri11] Hedvig Hricak. “Oncologic imaging: a guiding hand of personalized cancer care”. en. In: *Radiology* 259.3 (June 2011), pp. 633–640.

[Hua+13] Yuhui Huang et al. "Vascular normalization as an emerging strategy to enhance cancer immunotherapy". en. In: *Cancer Res.* 73.10 (May 2013), pp. 2943–2948.

[Hwa+16] Thomas J Hwang et al. "Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results". en. In: *JAMA Intern. Med.* 176.12 (Dec. 2016), pp. 1826–1833.

[IRT07] MC Iannuzzi, BA Rybicki, and AS Teirstein. "Sarcoidosis". In: *J Engl J Med* 357 (2007), pp. 2153–2165.

[Ion+93] Y Ionov et al. "Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis". en. In: *Nature* 363.6429 (June 1993), pp. 558–561.

[Ish+92] Y Ishida et al. "Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death". en. In: *EMBO J.* 11.11 (Nov. 1992), pp. 3887–3895.

[JM] Mika S Jain and Tarik F Massoud. *Predicting tumour mutational burden from histopathological images using multiscale deep learning*.

[Joh+16] Douglas B Johnson et al. "Fulminant Myocarditis with Combination Immune Checkpoint Blockade". en. In: *N. Engl. J. Med.* 375.18 (Nov. 2016), pp. 1749–1755.

[JPP18] Dennis Jones, Ethel R Pereira, and Timothy P Padera. "Growth and Immune Evasion of Lymph Node Metastasis". en. In: *Front. Oncol.* 8 (Feb. 2018), p. 36.

[JD19] Shweta Joshi and Donald L Durden. "Combinatorial Approach to Improve Cancer Immunotherapy: Rational Drug Design Strategy to Simultaneously Hit Multiple Targets to Kill Tumor Cells and to Activate the Immune System". en. In: *J. Oncol.* 2019 (Feb. 2019), p. 5245034.

[Jua+18] Tiffany M Juarez et al. “Understanding the brain tumor microenvironment: Considerations to applying systems biology and immunotherapy”. en. In: *International Journal of Neurooncology* 1.1 (Jan. 2018), p. 25.

[Kat+19] Jakob Nikolas Kather et al. “Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer”. en. In: *Nat. Med.* 25.7 (July 2019), pp. 1054–1056.

[Kel+19] Christopher J Kelly et al. “Key challenges for delivering clinical impact with artificial intelligence”. en. In: *BMC Med.* 17.1 (Oct. 2019), p. 195.

[Ker+15] Keith M Kerr et al. “Programmed Death-Ligand 1 Immunohistochemistry in Lung Cancer: In what state is this art?” en. In: *J. Thorac. Oncol.* 10.7 (July 2015), pp. 985–989.

[Kho+20] Mohammadhadi Khorrami et al. “Changes in CT radiomic features associated with lymphocyte distribution predict overall survival and response to immunotherapy in non–small cell lung cancer”. In: *Cancer Immunology Research* 8.1 (2020), pp. 108–119.

[Kic+16] Philipp Kickingereder et al. “Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models”. en. In: *Radiology* 280.3 (Sept. 2016), pp. 880–889.

[Kim+13] Sae Byol Kim et al. “Ground-glass opacity in lung metastasis from breast cancer: a case report”. In: *Tuberculosis and Respiratory Diseases* 74.1 (2013), pp. 32–36.

[Kim19] Kwang Gi Kimg. “Deep Learning”. In: *Healthc. Inform. Res.* 22.4 (2019), pp. 351–354.

[Kin+17] Paul Kinahan et al. *Data from ACRIN-FLT-Breast*. 2017.

[Kin+18] Paul Kinahan et al. *Data from ACRIN-FMISO-Brain*. 2018.

[Kin+19] Paul Kinahan et al. *Data from the ACRIN 6668 Trial NSCLC-FDG-PET*. 2019.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[Kir+17] Margarita Kirienko et al. "Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery". en. In: *Eur. J. Nucl. Med. Mol. Imaging* (Sept. 2017).

[Kis+16] Pavel Kisilev et al. "Semantic description of medical image findings: structured learning approach". In: *Proceedings of the Britisch Machine Vision Conference* (2016), pp. 171.1–171.11.

[Kle+10] Stefan Klein et al. "elastix: a toolbox for intensity-based medical image registration". en. In: *IEEE Trans. Med. Imaging* 29.1 (Jan. 2010), pp. 196–205.

[Kni+19] Helge C Kniep et al. "Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type". en. In: *Radiology* 290.2 (Feb. 2019), pp. 479–487.

[Al-+18] Yousef Al-Kofahi et al. "A deep learning-based algorithm for 2-D cell segmentation in microscopy images". en. In: *BMC Bioinformatics* 19.1 (Oct. 2018), p. 365.

[KJ14] Michael D Kuo and Neema Jamshidi. "Behind the numbers: Decoding molecular phenotypes with radiogenomics- guiding principles and technical considerations". In: *Radiology* 270.2 (2014), pp. 320–325.

[Kur+15] Karen A Kurdziel et al. *Data From NaF_PROSTATE*. 2015.

[Lar+18] James Larkin et al. "Overall Survival in Patients With Advanced Melanoma Who Received Nivolumab Versus Investigator's Choice Chemotherapy in CheckMate 037: A Randomized, Controlled, Open-Label Phase III Trial". en. In: *J. Clin. Oncol.* 36.4 (Feb. 2018), pp. 383–390.

[LKA96] D R Leach, M F Krummel, and J P Allison. “Enhancement of antitumor immunity by CTLA-4 blockade”. en. In: *Science* 271.5256 (Mar. 1996), pp. 1734–1736.

[Li+16] Hui Li et al. “MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays”. en. In: *Radiology* 281.2 (Nov. 2016), pp. 382–391.

[Li+20] Menglei Li et al. “A clinical-radiomics nomogram for the preoperative prediction of lymph node metastasis in colorectal cancer”. en. In: *J. Transl. Med.* 18.1 (Jan. 2020), p. 46.

[Li+18] Qian Li et al. “Comparison Between Radiological Semantic Features and Lung-RADS in Predicting Malignancy of Screen-Detected Lung Nodules in the National Lung Screening Trial”. en. In: *Clin. Lung Cancer* 19.2 (Mar. 2018), 148–156.e3.

[Lib+11] Arthur Liberzon et al. “Molecular signatures database (MSigDB) 3.0”. en. In: *Bioinformatics* 27.12 (June 2011), pp. 1739–1740.

[Lin+17] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[LG16] Helena Linardou and Helen Gogas. “Toxicity management of immunotherapy for patients with metastatic melanoma”. In: *Annals of translational medicine* 4.14 (2016).

[Liu+20] Xiaoxuan Liu et al. *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. 2020.

[Liu+19] Zhenyu Liu et al. "The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges". en. In: *Theranostics* 9.5 (Feb. 2019), pp. 1303–1322.

[Loh+18] Philipp Lohmann et al. "Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis". en. In: *Neuroimage Clin* 20 (Aug. 2018), pp. 537–542.

[Lov+18] Pierre Lovinfosse et al. "FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer". In: *European Journal of Nuclear Medicine and Molecular Imaging* 45.3 (2018), pp. 365–375.

[LN19] Siew-Kee Low and Yusuke Nakamura. "The road map of cancer precision medicine with the innovation of advanced cancer detection technology and personalized immunotherapy". en. In: *Jpn. J. Clin. Oncol.* 49.7 (July 2019), pp. 596–603.

[Lu+19] Victor M Lu et al. "Concurrent versus non-concurrent immune checkpoint inhibition with stereotactic radiosurgery for metastatic brain disease: a systematic review and meta-analysis". In: *J. Neurooncol.* 141.1 (Jan. 2019), pp. 1–12.

[Luc+19] C Luchini et al. "ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach". en. In: *Ann. Oncol.* 30.8 (Aug. 2019), pp. 1232–1243.

[Lüt+20] Susanne Lütje et al. "Immune Checkpoint Imaging in Oncology: A Game Changer Toward Personalized Immunotherapy?" en. In: *J. Nucl. Med.* 61.8 (Aug. 2020), pp. 1137–1144.

[Lv+18] J Lv et al. “Comparison of CT radiogenomic and clinical characteristics between EGFR and KRAS mutations in lung adenocarcinomas”. en. In: *Clin. Radiol.* 73.6 (June 2018), 590.e1–590.e8.

[Lyl+16] L T Lyle et al. “Alterations in pericyte subpopulations are associated with elevated blood–tumor barrier permeability in experimental brain metastasis of breast cancer”. In: *Clin. Cancer Res.* (2016).

[Ma+16] Weijie Ma et al. “Current status and perspectives in translational biomarker research for PD-1/PD-L1 immune checkpoint blockade therapy”. en. In: *J. Hematol. Oncol.* 9.1 (May 2016), p. 47.

[MSG17] Laura Maciejko, Munisha Smalley, and Aaron Goldman. “Cancer Immunotherapy and Personalized Medicine: Emerging Technologies and Biomarker-Based Approaches”. en. In: *J. Mol. Biomark. Diagn.* 8.5 (Sept. 2017).

[Mah+19] Faisal Mahmood et al. “Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images”. en. In: *IEEE Trans. Med. Imaging* PP (July 2019).

[MK16] Artem B Mamonov and Jayashree Kalpathy-Cramer. *Data From QIN GBM Treatment Response*. 2016.

[Man+13] Subramani Mani et al. “Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy”. In: *J. Am. Med. Inform. Assoc.* 20.4 (2013), pp. 688–695.

[Man11] Matthew Mansh. “Ipilimumab and cancer immunotherapy: a new hope for advanced stage melanoma”. In: *The Yale journal of biology and medicine* 84.4 (2011), p. 381.

[Mar+19] Leigh Marcus et al. "FDA approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors". In: *Clinical Cancer Research* 25.13 (2019), pp. 3753–3758.

[Mar+14] Guilherme Nader Marta et al. "Treatment priorities in oncology: do we want to live longer or better?" en. In: *Clinics* 69.8 (Aug. 2014), pp. 509–514.

[May+20] Marius E Mayerhoefer et al. "Introduction to radiomics". In: *J. Nucl. Med.* 61.4 (2020), pp. 488–495.

[Maz+19] Maciej A Mazurowski et al. "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI". In: *J. Magn. Reson. Imaging* 49.4 (2019), pp. 939–954.

[Maz+20] Giulia Mazzaschi et al. "Integrated CT imaging and tissue immune features disclose a radio-immune signature with high prognostic impact on surgically resected NSCLC". en. In: *Lung Cancer* 144 (June 2020), pp. 30–39.

[McG+16] Nicholas McGranahan et al. "Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade". en. In: *Science* 351.6280 (Mar. 2016), pp. 1463–1469.

[McK+20] William B McKean et al. "Biomarkers in Precision Cancer Immunotherapy: Promise and Challenges". en. In: *Am Soc Clin Oncol Educ Book* 40 (May 2020), e275–e291.

[McN+07] M F McNitt-Gray et al. *The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation*. 2007.

[McQ+95] R P McQuellon et al. "Patient preferences for treatment of metastatic breast cancer: a study of women with early-stage breast cancer". en. In: *J. Clin. Oncol.* 13.4 (Apr. 1995), pp. 858–868.

[Mek+18] Ahmed Mekki et al. “Detection of immune-related adverse events by medical imaging in patients treated with anti-programmed cell death 1”. en. In: *Eur. J. Cancer* 96 (June 2018), pp. 91–104.

[Men+15] Xiangjiao Meng et al. “Predictive biomarkers in PD-1/PD-L1 checkpoint blockade immunotherapy”. en. In: *Cancer Treat. Rev.* 41.10 (Dec. 2015), pp. 868–876.

[Mic+19] O Michielin et al. “Cutaneous melanoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up”. In: *Ann. Oncol.* 30.12 (2019), pp. 1884–1901.

[Al-+19] Zahraa Al-Milaji et al. *Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H&E images*. 2019.

[MS05] Wallace T Miller Jr and Rosita M Shah. “Isolated diffuse ground-glass opacity in thoracic CT: causes and clinical presentations”. In: *American Journal of Roentgenology* 184.2 (2005), pp. 613–622.

[Möl+20] Miriam Möller et al. “Blood Immune Cell Biomarkers in Patient With Lung Cancer Undergoing Treatment With Checkpoint Blockade”. In: *J. Immunother.* 43.2 (2020), pp. 57–66.

[Mot+19] Robert J Motzer et al. “Nivolumab plus ipilimumab versus sunitinib in first-line treatment for advanced renal cell carcinoma: extended follow-up of efficacy and safety results from a randomised, controlled, phase 3 trial”. en. In: *Lancet Oncol.* 20.10 (Oct. 2019), pp. 1370–1385.

[Mot+20] Robert J Motzer et al. “Survival outcomes and independent response assessment with nivolumab plus ipilimumab versus sunitinib in patients with advanced renal cell carcinoma: 42-month follow-up of a randomized phase 3 clinical trial”. en. In: *J Immunother Cancer* 8.2 (July 2020).

[Mu+18] Wei Mu et al. *Radiomic biomarkers from PET/CT multi-modality fusion images for the prediction of immunotherapy response in advanced non-small cell lung cancer patients*. 2018.

[Mu+20] Wei Mu et al. "Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy". en. In: *Eur. J. Nucl. Med. Mol. Imaging* 47.5 (May 2020), pp. 1168–1182.

[MWK15] Peter Muzi, Michelle Wanner, and Paul Kinahan. *Data From RIDER Lung PET-CT*. 2015.

[Nai+18] Arjun Nair et al. "Variable radiological lung nodule evaluation leads to divergent management recommendations". In: *European Respiratory Journal* 52.6 (2018).

[Nar+20] Valerio Nardone et al. "Radiomics predicts survival of patients with advanced non-small cell lung cancer undergoing PD-1 blockade using Nivolumab". en. In: *Oncol. Lett.* 19.2 (Feb. 2020), pp. 1559–1566.

[NO19] Faria Nasim and David E Ost. "Management of the solitary pulmonary nodule". In: *Current opinion in pulmonary medicine* 25.4 (2019), pp. 344–353.

[Nay+17] Peter Naylor et al. *Nuclei segmentation in histopathology images using deep neural networks*. 2017.

[Ner+20] Emanuele Neri et al. "Involvement of radiologists in oncologic multidisciplinary team meetings: an international survey by the European Society of Oncologic Imaging". en. In: *Eur. Radiol.* (Aug. 2020).

[Nie+19] Ke Nie et al. "NCTN Assessment on Current Applications of Radiomics in Oncology". en. In: *Int. J. Radiat. Oncol. Biol. Phys.* 104.2 (June 2019), pp. 302–315.

[NHH19] Mizuki Nishino, Hiroto Hatabu, and F Stephen Hodi. "Imaging of Cancer Immunotherapy: Current Approaches and Future Directions". en. In: *Radiology* 290.1 (Jan. 2019), pp. 9–22.

[Nis+15] Mizuki Nishino et al. "Cancer immunotherapy and immune-related response assessment: the role of radiologists in the new arena of cancer treatment". In: *European journal of radiology* 84.7 (2015), pp. 1259–1268.

[Nis+13] Mizuki Nishino et al. "Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements". en. In: *Clin. Cancer Res.* 19.14 (July 2013), pp. 3936–3943.

[OCo+17] James P B O'Connor et al. "Imaging biomarker roadmap for cancer studies". en. In: *Nat. Rev. Clin. Oncol.* 14.3 (Mar. 2017), pp. 169–186.

[Ohs+17] Shinichiro Ohshimo et al. "Differential diagnosis of granulomatous lung disease: clues and pitfalls: Number 4 in the Series "Pathology for the clinician" Edited by Peter Dorfmüller and Alberto Cavazza". In: *European Respiratory Review* 26.145 (2017), p. 170012.

[Oka+15] Hideho Okada et al. "Immunotherapy response assessment in neuro-oncology: a report of the RANO working group". en. In: *Lancet Oncol.* 16.15 (Nov. 2015), e534–e542.

[Ols+16] Randal S Olson et al. "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science". In: *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. GECCO '16. Denver, Colorado, USA: ACM, 2016, pp. 485–492.

[Opi+20] Mark P van Opijnen et al. "The impact of current treatment modalities on the outcomes of patients with melanoma brain metastases: A systematic review". en. In: *Int. J. Cancer* 146.6 (Mar. 2020), pp. 1479–1489.

[OGB20] Ohad Oren, Bernard J Gersh, and Deepak L Bhatt. "Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints". In: *The Lancet Digital Health* 2.9 (2020), e486–e488.

[Org+79] World Health Organization et al. *WHO handbook for reporting results of cancer treatment*. World Health Organization, 1979.

[Ort+17] R Ortiz-Ramón et al. "A radiomics evaluation of 2D and 3D MRI texture features to classify brain metastases from lung cancer and melanoma". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2017, pp. 493–496.

[Ort+18] Rafael Ortiz-Ramón et al. "Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study". en. In: *Eur. Radiol.* 28.11 (Nov. 2018), pp. 4514–4523.

[Ove+13] Michael J Overman et al. "Use of research biopsies in clinical trials: are risks and benefits adequately discussed?" en. In: *J. Clin. Oncol.* 31.1 (Jan. 2013), pp. 17–22.

[Pai+19] Sara I Pai et al. "Comparative analysis of the phase III clinical trials of anti-PD1 monotherapy in head and neck squamous cell carcinoma patients (CheckMate 141 and KEYNOTE 040)". en. In: *J Immunother Cancer* 7.1 (Apr. 2019), p. 96.

[PGW95] HS Pandha, H Griffiths, and J Waxman. "Sarcoidosis and cancer". In: *Clinical Oncology* 7.5 (1995), pp. 277–278.

[PMK20] Nikolaos Papanikolaou, Celso Matos, and Dow Mu Koh. "How to develop a meaningful radiomic signature for clinical use in oncologic patients". In: *Cancer Imaging* 20.1 (2020), pp. 1–10.

[Par+20a] Changhee Park et al. “Tumor immune profiles noninvasively estimated by FDG PET with deep learning correlate with immunotherapy response in lung adenocarcinoma”. en. In: *Theranostics* 10.23 (Aug. 2020), pp. 10838–10848.

[Par+05] JES Park et al. “The HRCT appearances of granulomatous pulmonary disease in common variable immune deficiency”. In: *European journal of radiology* 54.3 (2005), pp. 359–364.

[Par+20b] Kye Jin Park et al. “Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma”. en. In: *Eur. Radiol.* 30.10 (Oct. 2020), pp. 5392–5403.

[Par+20c] Kye Jin Park et al. “Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma”. en. In: *Eur. Radiol.* 30.10 (Oct. 2020), pp. 5392–5403.

[Par+15] Chintan Parmar et al. “Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer”. en. In: *Front. Oncol.* 5 (Dec. 2015), p. 272.

[PU19] Luigi Pasini and Paola Ulivi. “Liquid Biopsy for the Detection of Resistance Mechanisms in NSCLC: Comparison of Different Blood Biomarkers”. In: *J. Clin. Med. Res.* 8.7 (2019), p. 998.

[Pat+19] Pradnya Patil et al. *A combination of intra- and peritumoral features on baseline CT scans is associated with overall survival in non-small cell lung cancer patients treated with immune checkpoint inhibitors: a multi-agent multi-site study*. 2019.

[PPT19a] Madhavi Patnana, Sapna Patel, and Anne Tsao. *Anti-PD-1 Immunotherapy Melanoma Dataset*. 2019.

[PPT19b] Madhavi Patnana, Sapna Patel, and Anne S Tsao. *Data from Anti-PD-1 Immunotherapy Lung*. 2019.

[Pee+18] Jan C Peeken et al. “Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients”. In: *Strahlenther. Onkol.* 194.6 (2018), pp. 580–590.

[Pen20] Lihong Peng. “Peripheral blood markers predictive of outcome and immune-related adverse events in advanced non-small cell lung cancer treated with PD-1 inhibitors”. In: *Cancer Immunol. Immunother.* (2020), p. 10.

[Pen+18] Luke Peng et al. “Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics”. In: *International Journal of Radiation Oncology*Biology*Physics* 102.4 (Nov. 2018), pp. 1236–1243.

[Pet+19] Fausto Petrelli et al. “Combination of radiotherapy and immunotherapy for brain metastases: A systematic review and meta-analysis”. en. In: *Crit. Rev. Oncol. Hematol.* 144 (Dec. 2019), p. 102830.

[Pet+] Mila P Petrova et al. “Neutrophil to lymphocyte ratio as a potential predictive marker for treatment with pembrolizumab as a second line treatment in patients with non-small cell lung cancer”. In: *Biosci. Trends* (), p. 8.

[PCW19] Sean P Pitroda, Steven J Chmura, and Ralph R Weichselbaum. “Integration of radiotherapy and immunotherapy for treatment of oligometastases”. en. In: *Lancet Oncol.* 20.8 (Aug. 2019), e434–e442.

[Pla+18] D Planchard et al. “Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up”. en. In: *Ann. Oncol.* 29.Suppl 4 (Oct. 2018), pp. iv192–iv237.

[Pol+20] Giulia Polverari et al. “18F-FDG Pet Parameters and Radiomics Features Analysis in Advanced Nsclc Treated with Immunotherapy as Predictors of Therapy Response and Survival”. en. In: *Cancers* 12.5 (May 2020).

[Pop+05] Whitney B Pope et al. “MR imaging correlates of survival in patients with high-grade gliomas”. In: *AJNR Am. J. Neuroradiol.* 26.10 (2005), pp. 2466–2474.

[PT17] Zoran B Popovic and James D Thomas. “Assessing observer variability: A user’s guide”. In: *Cardiovascular Diagnosis and Therapy* 7.3 (2017), pp. 317–324.

[Pra+16] Prateek Prasanna et al. “Radiomic features from the peritumoral brain parenchyma on treatment-naıve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings”. en. In: *Eur. Radiol.* (Oct. 2016).

[Qia+19] Zenghui Qian et al. “Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers”. en. In: *Cancer Lett.* 451 (June 2019), pp. 128–135.

[Rah+19] Paul Rahul et al. “Explaining Deep Features Using Radiologist-Defined Semantic Features and Traditional Quantitative Features”. In: *Tomography (Ann Arbor, Mich.)* 5.1 (2019), pp. 192–200.

[Ram+20] Santiago Ramón y Cajal et al. “Clinical implications of intratumor heterogeneity: challenges and opportunities”. In: *J. Mol. Med.* 98.2 (Feb. 2020), pp. 161–177.

[Ran+18] Aman Rana et al. *Computational Histological Staining and Destaining of Prostate Core Biopsy RGB Images with Generative Adversarial Neural Networks*. 2018.

[Rec+20] Michael P Recht et al. “Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations”. In: *Eur. Radiol.* 30.6 (2020), pp. 3576–3584.

[Rei06] Jerome M Reich. “Neoplasia in the etiology of sarcoidosis”. In: *European Journal of Internal Medicine* 17.2 (2006), pp. 81–87.

[Rid+16] Carole A Ridge et al. “Differentiating between subsolid and solid pulmonary nodules at CT: Inter- and intraobserver agreement between experienced thoracic radiologists”. In: *Radiology* 278.3 (2016), pp. 888–896.

[Rio+17] Emmanuel Rios Velazquez et al. “Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer”. en. In: *Cancer Res.* 77.14 (July 2017), pp. 3922–3930.

[Riv+20] Samantha Cruz Rivera et al. *Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension*. 2020.

[Riz+15] Naiyer A Rizvi et al. “Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer”. en. In: *Science* 348.6230 (Apr. 2015), pp. 124–128.

[Riz+16] Stefania Rizzo et al. “CT Radiogenomic Characterization of EGFR, K-RAS, and ALK Mutations in Non-Small Cell Lung Cancer”. en. In: *Eur. Radiol.* 26.1 (Jan. 2016), pp. 32–42.

[Riz+18] Stefania Rizzo et al. “Radiomics: the facts and the challenges of image analysis”. In: *European Radiology Experimental* 2.1 (2018).

[Roc+19] Danilo Rocco et al. “The role of combination chemo-immunotherapy in advanced non-small cell lung cancer”. en. In: *Expert Rev. Anticancer Ther.* 19.7 (July 2019), pp. 561–568.

[Rog+20] William Rogers et al. “Radiomics: from qualitative to quantitative imaging”. In: *Br. J. Radiol.* 93.1108 (Apr. 2020), p. 20190948.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. en. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer, Cham, Oct. 2015, pp. 234–241.

[Ros+19] Javier Ros et al. *Review of immunogenomics and the role of tumor mutational burden as a biomarker for immunotherapy response*. 2019.

[Ros+15] Andrew B Rosenkrantz et al. “Clinical utility of quantitative imaging”. en. In: *Acad. Radiol.* 22.1 (Jan. 2015), pp. 33–49.

[Ros+18] Samuel Rosner et al. “Peripheral blood clinical laboratory variables associated with outcomes following combination nivolumab and ipilimumab immunotherapy in melanoma”. In: *Cancer Med.* 7.3 (2018), pp. 690–697.

[Rot+15] Holger Roth et al. *A new 2.5 D representation for lymph node detection in CT*. 2015.

[Rul+19] Eliana Rulli et al. *The impact of targeted therapies and immunotherapy in melanoma brain metastases: A systematic review and meta-analysis*. 2019.

[Run+19] Francesco Rundo et al. “Advanced deep learning embedded motion radiomics pipeline for predicting anti-PD-1/PD-L1 immunotherapy response in the treatment of bladder cancer: Preliminary results”. In: *Electronics (Switzerland)* 8.10 (2019).

[RN03] Stuart Jonathan Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. en. Prentice Hall/Pearson Education, 2003.

[RK09] Aaron M Rutman and Michael D Kuo. "Radiogenomics: Creating a link between molecular diagnostics and diagnostic imaging". In: *Eur. J. Radiol.* 70.2 (2009), pp. 232–241.

[Sab+19] Nabil F Saba et al. "Nivolumab versus investigator's choice in patients with recurrent or metastatic squamous cell carcinoma of the head and neck: Efficacy and safety in CheckMate 141 by age". en. In: *Oral Oncol.* 96 (Sept. 2019), pp. 7–14.

[Sal+18] Joel Saltz et al. "Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images". en. In: *Cell Rep.* 23.1 (Apr. 2018), 181–193.e7.

[SC18] Christine R Sanderson and David C Currow. "Palliative care meets immunotherapy: what happens as cancer paradigms change?" en. In: *BMJ Support. Palliat. Care* 8.4 (Dec. 2018), pp. 431–432.

[Sca+19] Lisa Scarpace et al. *Data From REMBRANDT*. 2019.

[SP19] Kathleen Schmainda and Melissa Prah. *Data from Brain-Tumor-Progression*. 2019.

[SS16] Ton N Schumacher and Wouter Scheper. "A liquid biopsy for cancer immunotherapy". en. In: *Nat. Med.* 22.4 (Apr. 2016), pp. 340–341.

[Sch+16] Lawrence H Schwartz et al. "RECIST 1.1—Update and clarification: From the RECIST committee". In: *Eur. J. Cancer* 62 (July 2016), pp. 132–137.

[Seg+07] Eran Segal et al. "Decoding global gene expression programs in liver cancer by noninvasive imaging". In: *Nature biotechnology* 25.6 (2007), pp. 675–680.

[Sev+17] Eva M Sevick-Muraca et al. "Moonshot acceleration factor: medical imaging". In: *Cancer Research* 77.21 (2017), pp. 5717–5720.

[Sey+17a] Lesley Seymour et al. *iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics*. 2017.

[Sey+17b] Lesley Seymour et al. *iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics*. 2017.

[Sha+19] Lingdao Sha 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. In: *J. Pathol. Inform.* 10 (July 2019), p. 24.

[Sha+20] Xue Sha et al. "Discrimination of mediastinal metastatic lymph nodes in NSCLC based on radiomic features in different phases of CT imaging". en. In: *BMC Med. Imaging* 20.1 (Feb. 2020), p. 12.

[Sha+16] Nameeta Shah et al. *Data from Ivy GAP*. 2016.

[Sha+13] Denis P Shamonin et al. "Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease". en. In: *Front. Neuroinform.* 7 (2013), p. 50.

[Shi+15] Brianh Shirts et al. "Clinical laboratory analytics: Challenges and promise for an emerging discipline". In: *J. Pathol. Inform.* 6.1 (2015), p. 9.

[Sid+17] Parveen Sidhu et al. "Radiological manifestations of immune-related adverse effects observed in patients with melanoma undergoing immunotherapy". In: *Journal of Medical Imaging and Radiation Oncology* 61.6 (2017), pp. 759–766.

[SVZ13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[Sir+16] Korsuk Sirinukunwattana et al. “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images”. en. In: *IEEE Trans. Med. Imaging* 35.5 (May 2016), pp. 1196–1206.

[Siu+17] L L Siu et al. “Challenges and opportunities in adapting clinical trial design for immunotherapies”. In: (2017).

[Sle+90] M L Slevin et al. “Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses, and general public”. en. In: *BMJ* 300.6737 (June 1990), pp. 1458–1460.

[Smi+20] Annabel Smith et al. “Duration of immunotherapy - should we continue ad infinitum?” en. In: *Intern. Med. J.* 50.7 (July 2020), pp. 865–868.

[Smi15] Clark K Smith K. *Data From CT_COLONOGRAPHY*. 2015.

[SMH19] Alexandra Snyder, Michael P Morrissey, and Matthew D Hellmann. “Use of Circulating Tumor DNA for Cancer Immunotherapy”. en. In: *Clin. Cancer Res.* 25.23 (Dec. 2019), pp. 6909–6915.

[Soy+18] Aixa E Soyano et al. “Peripheral blood biomarkers correlate with outcomes in advanced non-small cell lung Cancer patients treated with anti-PD-1 antibodies”. In: *Journal for ImmunoTherapy of Cancer* 6.1 (2018), p. 129.

[SDL16] Lavinia Spain, Stefan Diem, and James Larkin. “Management of toxicities of immune checkpoint inhibitors”. In: *Cancer treatment reviews* 44 (2016), pp. 51–60.

[SM19] Mary H Stanfill and David T Marc. “Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management”. en. In: *Yearb. Med. Inform.* 28.1 (Aug. 2019), pp. 56–64.

[Ste+21] Stefano Trebeschi, Thi Dan Linh Nguyen-Kim et al. "AI-driven identification of prognostic response patterns to immunotherapy of melanoma brain metastases". In: *Submitted for publication.* (2021).

[Ste+20] Julie E Stein et al. *Pan-Tumor Pathologic Scoring of Response to PD-(L)1 Blockade*. 2020.

[Sub+05] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (Oct. 2005), pp. 15545–15550.

[Sun+20a] Kai-Yu Sun et al. "CT-based radiomics scores predict response to neoadjuvant chemotherapy and survival in patients with gastric cancer". en. In: *BMC Cancer* 20.1 (May 2020), p. 468.

[Sun+18] Roger Sun 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". In: *Lancet Oncol.* 19.9 (2018), pp. 1180–1191.

[Sun+20b] Zongqiong Sun et al. *Radiomics study for predicting the expression of PD-L1 in non-small cell lung cancer based on CT images and clinicopathologic features*. 2020.

[Suo+16] Kathleen C Suozzi et al. "Immune-related sarcoidosis observed in combination ipilimumab and nivolumab therapy". In: *JAAD case reports* 2.3 (2016), pp. 264–268.

[Tan+19] Xianzheng Tan et al. "Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma". en. In: *Eur. Radiol.* 29.1 (Jan. 2019), pp. 392–400.

[Tan+18] Chad Tang et al. "Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer". en. In: *Sci. Rep.* 8.1 (Jan. 2018), p. 1922.

[Tar+16] Francesca Tartari et al. “Economic sustainability of anti-PD-1 agents nivolumab and pembrolizumab in cancer patients: Recent insights and future challenges”. en. In: *Cancer Treat. Rev.* 48 (July 2016), pp. 20–24.

[Ten+18] Feifei Teng et al. “Progress and challenges of predictive biomarkers of anti PD-1/PD-L1 immunotherapy: A systematic review”. en. In: *Cancer Lett.* 414 (Feb. 2018), pp. 166–173.

[TM20] Rajat Thawani and Syed Atif Mustafa. “The future of radiomics in lung cancer”. In: *The Lancet Digital Health* 2.3 (2020), e103.

[The+00] Patrick Therasse et al. “New Guidelines to Evaluate the Response to Treatment in Solid Tumors”. en. In: *J. Natl. Cancer Inst.* 92.3 (Feb. 2000), pp. 205–216.

[TBS93] S N Thibodeau, G Bren, and D Schaid. “Microsatellite instability in cancer of the proximal colon”. en. In: *Science* 260.5109 (May 1993), pp. 816–819.

[Tie20] Jeanne Tie. “Tailoring immunotherapy with liquid biopsy”. In: *Nature Cancer* 1.9 (Sept. 2020), pp. 857–859.

[Tim+20] Janita E van Timmeren et al. “Radiomics in medical imaging-“how-to” guide and critical reflection”. en. In: *Insights Imaging* 11.1 (Aug. 2020), p. 91.

[Tir+15] Sree Harsha Tirumani et al. “Radiographic Profiling of Immune-Related Adverse Events in Advanced Melanoma Patients Treated with Ipilimumab”. en. In: *Cancer Immunol Res* 3.10 (Oct. 2015), pp. 1185–1192.

[Tre+20] Stefano Trebeschi et al. “Deep learning distinguishing pulmonary progression from pulmonary sarcoid-like lesions in immunotherapy-treated melanoma patients”. In: *British Journal of Cancer, accepted for publication* (2020).

[Tre+21a] Stefano Trebeschi et al. “Development of a prognostic AI-monitor for metastatic urothelial cancer patients receiving immunotherapy”. In: *Submitted for publication.* (2021).

[Tre+19] Stefano Trebeschi et al. “Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers”. In: *Ann. Oncol.* 30.6 (2019), pp. 998–1004.

[Tre+21b] Stefano Trebeschi et al. “Prognostic value of deep learning mediated treatment monitoring in lung cancer patients receiving immunotherapy”. In: *Frontiers in Oncology, accepted for publication* (2021).

[Tun+] Ilke Tunali et al. *Hypoxia-related radiomics predict immunotherapy response: A multi-cohort study of NSCLC*.

[Tun+19] Ilke Tunali et al. “Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report”. In: *Lung Cancer* 129.January (2019), pp. 75–79.

[TS16] Samra Turajlic and Charles Swanton. “Metastasis as an evolutionary process”. en. In: *Science* 352.6282 (Apr. 2016), pp. 169–175.

[Tur+16] Riku Turkki et al. “Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples”. en. In: *J. Pathol. Inform.* 7 (Sept. 2016), p. 38.

[Twy+15] Christina Twyman-Saint Victor et al. “Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer”. en. In: *Nature* 520.7547 (Apr. 2015), pp. 373–377.

[U Sa] U S Food And. *U.S. Food and Drug Administration. Drugs (KEYTRUDA Label).* `https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125514s014lbl.pdf`. Accessed: 2018-9-5.

[U Sb] U S Food And. *U.S. Food and Drug Administration. Drugs (OPDIVO Label)*. https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125554s055lbl.pdf. Accessed: 2018-9-5.

[Ume+20] Yoshie Umemura et al. *DCE-MRI perfusion predicts pseudoprogression in metastatic melanoma treated with immunotherapy*. 2020.

[Ung+16] Joseph M Unger et al. "The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies". en. In: *Am Soc Clin Oncol Educ Book* 35 (2016), pp. 185–198.

[US] US Food And Drug Administation. *Drug approval package. Bidil (Isosorbide Dinitrate and Hydralazine Hydrochloride) Tablets*. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2005/020727_s000_BidilTOC.cfm. Accessed: 2020-12-17.

[Vai+20] Pranjal Vaidya et al. "Novel, non-invasive imaging approach to identify patients with advanced non-small cell lung cancer at risk of hyperprogressive disease with immune checkpoint blockade". en. In: *J Immunother Cancer* 8.2 (Oct. 2020).

[Val+15] Martin Vallières et al. *A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities*. 2015.

[Val+17] Martin Vallières et al. *Data from Head-Neck-PET-CT*. 2017.

[Van+15] Sarah J Van Riel et al. "Observer variability for classification of pulmonary nodules on low-dose CT Images and its effect on nodule management1". In: *Radiology* 277.3 (2015), pp. 863–871.

[Vee+20] Harini Veeraraghavan et al. “Machine learning-based prediction of microsatellite instability and high tumor mutation burden from contrast-enhanced computed tomography in endometrial cancers”. en. In: *Sci. Rep.* 10.1 (Oct. 2020), p. 17769.

[Ver+18] Vivek Verma et al. “A systematic review of the cost and cost-effectiveness studies of immune checkpoint inhibitors”. en. In: *J Immunother Cancer* 6.1 (Nov. 2018), p. 128.

[VS13] Liza C Villaruz and Mark A Socinski. “The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement”. en. In: *Clin. Cancer Res.* 19.10 (May 2013), pp. 2629–2636.

[Vok+18] E E Vokes et al. “Nivolumab versus docetaxel in previously treated advanced non-small-cell lung cancer (CheckMate 017 and CheckMate 057): 3-year update and outcomes in patients with liver metastases”. en. In: *Ann. Oncol.* 29.4 (Apr. 2018), pp. 959–965.

[Voo+17] Khinh Ranh Voong et al. “Beyond PD-L1 testing-emerging biomarkers for immunotherapy in non-small cell lung cancer”. In: *Annals of translational medicine* 5.18 (2017).

[Vou20] Ioannis A Voutsadakis. “Prediction of Immune checkpoint inhibitors benefit from routinely measurable peripheral blood parameters”. In: *Chinese Clinical Oncology* 9.2 (2020), pp. 19–19.

[Wan+20] Liansheng Wang et al. *A novel approach combined transfer learning and deep learning to predict TMB from histology image*. 2020.

[WN11] Yi-Xiang J Wang and Chin K Ng. “The impact of quantitative imaging in medicine and surgery: Charting our course for the future.” In: *Quant. Imaging Med. Surg.* 1.1 (2011), pp. 1–3.

[Wan+19] Zhijie Wang et al. "Assessment of Blood Tumor Mutational Burden as a Potential Biomarker for Immunotherapy in Patients With Non-Small Cell Lung Cancer With Use of a Next-Generation Sequencing Cancer Gene Panel". en. In: *JAMA Oncol* 5.5 (May 2019), pp. 696–702.

[Web+15] Jeffrey S Weber et al. "Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial". en. In: *Lancet Oncol.* 16.4 (Apr. 2015), pp. 375–384.

[Wee+19] Leonard Wee et al. *Data from NSCLC-Radiomics-Interobserver1*. 2019.

[Wee+98] J C Weeks et al. "Relationship between cancer patients' predictions of prognosis and their treatment preferences". en. In: *JAMA* 279.21 (June 1998), pp. 1709–1714.

[Wet+02] Stephan G Wetzel et al. "Relative cerebral blood volume measurements in intracranial mass lesions: Interobserver and intraobserver reproducibility study". In: *Radiology* 224.3 (2002), pp. 797–803.

[Whi08] T L Whiteside. "The tumor microenvironment and its role in promoting tumor growth". en. In: *Oncogene* 27.45 (Oct. 2008), pp. 5904–5912.

[Wol+15] Jedd D Wolchok et al. "Efficacy and safety results from a phase III trial of nivolumab (NIVO) alone or combined with ipilimumab (IPI) versus IPI alone in treatment-naive patients (pts) with advanced melanoma (MEL) (CheckMate 067)". In: *J. Clin. Orthod.* 33.18_suppl (June 2015), LBA1–LBA1.

[Wol+09] Jedd D Wolchok et al. "Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria". en. In: *Clin. Cancer Res.* 15.23 (Dec. 2009), pp. 7412–7420.

[Woo19] Marcus Woo. “An AI boost for clinical trials”. en. In: *Nature* 573.7775 (Sept. 2019), S100–S102.

[Wu09] Anna M Wu. “Antibodies and antimatter: the resurgence of immuno-PET”. en. In: *J. Nucl. Med.* 50.1 (Jan. 2009), pp. 2–5.

[Wu+19] Wei Wu et al. “Comparison of prediction models with radiological semantic features and radiomics in lung cancer diagnosis of the pulmonary nodules: a case-control study”. en. In: *Eur. Radiol.* 29.11 (Nov. 2019), pp. 6100–6108.

[WH18] Yuxin Wu and Kaiming He. “Group Normalization: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII”. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Vol. 11217. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 3–19.

[Yan+18] Ke Yan et al. “DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning”. en. In: *J Med Imaging (Bellingham)* 5.3 (July 2018), p. 036501.

[Yin08] Jennifer G Dy Ying Cui. “Orthogonal Principal Feature Selection”. In: (2008).

[Yip+17a] Stephen S F Yip et al. *Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer*. Tech. rep. 2017.

[Yip+17b] Stephen SF Yip et al. “Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer”. In: *Scientific reports* 7.1 (2017), pp. 1–11.

[Yor+19] Afua A Yorke et al. *Pelvic Reference Data [Dataset]*. 2019.

[Yu+19] Y Yu et al. *Predictors of Early Response to Immunotherapy in Head and Neck Cancer: A Secondary Clinical and Radiomic Analysis of a Prospective Randomized Trial with Nivolumab*. 2019.

[ZF14] Matthew D Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 818–833.

[ZWZ17] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. *Self supervised deep representation learning for fine-grained body part recognition*. 2017.

[Zha+18] Zijian Zhang et al. "A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images". en. In: *Eur. Radiol.* 28.6 (June 2018), pp. 2255–2263.

[Zha+19a] Zizhao Zhang et al. "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning". In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 236–245.

[Zha+19b] Pengfei Zhao et al. "Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-PD-1/PD-L1 immunotherapy efficacy". en. In: *J. Hematol. Oncol.* 12.1 (May 2019), p. 54.

[Zha+20] Shengyu Zhao et al. "Unsupervised 3D End-to-End Medical Image Registration With Volume Tweening Network". en. In: *IEEE J Biomed Health Inform* 24.5 (May 2020), pp. 1394–1404.

[Zhe+20] Xueyi Zheng et al. *Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer*. 2020.

[Zho+20] Xiaoxiang Zhou et al. “Are immune-related adverse events associated with the efficacy of immune checkpoint inhibitors in patients with cancer? A systematic review and meta-analysis”. en. In: *BMC Med.* 18.1 (Apr. 2020), p. 87.

[Zit+17] Federica Zito Marino et al. “Are tumor-infiltrating lymphocytes protagonists or background actors in patient selection for cancer immunotherapy?” en. In: *Expert Opin. Biol. Ther.* 17.6 (June 2017), pp. 735–746.

[Zuh+21] Zuhir Bodalal, Stefano Trebeschi et al. “The future of artificial intelligence applied to immunotherapy trials”. In: *Neoadjuvant Immunotherapy Treatment of Localized Genitourinary Cancers: Multidisciplinary Management*. Book chapter accepted for publication. Springer, 2021.

Valorisation

Relevance

Cancer immunotherapy is becoming a standard treatment for advanced-stage cancer patients. While results from numerous clinical trials show significantly higher response and survival rate compared to the standard chemoradiotherapy, there is still a number of patients who do not benefit from starting (or continuing) the treatment. In the meanwhile, patients can experience treatment side-effects, normally manifested as auto-immune and inflammatory disorders, such as dermatitis, pneumonitis, colitis, hepatitis, and granulomatous disease [Zho+20]. While the impact of treatment side-effects on the response is still under investigation, with meta-analyses pointing to them as positive prognostic factors [Zho+20], it is well known that the cost of immune checkpoint inhibitors can pose a significant strain on hospital resources, and public resources in general, especially in Europe where access to healthcare services are guaranteed, regulated and subsidised by the government. Immunotherapy is known to cost up to ten times more than chemotherapeutic options. The average immunotherapeutic treatment costs in the order of the hundreds of thousands of euros, compared to the tens of thousands needed for traditional chemotherapy [Ver+18]. In patients that are not likely to derive any clinical benefit from immunotherapy, these resources could be reallocated for other therapeutic options. In this thesis, we propose the use of artificial intelligence on already-available routine clinical imaging. By enhancing the analytic process with AI, we reach an accurate, and cost-effective marker that, by virtue of being trained on routine imaging, can scale well from academic centers to peripheral hospitals.

Target population

This thesis has a wide target population. It includes **radiologists** who are already using medical imaging as a fundamental tool in their diagnostic process, and who will likely be in charge of the new technologies that will be developed out of this thesis and in the field of artificial intelligence in general. Some of the techniques described and proposed here, such radiomics analysis of cancer lesions (Chapters 2 and 3), or whole-body prognostic monitoring (Chapters 4, 5, and 6), offer a new way to analyze radiological scans. Radiologists will be at the forefront of the implementation (Chapter 7 and 8). Their contribution will be vital in developing the interaction of these technologies within the clinical team.

Radiologists will also play a role when these technologies will enter the **multidisciplinary tumor board**. As mentioned in Chapters 7 and 9, the envisioned usage of the AI methods developed within this thesis reflects this point. We envision our system to be an interactive window into the status of the patient, their potential risks for survival, and suggestions for most likely response to different treatment options. Treatment planning will be defined according to this data and the input of the clinicians in the board. Members of the tumour board will be essential in the study of the interaction between AI and clinicians, and the development of a coordinated approach for its development and implementation.

In this, **tech companies**, specifically the ones that are involved in the development of AI-based healthcare solutions will also play a role. These entities possess the know-how of the actual implementation process, which should cover multiple practical aspects, from the hardware requirement to the steps for approval from the healthcare regulators (e.g. the European Medicines Agency [Coh+20]).

The methodology presented in this thesis targets mostly **cancer patients with advanced disease**, namely spread outside of the original location and throughout the body. In the majority of the cases, the

main clinical intent in these patients is palliative, in other words, to improve quality of life more than survival. Even in immunotherapy, where we observed higher survival rates and even some complete responses, palliative intent remains the primary focus [SC18]. AI-based technologies, such as the one presented, aside from suggesting the most effective treatment which would in turn improve response and overall survival, could also help pin-point localized conditions of risk, which could impair quality of life, and give the clinician indications to address the issue. In summary, the overall expected outcome from employing AI-based evidence decision-making is improved survival and improved quality of life.

To the **scientific community**, we provide evidence not only of the efficacy of AI-mediated pipelines, but also the success of multidisciplinarity in research, with all the work published in this thesis relying on the joint, coordinated work of experts from different fields.

Innovation and future

The results of this thesis show how prognostic and predictive factors can be found by artificial intelligence-based image analysis on routine radiological scans of cancer patients receiving immunotherapy. We filled the knowledge gap, and demonstrated a link between imaging-derived morphological features to its biological profile and treatment outcome in immunotherapy, being one of the first investigations to do so. We further identified the limitation of the classical image analysis pipeline that we employed in the initial study, which involves a radiologist identifying a region of interest to analyse, and improved it to a new analytic pipeline, termed prognostic monitoring, which does not require manual input, which is not limited to static baseline imaging, and which utilises the entire whole-body scan. This represents a shift from the image analysis pipelines that have been proposed so far, not for its full automatic aspect, but rather because we shift the identification of factors to the AI, removing any bias that may come from

human analytics. In other words, we are not developing a pipeline which automates the manual work already performed in the clinics, but rather developing something that was not even present in the clinics, i.e. tracking of all morphological changes in the body. We are not employing AI to reduce the information included in a whole-body scan to a limited set of factors easily understandable by humans, such as change in tumor size, but we rather let the AI-algorithm run independently to determine what are the factors that correlate with worse prognosis. We are not training an AI-model to imitate us, or think like us (i.e. smaller tumor, better outcomes), rather we let the AI-algorithm draw its own conclusions from the data.

We envision more studies following this idea, as it would produce novel solutions, which are complementary to, and not disruptive of, the current clinical world. The AI-algorithms developed in this study are still prototypes. Future propostive will be to extend them to a fully functional medical device, tested on pan-cancer and pan-treatment clinical trials, and approved by the appropriate regulatory agencies.

Summary

Chapter 2. Lesion response prediction to immunotherapy.

In this chapter, we aimed to link imaging-derived, *radiomics* features with outcome, lesion-wise. Our findings suggest associations between radiomics features and immunotherapy response. Lesions that are more likely to respond to immunotherapy typically present with more heterogeneous morphological profiles with non-uniform density patterns and compact borders. Moreover, a machine learning model is provided that could be used within the context of lesion response to treatment, patient treatment response, and response pattern characterization.

Chapter 3. Lesion diagnosis to therapy-induced lung disease.

Aim of this chapter was to apply artificial intelligence analytic pipelines on routine medical imaging for the diagnosis of sarcoid-like granulomatous lesions induced by novel cancer immunotherapeutic agents. We found significant performance in the diagnosis of sarcoid-like granulomatous lesions, while simultaneously significantly improving the performance of the original screening network for the diagnosis of pulmonary metastases. Moreover, the network was able to distinguish between sarcoid-like granulomatous lesions and non-specific post-infection granulomas. Diagnostic signatures were also found to possess prognostic relevance.

Chapter 4. Prognostic value of chest-imaging monitoring.

In this chapter, we aimed to investigate the potential prognostic value of AI-mediated monitoring in NSCLC patients receiving PD-1 blockade. We hypothesized the existence of quantitative imaging features describing a set of gross morphological changes happening during treatment that hold prognostic information. Based on image-to-image registration, we develop a deep learning algorithm for the detection

of changes between serial imaging of the same patient. Our results demonstrate the existence of such factors (as described by the AI on imaging), that are tumor-related, such as nodal, lung and bone lesions, as well as non-tumor related, such as pleural effusions, atelectasis and non-specific consolidations.

Chapter 5. Whole-body imaging-based prognostic monitoring.

In this chapter, we investigated the prognostic information of AI-derived whole-body imaging monitoring markers in advanced urothelial cancer receiving checkpoint inhibitors. We hypothesised that quantitative AI-derived features describing morphological changes happening during the course of treatment could hold prognostic information. To this end, we designed and implemented a prognostic AI-monitor (PAM), based on the prototype design of Chapter 4, and extended to handle heterogeneous datasets and abdominal imaging. Our findings demonstrate that PAM is complementary to existing monitoring methods, while reaching comparable or superior accuracy. We argue that this could be the result of PAM's ability to analyze the whole body, including non-target cancer lesions and non-cancer lesions.

Chapter 6. Prognostic response patterns in brain imaging.

In this chapter, we present an expansion of the PAM analytic pipeline to brain imaging of BM patients receiving immunotherapy. Our results demonstrate that PAM can be extended to imaging modalities beyond CT, and be used to capture prognostic response patterns that are unique and complementary to a wide range of different brain-specific markers, currently used in the clinics.

Chapter 7. The future of artificial intelligence immunotherapy trials.

Clinical trials serve as a barrier of entry for new interventions and treatments prior to implementation in routine clinical practice. At its essence, the primary role of a clinical trial is to monitor a patient longitudinally using the diagnostic disciplines (radiology, pathology, and laboratory medicine) to assess clinical outcomes. As the diagnostic

fields have begun to fully digitalise, large volumes of data are being generated per patient - creating a ripe environment for the implementation of AI. In this chapter, we will explore how AI has been applied in each of these diagnostic disciplines and discuss how this may influence clinical trials in the future.

Chapter 8. Towards integrated healthcare.

Medical imaging is a vital part of the clinical decision making process, especially in an oncological setting. Radiology has experienced a great wave of change and the advent of quantitative imaging has provided a unique opportunity to analyze patient images objectively. Leveraging AI, there is increased potential for synergy between physicians and computer networks – via computer aided diagnosis (CAD), computer aided prediction of response (CARP), and computer aided biological profiling (CABP). The ongoing digitalization of other specialties further opens the door for even greater multidisciplinary integration. In this chapter, we envision the development of an integrated system composed of an aggregation of sub-systems interoperating with the aim of achieving an overarching functionality (in this case better CAD, CARP, and CABP). This will require close multidisciplinary cooperation between the clinicians, biomedical scientists, and (bio)engineers as well as an administrative framework where the departments will operate not in isolation but in successful harmony.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my protomors, **Prof. Regina Beets-Tan** and **Prof. Hugo Aerts**, for their continuous support of my PhD research, for their time, patience, motivation, and for all knowledge I have acquired from them in the last five years. Their guidance helped me in the time of research and writing of this thesis, as well as in the planning and development of my own research line, which I will continue pursuing after my PhD. Their mentorship helped me to become an independent scientist.

Besides my protomots, I would like to thank the rest of my thesis committee. This includes the members of the assessment committee, Prof. Dirk de Ruysscher, **Prof. Paul Hofman**, **Prof. Paul Baas**, **Prof. Klaus Maier-Hein** and **Dr. Peters**, as well as the PhD assessment committee from the Oncology Graduate School of Amsterdam, **Prof. Uulke van der Heide**, **Prof. Koen Hartemink**, and **Dr. Wouter Vogel**. Their insightful comments and availability during my PhD trajectory allowed me to strengthen my thesis, and my research in general.

My sincere thanks also goes to the oncologists we have had the pleasure and honor to work with in the last five years, **Prof. John Haanen**, **Prof. Christian Blank**, **Prof. Egbert Smit**, and **Prof. Michiel van der Heijden**, and their collaborators and lab members, and our own clinicians within the department of radiology, the radiologists, and the technicians, for their immense help, support, and guidance.

I would also like to extend a word of gratitude to our Swiss and American collaborators and co-workers from the University Hospital of Zurich and the Harvard Medical School, with whom we have instantiated a solid and fruitful collaboration, which I'm sure will be strengthened in the years to come.

I thank my fellow **lab mates and colleagues** from the department of Radiology, and the immunotherapy imaging team, who surrounded

me in the last half decade. Thank you for your company, advice, help and support, making every day at the office a pleasant experience. I could have not asked for better colleagues. A special thank to my direct colleagues of the immunotherapy team **Dan Linh**, **Zuhir**, and **Teresa** for the time taken in answering all my questions about radiology, immunology, oncology, and biology, for the brainstorming on our revolutionary ideas, the late hours working, and the laughter. I could have not made it without you.

Last but not least, I would like to thank **my friends** here in Amsterdam, who made my stay so far one of the best experiences of my life, and who enabled me to discover myself, and to grow. I would like to extend my gratitude to friends, teachers, and co-workers I left in Germany and Italy. I would like to thank **my family**, especially my sister, and their support during my childhood and (rebellious) adolescence.

To all the amazing, diverse, and crazy people in my life, thank you.

Published work

Stefano Trebeschi et al. "Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers". In: *Ann. Oncol.* 30.6 (2019), pp. 998–1004

Stefano Trebeschi et al. "Deep learning distinguishing pulmonary progression from pulmonary sarcoid-like lesions in immunotherapy-treated melanoma patients". In: *British Journal of Cancer, accepted for publication* (2020)

Stefano Trebeschi et al. "Prognostic value of deep learning mediated treatment monitoring in lung cancer patients receiving immunotherapy". In: *Frontiers in Oncology, accepted for publication* (2021)

Stefano Trebeschi et al. "Development of a prognostic AI-monitor for metastatic urothelial cancer patients receiving immunotherapy". In: *Submitted for publication.* (2021)

Stefano Trebeschi, Thi Dan Linh Nguyen-Kim et al. "AI-driven identification of prognostic response patterns to immunotherapy of melanoma brain metastases". In: *Submitted for publication.* (2021)

Zuhir Bodalal, Stefano Trebeschi et al. "The future of artificial intelligence applied to immunotherapy trials". In: *Neoadjuvant Immunotherapy Treatment of Localized Genitourinary Cancers: Multidisciplinary Management*. Book chapter accepted for publication. Springer, 2021

Zuhir Bodalal, Stefano Trebeschi, and Regina Beets-Tan. "Radiomics: a critical step towards integrated healthcare". In: *Insights into imaging* 9.6 (2018), pp. 911–914

About the author

Stefano Trebeschi was born on the 20th of June, 1991, in Desenzano del Garda, Italy. He grew up in Cavriana, Italy. At the age of 19, he moved to South Tyrol, where he started his bachelor degree in computer science and engineering at the Free University of Bolzano-Bozen. During this period, he got involved in the research on software engireering, within the Department of Computer Science. He graduated in 2013, supervised by Dr. Ilenia Fronza and Dr. Andrea Janes, with a thesis on data analytics for software development monitoring.

In 2013 he moved to Munich, Germany, to start his master in computer science at the Technical University of Munich. As part of his master program, he got involved in a series of projects in medical imaging, and machine learning, both within the Department of Computer Science, as well as the Department of Neurology at the Klinikum Rechts der Isar. Here he graduated in 2016, under the supervision of Dr. Alexander Valentinitsch, Prof. dr. Jan Kirschke (geb. Bauer), and Prof. dr. Bjoern Menze, for the identification of osteoporosis patients at risk of vertebral fracture. He moved to Amsterdam, the Netherlands, on the same year to start his PhD at the Netherlands Cancer Institute, under the supervision of Prof. dr. Regina Beets-Tan and Prof. Hugo Aerts, from the Harvard Medical School.

During his PhD, Stefano has focused his research line in imaging of cancer patients receving immunotherapy, on which he has authored six papers, a book chapter, and talks at the European Society of Radiology (ESR), European Hematology Association (EHA), and European Society of Medical Imaging Informatics (EUSOMII). In these five years, he has supervised internships and theses from master students of the University of Twente, among other national and international coollaborations.