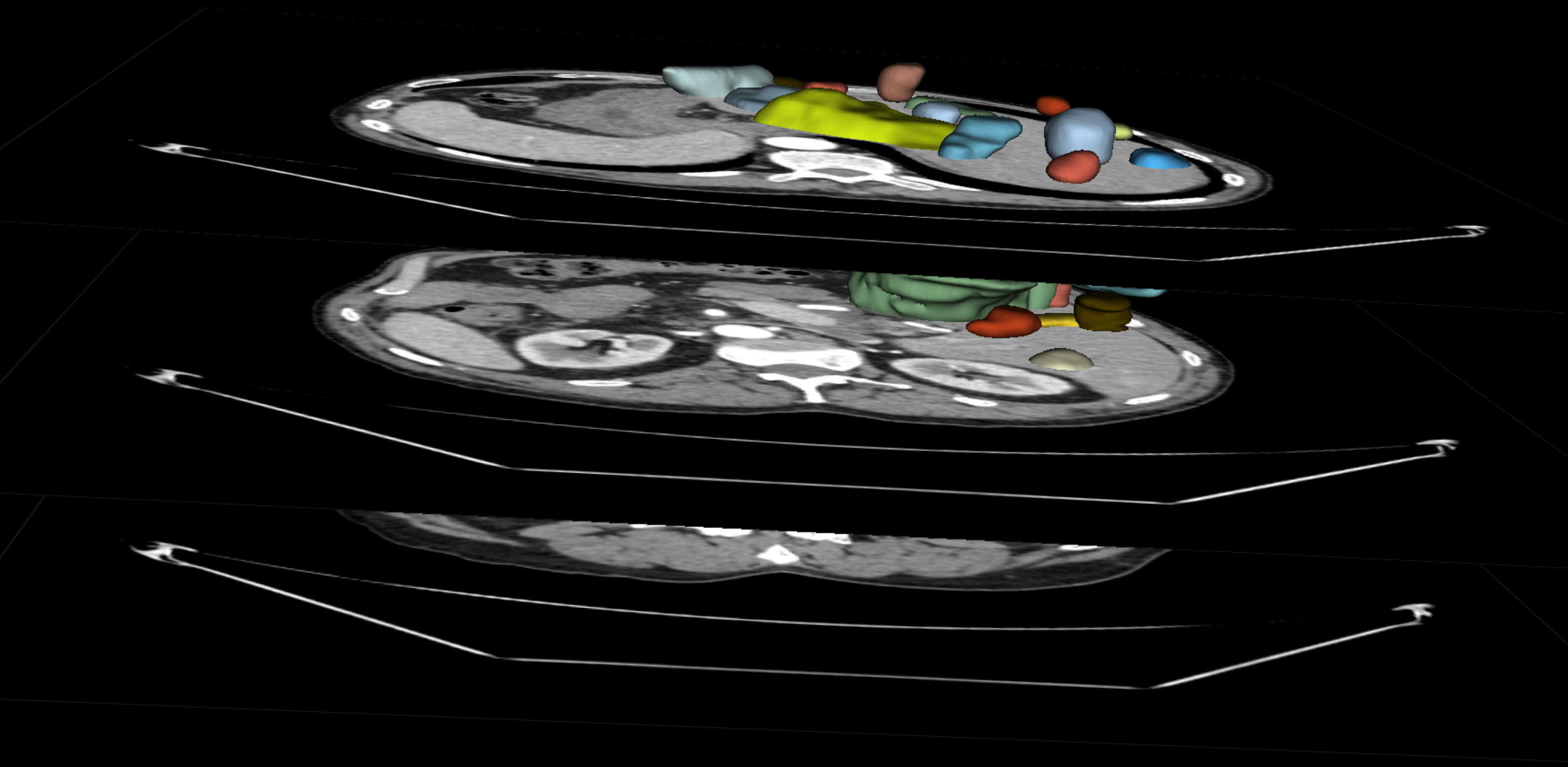
Prediction of Response to Immune Checkpoint Inhibitors in Solid Tumours using CT-based Biomarkers

Biomedical Engineering - Medical Physics

B.E.J. Gielen



TU Delft

Prediction of Response to Immune Checkpoint Inhibitors in Solid Tumours using CT-based Biomarkers

Biomedical Engineering - Medical Physics

Thesis report

by

B.E.J. Gielen

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on June 22, 2023 at 14:00

Thesis committee:

Chair:	F.M. Vos
Supervisors:	M. Ligero Hernandez
	R. Perez-Lopez
External examiner:	J. Hernandez Tamames
Place:	Faculty of Mechanical, Maritime and Materials Engineering, Delft
Project Duration:	September, 2022 - June, 2023
Student number:	5182794

Vall d’Hebron Institute of Oncology, Barcelona · Delft University of Technology, Delft

Prediction of Response to Immune Checkpoint Inhibitors in Solid Tumours using CT-based Biomarkers

Bente Gielen[1,2]

Abstract
Immune checkpoint inhibitors (ICIs) have revolutionized cancer treatment by harnessing the immune system's ability to target cancer cells. However, only a subset of patients clinically benefit from ICIs, highlighting the need for predictive biomarkers. Currently, FDA-approved biomarkers such as PD-L1 expression and mismatch repair deficiency have limited efficacy and require invasive tumour biopsies. An alternative approach involves the use of radiomics, which leverages quantitative analysis of medical images to extract a large number of imaging features. Unlike biopsies, radiomics analysis is non-invasive and provides insights into tumour heterogeneity at a whole-tumour level. In this study, we aimed to predict clinical benefit in patients treated with ICI therapy using radiomic features extracted from baseline Computed Tomography (CT) images. We analysed a data set of 447 patients with 13 different primary tumour types. Five aggregation methods were employed to combine features from lesion level to patient level. The so-called radiomics standard pipeline, LASSO and logistic regression was used for feature selection and classification. Additionally, we explored the impact of primary tumour location and developed tumour-specific models. The best performance was achieved in the case of bladder cancer ($n = 53$, AUC: 0.717) when using the *all lesions* per patient for feature aggregation, using the *largest lesion* as feature aggregation yielded better results for the rest of the analysed cohorts: the whole cohort ($n = 447$, AUC: 0.634), thoracic cancer ($n = 108$, AUC: 0.741), skin cancer ($n = 79$, AUC: 0.766), and lower gastrointestinal cancer ($n = 64$, AUC: 0.794). Interestingly, better results were obtained when using tumour-specific models. These results may indicate the importance of distinguishing between different tumour types when predicting response to ICIs. In order to enhance the accuracy of predicting responses to ICIs, future research should focus on investigating tumour-specific strategies, examining the potential benefits of incorporating additional clinical, genomics, and immunohistochemistry data and the use of deep learning techniques in larger, more representative cohorts.

Keywords
Immune Checkpoint Inhibitors — Computed Tomography — Radiomics – Machine Learning

[1] *Master Biomedical Engineering, Delft University of Technology, Delft, The Netherlands*
[2] *Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain*

1. Introduction

Immune checkpoint inhibitors (ICIs) exploit the natural ability of the immune system to eliminate cancer cells [1]. During tumourigenesis, the immune system detects and eliminates malignant cells with T-lymphocytes (T-cells) and natural-killer (NK) cells. Tumour cells express neoantigens that are recognised by the immune system and presented to T-cells, which then eliminate malignant cells. This system is visualized in Fig. 1a). However, the immune system has an extra control mechanism to prevent autoimmune responses: immune checkpoints [2, 3]. Immune checkpoints are naturally present brakes that inhibit T-cell activity. This is crucial under physiologically normal conditions. PD-1 checkpoints and PD-L1 ligands block the T-cells so that they cannot induce cell apoptosis, see Fig. 1b). The expression of immune checkpoint ligand on the tumour membrane, such as PD-L1, is key for immune resistance and tumour progression. The blockade of immune checkpoints on the tumours cells using ICIs attempts to prevent immune resistance (Fig. 1c).

Six ICIs have been approved by the FDA and became standard of care in clinical practice for some tumour types such as melanoma and lung cancer [4, 5, 6]. However, only a low percentage, between 15 % and 25 %, of patients respond adequately to treatment [7, 8]. Therefore, there is a need for biomarkers that can stratify patients between responders and non-responders. In various research fields, biomarkers that can accurately predict patients' response to ICI treatment are explored. These biomarkers include factors like tumour infiltrating lymphocytes (TILs), microsatellite instability (MSI), PD-L1 expression, as well as clinical variables such as lactate dehydrogenase (LDH) and liver involvement [9, 10, 11]. However, there remains an important gap in effectively char-

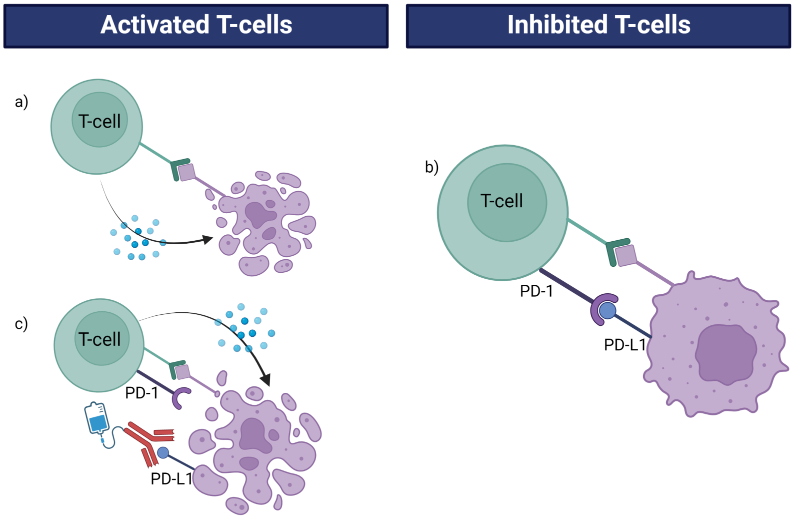


Figure 1. (a) When the T-cell receptor is able to bind to the antigen on the cancer cells, the T-cell will be activated and secretes granzymes and perforines which induce cell death. (b) The T-cell can be inhibited when the PD-1 receptor binds to the PD-L1 receptor of the cancer cell. This way the T-cell will not be able to induce cell death. (c) Immune checkpoint inhibitors can address this challenge by binding to PD-L1 or PD-1: in this case PD-L1. This leads to T-cell activation and subsequently, cell death.

acterizing patients' responsiveness at baseline.

Currently, the Food and Drug Administration (FDA) has approved two biomarkers: PD-L1 expression and mismatch repair deficiency. However, relying solely on these two biomarkers does not provide sufficient clinical predictive efficacy. In addition, both biomarkers require a tumour sample obtained through biopsy. Biopsies are invasive, not always easily obtainable, and do not represent the heterogeneity of the whole disease [12]. Using Computed Tomography (CT) based biomarkers has potential to overcome these limitations.

A promising approach for developing CT-based biomarkers is the application of radiomics. Radiomics is a quantitative approach to image analysis. It relies on extraction of large numbers of shape, edge, and texture features from medical images. The underlying hypothesis of radiomics is that disease specific pixel patterns can be identified that may not be detectable by an expert's eye, as such we might extract valuable information from medical images. The use of radiomics could therefore contribute to developing personalised therapy [13, 14, 15, 16]. Radiomics analysis uses already collected imaging data from the everyday clinical practice and, therefore it does not result in extra work load. Radiomics analysis offers a non-invasive approach to understanding the heterogeneity of a disease at both the whole-tumor and disease burden level. However, it comes with the potential drawback of having multiple sources of information from the same patient. Aggregating features from a lesion level to perform predictions at a patient level is particularly challenging, and no consensus method is defined [17].

In general, we can divide the radiomics workflow in four different steps. These steps involve (i) image acquisition, processing and segmentation, (ii) feature extraction, (iii) feature selection to identify the most predictive features, and (iv) model building, often entailing a classification problem. In my literature review that I previously performed, I observed that a range of machine learning (ML) algorithms have been employed for feature selection and classification tasks in the context of ICI treatment.

The objective of this project was to predict response to ICIs using CT scans. To achieve this, various aggregation techniques were employed to aggregate the data obtained from the lesion level to the patient level. Additionally, we tested two workflows to find the most optimal feature selector and classification model. One is the most used workflow found in the literature, defined as the radiomics standard pipeline. The other is the the recently published Workflow for Optimal Radiomics Classification (WORC). WORC provides cutting-edge pipelines for automatic optimisation of the radiomics workflow, testing several feature selectors and classifiers [18].

2. Methods

2.1 End-point

The target end-point in this study is clinical benefit. Clinical benefit is either defined as complete response (CR), partial response (PR) or stable disease (SD) after 5 months of treatment, according to the RECIST 1.1 guidelines [19].

2.2 Data set

The data set includes 447 patients who received treatment with ICIs. The patients were given either a single ICI therapy or a combination of multiple ICIs. The patients analysed in this study had advanced cancer including different primary tumour types. (e.g. thoracic, skin, lower gastrointestinal, and bladder cancers).

2.3 Imaging preprocessing and feature extraction

Radiomic features were extracted from the baseline contrast-enhanced CT images, which were acquired using scanners from Siemens, Philips, or GE Healthcare. All images had slice thickness $\leq$ 5 mm and were reconstructed using soft or standard convolution kernels. An experienced radiologist delineated lesions on baseline scans using 3D Slicer (version 4.11.20210226.). From the segmentations, 107 features were extracted per lesion using PyRadiomics software (version 3.0.1) for Python (version 3.7.16), compliant with the Image Biomarker Standardization Initiative guidelines [20, 21]. The features included, 18 first order (FO) statistics, 14 shape-based (SB) features, 24 Gray level co-occurrence matrix (GLCM) features, 16 Gray level run length matrix (GLRLM) features, 16 Gray level size zone matrix (GLSZM) features, five neighbouring grey tone difference matrix (NGTDM) features, and 14 Gray level dependence matrix (GLDM) features. Features were extracted using the default setting of PyRadiomics, that is, a kernel size of 3 mm (radius 1 mm) and fixed bin width of 25HU. Images were resampled to 1x1x1mm^3 voxel size using B-spline interpolation, which has been demonstrated to be robust and the standard for PyRadiomics analysis [22].

2.4 Feature aggregation

In many cases, patients have multiple lesions. Therefore various methods were explored to identify the most effective approach for aggregating data from lesion level to patient level. As illustrated in Fig. 2, five aggregation methods were employed, including *all lesions*, *by lesion type*, *largest lesion*, *mean*, and *weighted average*.

We consider $\vec{f}$ as the vector containing radiomic features (f_i) extracted from each lesion. Let N be the total number of baseline lesions segmented for every patient and V_i the volume of one lesion. The following aggregation methods were compared to build the predictive model.

(1) *All lesions*: each feature vector coming from one lesion ($\vec{f}$) was considered as an individual input.

$$\vec{f} = \vec{f}_{lesion}\,; \quad lesion = 1,..,N \tag{1}$$

(2) *By lesion type*: in this aggregation method, we consider four different types of lesions: lung, liver, node, and other. We calculate the average feature for each lesion type. As such, every patient can have a maximum of four $\vec{f}$ for each lesion, f_{lung}, f_{liver}, f_{node}, and f_{other}. Each of them represented as f_x. Each f_x represents the average of the features from the corresponding lesion type. In Eq. 2, x stands for lung, liver, node or other.

$$\vec{f} = f_x\,; \quad f_x = \frac{\sum_i^N f_x^{(i)}}{N_x} \tag{2}$$

(3) *Largest lesion*: only the largest lesion was considered for each patient. Thus, every patient's feature vector, $\vec{f}$, consisted of the 107 feature values extracted from the largest lesion at baseline.

$$\vec{f} = \vec{f}_{V_{max}}, \; V_{max} = max(V_i)\,; \; i = 1,...,N \tag{3}$$

(4) *Mean of all lesions*: all lesions were considered for every patient. Thus, every patient's feature vector consisted of 107 feature values extracted from all lesions: the values of the features were averaged.

$$\vec{f} = \frac{\sum_i^N f_i}{N} \tag{4}$$

(5) *Weighted average of all lesions*: all lesions were considered for each patient. Therefore, the feature vector for every patient, $\vec{f}$, consisted of 107 feature values extracted from all lesions: the values of the features were averaged based on a weighted proportion of total volume at baseline. Therefore, feature values coming from larger metastases had more weight (ω) in the model.

$$\vec{f} = \sum_i^N \omega_i f_i\,; \quad \omega_i = \frac{V_i}{\sum_i^N V_i} \tag{5}$$

2.5 Radiomics standard pipeline

The radiomics standard pipeline was based on the workflow that is most frequently found in the literature for predicting response to ICIs. The standard pipeline included PyRadiomics, least absolute shrinkage and selection operator (LASSO) and logistic regression, as a feature extractor, feature selector, and classifier, respectively.

LASSO is a valuable technique for analysing high dimensional data sets. Its main purpose is to identify the most important features for predicting a target variable while eliminating less significant ones. This is achieved by incorporating a penalty term into the regression equation, which encourages the coefficients of less important features to be set to zero, effectively removing them from the model. Several studies have highlighted the benefits of using LASSO to predict response, particularly in high-dimensional data regression. LASSO's efficacy in handling high-dimensional data helps mitigate the risk of overfitting, making it a valuable tool in data analysis and modelling [23, 24].

2.5.1 Tumour specific analysis

The total studied cohort contains 13 different primary tumour types. The relation between primary tumour and the phenotype by means of radiomics features and on patient response pattern was evaluated. We explored adding the primary tumour location into the feature set as an encoded label. In this way we tried to compensate for changes in phenotype among the different tumour types.Therefore, we developed two different models for every aggregation method, one using radiomics features (RF) alone and the other one using RF combined with primary tumour (PT) location.

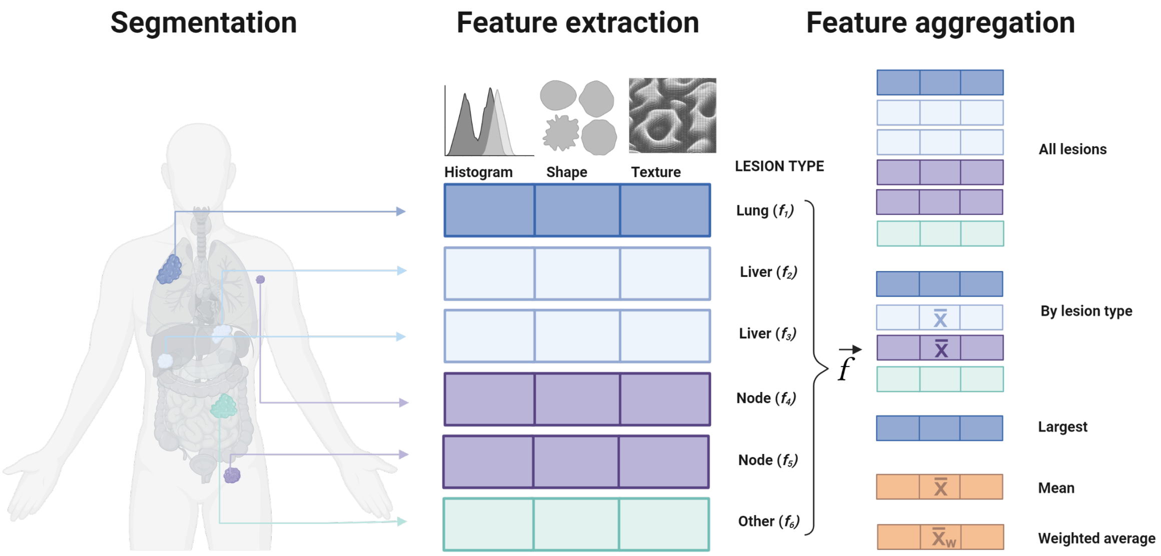


Figure 2. Various methods for feature aggregation from lesion level to patient level. The use of all lesions, aggregation per lesion type, largest lesion, mean of all lesions and the weighted average across all lesions has been evaluated.

Furthermore, we investigated the potential usefulness of developing tumour-specific models in contrast to pancancer approaches. Specifically, we developed five models: a comprehensive model that used the entire cohort and four specialised models for thoracic ($n = 108$), skin ($n = 79$), lower gastrointestinal ($n = 64$), and bladder tumours ($n = 53$).

2.5.2 Experimental approach

The experimental set-up is illustrated in Fig. 3. The entire data set is initially split into a training set that comprises 75% of the data and a test set that comprises the remaining 25%. The split is stratified based on the clinical benefit, ensuring balanced data. For integrity of the experimental results, when the aggregation methods *all lesions* and *by lesion type* are used, it was ensured that features of lesions from the same patient are exclusively present in either the training or test set.

Experiment 1 Experiment 1 is visualised in purple in Fig. 3. The training set is used to train a model using a five-fold cross-validation approach. In each fold, the optimal regularisation penalty term, λ, is determined through a grid search, resulting in five different models. For robustness of the models, the train-test split is repeated 100 times using different data distributions, ensuring that one split does not become biased towards a specific distribution of the data. This nested cross-validation procedure leads to a total of 500 different models generated. Experiment 1 will be executed for all aggregation methods to find the optimal aggregation method using the whole cohort. The best-performing aggregation methods will then be used to make tumour-specific models.

Experiment 2 The second experiment, illustrated in green (Fig. 3) is applied to the best performing models in terms of feature aggregation methods for the whole cohort and the four tumour specific models. The most selected features across all folds are counted for these models during Experiment 1, resulting in a top 10 list of the most selected features. Subsequently, the model is retrained another 100 times on the train set using only these top 10 features and evaluated on the test set, for different train-test splits. Since we generated 500 different models in Experiment 1, each feature had the potential to be selected up to 500 times. This was done to determine the improvement achieved by utilising only top 10 features. Initially, top 10 features were extracted from the best performing models. To investigate the impact of the number of features in the model, we evaluated the performance using the top 5 and 15 features as well by repeating Experiment 2.

The area under the curve (AUC) and 95% CI were determined from the receiver operating characteristic curve for both Experiment 1 and 2.

2.6 WORC

To assess the potential improvement of the radiomics standard pipeline, we employed WORC (version 3.6.2), an open-source toolbox designed to optimise radiomic workflows automatically using automated ML through random search and ensembling [18]. Within WORC, a total of 564 quantitative features were extracted from segmented lesions. The optimisation process involved the testing of different feature selection and classification algorithms.

The workflows were constructed by randomly sampling algorithms, denoted as A^*, along with their associated hyperparameter sets, denoted as λ^*. The sampling process is repeated 1000 times, resulting in the generation of multiple workflows. To rank these workflows, the performance is based on the F1 score of the validation data set. The final ensemble consisted of the top 100 workflows, which is then validated in the test set.

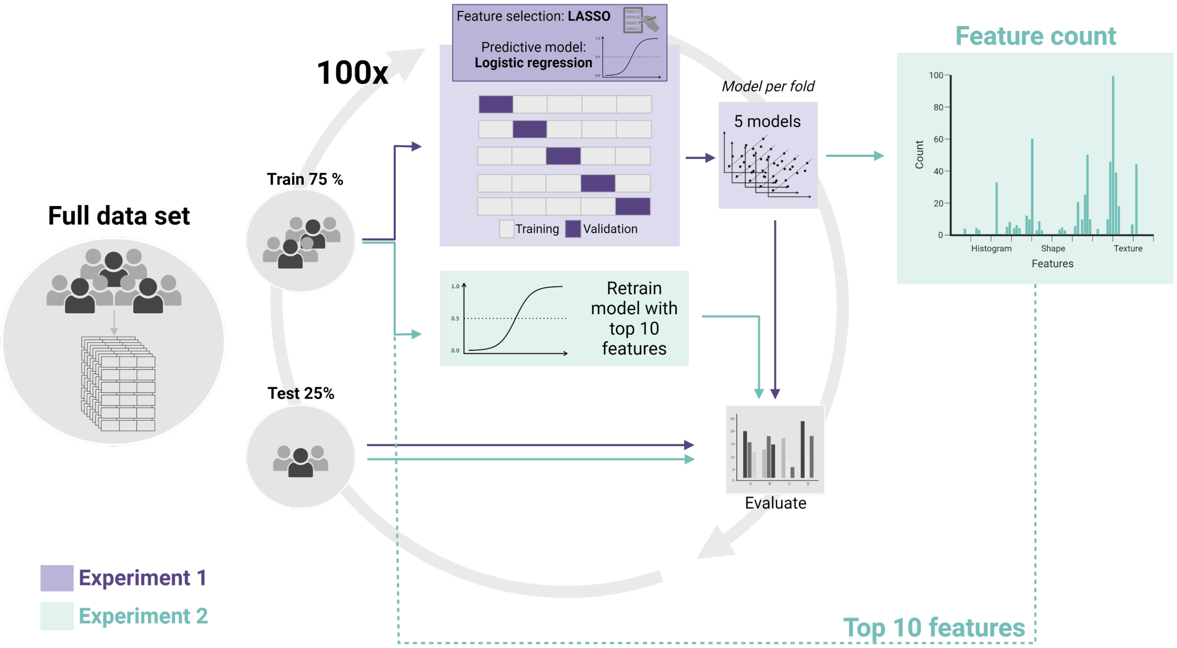


Figure 3. The figure illustrates the experimental setup used to evaluate feature aggregation methods and investigate the impact of tumour-specific models. The data set is divided into training (75%) and test sets (25%) and the models are trained using five-fold cross-validation, resulting in five models. The train-test split is repeated 100 times. For experiment 1 this results in a total of 500 models due to the 5-fold cross validation. All generated models are consequently evaluated on the test set. Experiment 2 examines the top 10 frequently selected features, retraining models, and evaluating their performance. AUC and 95% CI are calculated from the receiver operating characteristic curve for both experiments.

The data set was divided into an 80% training set and a 20% testing set. Each workflow was optimised using the training set within a 5-fold cross-validation framework. This entire process was repeated 100 times.

2.6.1 Experimental approach

WORC is an enclosed package and therefore its complexity limits the capacity of personalizing the data entry. It does not allow to test feature aggregation methods, therefore, the use of *by lesion*, *mean of all lesions* or *weighted average* was excluded when using WORC. As the amount of data was too large when using all lesions ($n = 2197$) from the entire cohort, we only tested the *largest lesion* aggregation method when using the whole cohort. When exploiting the tumour specific models, we could test the *all lesions* and *largest lesion* methods. The performance of WORC will be compared with the performance obtained in Experiment 1 when using the standard pipeline as this approach is most comparable.

2.7 Association with PFS

We are predicting a binary classification, specifically whether patients experience clinical benefit or not. As a result, we do not differentiate between patients who just experienced clinical benefit (i.e. predicting progression-free survival (PFS) of 5 months) compared to patients who had a very long PFS ($>> 5$ months). PFS is the amount of time between the start of the treatment to the first occurrence of disease progression or death. Therefore, to gain more insight into how well our predictions divide the population in responders and non-responders, we employ Kaplan-Meier analysis. We evaluated the performance of the models in predicting PFS using both the entire cohort and four specific tumour cohorts. The prediction scores for the best performing models in terms of aggregation method and amount of features (5, 10 or 15) were obtained from the test sets of Experiment 2. The prediction score (S) will be calculated by averaging all prediction scores (s_i) obtained for every patient.

PFS analysis was performed using the Kaplan-Meier method after score dichotomization. Patients who did not reach the end-point were censored to the last follow-up date. The survival analysis and significance assessment were performed using the R software (version 4.2.2) programming language. The survival outcomes were compared between the groups based on the predicted outcome. To assess the significance of the observed differences, the log-rank test was employed.

3. Results

3.1 Data set

Table 1 provides an overview of the clinical characteristics. The total cohort consists of 447 patients with a total of 2197 lesions. The median PFS is 3.5 months (95%CI: 2.8-4.1).

Table 1. Cohort description

Whole cohort (*n=447*)	
Primary tumour type	
Thoracic	108
Skin	79
Lower gastrointestinal	64
Bladder	53
Breast	39
Head and neck	30
Female pelvis	20
Endocrine	17
Upper gastrointestinal	13
Renal	12
Hepatobilary	5
Bone	5
Penile	2
Clinical benefit	
Yes	177
No	270
Progression free survival	
Median [95%CI]	3.5 [2.8-4.1] months
Number of lesions	
n	2197
Mean lesions per patient	
Mean [range]	4.91 [1-37]

Among the patients, 177 experienced clinical benefit, while 270 did not.

3.2 Feature aggregation

The results in Fig. 4 illustrate the AUC values along with their corresponding 95% CI for the various aggregation methods used in this study. Two sets of input data were used: RF alone and RF combined with the PT location. Incorporating the PT location as an input led to an improvement in AUC for all aggregation methods. Among the aggregation methods tested, the most effective aggregation methods were *all lesions* and *largest lesion* when using RF combined with PT. The *all lesions* method yielded an AUC of 0.582 (95% CI: 0.575-0.589) and *largest lesion* achieved an AUC of 0.612 (95% CI: 0.606-0.619).

3.3 Tumour specific models

The most effective aggregation methods, namely *all lesions* and *largest lesion*, were employed to develop tumour-specific models for thoracic cancers ($n = 108$), skin cancer ($n = 79$), lower gastrointestinal cancer ($n = 64$), and bladder cancer ($n = 53$). The obtained AUC values are shown in Fig. 5. Both the results for Experiment 1 and Experiment 2 are shown. In all cases, Experiment 2 demonstrated improved results compared to Experiment 1.

For Experiment 2, the whole cohort (AUC: 0.634), thoracic cancer (AUC: 0.741), skin cancer (AUC: 0.766) and lower gastrointestinal cancer (AUC: 0.794) obtained the best AUC when using *largest lesion*, only in the case of bladder cancer (AUC: 0.717) using *all lesions* yielded better results.

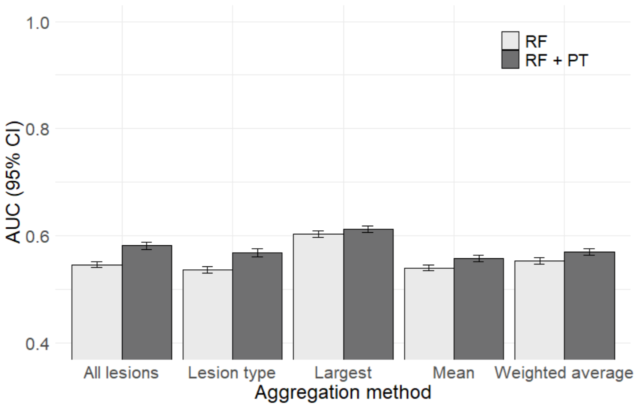


Figure 4. AUC values with corresponding 95% CI for the various aggregation methods using two sets of input data: RF alone and RF combined with PT location. The most effective aggregation methods were *all lesions* and *largest lesion* when using RT combined with PT. The *all lesions* method yielded an AUC of 0.582 (95% CI: 0.575-0.589) and *largest lesion* achieved an AUC of 0.612 (95% CI: 0.606-0.619)

3.4 Feature analysis

From Experiment 1, we extracted the top 10 features from the best performing models illustrated in Figure 5. The amount of features that were selected during every fold was 11.15 (SD: 2.039). In Fig. 6 we evaluated the effect of using 5, 10 and 15 features. For the whole cohort and bladder tumours, using 5 features was the most optimal (AUC: 0.653, and 0.760, respectively). When using thoracic, skin and lower gastrointestinal tumours the prediction was most accurate using 10 features (AUC: 0.741, 0.766, and 0.794, respectively).

3.5 WORC

The AUC values obtained for the use of WORC are shown in Fig. 7. The results of WORC are compared with the best performing models of experiment 1 when using the whole cohort and tumour-specific models. In all cases the standard radiomics pipeline outperformed the performance of WORC.

3.6 Association with PFS

For the best performing models we obtained the prediction scores which were used to stratify patients into responders and non-responders. The best performing models were for the whole cohort using *largest lesion* and 5 features, for the thoracic tumour model using *largest lesion* and 10 features, for the skin tumour model using *largest lesion* and 10 features, for the lower gastrointestinal tumour model using *largest lesion* and 10 features, and for the bladder tumour model using *all lesions* and 5 features. The survival outcomes were compared between the two groups and the resulting Kaplan-Meier curves are shown in Fig. 8. The whole cohort and thoracic

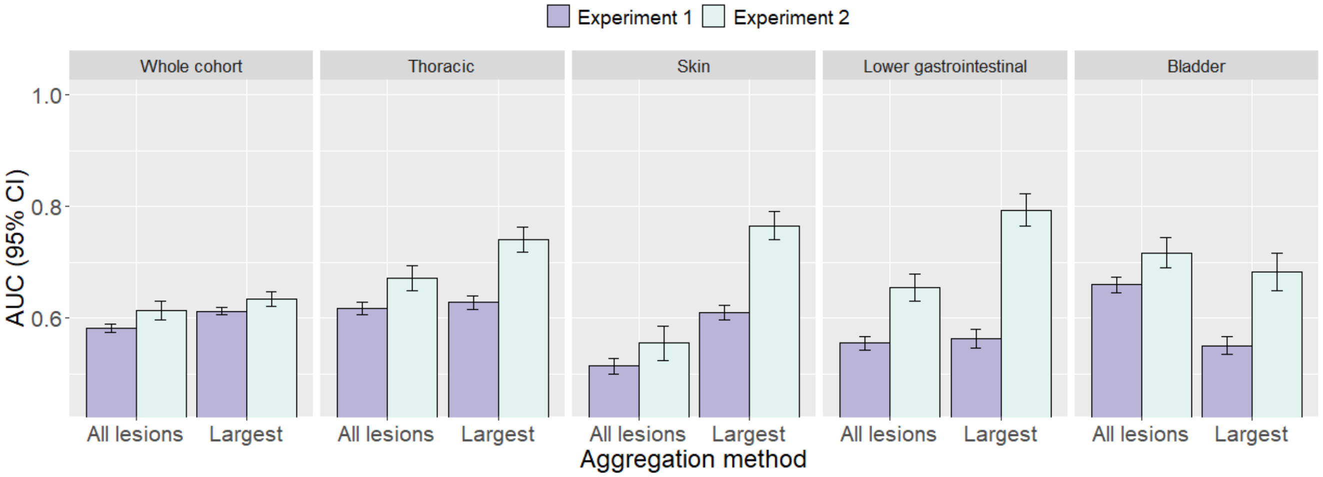


Figure 5. AUC values with corresponding 95% CI for Experiment 1 and 2 using the whole cohort ($n = 447$) and for tumour-specific models, thoracic cancers ($n = 108$), skin cancer ($n = 79$), lower gastrointestinal cancer ($n = 64$), and bladder cancer ($n = 53$). When using the whole cohort radiomic features were combined with the primary tumour location.

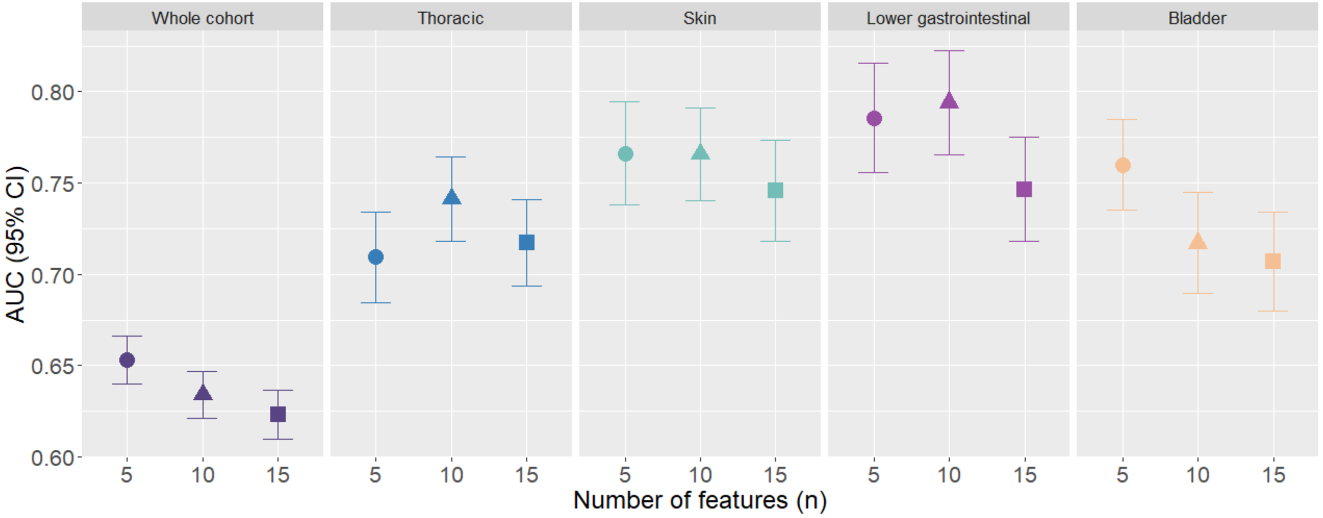


Figure 6. AUC values with corresponding 95% CI from Experiment 2 using the top 5, 10 and 15 selected features in the model.

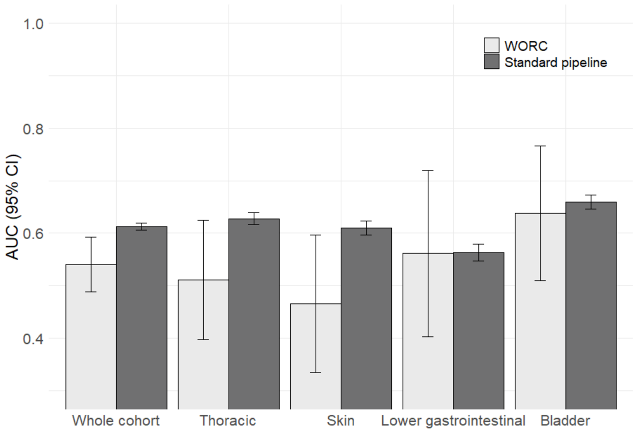


Figure 7. AUC values with corresponding 95% CI for the standard pipeline compared to WORC. For the whole cohort, thoracic cancer, skin cancer and lower gastrointestinal cancer *largest lesion* is used as aggregation method. For bladder cancer *all lesions* is used. The standard workflow for the whole cohort incorporated PT location.

tumour model showed significant differences in survival outcomes between responders and non-responders, with p-values of less than 0.01. The skin tumour model showed a statistically significant result, with a p-value of 0.0059. Similarly, the lower gastrointestinal model demonstrated a significant difference in survival outcomes, yielding a p-value of 0.01. However, the bladder tumour model did not exhibit a significant distinction between the two groups, with a p-value of 0.27.

4. Discussion

Our main objective was to predict response to ICI therapy using CT based biomarkers. To achieve this objective, three subobjectives were identified. Firstly, we aimed to determine the most effective approach for aggregating features from the lesion level to the patient level. Secondly, we investigated the impact of the primary tumour in our model. We added the primary tumour location as an encoded label and developed four tumour-specific models. Lastly, we sought to identify the most optimal radiomic workflow for creating radiomic signatures, a radiomics standard workflow and WORC.

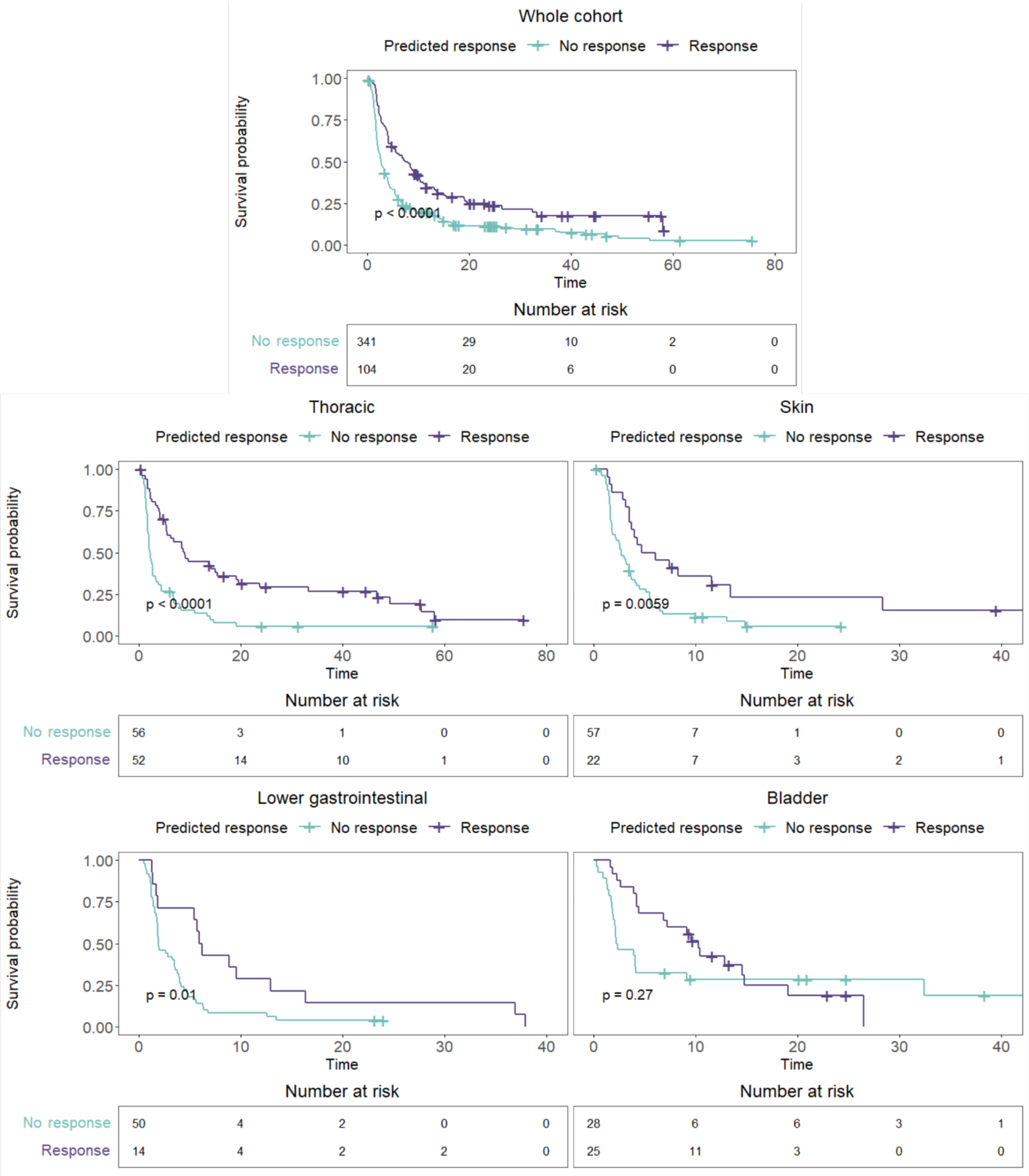
Regarding feature aggregation, the signature achieved the highest performance when using features from the largest lesion only. This finding aligns with previous research conducted by Montagnon *et al.* [25], as they found that using *largest lesion* in combination with the amount of lesions provided the highest concordance index for prediction of disease-free survival. Interestingly, the authors found that using *mean* as aggregation method yielded the highest performance when predicting recurrence. However, we found that using aggregation methods in which we average over lesions (*by lesion*, *weighted averaged* and *mean*) does not result in good performances. We believe that averaging over features loses crucial information that can be kept when not averaging beforehand.

Predicting response to any therapy is a highly complex task, and unfortunately, we did not achieve satisfactory results when utilising the pancancer approach. Despite the advantage of having access to a larger volume of data, incorporating more data did not yield improved prediction performance. Potentially the heterogeneity in the data made it complex to model. It might be that the response prediction for ICIs is influenced by the specific tumour type. This was also observed when incorporating the PT location as an additional feature as this also improved the predictive capacity. When we focused on tumour-specific models, we observed more promising results. These results indicate the importance of distinguishing between different tumour types when attempting to predict response to ICIs. By tailoring our models to specific tumour types, we were able to uncover more meaningful patterns that contribute to response prediction. Future research should continue to explore tumour-specific approaches to improve the precision of response prediction for ICIs.

Interestingly, despite the utilisation of tumour-specific models that improved our results, we were unable to achieve comparable outcomes to previously published papers that followed similar approaches, as mentioned by [24, 26, 27, 28]. This finding suggests that the generalisability of using radiomics for predicting response to ICI therapy is limited. Consequently, future research should not only concentrate on employing tumour-specific models but also focus on incorporating larger cohorts with multi-center data to obtain more representative and reliable results.

Using WORC, which incorporates random searches for selecting feature selectors and classifiers, did not surpass the performance of the standard radiomics workflow. Based on these findings, it is unlikely that employing alternative ML approaches would significantly enhance the results. Nevertheless, there is still an opportunity to investigate the potential of deep learning (DL) in predicting ICI response. In a recent study conducted by Zhao *et al.* [29], they investigated the application of a multilayer perceptron (MLP) to predict response in patients with advanced breast cancer. Their findings were highly promising, as the MLP demonstrated exceptional performance in distinguishing between responders and non-responders. These results show the potential value of utilising a MLP in tumour-specific cohorts.

To gain more insight into how well our predictions divided the population in responders and non-responders, we employed Kaplan-Meier analysis. In four out of five models, we observed a significant difference in PFS between the two populations based on our predictions. However, the bladder model did not show any significant differences. This suggests that even if the predictive accuracy of individual models might not be very high, they are still capable of capturing meaningful patterns that can effectively divide the population into responders and non-responders. To further enhance the accuracy and effectiveness of predicting immunotherapy re-



Number at risk (Whole cohort)

	0	20	40	60	80
No response	341	29	10	2	0
Response	104	20	6	0	0

Number at risk (Thoracic)

	0	20	40	60	80
No response	56	3	1	0	0
Response	52	14	10	1	0

Number at risk (Skin)

	0	10	20	30	40
No response	57	7	1	0	0
Response	22	7	3	2	1

Number at risk (Lower gastrointestinal)

	0	10	20	30	40
No response	50	4	2	0	0
Response	14	4	2	2	0

Number at risk (Bladder)

	0	10	20	30	40
No response	28	6	6	3	1
Response	25	11	3	0	0

Figure 8. Kaplan-Meier PFS curve analysis. The analysis was done using the the whole cohort, and tumour-specific analysis for thoracic, skin, lower gastrointestinal and bladder cancers.

sponse, it might be beneficial to integrate various types of data, such as radiomics, histopathologic, and genomic data. Vanguri *et al.* [30] already conducted a study to determine the value of using multimodal features to improve prediction of immunotherapy response in patients with advanced non-small cell lung cancer (NSCLC). They combined medical imaging, histopathologic, and genomic features to create a predictive model for immunotherapy response. The results showed that the multimodal model outperformed the unimodal measures reported in the study. Therefore, future research should prioritize the integration of multimodal data to further improve the accuracy and effectiveness of predicting immunotherapy response.

Our study had several limitations. First, the use of tumour-specific cohorts resulted in a decreased number of patients, which may affect the generalisability of our findings. Secondly, it should be noted that the findings of Experiment 2 are influenced by bias due to the incorporation of features derived from the outcomes of Experiment 1, where the entire data set had already been observed. However, the classification models developed in Experiment 2 were entirely new and did not depend on Experiment 1. Ideally, it would have been preferable to evaluate Experiment 2 on a completely independent data set. Unfortunately, again due to the limited data availability of tumour-specific cohorts, we were unable to pursue this approach. Third, in Experiment 2, we examined the effectiveness of the top 10 features that were selected the most frequently. A similar approach was done by Shahzadi *et al.* [31] in which they selected the top 5 features. Radiomic features are known to be highly correlated, which means that certain features can be used interchangeably without affecting predictive performance and therefore different features can be randomly selected in different splits [32]. Due to this, the final features incorporated in the models when using the most selected features might be highly correlated. Shahzadi *et al.* added an additional feature selection method namely, if features showed a Spearman correlation > 0.5, only the feature with the highest cumulative occurrence was considered. This would be a good addition in the final feature selection of our approach to avoid the problem of incorporating highly correlated features in the model. Lastly, when using feature aggregation method *all lesions*, it was assumed that all lesions within a given patient would exhibit a uniform response. This assumption implied that applying the same treatment would yield identical effects on all lesions of the patient. Nonetheless, clinical studies have revealed that in 8% to 14% of cases exhibit a mixed response [33, 34]. This indicates that certain lesions may respond favourably to a specific treatment while others may not, resulting in a heterogeneous response. In this project, we disregarded the occurrence of mixed responses. For future research, it could be essential to consider the presence of mixed responses and incorporate strategies to account for this variability in treatment outcomes.

In conclusion, our study aimed to identify CT-based biomarkers that can predict response to ICI therapy. Among the five aggregation methods tested, we found that utilising *largest lesion* obtained the best results. Additionally, tumour-specific model obtained more predictive capacity compared to a pancancer model. Furthermore, the identified radiomics standard workflow obtained the best results which indicates that using LASSO as feature selector and logistic regression as classifier proves to be an effective strategy. However, to validate these findings, data from larger tumour-specific cohorts is necessary. Looking ahead, the potential of DL, specifically MLP, and multimodel approaches that combine medical images, immunohistochemistry slides and genomics data could be explored to improve predictive capacities ideally resulting in patient specific medicine.

References

[1] Parkin J, Cohen B. An overview of the immune system. Lancet (London, England). 2001 6;357:1777-89. Available from: https://pubmed.ncbi.nlm.nih.gov/11403834/.

[2] Pham T, Roth S, Kong J, Guerra G, Narasimhan V, Pereira L, et al. An Update on Immunotherapy for Solid Tumors: A Review. Annals of surgical oncology. 2018 10;25:3404-12. Available from: https://pubmed.ncbi.nlm.nih.gov/30039324/.

[3] Loose D, Wiele CVD. The immune system and cancer. Cancer biotherapy radiopharmaceuticals. 2009 6;24:369-76. Available from: https://pubmed.ncbi.nlm.nih.gov/19538060/.

[4] Jardim DL, Gagliato DDM, Giles FJ, Kurzrock R. Analysis of Drug Development Paradigms for Immune Checkpoint Inhibitors. Clinical cancer research : an official journal of the American Association for Cancer Research. 2018 4;24:1785-94. Available from: https://pubmed.ncbi.nlm.nih.gov/29212781/.

[5] Postow MA, Chesney J, Pavlick AC, Robert C, Grossmann K, McDermott D, et al. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. The New England journal of medicine. 2015 5;372:2006-17. Available from: https://pubmed.ncbi.nlm.nih.gov/25891304/.

[6] Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. The New England journal of medicine. 2015 7;373:123-35. Available from: https://pubmed.ncbi.nlm.nih.gov/26028407/.

[7] Pilard C, Ancion M, Delvenne P, Jerusalem G, Hubert P, Herfs M. Cancer immunotherapy: it's time to better predict patients' response. British journal of cancer. 2021 9;125:927-38. Available from: https://pubmed.ncbi.nlm.nih.gov/34112949/.

[8] Ventola CL. Cancer Immunotherapy, Part 3: Challenges and Future Trends. Pharmacy and Therapeutics. 2017 8;42:514. Available from: /pmc/articles/PMC5521300//pmc/articles/PMC5521300/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5521300/.

[9] Dercle L, Ammari S, Roblin E, Bigorgne A, Champiat S, Taihi L, et al. High serum LDH and liver metastases are the dominant predictors of primary cancer resistance to anti-PD(L)1 immunotherapy. European Journal of Cancer. 2022 12;177:80-93.

[10] Howitt BE, Strickland KC, Sholl LM, Rodig S, Ritterhouse LL, Chowdhury D, et al. Clear cell ovarian cancers with microsatellite instability: A unique subset of ovarian cancers with increased tumor-infiltrating lymphocytes and PD-1/PD-L1 expression. OncoImmunology. 2017 2;6. Available from: https://www.tandfonline.com/doi/abs/10.1080/2162402X.2016.1277308.

[11] Howitt BE, Shukla SA, Sholl LM, Ritterhouse LL, Watkins JC, Rodig S, et al. Association of Polymerase e–Mutated and Microsatellite-Instable Endometrial Cancers With Neoantigen Load, Number of Tumor-Infiltrating Lymphocytes, and Expression of PD-1 and PD-L1. JAMA Oncology. 2015 12;1:1319-23. Available from: https://jamanetwork.com/journals/jamaoncology/fullarticle/2383137.

[12] Shum B, Larkin J, Turajlic S. Predictive biomarkers for response to immune checkpoint inhibition. Seminars in cancer biology. 2022 2;79:4-17. Available from: https://pubmed.ncbi.nlm.nih.gov/33819567/.

[13] Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology. 2013 10;269:8-15. Available from: https://pubmed.ncbi.nlm.nih.gov/24062559/.

[14] van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging. 2020 12;11. Available from: /pmc/articles/PMC7423816//pmc/articles/PMC7423816/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7423816/.

[15] Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, et al. Introduction to Radiomics. Journal of nuclear medicine : official publication, Society of Nuclear Medicine. 2020 4;61:488-95. Available from: https://pubmed.ncbi.nlm.nih.gov/32060219/.

[16] Fave X, Zhang L, Yang J, MacKin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific reports. 2017 12;7. Available from: https://pubmed.ncbi.nlm.nih.gov/28373718/.

[17] Sun R, Henry T, Laville A, Carré A, Hamaoui A, Bockel S, et al. Imaging approaches and radiomics: toward a new era of ultraprecision radioimmunotherapy? Journal for Immunotherapy of Cancer. 2022 7;10:4848. Available from: /pmc/articles/PMC9260846//pmc/articles/PMC9260846/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9260846/.

[18] Starmans MPA, van der Voort SR, Phil T, Timbergen MJM, Vos M, Padmos GA, et al. Reproducible radiomics through automated machine learning validated on twelve clinical applications. 2021 8. Available from: https://arxiv.org/abs/2108.08618v2.

[19] Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria

in solid tumours: Revised RECIST guideline (version 1.1). Available from: https://pubmed.ncbi.nlm.nih.gov/19097774/.

[20] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 5;295:328-38. Available from: https://pubmed.ncbi.nlm.nih.gov/32154773/.

[21] Griethuysen JJMV, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer research. 2017 11;77:e104-7. Available from: https://pubmed.ncbi.nlm.nih.gov/29092951/.

[22] Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. European Radiology. 2021 3;31:1460-70. Available from: https://link.springer.com/article/10.1007/s00330-020-07174-0.

[23] Jazieh K, Khorrami M, Saad A, Gad M, Gupta A, Patil P, et al. Novel imaging biomarkers predict outcomes in stage III unresectable non-small cell lung cancer treated with chemoradiation and durvalumab. Journal for immunotherapy of cancer. 2022 3;10. Available from: https://pubmed.ncbi.nlm.nih.gov/35256515/.

[24] Park KJ, Lee JL, Yoon SK, Heo C, Park BW, Kim JK. Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma. European radiology. 2020 10;30:5392-403. Available from: https://pubmed.ncbi.nlm.nih.gov/32394281/.

[25] Montagnon E, Elforaici M, Montréal P, Montreal FR, Hub AI, Canada E, et al. Radiomics analysis of baseline computed tomography to predict oncological outcomes in patients treated for resectable colorectal cancer liver metastasis. Available from: https://doi.org/10.21203/rs.3.rs-2762043/v1.

[26] Tunali I, Gray JE, Qi J, Abdalah M, Jeong DK, Guvenis A, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. Lung cancer (Amsterdam, Netherlands). 2019 3;129:75-9. Available from: https://pubmed.ncbi.nlm.nih.gov/30797495/.

[27] Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Pizzi AD, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Annals of oncology : official journal of the European Society for Medical Oncology. 2019 6;30:998-1004. Available from: https://pubmed.ncbi.nlm.nih.gov/30895304/.

[28] Khorrami M, Prasanna P, Gupta A, Patil P, Velu PD, Thawani R, et al. Changes in CT Radiomic Features Associated with Lymphocyte Distribution Predict Overall Survival and Response to Immunotherapy in Non-Small Cell Lung Cancer. Cancer immunology research. 2020;8:108-19. Available from: https://pubmed.ncbi.nlm.nih.gov/31719058/.

[29] Zhao J, Sun Z, Yu Y, Yuan Z, Lin Y, Tan Y, et al. Radiomic and clinical data integration using machine learning predict the efficacy of anti-PD-1 antibodies-based combinational treatment in advanced breast cancer: a multicentered study. Journal for ImmunoTherapy of Cancer. 2023 5;11:e006514. Available from: https://jitc.bmj.com/content/11/5/e006514https://jitc.bmj.com/content/11/5/e006514.abstract.

[30] Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. Nature Cancer. 2022 10;3:1151-64. Available from: https://pubmed.ncbi.nlm.nih.gov/36038778/.

[31] Shahzadi I, Zwanenburg A, Lattermann A, Linge A, Baldus C, Peeken JC, et al. Analysis of MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. Scientific Reports 2022 12:1. 2022 6;12:1-15. Available from: https://www.nature.com/articles/s41598-022-13967-8.

[32] Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Physics in medicine and biology. 2016 6;61:R150-66. Available from: https://pubmed.ncbi.nlm.nih.gov/27269645/.

[33] Morinaga T, Inozume T, Kawazu M, Ueda Y, Sax N, Yamashita K, et al. Mixed Response to Cancer Immunotherapy is Driven by Intratumor Heterogeneity and Differential Interlesion Immune Infiltration. Cancer Research Communications. 2022;2(7):739-53. Available from: https://pubmed.ncbi.nlm.nih.gov/36923281/.

[34] Tazdait M, Mezquita L, Lahmar J, Ferrara R, Bidault F, Ammari S, et al. Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: comparison of RECIST 1.1, irRECIST and iRECIST criteria. European Journal of Cancer. 2018;88:38-47. Available from: https://pubmed.ncbi.nlm.nih.gov/29182990/.