Pre-operative radiomics model for prognostication in resectable pancreatic adenocarcinoma: multi-institutional development and external validation

Thesis submitted for the degree of Doctor of Medicine (M.D.) at the National University of Ireland, Galway

By

Gerard Michael Healy,
MB BCh BAO, MRCPI, FFRRCSI, EBIR

Internal supervisor: Professor Peter McCarthy, School of Medicine, National University of Ireland, Galway, Ireland.

External supervisor: Professor Masoom Haider,
Department of Medical Imaging, University of Toronto, Canada

Original submission: June 2021
Revision: January 2022

Table of contents

Table of contents

Table of contents

Declaration

This thesis is my own work and I have not obtained a degree in this University, or elsewhere, on the basis of this research. I and the members of the Radiomics and Machine Learning Lab at the Lunenfeld Tanenbaum Research Institute declare no relationships with any companies, whose products or services may be related to the subject matter of the thesis.

Funding:

Dr Gerard M Healy received a scholarship from the Faculty of Radiologists, Royal College of Surgeons in Ireland ($86,000 Canadian) and also received funding from the Radiomics and Machine Learning Lab at the Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, Toronto ($4,300 Canadian). The lab is funded in part by the Ontario Institute for Cancer Research (OICR).

Informed consent: Written informed consent was waived by the Institutional Review Board.

Ethical approval: Institutional Review Board approval was obtained.

Dedication

I would like to dedicate this work to my wife Aoibhinn and our daughter Féile. My work on this MD was only possible because of their support while living through the COVID-19 pandemic in Toronto. I would also like to thank my parents for their support and encouragement throughout my career to date.

Projects like this require a large team and therefore, I would like to thank my colleagues at the Radiomics and Machine Learning Lab at Mount Sinai Hospital, Dr. Emmanuel Salinas Miranda, Dr. Dominik Deniffel, Xin Dong and Dr. Rahi Jain, for their contributions and I would like to thank our lab PI, Professor Masoom Haider, for his leadership and guidance. We also received support from other disciplines at the University Health Network / University of Toronto; Ms. Ayelet Borgida (PanCuRx Translational Research Initiative, Ontario Institute for Cancer Research), Professor Wei Xu (Biostatistics), Dr Robert Grant (Medical Oncology) and Dr Steven Gallinger (Hepatobiliary surgery). On the Irish side, I would like to than Dr David Ryan, Prof. Ronan Ryan and Prof Jonathan Dodd from the department of Radiology, St Vincent's University Hospital, Dublin and Professor McCarthy, my internal supervisor from NUI Galway. Finally, I would like to thank the members and board of the faculty of Radiologists, Royal College of Surgeons in Ireland, who part funded my clinical-research fellowship.

"Tell me and I forget. Teach me and I remember. Involve me and I learn"

Benjamin Franklin

Abstract

Objectives: In patients with resectable pancreatic ductal adenocarcinoma (PDAC), there are few strictly pre-operative prognostic biomarkers available to guide therapy decisions. Radiomics has demonstrated potential prognostic value but it lacks external validation. We aimed to develop and externally validate a pre-operative clinical-radiomic prognostic model for PDAC.

Methods: This was a retrospective international, multi-center study in patients with resectable PDAC who underwent pre-operative contrast-enhanced CT. Patients who received neoadjuvant therapy were excluded. The training cohort consisted of 352 patients who underwent CTs at five Toronto hospitals and subsequent resection at Toronto General Hospital, Toronto, Canada. The external test cohort consistent of 215 patients who underwent resection at a St Vincent's University Hospital, Dublin, following pre-operative CTs performed at 34 Irish hospitals. Segmentation was performed using 3d Slicer v 4.11.2. Then 116 radiomic features were extracted using the PyRadiomics 3.0 library. Pre-operative Cox proportional hazard models incorporated (a) clinical factors (clinical), (b) clinical plus radiomics features (clinical-radiomic) and (c) a post-operative model incorporating pathological findings (TNM), which served as the reference standard. Outcomes were overall (OS) and disease-free survival (DFS). Model discrimination and calibration were assessed using concordance index (C-index), calibration plots and mean calibration error. A previously validated statistical tool for batch-effect correction (Combat) was used in an attempt to mitigate the impact of variation in CT scanner protocols between the multiple study sites.

Results: In the validation cohort, the Radiomic signature was predictive of OS / DFS, with adjusted hazard ratios (HR) of 2.87 (95% CI: 1.40-5.87, p<0.001) / 5.28 (95% CI 2.35-11.86, p<0.001) respectively, along with age 1.02 (1.01-1.04, p=0.01) / 1.02 (1.00-1.04, p=0.03). No other clinical features were significantly associated with OS and DFS. Median OS was 22.9 versus 37 months (p=0.0092) and DFS 14.2 versus 29.8

(p=0.0023) for the high versus low-risk groups in the external cohort. Calibration was moderate in the external cohort, with mean absolute error 7% and 13% for OS at 3 and 5 years respectively. The clinical-radiomic model demonstrated better discrimination for OS in the external cohort (C-index 0.545, 95%: 0.543-0.546) than the clinical model alone (0.497 95% CI 0.496-0.499, p<0.001) or the post-operative TNM model (0.525 95% CI: 0.524-0.526, p<0.001). Implementation of Combat to mitigate the impact of multi-institutional variation in CT acquisition parameters did not improve discrimination results. In decision curve analysis, despite superior net benefit compared to clinical model, the clinical-radiomic model was not clinically useful for most threshold probabilities. TNM demonstrated the highest net benefit of the three models.

Conclusion: A pre-operative model containing clinical variables and radiomics significantly improved prognostication of patients with resectable PDAC compared to using clinical information alone and it generalized to a large external dataset. Performance was similar to using pathological data (TNM), which are only available post-operatively. Despite superior performance compared to the clinical model, discrimination and clinical utility are suboptimal. This likely reflects inherent limitations of radiomics for PDAC prognostication, when deployed in real-world settings. Future work should focus upon standardization of CT acquisition protocols.

Acronyms

AI – Artificial Intelligence.

AJCC - American Joint Committee on Cancer.

CDKN2A - Cyclin Dependent Kinase Inhibitor 2A.

CI – Confidence Intervals.

CMPA - Canadian Medical Protective Association (CMPA).

CNN – Convolutional Neural Network.

CT – Computed Tomography.

DCA – Decision Curve Analysis.

DICOM - Digital Imaging and Communications in Medicine.

DL – Deep Learning.

GDPR - General Data Protection Regulation (GDPR).

HR – Hazard Ratio.

HU – Hounsfield unit.

IBSI – Image Biomarker Standardization Initiative.

ISI - Image-to-surgery time interval (in days).

KRAS - Kirsten rat sarcoma 2 viral oncogene homolog.

LASSO - Least absolute shrinkage and selection operator.

ML – Machine Learning.

MPS - Medical Protection Society (MPS).

MVA – Multivariable analysis.

NCCN - National Comprehensive Cancer Network.

Acronyms

NRRD - Nearly Raw Raster Data.

OS – Overall Survival.

PACS – Picture archiving and communication system.

PDAC – Pancreatic Ductal Adenocarcinoma.

PV – Portal venous.

QIBA – Quantitative Imaging Biomarker Alliance.

RIS – Radiology information system.

RQS – Radiomics Quality Score.

SMAD4 - Mothers against decapentaplegic homolog 4.

SVUH – St Vincent's University Hospital, Dublin, Ireland.

TP53 – Tumour Protein 53.

TRIPOD - Transparent Reporting of multivariable prediction model for Individual Prognosis or Diagnosis statement.

UHN – University Health Network, Toronto, Canada.

UVA – Univariable analysis.

Chapter 1 – Introduction to Artificial Intelligence in Medical Imaging

"Images Are More than Pictures, They Are Data"

Gillies et al 2016 [1]

1.1 Current Radiology practice.

Radiology is the interpretation of medical imaging for the purposes of detection, diagnosis, follow-up and therapy of disease. This is performed by visual inspection of images by a radiologist, traditionally using printed images on radiology films, but more recently using digital images on computer screens. If we take the example of a tumour, the radiologist would use anatomical descriptions such as location, size, shape and relationship to surrounding structures. As radiology modalities have advanced, several additional methods for interrogation of body tissues have emerged, allowing the radiologist to comment on characteristics such as contrast enhancement pattern, metabolic activity (with nuclear medicine imaging) and microscopic attributes such the ability of water molecules to diffuse within a tissue on a Magnetic Resonance Imaging study. More recently, advanced computing and statistical techniques have evolved to potentially allow another level of tissue assessment. These techniques can potentially identify characteristics within the images which are not visible to the human eye, thereby providing additional biomarkers to aid the reporting Radiologist and Clinicians in decision making. The goal is to achieve the complete characterisation of tissues from radiology studies, therefore facilitating a 'non-invasive biopsy' [2] of tissues, with multiple advantages including patient safety and convenience, ease of repeatability to assess changes over time and sampling of the entire tissues/tumours rather than the small samples obtained from current needle biopsy techniques. These methods also hold the promise of a more personalised approach to medical care, a concept which is supported by multiple medical organisations worldwide, most notably the National

Institutes of Health (NIH) in the United States, which leads the Precision Medicine Initiative [3].

1.2 Artificial intelligence, machine learning and Radiomics.

'Any sufficiently advanced technology is indistinguishable from magic'

Arthur C. Clarke

There has been increased use of computers in medicine since at least the 1960s [4, 5], but there has been a significant shift towards research in Artificial intelligence (AI) over the past 15-20 years. AI is a general term, first coined by John McCarthy in 1956 [6], which describes computer systems that can perform tasks normally requiring human intelligence [7]. Machine Learning (ML) is a subdivision of AI, where computer systems learn from exposure to labelled data. An ML algorithm is 'trained' by exposure to such data and it creates a mathematical function (equation) to fit the inputs (for example patients age, sex and smoking status) to the outputs (for example survival time post treatment). It does this by making a prediction (guess) which it then compares to the truth (label) and the makes adjustments to the algorithm in order to narrow the gap between prediction and truth. This process is repeated multiple times, to optimise the performance. Once training has taken place in a large dataset, the ML algorithm can then make predictions when exposed to previously unseen and unlabelled test data, with varying degrees of accuracy. Therefore, ML is simply a statistical method which is automated and at large scale. Examples of frequently used ML methods are Support Vector Machines, Decision Trees and Random Forests, but many experts also consider more commonplace statistical techniques such as linear or logistic regression to be basic forms of machine learning.

The data labels for AI are also referred to as the ground truth. They may take many forms, for example medical images may be labelled with patient survival time,

pathological tumour grade or manual segmentations (contours) of organs performed by a radiologist. In medical imaging, the data to which the ML algorithms are exposed must be defined in some way by an intermediate step, for example certain characteristics of a tumour (such as mean CT attenuation) on an image must be defined by a human [8] or using a rule-based system before the ML algorithm can start training using this data. This is in contrast to Deep Learning (DL), which is a more complex subdivision of ML, where neural networks with multiple layers can learn from the unprocessed data (i.e., the raw CT images from a patient). This unprocessed data may be accompanied by a label, which is called supervised learning or it may be unlabelled, which is called unsupervised learning. DL algorithms in medical imaging identify the image characteristics relevant to the task at hand *de novo* during the training process, hence they construct their own custom image features and the algorithms have very limited ability to explain this decision-making process to humans, hence this technology has sometimes been termed a 'black box' method, which lacks transparency.

Radiomics is a less complex construct within the field of computer vision. The term was first proposed by Lambin *et al* in 2012 [9] and it combines Rad- (Radiology) with -omics, the latter meaning the mining of large volumes of data for precision medicine [10]. Radiomics is build upon the foundation of 'texture analysis' [5] and it involves extracting multiple predefined images characteristics, called radiomic features, from a medical image. These radiomic features are statistical representations which describe various facets of intensity, shape and texture of an image, typically of a particular body tissue, such as an organ or tumour. The features are defined by mathematical formulae. Once a large number of features are extracted from the image dataset, analysis is then performed (using classical statistics or ML) to identify a specific radiomic feature or group of features which are associated with an outcome of interest, such as a genetic profile or patient survival time. Hence, radiology images are converted from qualitative pictures into a dataset of quantitative imaging biomarkers, expressed as numbers on a continuous scale.

It is intended that radiomics should be performed on routine, standard of care images [1], thus potentially opening up the entire backlog of imaging available on the Picture Archiving and Communication Systems (PACS) of hospitals worldwide, subject to ethics oversight and approval. Multiple studies have demonstrated that radiomic features can act as biomarkers for underlying biological processes or patient outcomes in a variety of diseases [7], however there are no examples (to the best of our knowledge) where a radiomic method has been deployed for clinical use outside of a research setting. Radiomic research is expanding yearly, as evidenced by a PubMed search for 'Radiomics' for the years 2017, 2018, 2019 and 2020, which returns results of 254, 519, 898 and 1,474 publications respectively.

When approaching a task in computer vision, the decision to choose between different approaches (ML vs DL vs Radiomics) depends on the availability of labeled data, statistical methods appropriate for the domain and requirement for 'explainability' of the result. Some investigators have combined ML, Radiomics and DL methods together, for example a number of studies have fed radiomic features values extracted from medical images into a Random Forest ML algorithm [11-13] in order to classify outcomes (rather than using more classic linear statistics methods such as regression analysis), while other groups have combined radiomic and DL features together to predict outcomes [14, 15].

1.3. Data for AI: Training and validation.

In popular culture and the media, the focus is typically on the algorithm architectures used for AI (well known examples include ResNet and AlexNet), however it is important to remember that an AI method is only as good as the data upon which it is trained. The ideal training datasets for AI training are very large, multicenter, high quality, with accurate and relevant labels [16]. It has been shown that studies with larger population produce more reliable results and identify a significantly higher number of predictor features associated with the study

outcomes [17] however there are barriers to generating such databases, including ethical and legal barriers to accessing patient data [18, 19], we well as the significant labour required to curate and accurately label such datasets.

A key challenge for AI is generalizability i.e., the successful deployment of an algorithm or radiomics signature using an independent cohort of patients from a different clinic/hospital to where the algorithm was trained, which provides an unbiased assessment of the model performance. Guidelines advocate that novel algorithms should be validated in external test datasets prior to clinical implementation [20], however, less than 10% of AI studies currently report this level of testing [21] and there are many commercially available medical imaging AI products which lack such evidence [22]. This deficit in the literature is likely due to factors including (1) difficulty in obtaining suitable datasets, (2) bias from funders and journals towards novelty over the more mundane task of study replication/validation [23] and (3) the short term goals/deadlines in most research groups [23]. Many researchers therefore use internal validation methods, based on data from the same institution as the training data, using a hold-out test set (where they divide their patient dataset into a training set and a test set, typically using a ratio of 70:30) or temporal validation (where the test set is from the same source as the training data but from a different time period). More technically advanced options are cross-validation (e.g., leave one out or *k*-fold cross validation) or bootstrap resampling. While these advanced methods are better than split samples at counteracting overfitting, the algorithms are still being validated on the same data which was used for training and hence they cannot account for true differences between populations (difference demographics, disease prevalence, treatment practices etc.) [20]. Hence, interval validation methods can often report overoptimistic performance and therefore external validation is the gold standard [2, 24]. For those AI studies which have performed external validation, several have shown a considerable performance drop compared to the training data [25, 26]. A well publicised example of this phenomenon is the disappointing 2020 results from

the clinical deployment of Google's retina scanning algorithm for diabetic retinopathy; despite excellent performance in the lab (90% accuracy) and an approved European CE mark, performance was hampered during clinical deployment by simple factors such as poor ambient lighting in the room where the images were being acquired [27]. Therefore, all internal validation methods are considered inferior to external validation [20, 28] and the continued reporting of such internal performance metrics, particularly by large technology companies, has been derided as 'Silicon vally-dation' by medtech experts such as Dr Eric Topol [29].

1.4. Radiomics: Pipeline overview.

The general overview of a Radiomics study pipeline is as follows:

(1) Identify relevant imaging studies and associated clinical data.

(2) Segmentation of the area of interest: A region of interest is created by manual, semi-automatic or automated methods. The area of interest may be a tumour, organ or portion of an organ.

(3) Feature extraction.

(4) Feature selection/dimensionality reduction: using classic statistics, machine learning and/or neural networks.

(5) Build the model to predict the outcome of interest, for example patient survival.

(6) Test the model using internal and external validation methods.

1.5. Radiomics: Features.

Analysis of texture in images for the purposes of classification initially emerged in the 1970's engineering literature [30] but medical applications were not considered at that time. The technology has matured over the decades since and there are now 120 radiomic features for medical imaging which have undergone robust

standardization and reproducibility assessment by the Imaging Biomarker Standardisation Initiative (IBSI) [31]. These features are defined in a library which is freely available online [32]. Their number can be expanded into more than a thousand variables by the application of filters, for example a square root filter, which converts the image intensity values into their square root values before feature extraction. However, these filtered features have not been standardized by the IBSI [31] and the scientific rationale for applying these filters is not clear.

Radiomic features are grouped into those relating to: (1) shape (n=26), (2) distribution of intensities throughout the region of interest (called first order features, n=19) and (3) relationship of voxel intensities to each other in space (called second order features, n=75). Examples of shape features include maximum-2D-diameter and perimeter-to-surface-ratio. First order features include median, maximum and minimum attenuation. An example of a second order radiomic feature is 'coarseness', which is part of the Neighbouring Gray Tone Difference Matrix (NGTDM) family of second order features. NGTDM_ coarseness is defined as a 'measure of average difference between the center voxel and its neighbourhood' and is an indication of the spatial rate of change. A higher value indicates a 'lower spatial change rate and a locally more uniform texture' [31]. It has the following formula, reproduced from the Pyradiomics library [33]:

$$Coarseness = \frac{1}{\sum_{i=1}^{N_g} p_i s_i}$$

$\sum_{i=1}^{Ng} = \text{Sum of } PiSi, \text{from values } i = 1 \text{ to } i = Ng$. *i*=grey level value. N_g is the number of discrete gray levels. X_{gl} = a set of segmented voxels. N_i is the number of voxels in X_{gl} with gray level *i*. N_v = the number of voxels with a valid region; at least 1 neighbor. P_i = the gray level probability which is equal to n_i/N_v. S_i = the sum of absolute differences for gray level *i*.

The first and second order features all depend upon the intensities of the pixels in the image. In computed tomography (CT) the pixel intensity is called the gray level and it is measured in Hounsfield units (HU). The HU value depends upon the density of the material (e.g. bone vs soft tissue) but also on CT acquisition parameters used during that CT scan, such as tube voltage (kVp) [34]. The gray level values are calibrated internally (water is 0 HU and air is -1000 HU) and this calibration is checked regularly on every scanner as part of routine quality control. This reduces (but does not eliminate) the impact of variation in acquisition parameters upon the HU pixel values. Such a robust internal calibration does not exist for Magnetic Resonance Imaging (MRI), Ultrasound or Positron Emission Tomography (PET) studies. Hence, computed Tomography (CT) is the most common modality used in radiomics [17] and the fact that CT images are generated based upon a calibration to water using Hounsfield Units is an advantage in terms of correlating with underlying biology processes, compared to modalities which lack such calibration [5].

1.6. Radiomics: Biological basis.

One of the most important goals in radiomics is linkage between the novel biomarker(s) and an underlying biological process [35], for example a particular radiomic features (or signature of features grouped together) may be associated with improved survival for patients with pancreatic cancer, but we would like to know if there is a biological meaning of this feature, perhaps correlating to a particular genetic mutation in the tumour which could be targeted for therapy [2]. Prior to the advent of radiomics, two relevant papers in this area by Segal *et al* in 2007 [36] and Diehn *et al* in 2008 [37] identified significant associations between the imaging appearances of tumours and the underlying genetic oncologic mutations, although in contrast to radiomics, they used humans (rather than computer vision) to identify sematic features on the image, such as the degree of contrast enhancement or the presence of necrosis. For example, in the Diehn *et al* study,

they identified an association between brain tumour contrast enhancement and the expression of genes related to angiogenesis and tumour hypoxia (VEGF, ADM etc.)[37].

Following the advent of radiomics circa 2012, numerous associations between radiomic signatures and tumour biology have been identified across a range of disease states including lung [38], brain [39], breast [40], liver [41] and pancreatic cancer [42, 43]. Several studies have defined radiomic models which can predict patient survival and subsequently identified the underlying biological meaning of this prediction, for example a study in Non-small cell lung cancer in which a clinical-radiomic model was developed to predict survival, was found to correlate with hypoxia-related carbonic anhydrase, a glycoprotein induced by hypoxia [44]. Conversely, some studies have started out with a biological attribute known to influence survival and then created a radiomic model to predict this biological characteristic. This approach was taken in a study of glioblastoma, where they created a model to predict tissue hypoxia. They then confirmed that this radiomic signature could predict survival [45]. Some authors have even attempted to create artificial CT images from digital pathology slides, in order to explore the biological connection between imaging appearance and histology, with moderate success [46]. Such biological correlation can provide confidence in the results of radiomics studies [35]. In addition, the potential for non-invasive identification of biological attributes may be helpful in novel targeted drug trials in future.

1.7. Radiomics: Weaknesses.

"Radiomics and ML are still in their infancy for healthcare and often not yet ready for use in daily clinical practice"

Cuocolo and Imbriaco 2021 [22]

Despite considerable success in radiomics research over the past decade, there remain several weaknesses within this field which need to be solved before the 'translation gap' between research and routine clinical practice can be narrowed [47, 48]. Two 2020 reviews of publication quality in radiomics concluded that studies are generally of moderate-to-poor quality [17, 49] and these reviews identified issues ranging from poor reporting of study objective in the manuscripts, to more significant issues such as lack of external validation [22]. There are specific factors which makes it difficult to perform radiomic studies consistently [2] and I will address some major issues below:

1.7.1. Lack of standardization in CT protocols.

Large volumes of quality input data are key for radiomics research [2, 50] but the lack of standardization in the acquisition parameters used to acquire medical imaging causes problems [51]. Ideally, standard-of-care imaging should be used in radiomics, rather than controlled experimental data, in order to ensure the methods are widely applicable to other centres however this means that technical variation is common [52]. The utility of radiomic models to analyze medical images is sensitive to multiple technical imaging acquisition parameters [1,2][51]. The different appearances of a tissue depending upon these parameters do not typically affect the qualitative interpretation of clinical radiology examinations by experienced radiologists [10] but they can cause significant issues for interpretation by artificial intelligence systems. For example, altering the pitch factor or reconstruction kernel when acquiring CT images of a phantom can significantly reduce the reproducibility of radiomics features values in a test-retest analysis [53].

Several normalization methods have been proposed to address this issue, including grey scale discretization [54], denoising filters [55] and methods to compensate for different CT reconstruction kernels [56] or slice thickness [57] using convolutional neural networks (deep learning). Some groups have experimented with general adversarial networks for CT image normalisation [58]. One of the most promising methods is a batch-effect correction method called 'ComBat', which was originally developed to adjust for the batch-effect in genetics studies (i.e., taking into account that the data can be grouped into batches such as the technician who performed the experiment or the CT scanner used, to control for these variables). This method has been shown to successfully control for CT technical parameters in phantom studies and it is easy to use, since it is available as a package for the statistics program 'R' [59]. Multiple radiomic studies have found that Combat improved their results [60-62]. Such normalization techniques are not required for deep learning approaches, since they make these adjustments automatically [10].

Inter-observer variation in tumour contouring is another factor which is often raised in the literature as a potential confounder [3], however differences in scan parameters have been shown to have a larger impact on radiomic features values than differences in segmentation technique [63].

1.7.2. Lack of standardization in Radiomic software and features.

There are at least 14 different software packages available to perform radiomic analysis [12, 31, 64] and they vary in terms of feature definitions, feature name and extraction parameters (for example gray value discretization bin width). Therefore, the results from many early studies in this field are difficult to interpret. It is now recommended that standardized software packages are used [17] and considerable progress has been made in this regard since the formation of the Imaging Biomarker Standardization Initiative (IBSI) and the 2020 publication of a standardization paper by Zwanenburg *et al* [31]. This paper was available in pre-print form on arxiv.org since 2016. They defined a set of radiomics features, which are available in an online

library [32] and a recent study confirmed that compliance with the IBSI standards lead to the best reproducibility in radiomics features values [64].

1.8. Radiomics: Statistics, feature standardization and feature reduction.

The most commonly reported metric in studies which develop and/or validate risk prediction models is discrimination, also referred to as accuracy. This is defined as the ability of a model to separate individuals into groups with or without the disease/event (i.e., the probability that of two individuals, one with the disease/event and the other without the disease/event, will be correctly separated into high and low risk groups by the model) [65]. Discrimination depends on how much the predictors vary i.e., if there is minimal variance in predictor variables, it will be challenging to discriminate between cases [20]. A popular way to report discrimination is using the area under the curve (AUC) from a received operator curve analysis. This metric is typically used for binary outcomes [20, 66]. In a time-to-event analysis, discrimination is the probability that the model can identify which patient will experience the event (for example death or disease recurrence) sooner and concordance Index (C-Index) is a commonly used rank-order statistic used for this purpose. It is a development of AUC, adapted for time-to-event analysis. C-index has values ranging from 0.5-1, with a value of 0.5 indicating that the model performs no better than random chance at the given prediction task [67].

It has been suggested that AI studies suffer from an 'AI chasm' where they rely too much on reporting discrimination, whereas this is not the sole determinate of model performance [16]. Calibration is another measure which should always be reported in model development studies. This is defined as the agreement between the observed outcomes and predicted outcomes [68] (or between actual and predicted probabilities [67]). Discrimination is not affected by calibration. For example, two patients (A and B) may have probability of disease recurrence at one year of 0.2 and 0.8, respectively. A model may have good discrimination (i.e. it may correctly

identify that patient B is high risk and patient A is low risk), however if it is miscalibrated, it may report incorrect predicted probabilities, for example 0.02 and 0.08 for patient A and B respectively. You can see from this example that discrimination is not affected by calibration (i.e. patient B still has 4X the risk of patient A), but a doctor discussing the prognosis with patient B will grossely underestime the risk of disease recurrence as 0.08 rather than 0.8. Calibration depends both upon the algorithm itself and the population upon which it is being applied (i.e. demonstration of good calibration in one external dataset does not mean it will be well calibrated in all external datasets) [20]. One useful aspect of calibration is the models can be 'recalibrated' to a new population with relative ease, simply by adding an additional term to the regression equation.

In addition to discrimination and calibration, assessment of the clinical utility of predictive models has been advocated by the biostatistics community over the past 10-15 years [69]. When measuring clinical utility, efforts are undertaken to quantify the relative harm of false negative and false positive results, since in some clinical scenarios, a false negative results may be more harmful to a patient than a false positive (or visa versa) [70]. A popular method of assessing clinical utility is to report net benefit, which is a measure of true positives penalised for false negatives. This can be assessed over a range of risk thresholds using a method called decision curve analysis (DCA), which was developed by Vickers *et al* in 2006 [71]. The risk threshold is defined as 'the minimum probability of disease at which further intervention would be warranted'[70]. An example of a risk threshold would be: if a patient or clinician was asked at what cancer risk would they decide to perfrom a biopsy; if the biopsy had minimal side effects, they may decide to perfom it even at a low risk level (<5% predicted risk of cancer), but if the biopsy had higher side effects, they may decide not to biopsy it unless the patient's predicted risk was higher (>10%, >15% or >20%). In practical terms, it is difficult to pinpoint a precise threshold at which the patient/clinician will make the decision, therefore the decision curves are typically plotted for a clinically relevant range of thresolds, for example from 1-40%.

DCA was initially developed for binary outcomes, such as cancer vs no cancer on prostate biopsy, but has since been adapted for time-to-event analysis, where true and false positives are compared at a specific time point (for example 3 or 5 year survival) [70]. It has been acknowledged by the developers of DCA that it is not as intuitive to understand as a measure of discrimination [72], even to expert epidemiologists, however it has very popular in the medical literature and many of the of the highest ranked journals in medicine now endorse the use of DCA, including JAMA and the BMJ [72]. It must also be highlighted that like any statistical test, DCA can show over-optimistic results when internal validation methods are used, as opposed to true external validation [70] (discussed more in section 1.3. Data for AI: Training and validation).

Since recent guidelines advise the use of several metrics in model evaluation [2], it has been suggested that the level of detail which needs to be provided in the methodology section of an AI / radiomics study is more than would typically fit into a research manuscript, which has resulted in the increasing use of supplementary materials in such publications [73].

1.8.1. Feature standardization.

When radiomic features are extracted from an image, their values will range across very different scales, for example the feature 'original_shape_Elongation' may have values ranging from 0.27-0.99, while the feature 'original_firstorder_Energy' may range from 36,575-40,150,470. Therefore, a transformation is required to standardise the dataset prior to analysis and to ensure that results are comparable to other studies. Options for standardization include z-score, min-max and the whitening transformation from principal component analysis. All of these methods have been shown to improve AUC, sensitivity and accuracy (all of which are related metrics), however no one method emerged as the clear 'best' method [74]. Z-score is the most common used in the literature to date [75-78] and is therefore the methods which we have selected for our work. This method involves converting the actual features values into standard deviations from the mean, i.e. each features

value will be converted into a number which will indicate how many standard deviations above or below the mean it is.

1.8.2. Feature reduction

Radiomics is a type of 'high dimensional' data, meaning that there are a high number of predictors, often > 500, which is typically more than the total number of patients included in the study. This type of data is common to all 'omics' research fields including radiomics, genomics, proteomics and metabolomics, some of which can generate millions of variables to analyse. Therefore, complex statistical strategies are required to 'reduce' the features to a small set relevant for the prediction task at hand. This is often referred to a dimensionality reduction. Inclusion of a feature reduction step in radiomics analysis is advised by guidelines in order to reduce overfitting [2] and may be performed using supervised methods, for example least absolute shrinkage and selection operator (LASSO) regression, or unsupervised methods including principal component analysis. These methods generally function well to prevent overfitting, provided that the study population size is sufficiently large [79].

Multiple radiomic features may be highly correlated with each other, since many of them are minor derivations of each other, especially if filters are used (i.e., the square root etc.). This must be taken into account in the statistical analysis in radiomic studies, since many of the features reduction methods, including LASSO, do not address this issue. A major study in radiomics (by Aerts *et al*) successfully created a radiomic signature for survival prediction in both head and neck cancer and lung cancer [73] and this worked well despite a large variation in CT parameters / protocols, however a subsequent analysis identified that the four features included in the signature were highly correlated with tumour size, hence explaining the robustness of the signature [80]. This is a notable result, since tumour size is already known to influence patient survival and it has been reported by radiologists for decades. If radiomics is to have an additive impact upon survival prediction, it will

need to be independent of such pre-existing radiological attributes such as tumour size.

1.9. Chapter 1 references.

1. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data.* Radiology, 2016. **278**(2): p. 563-77.
2. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nat Rev Clin Oncol, 2017. **14**(12): p. 749-762.
3. Collins, F.S. and H. Varmus, *A New Initiative on Precision Medicine.* New England Journal of Medicine, 2015. **372**(9): p. 793-795.
4. May, M., *Eight ways machine learning is assisting medicine.* Nat Med, 2021. **27**(1): p. 2-3.
5. Rogers, W., et al., *Radiomics: from qualitative to quantitative imaging.* Br J Radiol, 2020. **93**(1108): p. 20190948.
6. Andresen, S.L., *John McCarthy: father of AI.* IEEE Intelligent Systems, 2002. **17**(5): p. 84-85.
7. *Oxford English Dictionary*. 2012: Oxford University Press.
8. Hale, A.T., et al., *Machine learning analyses can differentiate meningioma grade by features on magnetic resonance imaging.* Neurosurg Focus, 2018. **45**(5): p. E4.
9. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* Eur J Cancer, 2012. **48**(4): p. 441-6.
10. Mühlberg, A., et al., *The Technome - A Predictive Internal Calibration Approach for Quantitative Imaging Biomarker Research.* Sci Rep, 2020. **10**(1): p. 1103.
11. Chu, L.C., et al., *Utility of CT Radiomics Features in Differentiation of Pancreatic Ductal Adenocarcinoma From Normal Pancreatic Tissue.* AJR Am J Roentgenol, 2019. **213**(2): p. 349-357.
12. Chu, L.C., et al., *Diagnostic performance of commercially available vs. in-house radiomics software in classification of CT images from patients with pancreatic ductal adenocarcinoma vs. healthy controls.* Abdom Radiol (NY), 2020. **45**(8): p. 2469-2475.
13. E, L., et al., *Differentiation of Focal-Type Autoimmune Pancreatitis From Pancreatic Ductal Adenocarcinoma Using Radiomics Based on Multiphasic Computed Tomography.* J Comput Assist Tomogr, 2020. **44**(4): p. 511-518.
14. Li, K., et al., *Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma.* Am J Transl Res, 2019. **11**(7): p. 4491-4499.
15. Zhang, Y., et al., *Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images.* Sci Rep, 2021. **11**(1): p. 1378.
16. Kelly, C.J., et al., *Key challenges for delivering clinical impact with artificial intelligence.* BMC Med, 2019. **17**(1): p. 195.
17. Abunahel, B.M., et al., *Pancreas image mining: a systematic review of radiomics.* Eur Radiol, 2020. **31**(5): p. 3447-3467.
18. Gerke, S., T. Minssen, and G. Cohen, *Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare*, in *Artificial Intelligence in Healthcare*, A. Bohr and K. Memarzadeh, Editors. 2020, Elsevier. p. 295-336.

19. Adibuzzaman, M., et al., *Big data in healthcare - the promises, challenges and opportunities from a research perspective: A case study with a model database.* AMIA ... Annual Symposium proceedings. AMIA Symposium, 2018. **2017**: p. 384-392.
20. Vickers, A.J. and A.M. Cronin, *Everything you always wanted to know about evaluating prediction models (but were too afraid to ask).* Urology, 2010. **76**(6): p. 1298-301.
21. Kim, D.W., et al., *Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers.* Korean J Radiol, 2019. **20**(3): p. 405-410.
22. Cuocolo, R. and M. Imbriaco, *Machine learning solutions in radiology: does the emperor have no clothes?* Eur Radiol, 2021. **31**(6): p. 3783-3785.
23. Van Calster, B., et al., *Methodology over metrics: Current scientific standards are a disservice to patients and society.* J Clin Epidemiol, 2021.
24. Halligan, S., Y. Menu, and S. Mallett, *Why did European Radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting.* Eur Radiol, 2021. **31**(12): p. 9361-9368.
25. Zhang, C., et al., *Prediction of lymph node metastases using pre-treatment PET radiomics of the primary tumour in esophageal adenocarcinoma: an external validation study.* Br J Radiol, 2021. **94**(1118): p. 20201042.
26. Larue, R., et al., *Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer.* Acta Oncol, 2018. **57**(11): p. 1475-1481.
27. Beede, E., et al., *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy*, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020.
28. Esbensen, K. and P. Geladi, *Principles of Proper Validation: use and abuse of re-sampling for validation.* Journal of Chemometrics, 2010. **24**(3-4).
29. Topol, E., *Deep Medicine: How Artificial Intelligence can make healthcare Human again*. 2019: Basic Books.
30. Haralick, R.M., K. Shanmugam, and I. Dinstein, *Textural Features for Image Classification.* IEEE Transactions on Systems, Man, and Cybernetics, 1973. **SMC-3**(6): p. 610-621.
31. Zwanenburg, A., et al., *The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping.* Radiology, 2020. **295**(2): p. 328-338.
32. *The image biomarker standardisation initiative*. Available from: https://ibsi.readthedocs.io/en/latest/index.html.
33. van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.
34. Cropp, R.J., et al., *Scanner and kVp dependence of measured CT numbers in the ACR CT phantom.* J Appl Clin Med Phys, 2013. **14**(6): p. 4417.
35. Tomaszewski, M.R. and R.J. Gillies, *The Biological Meaning of Radiomic Features.* Radiology, 2021. **299**(2): p. E256.
36. Segal, E., et al., *Decoding global gene expression programs in liver cancer by noninvasive imaging.* Nat Biotechnol, 2007. **25**(6): p. 675-80.
37. Diehn, M., et al., *Identification of noninvasive imaging surrogates for brain tumor gene-expression modules.* Proc Natl Acad Sci U S A, 2008. **105**(13): p. 5213-8.

38. Grossmann, P., et al., *Defining the biological basis of radiomic phenotypes in lung cancer.* Elife, 2017. **6**: p. e23421.
39. Zhang, X., et al., *Optimizing a machine learning based glioma grading system using multi-parametric MRI histogram and texture features.* Oncotarget, 2017. **8**(29): p. 47816-47830.
40. Li, H., et al., *Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set.* NPJ Breast Cancer, 2016. **2**: p. 16012-.
41. Xu, X., et al., *Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma.* J Hepatol, 2019. **70**(6): p. 1133-1144.
42. Attiyeh, M.A., et al., *CT radiomics associations with genotype and stromal content in pancreatic ductal adenocarcinoma.* Abdom Radiol (NY), 2019. **44**(9): p. 3148-3157.
43. Kaissis, G., et al., *Image-Based Molecular Phenotyping of Pancreatic Ductal Adenocarcinoma.* Journal of Clinical Medicine, 2020. **9**(3): p. 724.
44. Tunali, I., et al., *Hypoxia-related radiomics predict immunotherapy response: A multi-cohort study of NSCLC.* bioRxiv, 2020: p. 2020.04.02.020859.
45. Beig, N., et al., *Radiogenomic analysis of hypoxia pathway is predictive of overall survival in Glioblastoma.* Sci Rep, 2018. **8**(1): p. 7.
46. Geady, C., et al., *Bridging the gap between micro- and macro-scales in medical imaging with textural analysis - A biological basis for CT radiomics classifiers?* Phys Med, 2020. **72**: p. 142-151.
47. Pinto Dos Santos, D., M. Dietzel, and B. Baessler, *A decade of radiomics research: are images really data or just patterns in the noise?* Eur Radiol, 2020.
48. Yip, S.S. and H.J. Aerts, *Applications and limitations of radiomics.* Phys Med Biol, 2016. **61**(13): p. R150-66.
49. Park, J.E., et al., *Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement.* European Radiology, 2020. **30**(1): p. 523-536.
50. Rizzo, S., et al., *Radiomics: the facts and the challenges of image analysis.* Eur Radiol Exp, 2018. **2**(1): p. 36.
51. He, L., et al., *Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule.* Sci Rep, 2016. **6**: p. 34921.
52. Fournier, L., et al., *Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers.* Eur Radiol, 2021. **31**(8): p. 6001-6012.
53. Berenguer, R., et al., *Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters.* Radiology, 2018. **288**(2): p. 407-415.
54. Larue, R., et al., *Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study.* Acta Oncol, 2017. **56**(11): p. 1544-1553.
55. Fernández Patón, M., et al., *MR Denoising Increases Radiomic Biomarker Precision and Reproducibility in Oncologic Imaging.* J Digit Imaging, 2021. **34**(5): p. 1134-1145.

56. Choe, J., et al., *Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses.* Radiology, 2019. **292**(2): p. 365-373.
57. Park, S., et al., *Deep Learning Algorithm for Reducing CT Slice Thickness: Effect on Reproducibility of Radiomic Features in Lung Cancer.* Korean J Radiol, 2019. **20**(10): p. 1431-1440.
58. Wei, L., Y. Lin, and W. Hsu, *Using a Generative Adversarial Network for CT Normalization and its Impact on Radiomic Features.* arXiv, 2020. **arXiv:2001.08741**.
59. Orlhac, F., et al., *Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics.* Radiology, 2019. **291**(1): p. 53-59.
60. Masson, I., et al., *Statistical harmonization can improve the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers.* Med Phys, 2021. **48**(7): p. 4099-4109.
61. Hotta, M., et al., *Prognostic value of (18)F-FDG PET/CT with texture analysis in patients with rectal cancer treated by surgery.* Ann Nucl Med, 2021.
62. Ligero, M., et al., *Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis.* Eur Radiol, 2021. **31**(3): p. 1460-1470.
63. Yamashita, R., et al., *Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation.* Eur Radiol, 2020. **30**(1): p. 195-205.
64. Fornacon-Wood, I., et al., *Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform.* Eur Radiol, 2020. **30**: p. 6241-6250.
65. Alba, A.C., et al., *Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature.* Jama, 2017. **318**(14): p. 1377-1384.
66. Pencina, M.J., R.B. D'Agostino, and J.M. Massaro, *Understanding increments in model performance metrics.* Lifetime Data Anal, 2013. **19**(2): p. 202-18.
67. Caetano, S.J., G. Sonpavde, and G.R. Pond, *C-statistic: A brief explanation of its construction, interpretation and limitations.* Eur J Cancer, 2018. **90**: p. 130-132.
68. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.
69. Steyerberg, E.W., et al., *Assessing the incremental value of diagnostic and prognostic markers: a review and illustration.* Eur J Clin Invest, 2012. **42**(2): p. 216-28.
70. Vickers, A.J., et al., *Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers.* BMC Med Inform Decis Mak, 2008. **8**: p. 53.
71. Vickers, A.J. and E.B. Elkin, *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.* Medical Decision Making, 2006. **26**(6): p. 565-574.
72. Vickers, A.J., B. van Calster, and E.W. Steyerberg, *A simple, step-by-step guide to interpreting decision curve analysis.* Diagnostic and Prognostic Research, 2019. **3**(1): p. 18.
73. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. **5**: p. 4006.
74. Haga, A., et al., *Standardization of imaging features for radiomics analysis.* J Med Invest, 2019. **66**(1.2): p. 35-37.

75. Chang, N., et al., *Development and multicenter validation of a CT-based radiomics signature for discriminating histological grades of pancreatic ductal adenocarcinoma.* Quant Imaging Med Surg, 2020. **10**(3): p. 692-702.
76. Li, K., et al., *Contrast-enhanced CT radiomics for predicting lymph node metastasis in pancreatic ductal adenocarcinoma: a pilot study.* Cancer Imaging, 2020. **20**(1): p. 12.
77. Chen, F., et al., *Radiomics-Assisted Presurgical Prediction for Surgical Portal Vein-Superior Mesenteric Vein Invasion in Pancreatic Ductal Adenocarcinoma.* Frontiers in Oncology, 2020.
78. Reinert, C.P., et al., *Complementary role of computed tomography texture analysis for differentiation of pancreatic ductal adenocarcinoma from pancreatic neuroendocrine tumors in the portal-venous enhancement phase.* Abdom Radiol (NY), 2020. **45**(3): p. 750-758.
79. Riley, R.D., et al., *Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small.* J Clin Epidemiol, 2021. **132**: p. 88-96.
80. Vallieres, M., D. Visvikis, and M. Hatt, *Dependency of a validated radiomics signature on tumor volume and potential corrections.* Journal of Nuclear Medicine, 2018. **59**(supplement 1): p. 640-640.

Chapter 2 - Introduction to Pancreatic Ductal Adenocarcinoma (PDAC)

2.1. Epidemiology of PDAC.

Pancreatic ductal adenocarcinoma (PDAC) is the most common cancer arising in the pancreas. It is an aggressive disease which is often quite advanced at the time of diagnosis and it has a high mortality-to-incidence ratio [1]. The median age at diagnosis is 72 and it is more common in men (incidence 11.9 cases per 100,000 per year) compared to women (8.7 cases per 100,000 per year)[2]. In the Republic of Ireland (henceforward referred to as 'Ireland'), it is the 9th most commonly diagnosed cancer, with an incidence of 11.64 per 100,000 population per year [2], compared to 13.5 per 100,000 in Canada [3]. Known risk factors for PDAC include advanced age, male sex, smoking, obesity, alcohol consumption and chronic pancreatitis [4, 5]. Current treatments are poor and the five year survival rates are 8.2% in Ireland [2]) and 8% in Canada [3]. Worldwide it is the 7th leading cause of cancer related deaths [3], ranking 5th in Ireland [2] and 4th in Canada [6], accounting for approximately 6% of all cancer related deaths in both countries. Treatment outcomes have improved slowly over the past 40 years (5-year survival has increased from 3.1% to 8% during this period [7]) and this is mainly attributed to developments in adjuvant chemotherapy [8].

2.2. Diagnosis of PDAC.

Pancreatic ductal adenocarcinoma is commonly asymptomatic in the early stages and this means that the disease is typically quite advanced at the time of diagnosis, for the majority of patients [9]. Symptoms tend to occur late in the disease and these include abdominal pain, back pain, nausea, yellow tinged skin/eyes (jaundice) and weight loss [10]. Less common presentations include gastrointestinal obstruction or bleeding.

In all patients with suspected PDAC, radiological imaging is the central component of investigation [11] (discussed in detail in the next section) and this may be supported by measurement of serum Ca19-9, a blood biomarker which is often elevated in PDAC. This is the only biomarker approved for use in PDAC by the United States Food and Drug Administration [4, 12]. However, the ultimate diagnosis of PDAC must be confirmed on pathology. This pathological confirmation should ideally be obtained prior to surgical resection, in order to rule out mimics of PDAC, such as autoimmune pancreatitis, however guidelines allow patients to proceed to surgery without pathology proof of diagnosis if the imaging is convincing and attempts at biopsy are not successful [10, 12, 13]. In patients who are not suitable for surgical resection, pathological diagnosis is mandatory prior to chemo/radiotherapy [10]. For patients with resectable disease (i.e., localised to the pancreas), sampling of the mass is usually performed by endoscopic fine needle aspiration or core biopsy. In patients with metastatic disease, the diagnosis is often made from percutaneous core biopsy of a metastatic deposit in the liver or peritoneum.

Screening of asymptomatic patients has been successful in some types of cancer, including breast, cervix and colon. Screening with radiological imaging has been investigated in PDAC, however it has not been proven beneficial in terms or mortality or morbidity [4, 14]. However, there is evidence for the use of CT or MRI in screening for PDAC of high-risk individuals, such as those with Peutz–Jeghers syndrome or patients with a BRCA 1 or 2 mutation [4, 12, 14]. Ca19-9 does not have a role in screening for this malignancy, due to the low positive predictive value [4]. Hence, screening is not recommended in asymptomatic individuals [14] and there are no formal screening programs for PDAC in any country.

2.3. Imaging of PDAC.

Radiology serves many purposes for patients with pancreatic ductal adenocarcinoma including diagnosis, assessment of resectability, staging, assessment of response to therapy and follow up post treatment. PDAC lesions may be visualised using a variety of non-invasive imaging modalities including ultrasound, computed tomography (CT), magnetic resonance imaging (MRI) and nuclear medicine studies with as Positron Emission Tomography (PET) [15]. However, contrast enhanced CT is the workhorse of PDAC imaging and is the recommended modality for assessment of patients with suspected or known PDAC due to its wide availability, high spatial resolution, rapid scanning time and excellent patient tolerance [10, 12, 13, 16].

2.3.1. CT pancreas.

CT of the pancreas is performed with a biphasic protocol, including arterial and venous phases. This protocol typically uses 500-700 mls of water as a negative oral contrast agent and acquires a late arterial (pancreatic) phase and a portal venous phase at 35-40 and 65-70 seconds respectively following injection of 150 mls non-ionic iodinated contrast at a rate of approximately 4mls/second. The pancreas is included in the arterial phase, while the entire abdomen is included in the portal venous phase. This protocol provides rapid and accurate local and distant staging of the tumour in a single scan. A newer 'split bolus' protocol, which acquires arterial and venous phase imaging in a single scan, is available and there is evidence that it gives similar results to biphasic protocol, but with less radiation [16].

The high spatial resolution afforded by modern multi-detector CT scanners with thin slice multi-planer reconstruction is ideal for local staging, which depends upon very strict criteria for tumour contact with local structures (for example, one of the criteria for resectability is ≤ 180 degrees of tumour contact with the portal vein, without vein contour irregularity [13]). Based upon this imaging, decision will be made to (1) treat the patient with upfront surgery (resectable patients), (2) treat the

patient with neoadjuvant chemotherapy +/- surgery (borderline resectable patients) (3) treat with chemo/radiotherapy alone (unresectable patients), (4) perform additional imaging for cases which are indeterminate on CT or (5) manage with supportive/palliative care approach. The criteria for resectability are further discussed in section 2.5. 'Treatment of PDAC'.

On CT, a typical PDAC lesion is hypoattenuating relative to the surrounding pancreas on the portal venous phase (Figure 1.) and this it due to reduced perfusion of intravenous contrast into the dense fibrous tumour compared to the surrounding tissue [15, 17]. There is commonly upstream pancreatic duct obstruction and/or distal pancreatic atrophy, which may sometimes be the only signs of disease, which is especially important in small isointense PDAC lesions which are not easily visible on CT [15]. Depending upon the location of the lesion, there may also be obstruction of the common bile duct, often resulting in the clinical presentation of jaundice and the radiological presentation of the 'double duct sign'.

Figure 1. Examples of pancreatic adenocarcinoma on contrast enhanced CT.

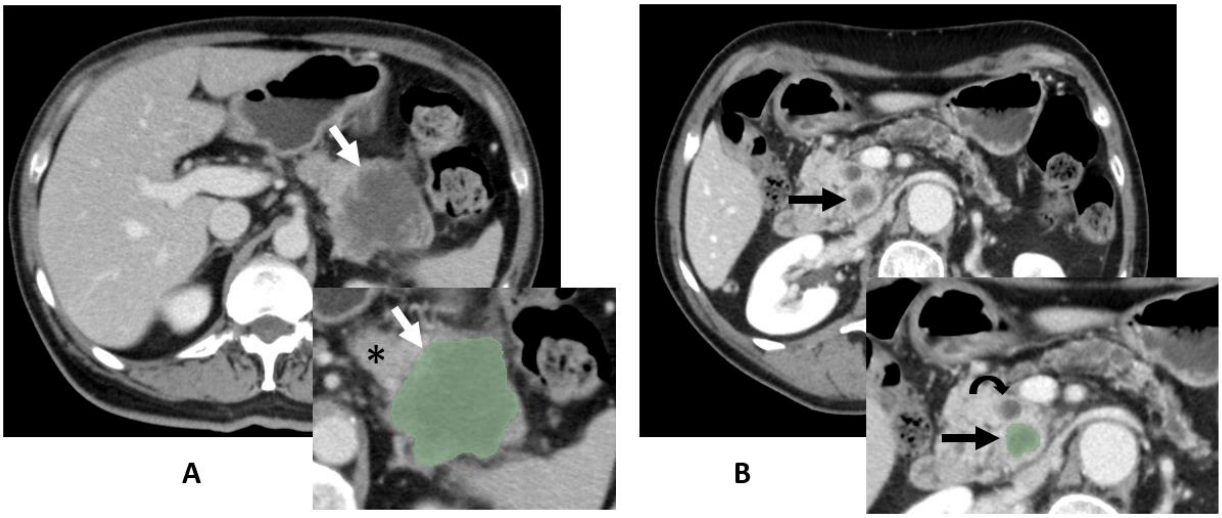


Figure 1 caption: Case examples from our study cohort. Axial slice images from contrast enhanced CT pancreas studies in the portal venous phase are presented, in patients with biopsy proven pancreatic ductal adenocarcinoma. In image A, this patient has a hypoattenuating 6cm pancreatic tail mass (white arrow). The insert image shows this lesion after segmentation. The normal pancreas is indicated by the black star (*). Image B is from a patient with a 2.5cm hypoattenuating pancreatic head mass (black arrow). The insert image shows this lesion post segmentation. The curves black arrow in the insert image points to the dilated common bile duct, located anterior to the mass on this axial slice.

2.3.2. Timing of Pre-operative CT pancreas.

The timing of pre-operative CT has been the subject of several studies [18-20]. In many patients with PDAC, there can be a delay from date of diagnosis to date of surgery because patients may require biliary drainage for cholangitis, while others may require more than one attempt at endoscopic sampling to achieve a conclusive pre-operative pathological diagnosis. Hence, this can delay surgery. If there is disease progression during this time interval, it can result in an aborted attempt at surgical resection if the surgeon identifies unexpected abdominal metastasis at the time of curative intent resection (so called 'open-and-close laparotomy). The rate of

such 'unexpected progression' is a key performance indicator in hepatobiliary surgery centers, which they try to minimise as much as possible. Raman *et al* from Johns Hopkins hospital in Baltimore, USA performed a study on this topic in 2015 and they identified that 25 days should be the maximum imaging-to-surgery interval (ISI) in order to avoid unexpected metastasis at time of curative intent surgery [18]. However, while prior work by our group confirmed this finding (patients with an ISI ≥ 25 days experienced unexpected disease progression at time of attempted resection in 17% of cases, vs. 6% in those with ISI < 25 days) we also found that the ISI does not influence overall patient survival [19]. The relationship is complex and there are indications that survival is likely dictated more by underlying disease biology, rather than the ISI [20]. There is some evidence that a longer ISI may actually prolong survival in those who undergo resection, since the patients with more aggressive biology are 'filtered' out by the longer waiting time [19].

There is certainly variation in the ISI internationally, with publications from Johns Hopkins stating that they now aim for an ISI of 14 days or less [18], whereas the waiting time tends to be longer in countries with publicly funded health systems, such as a study from Sweden looking at patients treated between 2008-2014 which reported a median ISI of 42 days (range 10 to 159) [21] and our prior work from Ireland reported a median ISI 32.5 days (IQR 35, range 0-254) [19]. There is no recommendation on this topic in international PDAC guidelines [13] and the majority of prior studies in PDAC CT Radiomic prognostication did not even report this metric [22-28]; two prior studies limited ISI to four weeks [29, 30], one reported mean ISI one month [31] and another reported '95% of resections ≤ 6 months from imaging' [32]. This is why we included the variable of image-to-surgery time interval (ISI) within the pre-operative clinical prognostic model for our study (discussed in chapter 4 – Hypothesis, Materials and Methods), in order to account for any difference between patients.

2.3.3. MRI pancreas.

MRI pancreas has comparable sensitivity and specificity to CT for the diagnosis and local staging of PDAC [15], but it is more expensive, the scans take longer, the availability of MRI is limited compared to CT and some patients cannot have MRIs due to issues including pacemakers and claustrophobia [13]. Therefore, guidelines advise that MRI is reserved for problem solving [15, 16], such as in patients with suspected PDAC who have no definite mass visible on CT, where MRI can pick up lesions which are CT isointense. MRI pancreas is also indicated in patients with an allergy to iodinated CT contrast [13]. There is evidence that routine use of MRI liver has superior diagnostic performance compared to CT in detecting liver metastasis in patients awaiting resection [33], however this has resource implications for many institutions where MRI slots are limited, so it has not been incorporated into practice guidelines [13].

2.3.4. Endoscopic Ultrasound and PET-CT

Endoscopic ultrasound (EUS) of the pancreas is more invasive compared to MRI or CT. Therefore, it is not typically used as an initial diagnostic test, but it is commonly used to guided fine needle aspiration or biopsy of a suspicious pancreatic mass or duct stricture. It can also be used for problem solving in cases where CT and MRI have not identified a definite mass. PET-CT has been shown in some studies to have high sensitivity for PDAC diagnosis and staging, however the evidence is not conclusive [15] and therefore, it has not been incorporated into routine pre-operative imaging algorithms in clinical practice. The 2019 National Comprehensive Cancer Network (NCCN) guidelines state that the role of PET-CT 'remains unclear' and that it is 'not a substitute for high-quality, contrast-enhanced CT' [13].

Considering the clear emphasis on contrast enhanced CT in the pre-operative workup of patients with PDAC, this modality was chosen for the present study.

2.4. Biology of PDAC and assessment with radiomics.

PDAC is a malignancy of the exocrine pancreas ductal epithelium, which is thought to arise from precursor lesions called pancreatic intraepithelial neoplasia. There are recognized modifiable risk factors for pancreatic cancer, such as obesity and alcohol use, however studies indicate that ~50% of PDAC cases begin with stochastic (random) errors which occur during DNA replication as part of normal cell division [5]. These cells than accumulate further mutations in a stepwise fashion, however there are four 'founder' mutations which are predominant: KRAS [in 90% of cases], CDKN2A, TP53 and SMAD4. These mutations facilitate the development of clonal expansion and invasion into adjacent pancreatic tissue [5]. Multiple additional mutations, including epigenetic alternations, are described in PDAC, however they are less frequent than the four founder mutations listed above [5].

When the tumour cells starts to invade adjacent pancreas tissue, a healing response in the surrounding pancreatic stroma will be initiated, but this process is manipulated by signals from the cancer cells which induce the supporting stromal tissues to promote further tumour expansion (i.e. continuous proliferative signals maintain activation of stromal mesenchymal cells long after the normal healing response would have finished) [34]. Hence, PDAC is characterised pathologically by a dense fibroblastic stroma [10] which consists of extracellular matrix, stromal vasculature and cancer-associated fibroblasts [35], all of which are recognised as a key factors in the pathogenesis of this disease [35]. Immunomodulation also plays a role in tumorigenesis, with suppression of T-cell function within the developing tumour mass [5]. Once these factors combine for form a PDAC mass, the final step in the process is development of metastasis. There is evidence that metastasis occurs early in PDAC [9] and it does not appear to require specific genetic mutations, rather the gain of function / loss of suppression provided by the four main founder mutations in PDAC are enough to facilitate this process, along with epigenetic modifications [5, 9].

Information about the biology of a PDAC lesion can be inferred from the appearance on pre-operative imaging. Radiomic models based on pre-operative imaging have been developed to predict tumour grade [30], molecular subtypes (quasi-mesenchymal vs. non-quasi-mesenchymal) [36] and p53 status [37], although none of these models are externally validated and hence, they have not yet been appropriately tested. Some of these studies do not report the actual features included in their models, such as the work by *Iwatate et al* who aimed to predict p53 and PD-L1 expression in PDAC using 1037 features extracted from original and expanded volumes of interest (hence 2074 total features) but did not report which were selected for the machine learning model [37]. Nonetheless, from overall review of the literature, it appears that two categories of Computed Tomography (CT) tumour characteristics are relevant to this work: tumour attenuation and heterogeneity.

2.4.1 Attenuation and tumour biology.

Well-differentiated PDAC tumours are associated with better survival and they are more commonly iso-attenuating (and hence difficult to see) on contrast enhanced CT imaging [15] compared to hypoattenuating tumours, which are associated with worse survival and earlier recurrence[38]. The reason for this association between tumor attenuation and biology has been investigated in several studies, most notably in a seminal 2014 paper from the MD Anderson Cancer Centre entitled 'Transport properties of pancreatic cancer describe gemcitabine delivery and response' [17]. In this study, the authors demonstrated that the volume of stromal tissue within the tumour correlated negatively with tumour enhancement on CT, hence the appearance on routine CT could be used to estimate the percentage of stromal tissue within the lesion. In this study, they also intravenously infused chemotherapy (gemcitabine) for 12 patients during their curative intent resection procedure, in order to assess how much of the infused chemotherapy was being delivered to the tumour. When they assessed the tumours in the pathology lab, they found that tumor gemcitabine incorporation was related to the presence of stromal

tissue and hence, could be predicted by tumour enhancement on the pre-procedure CT.

Figure 2. Attenuation of PDAC lesions. Case examples.

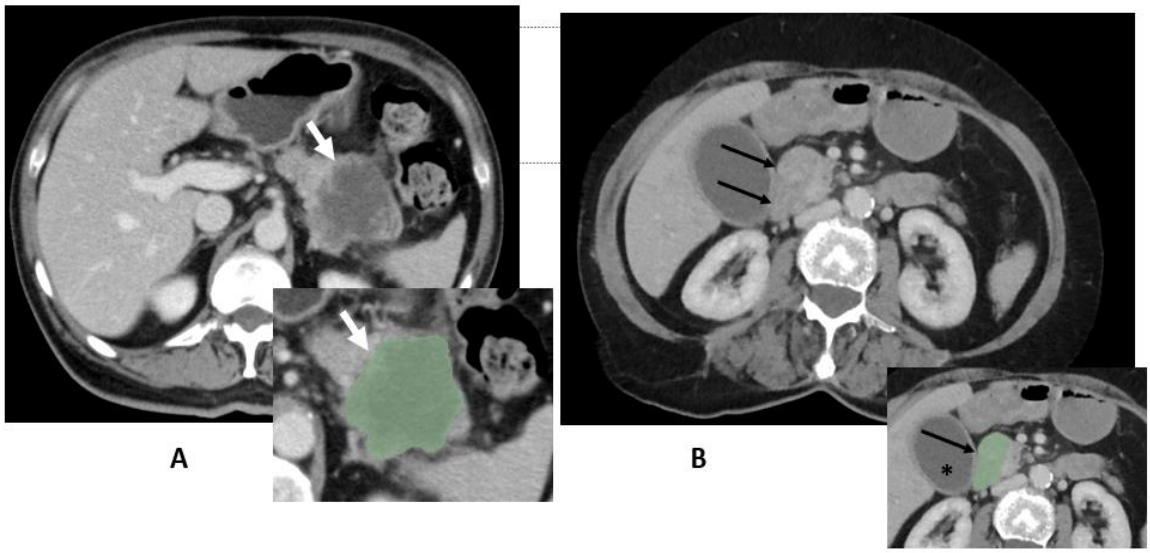


Figure 2 caption: Case examples from our study cohort. Axial slice images from contrast enhanced CT pancreas studies in the portal venous phase are presented, in patients with biopsy proven pancreatic ductal adenocarcinoma. Image A shows a 6cm hypoattenuating mass (white arrow) in the pancreatic tail with median attenuation 28 HU. The insert image shows the lesion post segmentation. Image B shows a mildly hypoattenuating/isoattenuating 3.5cm pancreatic head mass (black arrows) with median attenuation 73 HU. The insert image shows the lesion post segmentation and the star (*) denotes the distended gallbladder.

2.4.2. Heterogeneity and tumour biology.

It is well recognized that intra-tumour heterogeneity is common in PDAC, however the degree to which this reflects underlying genetic, epigenetic or histologic (epithelial vs stromal cells) differences in tumour tissue is unclear [5]. A study by

Kaisses et al [36] was able to use CT radiomics to categorise PDAC lesions into quasi-mesenchymal (QM) vs. non-quasi-mesenchymal (non-QM) types with an AUC of 0.92 and the most important features in their study quantified aspects of image heterogeneity (Entropy-/Energy-, Uniformity/Non-Uniformity and Correlation-/Variance-related features from the Pyradiomics library). However, the drawbacks of their study were the large number of predictor variables (161 selected from 1474 extracted) compared to the study population (181 training, 26 test) and the lack of external validation.

A recent publication on the topic of PDAC tumour heterogeneity by Grunwald *et* al from one of our partner labs at the University of Toronto [34] identified important categories of PDAC tumour micro environments (the tumour 'stroma') which can be identified on histology and correlate with patient survival and treatment response. Therefore, it is clear that heterogeneity in the PDAC tumour environment correlates with biological behaviour, but it is not yet clear whether this histological heterogeneity can be identified on pre-operative radiology imaging. This is the topic of another proposed study between our lab and that of Grunwald *et al* (discussed further in section 6.7. 'Next steps: Building upon our study results').

Figure 3. Heterogeneity of PDAC lesions. Case examples.

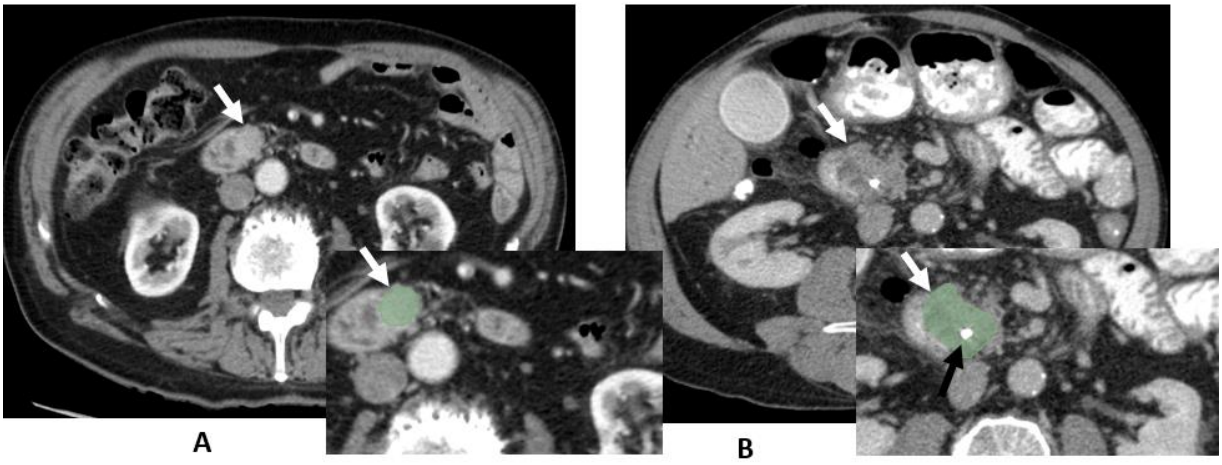


Figure 3 caption: Case examples from our study cohort. Axial slice images from contrast enhanced CT pancreas studies in the portal venous phase are presented, in patients with biopsy proven pancreatic ductal adenocarcinoma. Image A shows a 2.8cm homogenous isoattenuating mass (white arrow) in the pancreatic head with Joint-Entropy value of 2.6 and an ngtdm_Coarseness value of 0.048 (higher values of Joint_Entropy indicate more heterogeneity, whereas lower values of ngtdm_Coarseness indicate more heterogeneity). The insert image shows the lesion post segmentation. Image B shows a hypoattenuating and heterogenous 3.7cm pancreatic head mass (white arrows) with a Joint_Entropy value of 4.0 and ngtdm_Coarseness of 0.009. The insert image shows the lesion post segmentation and the black arrow points to a common duct stent.

2.5. Treatment of PDAC.

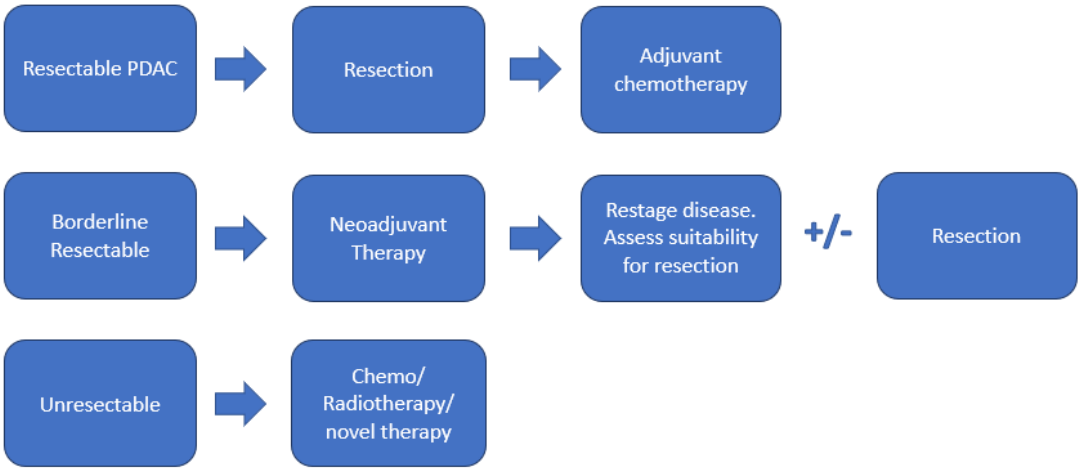
Complete surgical resection improves overall and disease-free survival in PDAC and it is the only potential chance of cure. Pancreatic resection was developed by Whipple in the 1930s [39]. The Whipple's procedure is a major operation which involves resection of the head of pancreas, common bile duct, gallbladder distal stomach, the first and second portions of the duodenum. Due to the complex anatomical relationships between the pancreas and surrounding vital structures,

pancreas resection is associated with significant patient morbidity, a 30-day mortality of approximately 5% [40, 41] and 90-day mortality up to 7.4% [42]. The median overall survival for patients who undergo resection for PDAC ranges between 24-28 months, compared to approximately 13 months for those do not undergo resection [19, 43]. Resection is recommended in patients who meet specific criteria, as defined by international consensus guidelines, such as from the National Comprehensive Cancer Network (NCCN) in the United states [13], namely: (1) no tumour contact with a major visceral artery (coeliac artery, superior mesenteric artery, common hepatic artery), no tumour contact with the superior mesenteric vein or portal vein, or <180 degree vein contact without vein contour irregularity and (2) no metastatic disease. These criteria were created to maximise the chance that patients will undergo a complete (R0) resection, meaning that there is a 1mm tumour free margin surrounding the resection specimen at pathological analysis. 70-80% of patients have advanced (stage III or IV) disease at the time of diagnosis [2, 3], meaning that they do not meet these criteria, and these patients are referred to as unresectable. There have been recent proposals to include biological features when selecting patients who are suitable for surgery, for example patients with CA19.9 > 500 iu/L may be considered unresectable regardless of anatomical staging [44], but this has not been incorporated into guidelines. Hence, patient selection based on local anatomy of the tumour and patient performance status remain the factors to consider at multidisciplinary discussion.

For those patients who undergo resection, post-operative (adjuvant) chemotherapy is advised for all [45] with either FOLFIRINOX (folinic acid, 5-fluorouracil, irinotecan and oxaliplatin) or Gemcitabine based regimes, as tolerated. In 2009, a third category was defined within the PDAC treatment algorithm, which is termed ‘borderline resectable’ PDAC and is an intermediate stage between the resectable and unresectable categories. There is now a strict and complex definition of this patient category in the guidelines, based on local tumour anatomy, for example including patients with tumours which demonstrate < 180 degree contact with the

superior mesenteric artery [45]. Borderline resectable PDAC accounts for up to 40% of all PDAC cases and half of those who are not suitable for upfront resection [46]. The treatment algorithm for these three patient categories is outlined below (Figure 4), however the borderline resectable and unresectable groups are not considered further in this study.

Figure 4. Treatment algorithm for Pancreatic Ductal Adenocarcinoma.



2.6. Neoadjuvant therapy in resectable PDAC.

As outlined in section 2.1, improving treatment outcomes in PDAC has been challenging, despite decades of research and advancement in medical practice, with only marginal gains achieved over the past 40 years. It has been proposed that administering neoadjuvant therapy to some or all patients with resectable PDAC may improve outcomes when compared to the current standard of care algorithm, where resectable patients proceed straight to upfront resection. This theory was initially based upon experience from other cancers, for example rectal cancer, where neoadjuvant therapy has led to improved survival [47] but also from the

successful use of neoadjuvant chemotherapy to improve survival in patients with borderline resectable PDAC [48]. Some large cancer care centres, such as MD Anderson in the United States [8], have adopted a strategy of offering neoadjuvant chemotherapy to all of their resectable PDAC patients, based upon the theory that all patients with PDAC have radiologically occult metastatic cancer at the time of diagnosis [8]. The proposed benefit is to ensure that all patients receive timely chemotherapy, since up to 50% of patients experience a delay to commencing adjuvant therapy while recovering from pancreas resection and some never recover sufficiently to receive treatment at all [49]. It has been shown to improve complete (R0) resection rate and disease-free survival (DFS) [8], however prospective trials have failed to demonstrated an overall survival benefit [49-52]. Therefore, it has been proposed that better patient selection for neoadjuvant therapy may improve response [8] but in order to do this, we require risk models to pre-operatively identify patients are the highest risk of recurrence disease and poor survival.

2.7. Predicting survival in PDAC.

Preoperative risk stratification is relevant in PDAC due to the debate about neoadjuvant therapy [53] and may also be relevant to decisions about the appropriateness of surgery in patients at high risk of early recurrence [54]. After resection, adjuvant chemotherapy is recommended for all patients [45] therefore, no further significant curative-intent care decisions are taken after this timepoint. The goal is to pre-operatively identify patients with aggressive tumour biology [55] but there are challenges of obtaining sufficient tissue for analysis from pre-operative pancreas biopsy, because a fine needle aspiration is typically performed via endoscopic ultrasound rather then a core biopsy. Therefore, discrimination with blood or imaging biomarkers is the primary focus of research [56]. However, very few published clinical or radiomic models have concentrated entirely on pre-operative data and none have undergone robust validation [25, 26, 31, 54, 57, 58].

For patients with radiologically resectable PDAC, there are several known clinical and pathological predictors of survival. The known predictors can be divided into pre-operative and post-operative variables (the latter group is larger):

- Pre-operative: advanced patient age [59], Ca 19.9 levels in blood [59], lymphocyte-neutrophil ratio in blood [57, 60], SPan-1 level in blood [54], patient symptoms at presentation [57], tumour location within the pancreas (Head of pancreas vs body/tail) [61], size of tumour measured on CT [54] and skeletal muscle index [22].
- Post-operative: pathologic T stage [62], histologic tumour grade [59, 63], lymphovascular invasion [62], perineural invasion [62], resection margin status (R0/R1) [59] and pathologic N stage [59]. Tumour size is also associated with prognosis, however it has been shown that even very small tumours (<0.5cm, classified as T1a on the current AJCC TNM system) can metastasis in up to 31% of cases [55], therefore it is not a linear relationship.

2.7.1. Pathologic vs imaging defined N stage.

Pathologic N stage is a strong predictor of prognosis in resectable PDAC [26]. While grossly enlarged nodes are visible on pre-operative imaging, it has been shown that many lymph nodes which appear normal to the Radiologist on pre-operative CT actually harbour micrometastasis. In studies looking specifically at the pre-operative prediction of lymph node status in PDAC, there are conflicting results as to whether lymph node status defined based upon short-axis size on pre-operative CT can [64] or cannot [65] accurately predict the presence of lymph node metastasis on pathology. Thus, pre-operative CT is poor at predicting lymph nodes status in PDAC [66] but we have included it in our study for the sake of completeness.

2.7.2. Ca 19.9.

Ca19.9 is a plasma biomarker which can be used to support the diagnosis of PDAC, help with treatment response evaluation and aid in prognostication [56]. It is the

only biomarker approved for use in PDAC by the United States Food and Drug Administration [4, 12]. Unfortunately, it does have several limitations, for example the sensitivity for PDAC diagnosis is only 70-74% and approximately 5-10% of the general population are non-secretors (Lewis (a-b-) phenotype), meaning that they cannot express Ca19.9 at all [67, 68] (interestingly, the less well known Span-1 antigen is not impacted by Lewis phenotype). Specificity is also compromised by the fact that levels can be elevated in other benign or malignant conditions including pancreatitis and cholangiocarcinoma. Nonetheless, it is considered the current best serum biomarker in this disease [67] and multiple studies have found it to be a significant pre-operative predictor of patient survival post resection, with hazard ratios (HR) for overall survival ranging from 1.37-1.86 in univariable (UVA) regression analysis [22, 62] and 1.1-2 in multivariable analysis (adjusted for other pre and post-operative clinical variables) [22, 69]. HRs for early disease recurrence are approximately 2 in UVA [57], although there is no agreed cut-point for ca19.9 to differentiate high vs low risk patients. It has also been shown that the addition of radiomics can improve upon the prognostic performance of ca19.9 [26, 32]. However, it must be acknowledged that there is some inconsistency in the literature regarding the prognostic ability of pre-operative Ca19.9 however, with a number of recent studies reporting that pre-operative Ca19.9 was not a significant predictor of survival [29, 30], including one study which included 205 resected patients and excluded Ca19.9 non-secretors [68].

Several other pre-operative biomarker candidates have been investigated, some with positive results including radiological markers such as pre-operative SUVmax or ADC quantification on PET-CT and MRI respectively, however none have been validated to date [56].

2.7.3. Prognostic models.

The gold standard prognostic system used for all cancers is the Tumour, Node, Metastasis (TNM) system developed by the American Joint Committee on Cancer (AJCC) and it is the primary system used on a day-to-day bases at pancreas cancer

clinics. The AJCC updates these systems for all cancers periodically. The 6th edition was published in 2002 and there were no changes for PDAC in the 7th edition published in 2010, however refinements were made in the 8th edition in 2016, with changes in the definition of T stage and addition of a new N stage (from No/N1 to N0/N1/N2) [70]. Despite the refinements, TNM remains suboptimal for PDAC, with reported C-indices for prognostication of 0.572-0.699 for the 8th edition [29, 62], which is disappointing considering that it has the benefit of pathological data (T stage, N stage) which is only available after the patient has undergone surgical resection. Two groups have developed clinical models for prediction of PDAC survival [62, 71], with reported C-indices of 0.65-0.7, however these have also incorporated data which is only available post-operatively (such as tumour margin status), hence they are of no use for pre-operative decision making. One recent publication created an entirely pre-operative clinical model for prediction of early disease recurrence reporting c-index 0.85 for early recurrence, but this has not been robustly validated and they chose an unusual outcome of recurrence with 162 days of surgery, without a good explanation as to why this time interval was chosen [57]. In practice, none of these models are used in day-to-day hepatobiliary or oncology clinics and there is clearly an unmet need in this domain.

2.8. Chapter 2 references.

1. Safi, S.A., et al., *Site of relapse of ductal adenocarcinoma of the pancreas affects survival after multimodal therapy.* BMC Surg, 2021. **21**(1): p. 110.
2. *National Cancer Registry Ireland. Pancreas Factsheet.* [cited 2021 February 3rd]; Available from: https://www.ncri.ie/sites/ncri/files/factsheets/Factsheet%20pancreas.pdf.
3. *Canadian Cancer Statistics 2017. Special topic: Pancreatic cancer*. 2017 February 3rd 2021]; Available from: https://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-Cancer-Statistics-2017-EN.pdf.
4. McGuigan, A., et al., *Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes.* World J Gastroenterol, 2018. **24**(43): p. 4846-4861.
5. Makohon-Moore, A. and C.A. Iacobuzio-Donahue, *Pancreatic cancer biology and genetics from an evolutionary perspective.* Nat Rev Cancer, 2016. **16**(9): p. 553-65.

6. Singh, S., et al., *An examination of the association between lifetime history of prostate and pancreatic cancer diagnosis and occupation in a population sample of Canadians.* PLOS ONE, 2020. **15**(2): p. e0227622.
7. Sun, H., et al., *Survival improvement in patients with pancreatic cancer by decade: a period analysis of the SEER database, 1981-2010.* Sci Rep, 2014. **4**: p. 6747.
8. Gaskill, C.E., et al., *History of preoperative therapy for pancreatic cancer and the MD Anderson experience.* J Surg Oncol, 2021. **123**(6): p. 1414-1422.
9. Haeno, H., et al., *Computational modeling of pancreatic cancer reveals kinetics of metastasis suggesting optimum treatment strategies.* Cell, 2012. **148**(1-2): p. 362-375.
10. Hidalgo, M., *Pancreatic cancer.* N Engl J Med, 2010. **362**(17): p. 1605-17.
11. Megibow, A., *Pancreatic Neoplasms*, in *Textbook of Gastrointestinal Radiology. Fourth Edition*, Gore and Levine, Editors. 2015, Elsevier.
12. Mizrahi, J.D., et al., *Pancreatic cancer.* Lancet, 2020. **395**(10242): p. 2008-2020.
13. Tempero, M.A., et al., *Pancreatic Adenocarcinoma, Version 2.2021, NCCN Clinical Practice Guidelines in Oncology.* J Natl Compr Canc Netw, 2021. **19**(4): p. 439-457.
14. Lucas, A.L. and F. Kastrinos, *Screening for Pancreatic Cancer.* JAMA, 2019. **322**(5): p. 407-408.
15. Elbanna, K.Y., H.J. Jang, and T.K. Kim, *Imaging diagnosis and staging of pancreatic ductal adenocarcinoma: a comprehensive review.* Insights Imaging, 2020. **11**(1): p. 58.
16. Kulkarni, N.M., et al., *White paper on pancreatic ductal adenocarcinoma from society of abdominal radiology's disease-focused panel for pancreatic ductal adenocarcinoma: Part II, update on imaging techniques and screening of pancreatic cancer in high-risk individuals.* Abdom Radiol (NY), 2020. **45**(3): p. 729-742.
17. Koay, E.J., et al., *Transport properties of pancreatic cancer describe gemcitabine delivery and response.* J Clin Invest, 2014. **124**(4): p. 1525-36.
18. Raman, S.P., et al., *Impact of the time interval between MDCT imaging and surgery on the accuracy of identifying metastatic disease in patients with pancreatic cancer.* AJR Am J Roentgenol, 2015. **204**(1): p. W37-42.
19. Healy, G.M., et al., *Preoperative CT in patients with surgically resectable pancreatic adenocarcinoma: does the time interval between CT and surgery affect survival?* Abdom Radiol (NY), 2018. **43**(3): p. 620-628.
20. Vasilyeva, E., et al., *Impact of surgical wait times on oncologic outcomes in resectable pancreas adenocarcinoma.* HPB (Oxford), 2020. **22**(6): p. 892-899.
21. Sanjeevi, S., et al., *Impact of delay between imaging and treatment in patients with potentially curable pancreatic cancer.* Br J Surg, 2016. **103**(3): p. 267-75.
22. Shi, H., et al., *Survival prediction after upfront surgery in patients with pancreatic ductal adenocarcinoma: Radiomic, clinic-pathologic and body composition analysis.* Pancreatology, 2021. **21**(4): p. 731-737.
23. Zhang, H., et al., *An empirical framework for domain generalization in clinical settings.* Proceedings of the Conference on Health, Inference, and Learning, 2021: p. 279–290.
24. Zhang, Y., et al., *CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging.* BMC Med Imaging, 2020. **20**(1): p. 11.
25. Li, K., et al., *Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma.* Am J Transl Res, 2019. **11**(7): p. 4491-4499.

26. Khalvati, F., et al., *Prognostic Value of CT Radiomic Features in Resectable Pancreatic Ductal Adenocarcinoma.* Sci Rep, 2019. **9**(1): p. 5449.
27. Kim, H.S., et al., *Preoperative CT texture features predict prognosis after curative resection in pancreatic cancer.* Sci Rep, 2019. **9**(1): p. 17389.
28. Yun, G., et al., *Tumor heterogeneity of pancreas head cancer assessed by CT texture analysis: association with survival outcomes after curative resection.* Sci Rep, 2018. **8**(1): p. 7226.
29. Xie, T., et al., *Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and TNM staging for survival estimation after curative resection.* European Radiology, 2020. **30**(5): p. 2513-2524.
30. Cassinotto, C., et al., *Resectable pancreatic adenocarcinoma: Role of CT quantitative imaging biomarkers for predicting pathology and patient outcomes.* Eur J Radiol, 2017. **90**: p. 152-158.
31. Eilaghi, A., et al., *CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma - a quantitative analysis.* BMC Med Imaging, 2017. **17**(1): p. 38.
32. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.
33. Jhaveri, K.S., et al., *Can preoperative liver MRI with gadoxetic acid help reduce open-close laparotomies for curative intent pancreatic cancer surgery?* Cancer Imaging, 2021. **21**(1): p. 45.
34. Grünwald, B.T., et al., *Spatially confined sub-tumor microenvironments in pancreatic cancer.* Cell, 2021. **184**(22): p. 5577-5592.e18.
35. Hosein, A., R. Brekken, and A. Maitra, *Pancreatic cancer stroma: an update on therapeutic targeting strategies.* Nature reviews gastroenterology and hepatology, 2020. **17**(487-505).
36. Kaissis, G., et al., *Image-Based Molecular Phenotyping of Pancreatic Ductal Adenocarcinoma.* Journal of Clinical Medicine, 2020. **9**(3): p. 724.
37. Iwatate, Y., et al., *Radiogenomics for predicting p53 status, PD-L1 expression, and prognosis with machine learning in pancreatic cancer.* Br J Cancer, 2020. **123**(8): p. 1253-1261.
38. Yoo, H.J., et al., *Tumor conspicuity significantly correlates with postoperative recurrence in patients with pancreatic cancer: a retrospective observational study.* Cancer Imaging, 2020. **20**(1): p. 46.
39. Hammond, N., et al., *Pancreas: Normal Anatomy and Exmaination Techniques*, in *Textbook of Gastrointestinal Radiology, Fourth Edition*, Gore and Levine, Editors. 2015, Elsevier.
40. Biswas, S., et al., *Outcome of Surgical Management of Pancreas Neoplasms in a Large Community Hospital.* Am Surg, 2020: p. 3134820951490.
41. Lam, M.B., et al., *Changes in Racial Disparities in Mortality After Cancer Surgery in the US, 2007-2016.* JAMA Netw Open, 2020. **3**(12): p. e2027415.
42. Swanson, R.S., et al., *The 90-day mortality after pancreatectomy for cancer is double the 30-day mortality: more than 20,000 resections from the national cancer data base.* Ann Surg Oncol, 2014. **21**(13): p. 4059-67.
43. Neoptolemos, J.P., et al., *Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-*

4): a multicentre, open-label, randomised, phase 3 trial. Lancet, 2017. **389**(10073): p. 1011-1024.
44. Heckler, M. and T. Hackert, *Surgery for locally advanced pancreatic ductal adenocarcinoma-is it only about the vessels?* J Gastrointest Oncol, 2021. **12**(5): p. 2503-2511.
45. *NCCN Cinical Practice Guidelines in Oncology : Pancreatic Adenocarcinoma*. 2021.
46. Dhir, M., et al., *Neoadjuvant treatment of pancreatic adenocarcinoma: a systematic review and meta-analysis of 5520 patients.* World journal of surgical oncology, 2017. **15**(1): p. 183-183.
47. Shivnani, A.T., et al., *Preoperative chemoradiation for rectal cancer: results of multimodality management and analysis of prognostic factors.* Am J Surg, 2007. **193**(3): p. 389-93; discussion 393-4.
48. Gemenetzis, G., et al., *Survival in Locally Advanced Pancreatic Cancer After Neoadjuvant Therapy and Surgical Resection.* Ann Surg, 2019. **270**(2): p. 340-347.
49. Müller, P.C., et al., *Neoadjuvant Chemotherapy in Pancreatic Cancer: An Appraisal of the Current High-Level Evidence.* Pharmacology, 2021. **106**(3-4): p. 143-153.
50. Schneider, M., J.P. Neoptolemos, and M.W. Büchler, *Commentary: Neoadjuvant treatment of resectable pancreatic cancer: Lack of level III evidence.* Surgery, 2020. **168**(6): p. 1015-1016.
51. Versteijne, E., et al., *Preoperative Chemoradiotherapy Versus Immediate Surgery for Resectable and Borderline Resectable Pancreatic Cancer: Results of the Dutch Randomized Phase III PREOPANC Trial.* J Clin Oncol, 2020. **38**(16): p. 1763-1773.
52. O'Reilly, E.M. and C. Ferrone, *Neoadjuvant or Adjuvant Therapy for Resectable or Borderline Resectable Pancreatic Cancer: Which Is Preferred?* J Clin Oncol, 2020. **38**(16): p. 1757-1759.
53. Eid, M., et al., *Current view of neoadjuvant chemotherapy in primarily resectable pancreatic adenocarcinoma.* Neoplasma, 2021. **68**(1): p. 1-9.
54. Ishido, K., et al., *Development of a Biomarker-Based Scoring System Predicting Early Recurrence of Resectable Pancreatic Duct Adenocarcinoma.* Ann Surg Oncol, 2022. **29**(2): p. 1281-1293.
55. Roalsø, M., J.R. Aunan, and K. Søreide, *Refined TNM-staging for pancreatic adenocarcinoma – Real progress or much ado about nothing?* European Journal of Surgical Oncology, 2020. **46**(8): p. 1554-1557.
56. Chang, J.C. and M. Kundranda, *Novel Diagnostic and Predictive Biomarkers in Pancreatic Adenocarcinoma.* Int J Mol Sci, 2017. **18**(3): p. 667.
57. Guo, S.W., et al., *A preoperative risk model for early recurrence after radical resection may facilitate initial treatment decisions concerning the use of neoadjuvant therapy for patients with pancreatic ductal adenocarcinoma.* Surgery, 2020. **168**(6): p. 1003-1014.
58. Palumbo, D., et al., *Prediction of Early Distant Recurrence in Upfront Resectable Pancreatic Adenocarcinoma: A Multidisciplinary, Machine Learning-Based Approach.* Cancers (Basel), 2021. **13**(19).
59. Hartwig, W., et al., *Pancreatic cancer surgery in the new millennium: better prediction of outcome.* Ann Surg, 2011. **254**(2): p. 311-9.
60. Bhatti, I., et al., *Preoperative hematologic markers as independent predictors of prognosis in resected pancreatic ductal adenocarcinoma: neutrophil-lymphocyte versus platelet-lymphocyte ratio.* Am J Surg, 2010. **200**(2): p. 197-203.

61. Sheikh, M., et al., *Survival features, prognostic factors, and determinants of diagnosis and treatment among Iranian patients with pancreatic cancer, a prospective study.* PLoS One, 2020. **15**(12): p. e0243511.
62. Xu, D., et al., *Prognostic Nomogram for Resected Pancreatic Adenocarcinoma: A TRIPOD-Compliant Retrospective Long-Term Survival Analysis.* World J Surg, 2020. **44**(4): p. 1260-1269.
63. Wasif, N., et al., *Impact of tumor grade on prognosis in pancreatic cancer: should we include grade in AJCC staging?* Annals of surgical oncology, 2010. **17**(9): p. 2312-2320.
64. Li, K., et al., *Contrast-enhanced CT radiomics for predicting lymph node metastasis in pancreatic ductal adenocarcinoma: a pilot study.* Cancer Imaging, 2020. **20**(1): p. 12.
65. Liu, P., et al., *Applying a radiomics-based strategy to preoperatively predict lymph node metastasis in the resectable pancreatic ductal adenocarcinoma.* J Xray Sci Technol, 2020. **28**(6): p. 1113-1121.
66. Loch, F.N., et al., *Accuracy of various criteria for lymph node staging in ductal adenocarcinoma of the pancreatic head by computed tomography and magnetic resonance imaging.* World J Surg Oncol, 2020. **18**(1): p. 213.
67. Haab, B.B., et al., *Definitive Characterization of CA 19-9 in Resectable Pancreatic Cancer Using a Reference Set of Serum and Plasma Specimens.* PLoS One, 2015. **10**(10): p. e0139049.
68. Park, J.K., et al., *Clinical significance and revisiting the meaning of CA 19-9 blood level before and after the treatment of pancreatic ductal adenocarcinoma: analysis of 1,446 patients from the pancreatic cancer cohort in a single institution.* PLoS One, 2013. **8**(11): p. e78977.
69. Barton, J.G., et al., *Predictive and prognostic value of CA 19-9 in resected pancreatic adenocarcinoma.* J Gastrointest Surg, 2009. **13**(11): p. 2050-8.
70. Shin, D.W. and J. Kim, *The American Joint Committee on Cancer 8th edition staging system for the pancreatic ductal adenocarcinoma: is it better than the 7th edition?* Hepatobiliary surgery and nutrition, 2020. **9**(1): p. 98-100.
71. Brennan, M.F., et al., *Prognostic nomogram for patients undergoing resection for adenocarcinoma of the pancreas.* Ann Surg, 2004. **240**(2): p. 293-8.

Chapter 3 - Evidence based Medicine Literature review: The use of Radiomics for survival prediction in Pancreatic Adenocarcinoma

"There is undoubtedly huge potential for machine learning to transform healthcare, but going 'from code to clinic' is the hard part"

May et al, 2021 [1]

An evidence- based medicine (EBM) review was conducted using standard EBM methodology [2], namely; Ask a focused question, Search for the evidence, Appraise this evidence, Apply this evidence to clinical practice and Evaluate the performance of the review.

3.1. Ask - Ask a focused question.

An EBM question was generated, using the PICO [2] format. This acronym stands for Population/Problem (P), Intervention/Exposure (I), Comparison (C) and Outcome (O).

- P: Adult patients with who have undergone complete resection for pancreatic adenocarcinoma.
- I: Radiomic analysis of pre-operative CT OR texture analysis.
- C: No comparison.
- O: Survival.

3.2. Search – Search for the evidence.

The literature review was conducted using EBM methodology on March 8th, 2021, utilising the PICO question. This search was performed using the following search

engines: PubMed, PubMed reminer, Trip database, Embase and BMJ best practice. The free text search terms were: Pancreatic adenocarcinoma, pancreatic cancer, radiomic, radiomics, survival and the MESH terms were: humans, pancreatic neoplasm/diagnostic imaging, pancreatic neoplasm, precision medicine, radiology, diagnostic imaging. This revealed 320 results. One review article was identified [3] and a backward citation search was performed, adding eight more publications which had been not been included thus far (these were missed because they were older studies which used the term 'texture analysis' rather than radiomics. This term was subsequently added to the EBM search terms). One reference was a conference paper [4], which was found to have been published subsequently [5]. After application of the following exclusion criteria, the list was then reduced to 11:

- Animal studies.
- Studies on pancreatic neoplasms other than ductal adenocarcinoma (e.g., neuroendocrine tumour, intraductal papillary mucinous neoplasm) or studies on patients with unresectable PDAC.
- Studies assessing imaging modalities other than contrast enhanced CT.
- Deep Learning studies, with no radiomics component.
- Delta radiomics studies (i.e., assessing change in radiomics over time).
- Studies dealing with diagnosis of PDAC vs alternative pancreatic pathology (i.e., differentiation of PDAC from autoimmune pancreatitis).

Section 3.1 and 3.2 References

1. May, M., *Eight ways machine learning is assisting medicine.* Nat Med, 2021. **27**(1): p. 2-3.
2. Lavelle, L.P., et al., *Evidence-based Practice of Radiology.* Radiographics, 2015. **35**(6): p. 1802-13.
3. Abunahel, B.M., et al., *Pancreas image mining: a systematic review of radiomics.* Eur Radiol, 2020. **31**(5): p. 3447-3467.
4. Attiyeh, M.A., et al., *Through the looking-mass: Preoperative survival prediction in pancreatic ductal adenocarcinoma (PDAC) by quantitative ct analysis*, in *70th Annual Cancer Symposium of the Society of Surgical Oncology*. 2017. p. 1-202.

5. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.

-

Table 1. Prior resectable PDAC CT Radiomic studies.

Study	**Population**	**Multi-institutional training data?**	**Follow up (Months)**	**RQS (score/36)**	**Contour method**	**Radiomic Software**	**IBSI compliant software?**	**Internal vs external validation**	**Overall model Performance**
Shi *et al*. 2021.	299. Training (210). Int validation set (89).	Single institution.	Median: 20.5.	14	PV+AP 3D	Artificial Intelligence Kit.	Unknown.**	Internal. Train/test split.	C-index 0.73 in validation for OS.
Zhang *et al.* 2021.	98. Training (68). Ext validation set (30)	Single institution.	Unknown.	8	PV 2D	PyRadiomics.	Yes.	External.	Two-year survival AUC 0.84
XIE *et al.* 2020.	220. Training (147). Int validation set (73).	Single institution.	Median: 17.4.	14	PV 3D	Mazda v 4.6.	No.	Internal. Train/test split.	C-Index 0.726 in validation for OS.

Study	**Population**	**Multi-institutional training data?**	**Follow up (Months)**	**RQS (score/36)**	**Contour method**	**Radiomic Software**	**IBSI compliant software?**	**Internal vs external validation**	**Overall model Performance**
Zhang *et al.* 2020.	98. Training (68). Ext validation set (30)	Single institution.	Unknown.	8	PV 2D	PyRadiomics.	Yes.	External.	C-Index 0.651 in validation for OS.
Li *et* al. 2019.	111. No train/test split.	Single institution.	Unknown.	11*	PV 2D	Matlab.	Yes.	Internal. Cross validation.	No overall performance metric
Khalvati *et al.* 2019.	98. Training (30). Ext validation set (68)	Single institution.	Unknown.	10	PV 2D	PyRadiomics.	Yes.	External.	HR 1.35 in validation for OS.

Kim *et al.*2019.	116. No train/test split.	Single institution.	Unknown.	-4	AP 3D	In-house software.	Unknown.**	No validation.	No overall performance metric
Study	**Population**	**Multi-institutional training data?**	**Follow up (Months)**	**RQS (score/36)**	**Contour method**	**Radiomic Software**	**IBSI compliant software?**	**Internal vs external validation**	**Overall model Performance**
Attiyeh *et al.* 2018.	161. Training (113). Int validation set (48)	Single institution.	Unknown.	5	PV 3D	Matlab.	Yes.	Internal. Train/test split.	C-index 0.74 in validation for OS.
Yun *et al.* 2018.	88. No train/test split.	Single institution.	Mean 26.3.	2*	AP 2D	In-house software.	Unknown.**	Internal. Cross validation.	No overall performance metric
Cassinotto *et al.* 2017.	99. No train/test split.	Multi.	Median 19.1.	-5	PV 2D	TexRAD.	Unknown.**	No validation.	No overall performance metric

Eilaghi *et al.* 2017.	30. No train/test split.	Single institution.	Unknown.	-5	PV 2D	Matlab.	Yes.	No validation.	No overall performance metric

RQS = Radiomics Quality score [1]. Int = internal. Ext = External. PV = Portal-venous phase. AP = Arterial phase. IBSI = Imaging Biomarker Standardization Initiative. AUC = Area under the curve of a received operator curve analysis. OS = Overall survival. HR = Hazard Ratio.

*Considering cross validation as 'validation on a dataset from the same institution' for RQS.

**Not enough information publicly available to determine IBSI compliance.

1. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nat Rev Clin Oncol, 2017. **14**(12): p. 749-762.

3.3. Appraise - Critical appraisal of the evidence.

A description of the 11 included studies is presented in table 1. All studies reported a positive outcome (i.e., they were successful at using radiomics to predict the outcome of interest). The population sizes ranged from 30-299, with mean population size of 129 (StDev 73.5). Training populations ranged from 30-210 patients.

A previously published scoring system called the Radiomics Quality Score (RQS) was used to formally assess the studies [1]. This consists of 16 questions, each with a numerical score and a free online calculator is available [2]. The total is scored out of 36 and it is possible to achieve a negative overall tally, since some areas are heavily penalised, such as lack of a validation method for study results, which receives a count of -5. Some RQS questions received a negative answer for every study, for example question 3 asked whether phantom studies have been performed on every scanner in the study. Likewise, prospective studies are awarded 7 points (out of 36), but there were no such studies in this review. The scores in this review are generally low, with mean RQS 5.3 / 36 (St Dev 7.3). The highest ranking studies according to the RQS, were from Xie *et al* 2020 [3] and *Shi et* al 2021 [4], both with scores of 14/36 (39%). These two studies ranked highest because they included feature reduction methods, reported both discrimination and calibration statistics, compared their results to a gold standard (TNM system) and assessed clinical utility using decision curve analysis. Some of the most important questions which discriminated between studies in this review are outlined below:

3.3.1. Feature reduction or adjustment for multiple testing (RQS Question 5):

There is a risk of overfitting when multiple predictor variables are used in a study. Overfitting is when a model fits exactly to its training data (in simple terms, it knows the training dataset too well, including any noise/random error in the data) and thus produces impressive results in training, but typically demonstrates a significant

performance drop on external validation. This effect can be mitigated by using a feature reduction step, to identify important predictor variables and discard the rest, prior to model building. Including such a step is rewarded in the RQS. The average number of extracted radiomic features across the 11 studies in this EBM review was 593, ranging from 4-2041. Three studies extracted less than 20 features. Six of 11 studies included a feature reduction step in the analysis. The use of a feature reduction method is advised by guidelines, however no single proven 'best' method has been demonstrated in the literature. The most common method across the reviewed studies (5/11 publications) was the Least absolute shrinkage and selection operator (LASSO). This is a form of penalized (regularized) regression, described as a penalized maximum likelihood shrinkage method [5], which reduces the values of the regression Beta coefficients of each feature. The optimum amount of shrinkage (called the lambda value) is determined by cross validation. Any predictors whose coefficients is reduced to zero can be eliminated. Hence, it is a useful technique for eliminating predictor variables and thus reducing the number of predictors under investigation. For example, the study by Xie *et al* [3] entered 186 features into LASSO, from which the five features with non-zero coefficients were selected for use in the prediction of patient survival.

3.3.2. Validation (RQS Question 12):

Lack of external validation is a major drawback of the reviewed studies. Three publications in the review reported external validation, all of which are from the same research group and they focused on survival prediction using a dataset of 98 patients from Toronto, consisting of 68 patients from one hospital for training and 30 patients (with 10% neoadjuvant therapy [6]) from a separate Toronto hospital for validation [7-9] (note that there is an overlap of 55 patients from the training cohort from those studies and the training cohort in the present study).

The remainder of the studies used internal validation methods such as a hold-out test set from the same institution as the training dataset (3/11 studies, table 1), cross validation (2/11) or no validation at all (3/11). Internal validation methodology

can often demonstrate overoptimistic model performance [10] (discussed more in section 1.3). Disappointingly, one of the earliest studies in this field, from 2017, had a cohort of 99 patients recruited from two hospitals located in Canada and France, yet the patients were all grouped together for analysis, missing the opportunity for external validation.

3.3.3. Comparison to Gold standard (RQS Question 13):

Comparison to a gold standard model, such as the TNM staging system, is recommended so that the incremental gain of using the novel biomarker can be accurately assessed relative to the currently available information used in the clinic [11, 12]. This can also to provide a benchmark for comparison to other novel markers/models. It is important to highlight that TNM is based upon pathological data and it is therefore not available until after the resection, hence it is not useful to make pre-operative decisions. Nonetheless, there is no established reference standard for pre-operative decision making in PDAC, therefore TNM is a reasonable model to include as gold standard, considering its ubiquity in clinical practice.

This comparison was performed in 3/11 of the reviewed studies, despite that fact that 8/11 studies reported the post-operative pathological data (node status, tumour size) which is used to model TNM.

3.3.4. Assessment of calibration. Assessment of clinical utility (RQS Questions 10 and 14):

Calibration and clinical utility are introducted in section 1.8 of this manuscript (Statistics, feature standardization and feature reduction). Two studies (Shi *et* al 2021 and Xie *et* al 2020) assessed clinical utility using decision curve analysis [3, 4]. These two studies also explicitly assessed calibration (using calibration plots), which is why these studies have the highest RQS results (14/36) overall. Two additional studies [8, 13] considered calibration within the integrated Brier score, which is a metric that combines calibration and discrimination, although they did not provide calibration plots.

3.3.5. Reproducability and feature extraction software:

While not included in the RQS, it is worth noting whether a radiomics study publishes the regression equation/formula used in their study (or their actual computer code, for a more complex model which cannot be expressed in an equation). This information is essential so that other groups can attempt to replicate their work. 2/11 studies (Shi *et* al 2021 and Xie *et* al 2020) published their model equations equations, so that their work can be replicated. The only drawback is that neither of these studies used IBSI compliant software for feature extraction (see table 1), therefore attempts to replicate their work would require use of the exact software and identical software version in order to control for this variable. An example of an equation, from the Xie *et al* paper [3] is as follows (note that the feature names are different to our study, since this study used the non-IBSI compliant 'Mazda' software package):

$$\begin{aligned}\mathrm{Rad-score} = {} & 0.07342984 \times \mathrm{S.1.0.Entropy} \\ & + \ -0.03562417 \times \mathrm{S.4.0.SumAverg} \\ & + \ -0.19247206 \times \mathrm{S.4.4.AngScMom} \\ & + \ -0.01506736 \times \mathrm{WavEnHL_s.2} \ + \ 0.4422464 \times \mathrm{WavEnLL_s.3}\end{aligned}$$

Of the remaining studies, 5/9 studies reported the radiomic features included in their models accompanied by the univariable or multivariable results (hazard ratios or regression coefficients) meaning that replication of their findings can be attempted, although only two of these used IBSI complicant software. 4/11 studies did not provide enough data to attempt replication.

3.3.5. Common radiomic features across reviewed studies:

While not included in the RQS, it is important to consider the actual radiomic features which were identified in the 11 studies, to search for commonly occurring features which are useful for PDAC prognostication. We can only do this for the 6 studies who used IBSI compliant software, four of which reported the radiomic features used in their models (Table 2). One of these included a mixture of

conventional radiomic and deep learning features [14]. The majority of features (8/9) categorised different aspects of image texture, however there were no overlap in features between any of the studies. Only one of these studies provided the univariable or multivariable coefficients for the selected features, allowing for potential replication of their results, however this paper was from our own lab. Hence, there are no options for external validation of a previously published study in this field, unless more data can be obtained from the study authors (emails were sent to the authors of all studies in this review, but no replied were received).

Table 2. Radiomics features identified as prognostic for PDAC in prior studies who used IBSI compliant software.

Study	**Features**
Zhang_2021[9]	Not provided
Zhang_2020[8]	- gradient_gldm_SmallDependenceEmphasis - gradient_glszm_SmallAreaEmphasis - original_glszm_LargeAreaLowGrayLevelEmphasis - wavelet.HLH_glszm_HighGrayLevelZoneEmphasis **Equation/code/formula not provided. No UVA/MV results.*
Li_2019[14]	70 conventional radiomics and 256 Deep Learning features were extracted. One Radiomic feature identified: shape feature: extent. **Equation/code/formula not provided. No UVA/MV results.*
Khalvati_2019[7]	- Original_glcm_SumEntropy - Squareroot_glcm_ClusterTendency

	**UVA and MVA results were provided.*
Attiyeh_2018[13]	Not provided.
Eilaghi_2017[6]	- Tumor dissimilarity - Inverse difference normalized **Equation/code/formula not provided. No UVA/MV results.*

3.4 Apply and Evaluate - Apply the evidence in practice and evaluation impact on patients.

Application of evidence in the EBM framework refers to the impact of this evidence upon patient management in the clinic and similarly, evaluation refers to evaluation your performance i.e., how effective we are at incorporating the new practice into our workflow/clinic, which can be assessed by departmental audit. From the above appraisal, it is clear that the quality of the data within the field of PDAC radiomics is not yet strong enough to influence clinical practice. Hence, to the best of our knowledge, no clinical-radiomic model is currently being used in clinical setting and we cannot evaluate the impact upon patient management or incorporate it into our clinical workflows.

3.5. Conclusion of the evidence-based literature review

Based upon the available literature, there is no conclusive evidence that CT Radiomics provides accurate and reliable pre-operative prognostication for patients with PDAC. The main issues with the literature are: Poor quality methodology (low RQS score), small population sizes, no robust external validation, lack of detail in the published manuscript to attempt replication of results and lack of any overlap

between prognostic radiomic features in studies to date. In addition, the reviewed studies have used highly curated patient cohorts with respect to CT technical parameters and incorporated post-operative pathological data, which is not available pre-operatively. There will be no translation to clinical practice unless the quality of this evidence can be improved, since decisions regarding neoadjuvant therapy, surgery etc. must be based upon solid evidence. As stated in the EBM literature, when such deficits in available data are identified, a research project may be undertaken to bridge this gap [15], which is why we embarked on the current study.

3.6. Chapter 3 references.

1. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nat Rev Clin Oncol, 2017. **14**(12): p. 749-762.
2. *Radiomics World*. Available from: https://www.radiomics.world/home.
3. Xie, T., et al., *Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and TNM staging for survival estimation after curative resection.* European Radiology, 2020. **30**(5): p. 2513-2524.
4. Shi, H., et al., *Survival prediction after upfront surgery in patients with pancreatic ductal adenocarcinoma: Radiomic, clinic-pathologic and body composition analysis.* Pancreatology, 2021. **21**(4): p. 731-737.
5. Steyerberg, E., *Clinical Prediction Models: A practical Approach to Development, Validation and Updating. Second Edition.* 2009: Springer.
6. Eilaghi, A., et al., *CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma - a quantitative analysis.* BMC Med Imaging, 2017. **17**(1): p. 38.
7. Khalvati, F., et al., *Prognostic Value of CT Radiomic Features in Resectable Pancreatic Ductal Adenocarcinoma.* Sci Rep, 2019. **9**(1): p. 5449.
8. Zhang, Y., et al., *CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging.* BMC Med Imaging, 2020. **20**(1): p. 11.
9. Zhang, Y., et al., *Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images.* Sci Rep, 2021. **11**(1): p. 1378.
10. Esbensen, K. and P. Geladi, *Principles of Proper Validation: use and abuse of re-sampling for validation.* Journal of Chemometrics, 2010. **24**(3-4).
11. *AJCC Cancer Staging Manual (7th ed)*, ed. M.B. Amin, et al. 2010, New York, NY: Springer.
12. Steyerberg, E.W., et al., *Assessing the incremental value of diagnostic and prognostic markers: a review and illustration.* Eur J Clin Invest, 2012. **42**(2): p. 216-28.

13. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.
14. Li, K., et al., *Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma.* Am J Transl Res, 2019. **11**(7): p. 4491-4499.
15. Lavelle, L.P., et al., *Evidence-based Practice of Radiology.* Radiographics, 2015. **35**(6): p. 1802-13.

Chapter 4 – Hypothesis, Materials and Methods

4.1. Hypothesis and aims

It is clear from the evidence-based medicine review that there are weaknesses in the PDAC CT radiomics literature, including small sample sizes, use of non-standardized features definitions / software, lack of rigorous statistical methods, as well as lack of external validation. Therefore, we designed the present study to address some of these issues, guided by the Radiomics Quality Score, the Image Biomarker Standardization Initiative and adhering to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [1]. The aims of this study are:

(1) To gather a large multi-center sample of patients with radiologically resectable pancreatic adenocarcinoma, who have undergone resection following suitable pre-operative CT imaging and include a large (>100 events) external validation cohort from a different country.
(2) To develop and externally validate a prognostic model, incorporating pre-operative clinical and radiomic variables, which predicts survival in resectable pancreatic adenocarcinoma, using methodology and software which adheres to the latest international standards for studies in radiomics.

4.2. Study design.

This was a multi-center retrospective international study, for the development and external validation of a prognostic CT radiomic model in PDAC. It was designed in accordance with the latest guidelines from the International Image Biomarker Standardization Initiative and the European Society of Radiology [2-5]. This

manuscript was written in accordance with reporting guidelines from the Transparent Reporting of multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [1].

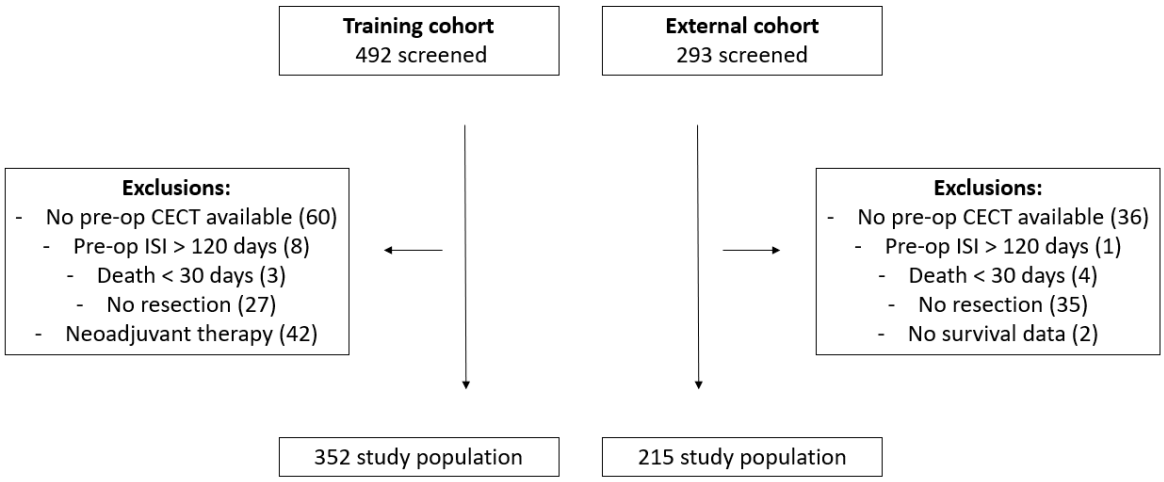
4.3. Ethical approval.

Ethics approval was granted by the Research Ethics Committees in Mount Sinai Hospital Toronto, Canada and St Vincent's University Hospital, Dublin, Ireland. The approval letters are included in the appendix. The main issue which arose during the ethics application related to the transfer of patient data out of Europe, within the context of the European Union General Data Protection Regulation (GDPR). This required consultation with the Canadian Medical Protective Association (CMPA), the Clinical Indemnity Scheme in Ireland and the Medical Protection Society (MPS). We also sought advise from Professor Yann Joly from the Centre for Genomics and Policy in McGill University in Montreal. Ultimately, approval was granted at both sites, allowing both sites to sign a data transfer agreement.

4.4. Patient datasets.

Patients were retrospectively enrolled in this study using the following inclusion criteria: adults who underwent resection for radiologically resectable PDAC and with suitable contrast enhanced pre-operative CT imaging, performed within 120 days prior to surgery. Patients who received neoadjuvant chemo/radiotherapy and/or those who died within 30 days of surgery were excluded (Figure 5).

Figure 5. Patient inclusion flowchart. CECT = Contrast enhanced CT. ISI = Imaging to Surgery time Interval.



4.4.1. Training (internal) dataset.

The internal cohort was retrospectively recruited from a prospective database at the Joint Department of Medical Imaging, University Health Network, Sinai Health Systems and Women's College Hospital, Toronto, Canada. This is a tertiary referral institution. The CTs in this cohort were performed at five hospitals (Toronto General Hospital, Toronto Western Hospital, Mount Sinai Hospital, Princess Margaret Cancer Center and Women's College hospital), all of which are served by the Joint Department of Medical Imaging. All resections were performed at one hospital (Toronto General Hospital) between 2005-2018, where treatment decisions are made in a multi-disciplinary team including surgery, oncology, radiation oncology and radiology, according to the NCCN guidelines [6]. These patients will be referred to as the training cohort in the remainder of this document.

4.4.2. Validation (external) dataset.

The external cohort was retrospectively recruited from a prospective database at the National Surgical Centre for Pancreatic Cancer, St Vincent's University Hospital

(SVUH), Dublin, Ireland. The CT scans were performed at 34 separate radiology departments throughout Ireland. All resections were performed at SVUH between 2010-2016. Treatment decisions were made in a multi-disciplinary team including surgery, oncology, pathology and radiology, according to the NCCN guidelines [6].

4.4.3. Makeup of training and validation datasets.

When designing this study, a decision had to be made about splitting the data (i.e. keep the internal and external datasets from Canada and Ireland separate, or mix them together and split the data in another manner, such 70:30% training:validation. This topic is discussed in more detail in section 1.3, Data for AI). The RQS promotes keeping the validation dataset separate from the training cohort, giving points for validation performed using data from distinct institutes. A recent position paper on Radiomics methodology from European Radiology [4] highlighted that researchers should aim to maximise the difference between training and validation datasets, saying that authors who combine data from multiple centres are 'missing a golden opportunity to evaluate their model properly' because non-random spitting has been shown to reduce overfitting [4]. We therefore decided to keep the Toronto and Irish datasets separate, since there are undoubtedly substantial differences between the populations (ethnicity, North American vs European hepatobiliary surgical practices) which would serve to provide a robust test for a model developed in one cohort and tested in the other.

4.4.4. Clinical data.

The following clinical data were collected in both cohorts: Patient age (on day of surgery), gender, Imaging-to-surgery time interval in days (ISI), pre-operative Ca19.9 result, presence of biliary stent, tumour location (head, body, tail), pathological T and N stages (as per AJCC TNM system 7th edition), pathological tumour size, tumour grade and differentiation, presence of perineural and/or lymphovascular invasion. Follow up was performed until 30th September 2020 and time to recurrence and/or death were recorded. For calculation of time to recurrence and/or death, date of surgery was considered day zero, which is consistent with the methodology of all

prior studies in this field [7-16]. Missing Ca19.9 data was handled by predictive mean matching imputation (a commonly used method in this field [17]), since imputation is preferred in the statistics literature to excluding cases with missing data when building prognostic models [18, 19].

4.5. Radiomics analysis pipeline: Segmentation and features extraction.

A flow chart of the radiomic pipeline is provided in figure 6. All studies were obtained in standard Digital Imaging and Communications in Medicine (DICOM) format. An open-source software package called DICOMsort was used to sort DICOM files by series number, so that the appropriate axial portal-venous series could be selected for analysis. An open-course software package called Dicomtocsv was then used to extract the following technical CT parameters from the DICOM header of every examination: PatientSex, StudyDescription, NumberOfReferences, SeriesNumber, SeriesDescription, Modality, KVP, Exposure, ExposureTime, XRayTubeCurrent, SpiralPitchFactor, SingleCollimationWidth, TotalCollimationWidth, DistanceSourceToPatient, SpacingBetweenSlices, PixelSpacing, SliceThickness, FilterType, Manufacturer, ManufacturerModelName, ReconstructionDiameter and ConvolutionKernel.

Figure 6. Flow chart of the data extraction and analysis steps.

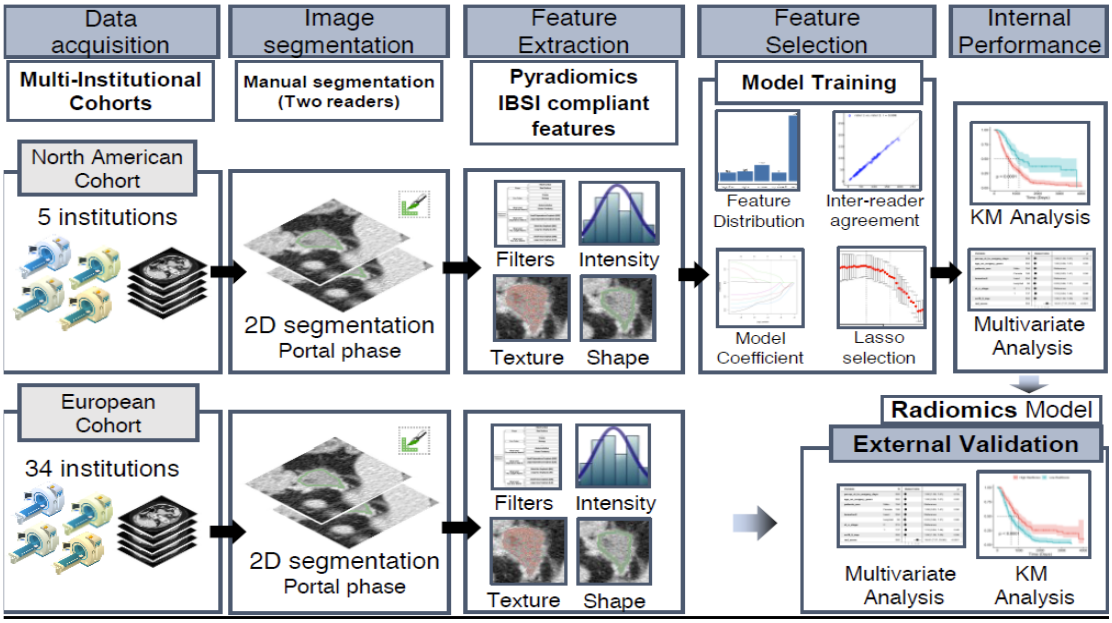


Figure 6: CT images were acquired from participating sites and manual 2D segmentation of the pancreatic tumour was performed. Radiomics features were extracted using the Pyradiomics library and prognostic models were developed on the training cohort. This model was then tested in the external cohort.

4.5.1. Segmentation

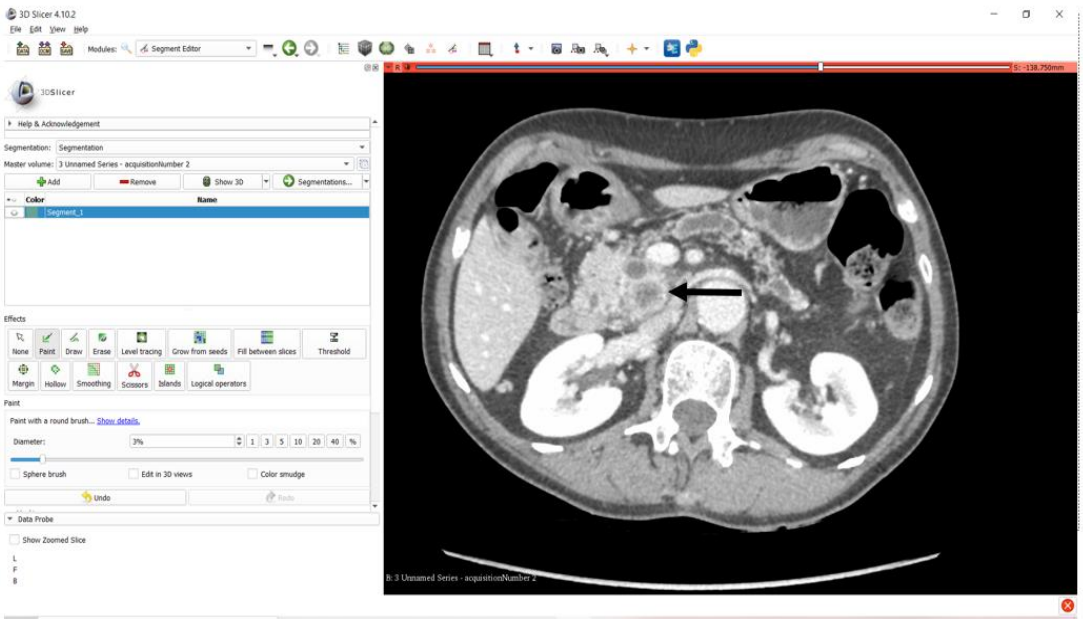
DICOM images were converted to Nearly Raw Raster Data (NRRD) files using 3D Slicer version 4.11.2, an open-source software package (https://www.slicer.org/). Segmentation was performed using the segmentation module of the draw tool in the editor module of 3D Slicer. The NRRD volume was first rotated to volume plain and then contours were manually drawn around the pancreatic tumours (Figure 7). Blood vessels and stents were avoided.

Segmentation was performed by two radiology fellows (each with 6 years of radiology experience) who were blinded to the patient's demographic, pathology and survival information. In the Portal venous (PV) phase, 2D segmentation was performed on the axial slice with the largest tumour diameter (Figure 7). Blood vessels and stents were avoided. Separate contours were first created independently by each fellow for 145 (41%) cases and these were used for inter-

rater reproducibility analysis (which is recommended by the RQS [2]). Then all cases were reviewed in consensus by the fellows, to create a mutually agreed upon 'reference' contour. A third radiologist (> 30 years experience) decided, if there was lack of consensus. These consensus contours were used for all radiomics analysis.

Figure 7. Segmentation example.

Screenshot A



Screenshot B

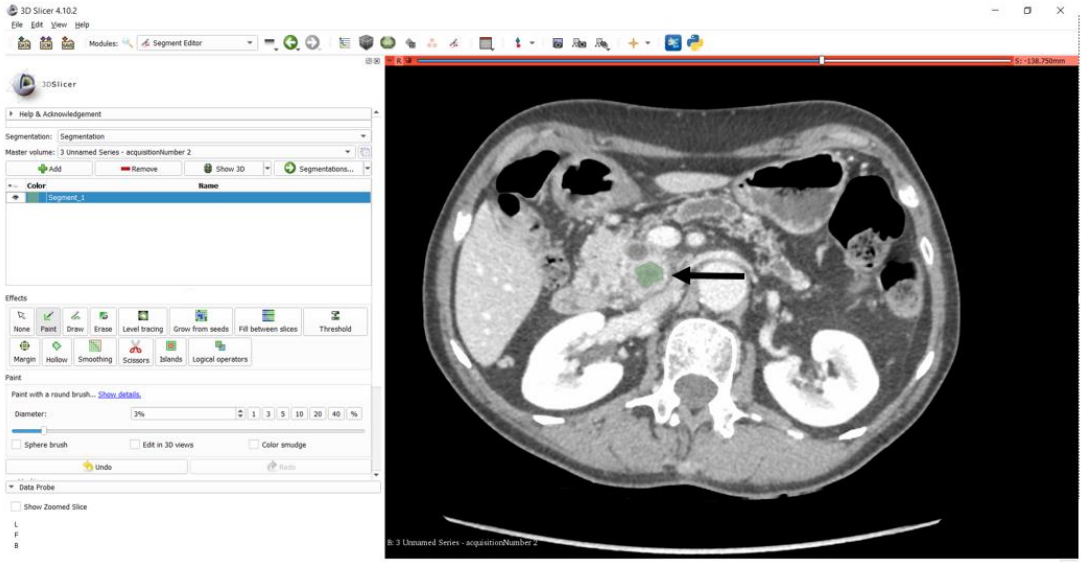


Figure 7 caption: Screenshot A shows an axial slice from a contrast enhanced CT at the level of the pancreatic tumour (arrow). In screenshot B, the tumour area has been segmented (arrow) in preparation for feature extraction. These images are from 3D Slicer, the open source DICOM segmentation software which was used for this study.

4.5.2. Lymph node classification

Each case was classified as positive/negative for lymphadenopathy based upon the presence of one or more lymph nodes with short axis diameter ≥ 1cm on the pre-operative CT study (designated CT-N-stage). This was performed in consensus by the two readers.

4.5.3. Feature extraction

Radiomic feature extraction was performed with the PyRadiomics platform, version 3.0 [20]. This software is compliant with the Image Biomarker Standardisation Initiative (IBSI) and it has been robustly validated using independent datasets [21, 22]. It is a free, open-source software based upon the programming language Python, which extracts 120 IBSI compliant radiomic features and has options to apply filters to the images prior to feature extraction, in order to expand this number beyond 120. These filters perform transformations of the pixel values prior to radiomic features extraction, for example a square root filter will calculate the

square root of each pixel value. We choose to focus on original (unfiltered) features for our primary analysis, because (1) the IBSI has thus far excluded image filters from their standardization guidelines [3], (2) to reduce the chance of overfitting due to high numbers of predictor variables compared to the number of patients in the training cohort and (3) because it is easier to interpret the biological meaning of original features compared to filtered features (for example, it is easier to interpret the meaning of the median attenuation value of a tumour image, rather than the median value of a tumour image where every pixel value has been preprocessed with an exponential filter). However, since several prior PDAC radiomic studies have included filtered images[23], we decided to include all available filters in a sub-analysis, in order to increase the comparability of our results to prior studies. The available filters in PyRadiomics are: square, square root, logarithm, exponential, gradient, LocalBinaryPattern2D and wavelet (low-high, high-low, high-high and low-low), thus expanding the total number of features to 1037. All pixels were normalized to 1mm prior to features extraction, a technique which has been shown in multiple studies to reduce variability in radiomic values [24-26].

4.6. Rad-score construction.

A Rad-score was developed in the training cohort for the prediction of OS. Radiomic features were first standardized using Z-transformation (using the feature means and standard deviations from training cohort to standardize both cohorts). Then the following steps were performed:

(1) In the first feature selection step, features with zero variance were removed. Univariable Cox Proportional Hazard (CPH) regression was then performed on the training cohort and significant features ($p<0.05$) were selected.
(2) In the next step, we removed features with Spearman correlation coefficient ≥ 0.8. To choose between highly correlated features, one feature was selected based upon robustness to outliers (i.e., median chosen over mean).

The decision to use Spearman vs Pearson correlation for this step was queried by reviewers for one of our publications arising from this work [27]. Either test is appropriate for our analysis, since we have large databases, and we are comparing continuous variables. However, Spearman is a rank correlation test, which is more robust, since it carries less assumptions and can handle continuous or categorical variables easily, even in small sample sizes [28]. Therefore, we decided that Spearman would be a more appropriate test for 'state-of-the-art' radiomics, if others may want to replicate our methodology.

(3) The final model was then developed with the least absolute shrinkage and selection operator (LASSO) including previously selected features and all possible two-way interaction terms. The LASSO hyperparameter regularization penalty (λ) was optimized with 5-fold cross validation. The range used for tuning was 0.0012-0.1353 and the optimal λ value was 0.0231. We used the glmnet package of R to perform LASSO. This package uses coordinate descent algorithm to cover the regularization parameter, λ, starting the algorithm from the λ value at which no predictor is selected. The use of optimization algorithm speeds up the optimal λ finding process and prevents the requirement to traverse all the potential values of λ. A numeric Rad-score was calculated by summing the selected feature values, weighted by their coefficients from LASSO (i.e. [feature value 1 x LASSO regression coefficient for feature 1] + [feature value 2 x LASSO regression coefficient for feature 2] + ...).

4.7. Clinical, Clinical-Radiomic and TNM Model construction

Three Cox proportional hazard models were fitted in the training cohort including (1) pre-operative clinical variables (clinical) (2) rad-score plus clinical variables (clinical-radiomic) and (3) pathological TNM classification from the surgical specimen. Model 1 and 2 cover the pre-operative scenario and model 3 served as the reference.

4.8. Feature harmonization.

It has been shown that variation in technical parameters influences the performance of radiomics [29]. In a multisite study, variation in technical parameters is common, for example different manufacturer of CT scanner used at different hospitals. Combat is a statistical method which is used to counteract this effect by realigning feature values after a dataset has been grouped into batches (for example grouped by CT manufacturer) [26, 30]. Combat was originally developed for use in genomics analysis, to adjust for 'batch-effect', where a particular group of experiments within a study may be different to the rest of the group because they were performed using a particular batch of reagents, or by a particular technician or at a certain lab temperature etc. This tool has since been adopted by the radiomics community where it has been shown to impact positively upon results in several cancer types [31, 32], however, to the best of our knowledge, it has not been used in PDAC CT radiomics to date. It is available freely from *Fortin et al* [33] who have developed implementations for R, Matlab and Python: https://github.com/Jfortin1/ComBatHarmonization. To identify batch effects in our cohort prior to Combat implementation we performed hierarchal clustering. Then, the following batching variables were attempted: manufacturer, slice thickness, reconstruction kernel, manufacturer+slice thickness and kernel+slice thickness. For batching variable, a new dataset was created, separately for the training and external cohorts and the radiomic pipeline described in the manuscript was performed.

4.9. Statistics.

The primary study endpoints were OS and DFS. Baseline variables were compared between cohorts using chi square and Kolmogorov-Smirnow tests for categorical and continuous variables, respectively.

Inter-rater agreement of segmentations and radiomics scores were assessed by Dice similarity coefficient and interclass correlation coefficient (ICC), respectively. The Sørensen–Dice similarity coefficient, also called the F1 score, is a 'spatial overlap index' and it can be used to characterise the amount of overlap between two image segmentations [34]. This metric is regularly used in comparing similarity between radiology image segmentations [34]. A value of 0 indicates that there is no overlap and a value of 1 indicates that the two segmentations are identical. ICC is a measure of agreement which is regularly used in radiomic inter-rater agreement analysis. It assesses the numeric radiomic features values compared between groups (in this case reader 1 and reader 2), rather than the segmentation images. Similar to Dice, values range from 0 (no agreement) to 1 (identical values).

Length of follow-up was calculated using the reverse Kaplan-Meier method (event coded as 0 and censoring coded as 1). The association of variables with OS and DFS in the training and external cohorts was assessed by multivariable Cox regressions analysis. To evaluate the risk stratification enabled by the Rad-score, the median value from the training cohort was used to dichotomize both the training and external cohort into high risk and low risk groups. Using median cut-off for dichotomization is recommended in the radiomics guidelines [2], unless a previously published cut-off is available (hence, using optimal cut-off techniques is discouraged). Differences in OS and DFS between risk groups were estimated using the Kaplan-Meier method and compared using the log-rank test. The discriminatory ability for OS and DFS was evaluated using Harrell's concordance index (c-index) and compared using t-tests. 95% confidence intervals (CI) were calculated based on 2000 bootstrap replicates. Model calibration was visually assessed using calibration

curves and quantified using mean absolute prediction error. Clinical utility was assessed by comparing net benefit of the models, graphed using decision curve analysis [35, 36]. Missing Ca19.9 and ISI data were substituted using predictive mean matching imputation [37]. Imputation models were built on the training cohort, then used to impute both cohorts. Statistics were performed using R v3.5.0 (R project for statistical computing). The 'R' packages used were: glmnet (for Lasso), survival (for Cox regression), pec and RMS (for calibration results), ggplot2 (for preparing plots) and github.com/ddsjoberg/dcurves (for decision curve analysis).

5.0. Chapter 4 references.

1. Collins, G.S., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement.* Br J Surg, 2015. **102**(3): p. 148-58.
2. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nat Rev Clin Oncol, 2017. **14**(12): p. 749-762.
3. Zwanenburg, A., et al., *The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping.* Radiology, 2020. **295**(2): p. 328-338.
4. Halligan, S., Y. Menu, and S. Mallett, *Why did European Radiology reject my radiomic biomarker paper? How to correctly evaluate imaging biomarkers in a clinical setting.* Eur Radiol, 2021. **31**(12): p. 9361-9368.
5. Fournier, L., et al., *Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers.* Eur Radiol, 2021. **31**(8): p. 6001-6012.
6. *NCCN Cinical Practice Guidelines in Oncology : Pancreatic Adenocarcinoma*. 2021.
7. Xie, T., et al., *Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and TNM staging for survival estimation after curative resection.* European Radiology, 2020. **30**(5): p. 2513-2524.
8. Shi, H., et al., *Survival prediction after upfront surgery in patients with pancreatic ductal adenocarcinoma: Radiomic, clinic-pathologic and body composition analysis.* Pancreatology, 2021. **21**(4): p. 731-737.
9. Zhang, Y., et al., *CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging.* BMC Med Imaging, 2020. **20**(1): p. 11.
10. Li, K., et al., *Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma.* Am J Transl Res, 2019. **11**(7): p. 4491-4499.
11. Khalvati, F., et al., *Prognostic Value of CT Radiomic Features in Resectable Pancreatic Ductal Adenocarcinoma.* Sci Rep, 2019. **9**(1): p. 5449.

12. Kim, H.S., et al., *Preoperative CT texture features predict prognosis after curative resection in pancreatic cancer.* Sci Rep, 2019. **9**(1): p. 17389.
13. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.
14. Yun, G., et al., *Tumor heterogeneity of pancreas head cancer assessed by CT texture analysis: association with survival outcomes after curative resection.* Sci Rep, 2018. **8**(1): p. 7226.
15. Cassinotto, C., et al., *Resectable pancreatic adenocarcinoma: Role of CT quantitative imaging biomarkers for predicting pathology and patient outcomes.* Eur J Radiol, 2017. **90**: p. 152-158.
16. Eilaghi, A., et al., *CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma - a quantitative analysis.* BMC Med Imaging, 2017. **17**(1): p. 38.
17. Xu, D., et al., *Prognostic Nomogram for Resected Pancreatic Adenocarcinoma: A TRIPOD-Compliant Retrospective Long-Term Survival Analysis.* World J Surg, 2020. **44**(4): p. 1260-1269.
18. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.
19. Steyerberg, E., *Clinical Prediction Models: A practical Approach to Development, Validation and Updating. Second Edition.* 2009: Springer.
20. van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.
21. *The image biomarker standardisation initiative*. Available from: https://ibsi.readthedocs.io/en/latest/index.html.
22. Fornacon-Wood, I., et al., *Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform.* Eur Radiol, 2020. **30**: p. 6241-6250.
23. Kaissis, G., et al., *Image-Based Molecular Phenotyping of Pancreatic Ductal Adenocarcinoma.* Journal of Clinical Medicine, 2020. **9**(3): p. 724.
24. Mackin, D., et al., *Harmonizing the pixel size in retrospective computed tomography radiomics studies.* PLoS One, 2017. **12**(9): p. e0178524.
25. Shafiq-Ul-Hassan, M., et al., *Voxel size and gray level normalization of CT radiomic features in lung cancer.* Sci Rep, 2018. **8**(1): p. 10545.
26. Ligero, M., et al., *Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis.* Eur Radiol, 2021. **31**(3): p. 1460-1470.
27. Healy, G.M., et al., *Pre-operative radiomics model for prognostication in resectable pancreatic adenocarcinoma with external validation.* Eur Radiol, 2021.
28. de Winter, J.C., S.D. Gosling, and J. Potter, *Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data.* Psychol Methods, 2016. **21**(3): p. 273-90.
29. Berenguer, R., et al., *Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters.* Radiology, 2018. **288**(2): p. 407-415.
30. Beaumont, H., et al., *Harmonization of radiomic feature distributions: impact on classification of hepatic tissue in CT imaging.* Eur Radiol, 2021. **31**(8): p. 6059-6068.

31. Masson, I., et al., *Statistical harmonization can improve the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers.* Med Phys, 2021. **48**(7): p. 4099-4109.
32. Saint Martin, M.J., et al., *A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study.* Magma, 2021. **34**(3): p. 355-366.
33. Fortin, J.P., et al., *Harmonization of multi-site diffusion tensor imaging data.* Neuroimage, 2017. **161**: p. 149-170.
34. Zou, K.H., et al., *Statistical validation of image segmentation quality based on a spatial overlap index.* Academic radiology, 2004. **11**(2): p. 178-189.
35. Vickers, A.J. and E.B. Elkin, *Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.* Medical Decision Making, 2006. **26**(6): p. 565-574.
36. Zhang, Z., et al., *Decision curve analysis: a technical note.* Annals of Translational Medicine, 2018. **6**(15): p. 19.
37. Little, R., *Missing-data adjustments in large surveys.* Journal of Busniess and Economic Statisics, 1988. **6**: p. 287-296.

Chapter 5 – Results

5.1. Training and external cohorts.

The training and external validation cohorts included 352 and 215 patients, respectively (figure 2, page 44). Median follow up was 58.6 and 61 months in training and external cohorts, with 249 and 151 deaths respectively. Median survival in the training and external cohorts was 25.03 and 26.87 months, which was slightly longer than prior PDAC CT radiomic studies, which ranged from 17.3-24 months [1, 2]. DFS information was available in 97% (343/352) of patients in the training cohort and 92% (198/215) of patients in the external cohort. Median DFS was 12.2 and 19 months respectively. Baseline characteristics of both cohorts are presented in table 3, showing that the external cohort had significantly longer ISI (median 29 days vs 22 days, p=0.0036), a higher proportion of biliary stents (55.8% vs 49.7%, p=0.042) and lower proportion of lymph node metastasis on pathology (63.7% vs 76.4%, p=0.002). There was also a significant difference in the distribution of pathological T stage between the two cohorts, with 7.9% T4 stage in the external cohort compared to 0.9% in the training cohort (p<0.001). CT parameters of both cohorts are presented in table 4, demonstrating (1) considerable heterogeneity in CT parameters throughout both cohorts and (2) significant differences in the distributions of CT parameters between the two cohorts.

Table 3. Baseline patient characteristics, compared between the training and external datasets.

		Training Cohort (n=352)	External Cohort (n=215)	p-value
Pre-operative clinical variables				
ISI (days): Median (Q1-Q3)		22 (24)	29 (31)	
Missing		0	34	0.0036
Patients with ISI > 60 days		25 (7.1%)	24 (11.2%)	0.095
Patients with ISI > 90 days		2 (0.6%)	3 (1.4%)	0.306
Ca 19.9 (kU/L): Median (Q1-Q3)		107 (411.5)	83 (174)	
Missing		130	140	0.065
Age - Median (Q1-Q3)		66 (14)	67 (12)	0.24
Sex (Male)		184 (52%)	124 (58%)	0.24
Tumour location (Head)		294 (84%)	190 (88%)	0.14
CT-N-stage (Positive)		137 (39%)	85 (40%)	0.95
Biliary stent (Positive)		175 (49.7%)	120 (55.8%)	0.042
Post-operative variables				
Pathologic T stage	1	17 (4.8%)	16 (7.4%)	
	2	59 (16.7%)	20 (9.3%)	
	3	273 (77.6%)	161 (74.8%)	
	4	3 (0.9%)	17 (7.9%)	
	Missing	0	1 (0.5%)	<0.001
Pathologic N stage	Positive	269 (76.4%)	137 (63.7%)	0.002
	Missing	0	1 (0.5%)	
ISI = Imaging to surgery time interval. Q1-Q3 = Interquartile range.				

Table 4. CT technical parameters compared between the two cohorts.

		Training (n=352)	External (n=215)	p
CT Manufacturer	Siemens	16 (4.5%)	103 (47.9%)	<0.0001
	Toshiba	288 (81.8%)	17 (7.9%)	
	GE	33 (9.4%)	54 (25.1%)	
	Philips	15 (4.3%)	26 (12.1%)	
	Unknown	0	15 (7%)	
Number of CT scanner models		26	24	<0.0001
Slice Thickness	≥5mm	71 (20.2%)	112 (52.1%)	<0.0001
	3-4mm	27 (7.7%)	22 (10.2%)	
	2-2.99mm	233 (66.2%)	19 (8.8%)	
	<2mm	17 (4.8%)	58 (27%)	
	Missing	3 (0.9%)	2 (0.9%)	
No. of Recon Kernels used		13	20	<0.0001
KVP	100	2 (0.6%)	9 (4.2%)	0.0005
	120	346 (98.3%)	185 (86%)	
	130	1 (0.3%)	5 (2.3%)	
	140	0	1 (0.5%)	

	Missing	3 (0.9%)	15	
Exposure (mAs)	<100	197 (56%)	72 (33.5%)	<0.0001
	100-199	85 (24.1%)	77 (35.8%)	
	200-299	38 (10.8%)	12 (5.6%)	
	≥300	6 (0.2%)	1 (0.5%)	
	Missing	26 (7.4%)	53 (24.7%)	
mAs = milliampere seconds.				

5.2. Rad-Score Development.

The mean DICE similarity coefficient for the 145 two reader contours was 0.66 (±0.26). ICC values for all features are presented in table 5. Of 116 extracted radiomics features, nine features with zero variance (constant values for all cases) were excluded. Following univariable Cox regression analysis, using the outcome of OS, 7 features were selected. Using a Spearman correlation coefficient cut-off ≥ 0.8, 4 features were selected (figure 8). All the possible two-way interaction terms were created for the selected features. LASSO was performed on this extended feature space.

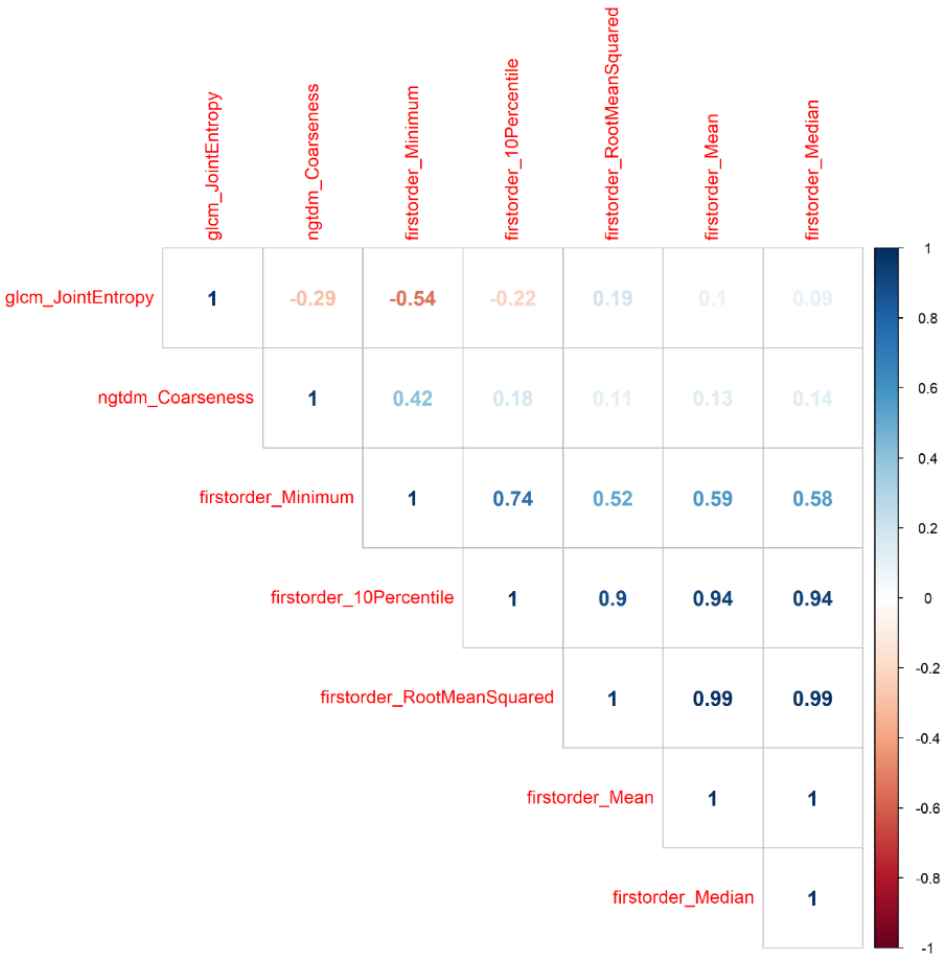
Table 5. Interclass correlation coefficient (ICC) results for the 145 double contour cases. The four features included in the Rad-Score are highlighted in grey.

Feature	ICC	Feature	ICC	Feature	ICC
shape_Elongation	0.371712	glcm_Correlation	0.627705	glszm_GrayLevelNonUniformity	0.83731
shape_MajorAxisLength	0.788304	glcm_DifferenceAverage	0.870147	glszm_GrayLevelNonUniformityNormalized	0.752996
shape_Maximum2DDiameterColumn	0.933547	glcm_DifferenceEntropy	0.892346	glszm_GrayLevelVariance	0.704487
shape_Maximum2DDiameterRow	0.965364	glcm_DifferenceVariance	0.549854	glszm_HighGrayLevelZoneEmphasis	0.473623
shape_Maximum2DDiameterSlice	0.802297	glcm_Id	0.886013	glszm_LargeAreaEmphasis	0.867205
shape_Maximum3DDiameter	0.932459	glcm_Idm	0.888312	glszm_LargeAreaHighGrayLevelEmphasis	0.786322
shape_MeshVolume	0.998478	glcm_Idmn	0.506721	glszm_LargeAreaLowGrayLevelEmphasis	0.824487
shape_MinorAxisLength	0.881003	glcm_Idn	0.629645	glszm_LowGrayLevelZoneEmphasis	0.492072
shape_Sphericity	0.916071	glcm_Imc1	0.675227	glszm_SizeZoneNonUniformity	0.829723
shape_SurfaceArea	0.982099	_glcm_Imc2	0.609499	glszm_SizeZoneNonUniformityNormalized	0.6275
shape_SurfaceVolumeRatio	0.974063	glcm_InverseVariance	0.54421	glszm_SmallAreaEmphasis	0.653158
shape_VoxelVolume	0.998605	glcm_JointAverage	0.445808	glszm_SmallAreaHighGrayLevelEmphasis	0.566184
firstorder_10Percentile	0.744635	glcm_JointEnergy	0.790809	glszm_SmallAreaLowGrayLevelEmphasis	0.46365
firstorder_90Percentile	0.735985	glcm_JointEntropy	0.817108	glszm_ZoneEntropy	0.671019
firstorder_Energy	0.596089	glcm_MCC	0.611328	glszm_ZonePercentage	0.846014

firstorder_Entropy	0.766121	glcm_MaximumProbability	0.758479	glszm_ZoneVariance	0.85992
firstorder_InterquartileRange	0.702731	glcm_SumAverage	0.445808	gldm_DependenceEntropy	0.598877
firstorder_Kurtosis	0.407557	glcm_SumEntropy	0.675455	gldm_DependenceNonUniformity	0.859418
firstorder_Maximum	0.640854	glcm_SumSquares	0.69776	gldm_DependenceNonUniformityNormalized	0.697648
firstorder_MeanAbsoluteDeviation	0.748937	glrlm_GrayLevelNonUniformity	0.88418	gldm_DependenceVariance	0.70639
firstorder_Mean	0.742761	glrlm_GrayLevelNonUniformityNormalized	0.766067	gldm_GrayLevelNonUniformity	0.892725
firstorder_Median	0.743192	glrlm_GrayLevelVariance	0.691853	gldm_GrayLevelVariance	0.694379
firstorder_Minimum	0.591252	glrlm_HighGrayLevelRunEmphasis	0.407265	gldm_HighGrayLevelEmphasis	0.396654
firstorder_Range	0.656799	glrlm_LongRunEmphasis	0.918148	gldm_LargeDependenceEmphasis	0.894683
firstorder_RobustMeanAbsoluteDeviation	0.73518	glrlm_LongRunHighGrayLevelEmphasis	0.321602	gldm_LargeDependenceHighGrayLevelEmphasis	0.285579
firstorder_RootMeanSquared	0.740996	glrlm_LongRunLowGrayLevelEmphasis	0.679033	gldm_LargeDependenceLowGrayLevelEmphasis	0.682379
firstorder_Skewness	0.633041	glrlm_LowGrayLevelRunEmphasis	0.482638	gldm_LowGrayLevelEmphasis	0.482152
firstorder_TotalEnergy	0.989443	glrlm_RunEntropy	0.679913	gldm_SmallDependenceEmphasis	0.851577
firstorder_Uniformity	0.74839	glrlm_RunLengthNonUniformity	0.852376	gldm_SmallDependenceHighGrayLevelEmphasis	0.633255
firstorder_Variance	0.702045	glrlm_RunLengthNonUniformityNormalized	0.86765	gldm_SmallDependenceLowGrayLevelEmphasis	0.462612
glcm_Autocorrelation	0.376815	glrlm_RunPercentage	0.892673	ngtdm_Busyness	0.781734
glcm_ClusterProminence	0.435666	glrlm_RunVariance	0.904747	ngtdm_Coarseness	0.510591

glcm_ClusterShade	0.566624	glrlm_ShortRunEmphasis	0.882632	ngtdm_Complexity	0.639672
glcm_ClusterTendency	0.658669	glrlm_ShortRunHighGrayLevelEmphasis	0.48284	ngtdm_Contrast	0.647604
glcm_Contrast	0.729356	glrlm_ShortRunLowGrayLevelEmphasis	0.443553	ngtdm_Strength	0.709696

Figure 8. Spearman coefficient results for the seven radiomic features which were selected using univariable cox regression analysis, as part of Rad-score building.



The final Rad-score equation comprised of 4 features and 3 two-way interaction terms:

$$Rad - Score = 0.06\,(glcm_{JointEntropy}) - 0.12\,(firstorder_{Median}) - 0.15\,(ngtdm_{Coarseness}) - 0.00\,(firstorder_{Minimum}) - 0.08\,(firstorder_{Median} * firstorder_{Minimum}) - 0.05\,(firstorder_{Median} * glcm_{JointEntropy}) - 0.08\,(glcm_{JointEntropy} * firstorder_{Minimum})$$

The interaction term plots are presented and discussed in figure 9. The coefficients of the other 3 two-way interaction terms were set to zero by LASSO and therefore, do not appear in the regression equation. It is important to note that glcm_JointEntropy and ngtdm_Coarseness have opposite signs within the Rad-score equation, despite the fact that they both measure texture heterogeneity. This is because higher values of glcm_JointEntropy and lower values of ngtdm_Coarseness are both associated with a more heterogenous texture, due to the way they are calculated from the image. The Rad-score equation was then used to calculate a numeric Rad-score for each patient (Figure 10).

Figure 9. Interaction plots for the two-way interaction terms included in Rad-score.

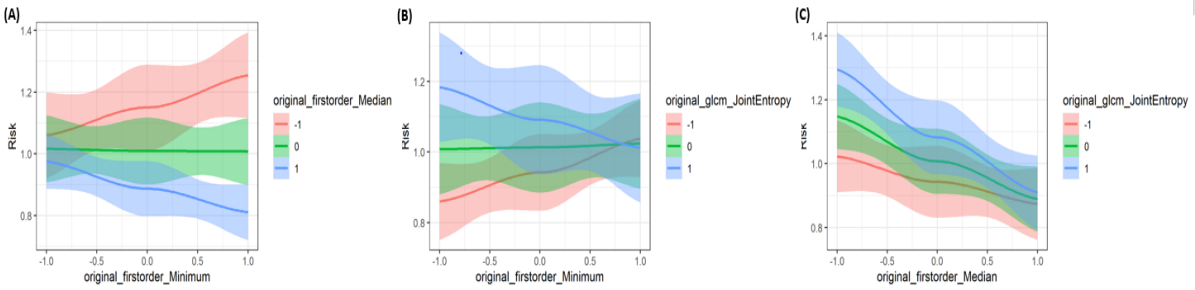


Figure 9 caption: Interaction plots comparing the risk score (probability of event, Y axis) compared with features values, stratified by interaction features values. These plots demonstrate that (A) the risk of death associated with increasing values of original_firstorder_minimum depends on the value of another feature, original_firstorder_median. While in patients with higher values of original_firstorder_median, the risk of death is negatively associated with increasing values

of original_firstorder_minimum, the opposite can be observed for those with lower values of original_firstorder_median. A similar interaction effect is observed between the features first_order_minimum and glcm_JointEntropy (B). For the features first_order_median and glcm_JointEntropy (C), the interaction effect is synergistic. With an increase in the value of first_order_median the risk of death is reduced regardless of the values of glcm_JointEntropy. However, the slope and, therefore, the impact on survival is greater at higher glcm_JointEntropy values compared with lower glcm_JointEntropy values.

Figure 10. Case Examples with calculated Rad-scores.

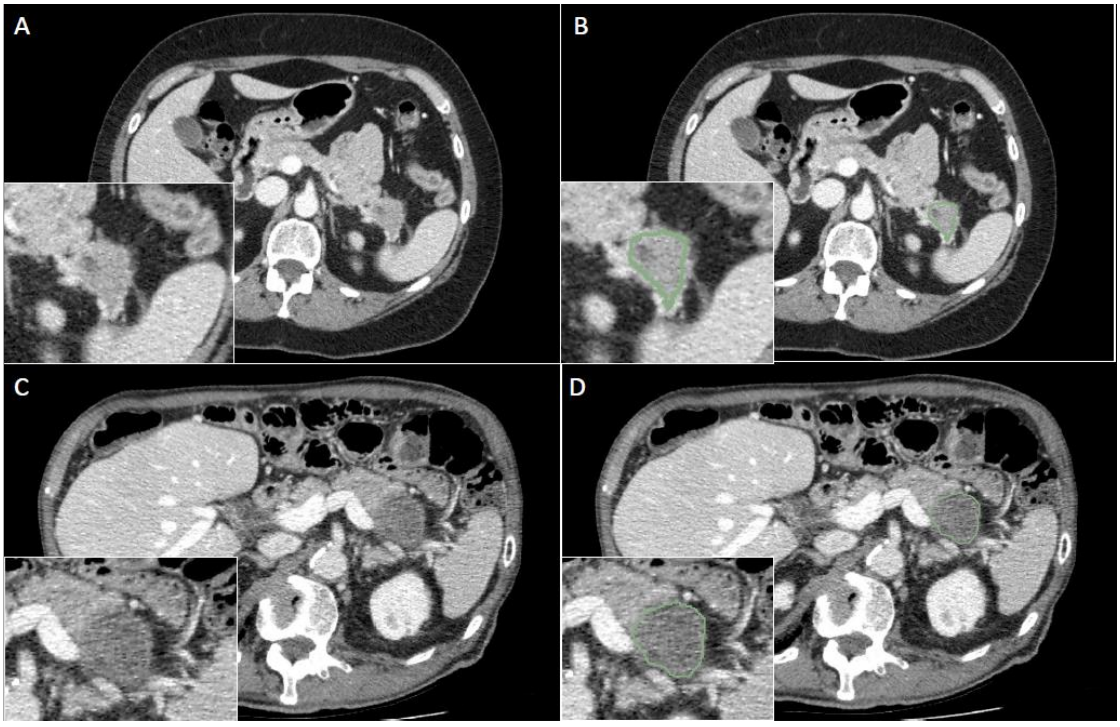


*Figure 10: Axial preoperative CT images of two patients with PDAC (**a, b**). Green contours represent segmented tumour (**b, d**). (a,b) 55 year-old patient with high attenuation (96 HU) and heterogenous pancreatic tail tumour measuring 30 mm. Calculated Rad-score was low (0.017345). The patient died 1140 days following resection. (**c, d**) 81 year-old patient with a low attenuation (36 HU) and homogenous tumour in the pancreatic tail measuring 34mm. Calculated Rad-score was high (0.5590). The patient died 280 days after the resection.*

5.3. Association of Rad-Score and clinical variables with OS and DFS.

At multivariable Cox regression in the training cohort, including pre-operative clinical variables and Rad-score, only Rad-score was significantly associated with OS and DFS with HRs of 3.78 (95% CI: 2.2-6.5, P<0.001) and 2.81 (95% CI: 1.67-4.71, P<0.001) respectively (table 6). At multivariable analysis in the external cohort, only Rad-score and age demonstrated significant associations with OS (Rad-score HR: 2.87, 95% CI: 1.40, 5.87, p<0.001; Age HR: 1.02 95% CI: 1.01, 1.04, p=0.01) and DFS (Rad-Score HR 5.28, 95% CI 2.35-11.86, p<0.001; Age HR: 1.02 95% CI: 1.00, 1.04, p=0.03). Univariable Cox regression analyses for OS in the training cohort for the four radiomic features and three interaction terms included in the Rad-score are presented in table 7.

Using the median Rad-score value from the training cohort (0.029) as the cut-point, the median OS for the groups with high versus low Rad-score were 17.8 versus 32 months in the training cohort (p<0.0001, figure 11) and 22.9 versus 37 months in the external cohort (p=0.0092, figure 9). For DFS, the high versus low Rad-score groups demonstrated median OS of 10.1 versus 15.2 months in the training cohort (p=0.00025) and 14.2 versus 29.8 months in external (p=0.0023, figure 11).

Table 6. Multivariable analysis for the association between clinical variables and Rad-score for the outcome of Overall Survival in the training and external cohorts.

Variable	Training cohort		External cohort	
Overall survival				
	HR (95% C.I)	p-value	HR (95% C.I)	p-value
ISI (days)	1.00 (1.00-1.01)	0.34	1.00 (0.99, 1.00)	0.37

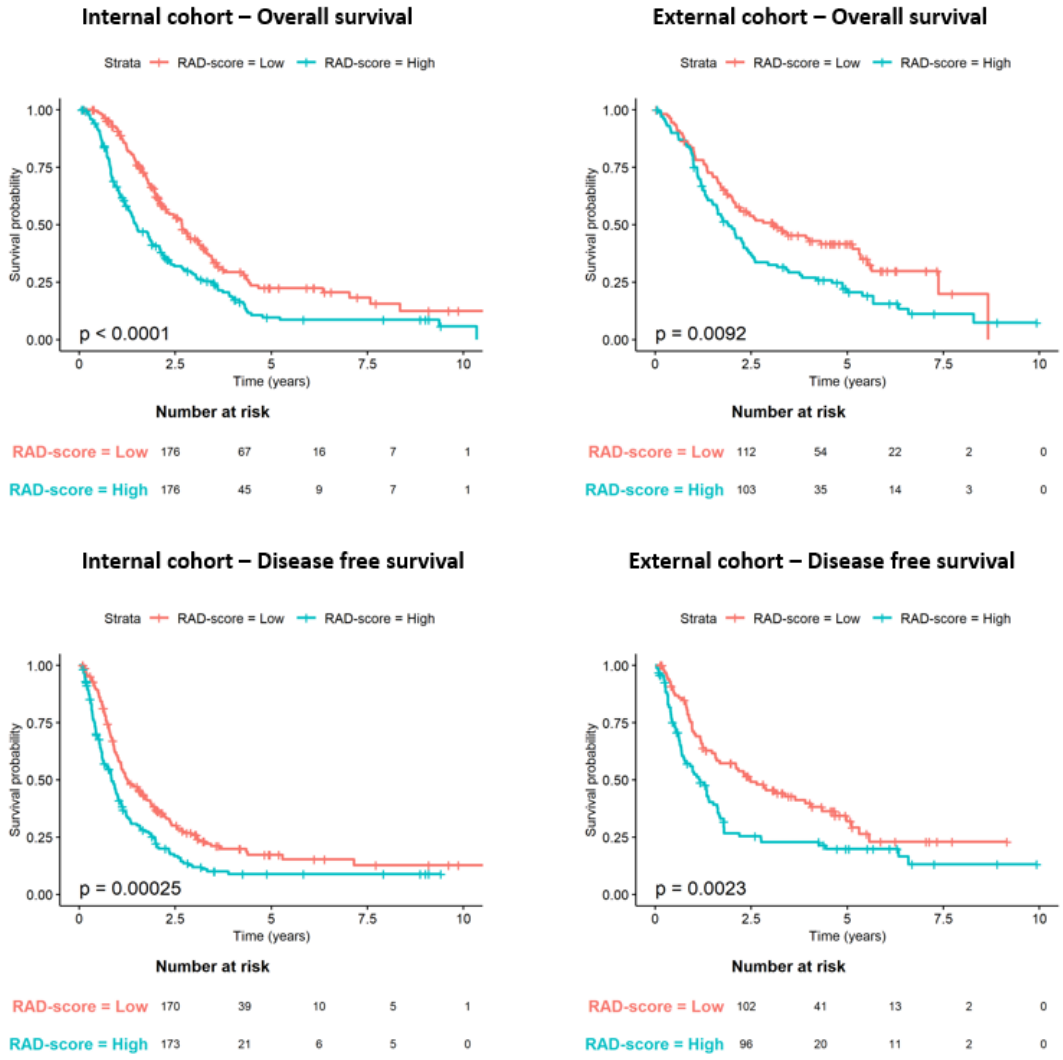
Ca 19.9 (kU/L)	1.00 (1.00-1.00)	0.87	1.00 (1.00, 1.00)	0.68
Age (years)	1.00 (0.98-1.01)	0.57	1.02 (1.01, 1.04)	0.01
Sex (Female)	1.03 (0.79-1.34)	0.85	0.98 (0.70, 1.38)	0.91
Tumour location (Body/Tail)	0.85 (0.60-1.21)	0.36	1.09 (0.65, 1.82)	0.76
CT-N-stage (positive)	1.09 (0.84-1.42)	0.52	1.12 (0.80, 1.56)	0.53
Rad-score	3.78 (2.2-6.5)	<0.001	2.87 (1.40, 5.87)	<0.001
Disease-free survival				
	HR (95% C.I)	p-value	HR (95% C.I)	p-value
ISI (days)	1.00 (1.00-1.01)	0.60	0.99 (0.98-1.00)	0.05
Ca 19.9 (kU/L)	1.00 (1.00-1.00)	0.72	1.00 (1.00-1.00)	0.46
Age (years)	0.99 (0.98-1.01)	0.36	1.02 (1.00-1.04)	0.03
Sex (Female)	1.05 (0.82-1.35)	0.69	0.79 (0.56-1.13)	0.20
Tumour location (Body/Tail)	1.13 (0.81-1.59)	0.47	1.44 (0.82-2.54)	0.20
CT-N-stage (positive)	1.07 (0.82-1.38)	0.63	0.90 (0.63-1.30)	0.59

Rad-score	2.81 (1.67-4.71)	<0.001	5.28 (2.35-11.86)	<0.001
Abbreviation: CA 19.9 = Carbohydrate Antigen; ISI = Imaging to surgery time interval. HR = Hazard ratio; CI = Confidence Interval; CT-N-stage = Nodal classification based on short axis diameter ≥ 1cm.				

Table 7. Univariable Cox proportional hazard analysis for the association between Rad-score features and Overall Survival in the training cohort.

Variable	Univariate	
	HR (95% C.I)	p-value
original_firstorder_Minimum	0.81 (0.71-0.92)	0.002
original_firstorder_Median	0.85 (0.76-0.96)	0.01
original_glcm_JointEntropy	1.14 (1.00-1.30)	0.05
original_ngtdm_Coarseness	0.81 (0.70-0.94)	0.004
original_firstorder_Median:original_firstorder_Minimum	0.95 (0.84-1.07)	0.36
original_firstorder_Median:original_glcm_JointEntropy	0.90 (0.79-1.02)	0.08
original_firstorder_Minimum:original_glcm_JointEntropy	0.91 (0.83-1.00)	0.05

Figure 11. Kaplan Meier analysis of Rad-score for overall survival and disease-free survival in the training and external cohorts.



5.4. Discrimination.

5.4.1. Training cohort.

Discriminatory ability for OS in the training cohort, as indicated by the c-indices, were 0.56 (95% CI: 0.559-561) for clinical, 0.626 (95% CI: 0.625-0.627) for clinical-radiomic and 0.583 (95% CI: 0.583-584) for TNM models (table 7). C-indices for DFS were 0.561 (95% CI: 0.56-562) for clinical, 0.603 (95% CI: 0.602-603) for clinical-radiomic and 0.594 (95% CI: 0.593-594) for TNM models. The TNM discrimination in our training cohort was similar to the performance in recent publications from Shi *et al* (c-index 0.59) and Xu *et al* [3] (c-index 0.572) but inferior to Xie *et al* (0.699), the latter likely due to the inclusion of patients with metastatic disease in their study (11% of their total cohort), since metastatic disease is a major predictor of survival in pancreas cancer. Patients with metastatic disease are do not meet NCCN criteria for resection [4] although the Xie *et al* paper does not detail what guidelines they follow at their institution.

5.4.2. External cohort.

Discriminatory ability for OS were 0.497 (95% CI: 0.496-0.499) for the clinical, 0.545 (95% CI: 0.543-0.546) for the clinical-radiomic and 0.525 (95% CI: 0.524-0.526) for the TNM model. C-indices for DFS were 0.472 (95% CI: 0.47-0.473) for clinical, 0.554 (95% CI: 0.552-0.556) for clinical-radiomic and 0.485 (95% CI: 0.484-0.486) for TNM (table 7). The clinical-radiomic model demonstrated significantly improved performance compared to the clinical model alone or TNM model for both OS and DFS (table 7).

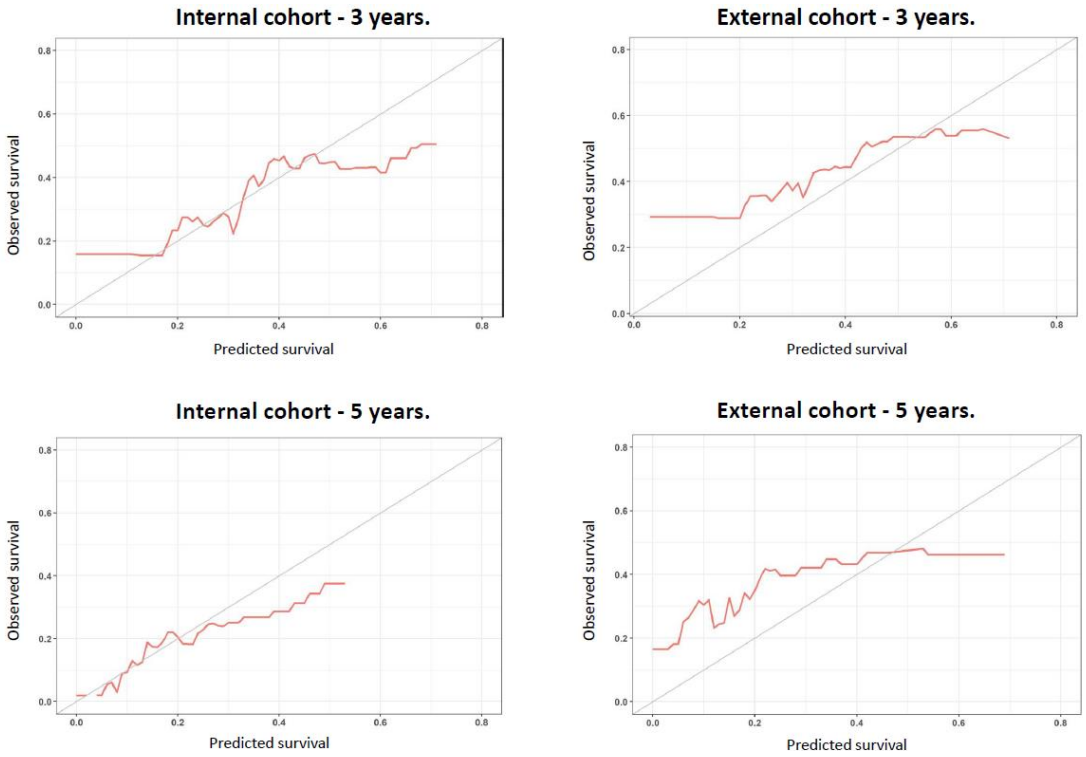
Table 8. Discrimination performance of the models for overall and disease-free survival

	C-index (95% CI) Training cohort	C-index (95% CI) External cohort
Outcome: Overall Survival		
Clinical Model	0.56 (0.559-0.561)	0.497 (0.496-0.499)
Rad-score	0.616 (0.615-0.617)*†	0.564 (0.562-0.565)*†
Clinical +Rad-score	0.626 (0.625-0.627)*†	0.545 (0.543-0.546)*†
AJCC TNM	0.583 (0.583-0.584)	0.525 (0.524-0.526)
Outcome: Disease Free Survival		
Clinical Model	0.561 (0.56-0.562)	0.472 (0.47-0.473)
Rad-score	0.593 (0.592-0.594)*	0.573 (0.572-0.574) *†
Clinical +Rad-score	0.603 (0.602-0.603)*†	0.554 (0.552-0.556)*†
AJCC TNM	0.594 (0.593-0.594)	0.485 (0.484-0.486)
* P-value < 0.001 compared to clinical model alone. † P-value < 0.001 compared to TNM.		

5.5. Calibration.

Calibration curves demonstrate good calibration in the training cohort for the clinical-radiomic model (Figure 12), with mean absolute prediction error of 3% and 2% for OS at 3 and 5 years, respectively. In the external cohort, calibration curves demonstrate moderate calibration (Figure 12), with mean absolute prediction error 7% and 13% for OS at 3 and 5 years.

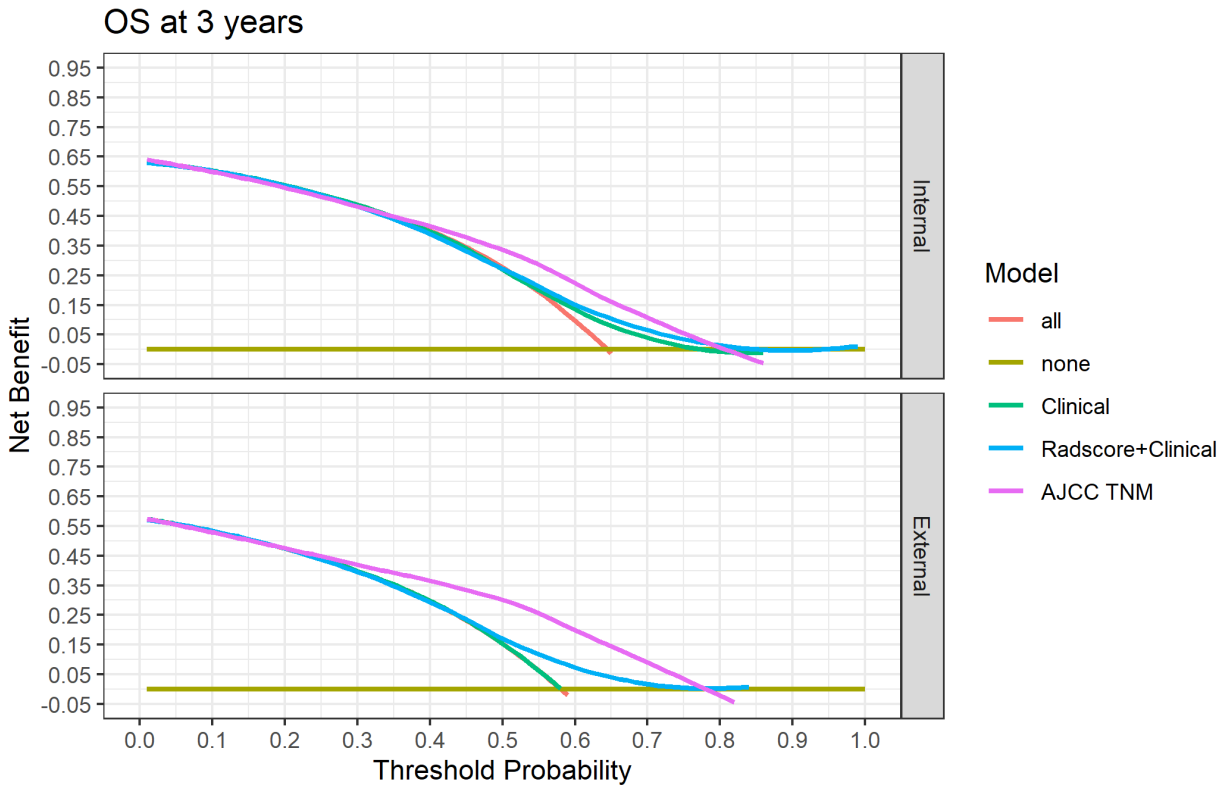
Figure 12. Calibration curves of the clinical and clinical-Radiomic models in the training and external cohorts at 3 and 5 years for overall survival. The y axis shows the observed overall survival probability, while the x axis shows the predicted overall survival probability.

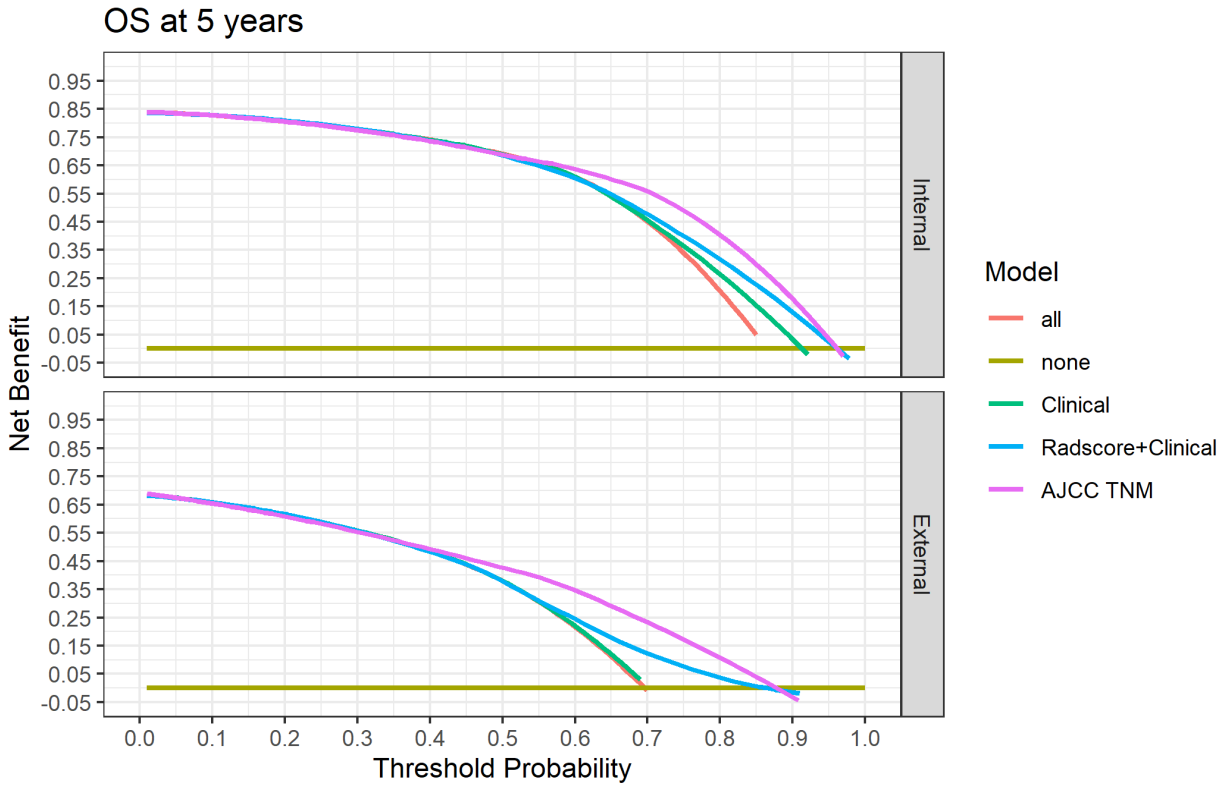


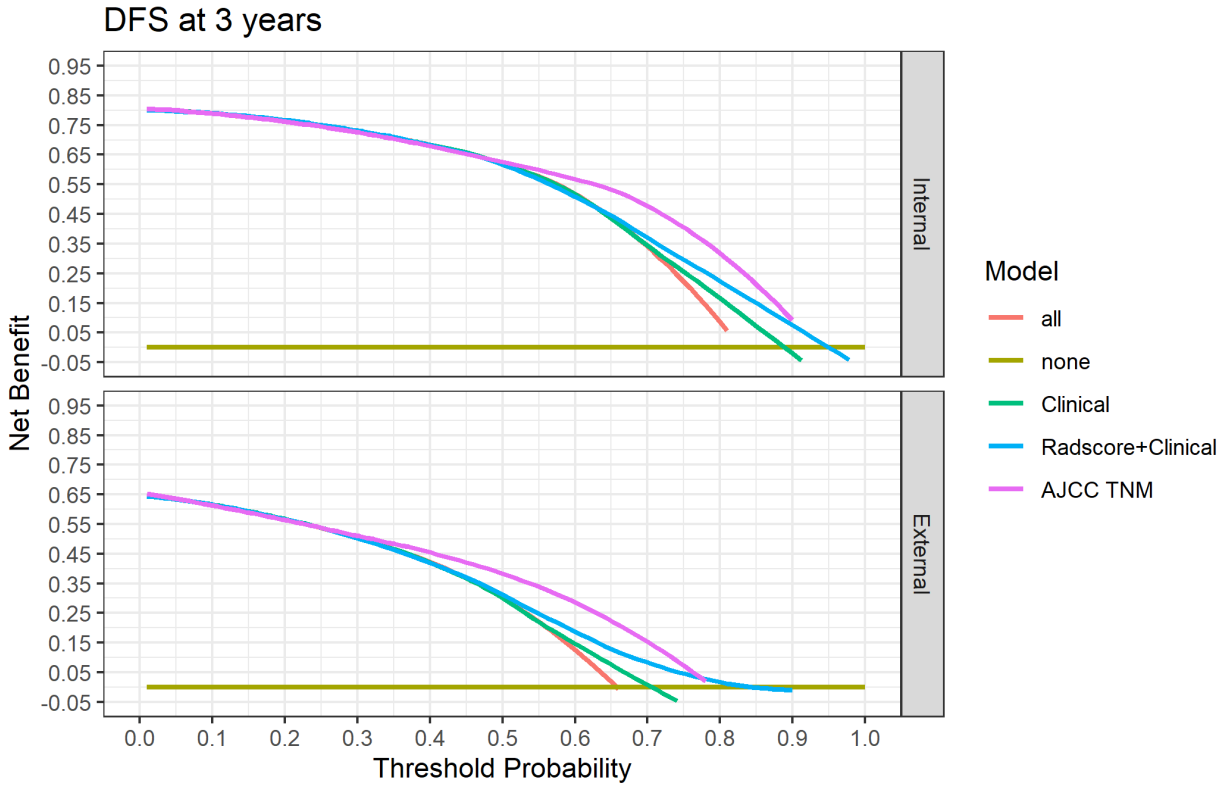
5.6. Decision Curve Analysis.

Decision curves are presented in figure 9 for the three- and five-year survival time points. These demonstrate that the clinical-radiomics model demonstrates a marginally higher net benefit compared to a clinical model alone, however they are both clinically harmful (i.e. crossing below the y=zero line) at higher risk thresholds, indicating that there is no clear clinical utility of these models.

Figure 13. Decision curve analysis of the clinical and clinical-radiomic models in the training and external cohorts at 3 and 5 years. DFS = Disease free survival.




OS at 5 years
Net Benefit
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
-0.05
Internal
External
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Threshold Probability
Model
all
none
Clinical
Radscore+Clinical
AJCC TNM


DFS at 3 years
Net Benefit
0.95
0.85
0.75
0.65
0.55
0.45
0.35
0.25
0.15
0.05
-0.05
Internal
External
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Threshold Probability
Model
all
none
Clinical
Radscore+Clinical
AJCC TNM

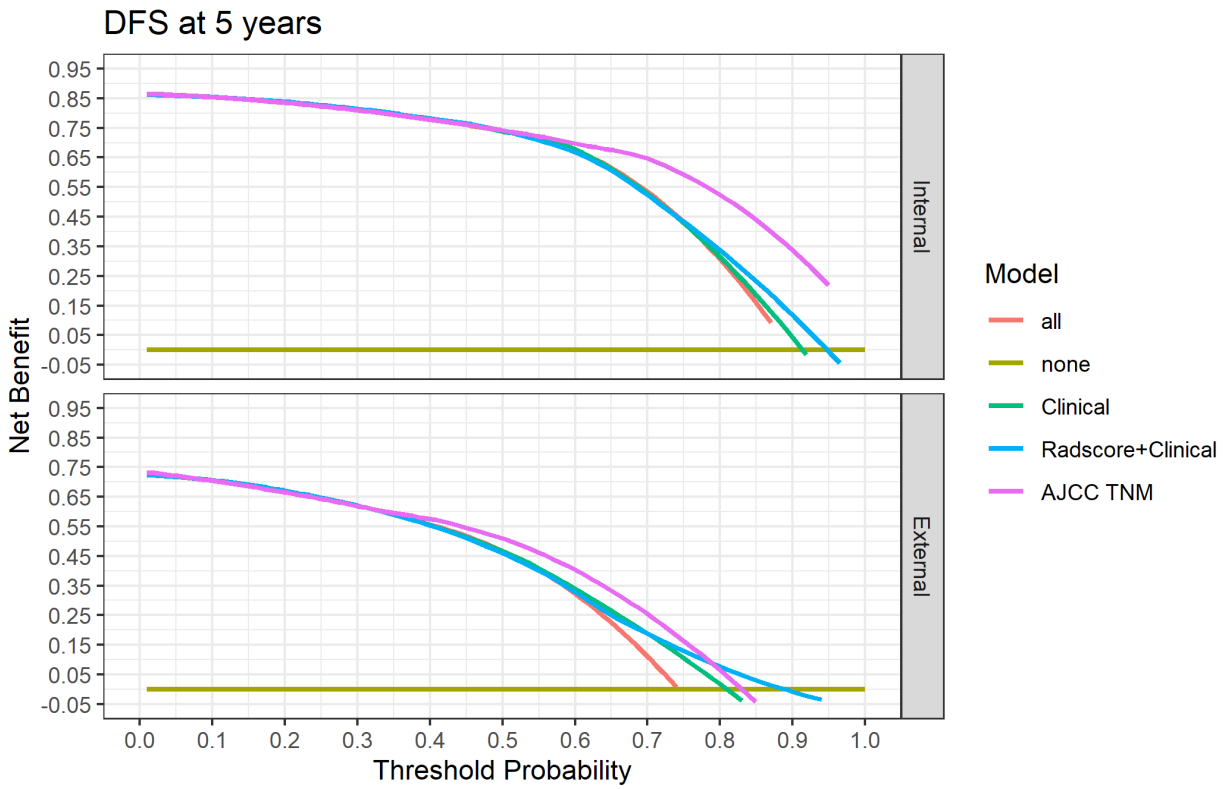


Figure 13 caption: Decision-curve analysis for the clinical and clinical-radiomic models for disease free and overall survival at 3 and 5 years. The y-axis displays the net benefit, which represents the proportion of true positive predictions given by the model accounting for the harms of false positive predictions. The x-axis shows the threshold probability, which in this analysis reflects the clinician's/patient's subjective judgment about what probability of disease recurrence would prompt initiation of treatment (Please note that the amount of benefit derived from the intervention/treatment is not taken into account in DCA – simply the ability of the model to classify patients into treat/no-treat groups). The 'none' line demonstrates no net benefit, since no treatment decisions are made and therefore the true and false positives are both zero, hence the net benefit = 0 at all threshold probabilities [5]. The 'all' line crosses the y axis at the prevalence of the event (approximately 0.63 for OS at 3 years in the internal cohort and 0.85 for OS at 5 years), hence it shows a net benefit for all thresholds below the prevalence, but net harm (dips below the y=0 point) for all thresholds above the prevalence.

At low-risk thresholds, none of the models demonstrate net benefit compared with a 'treat all' patients strategy, i.e., they are not clinically useful at low thresholds. The clinical-radiomics model demonstrated a marginally increased net benefit compared to the clinical model at high-risk thresholds in both cohorts. Net benefit of the clinical-radiomics model compared with default strategies ('treat all'/'treat none') in the external cohort was limited to only a small range of threshold probabilities (e.g., ~55-75% for 5-year OS in the external cohort). The reference TNM model demonstrates superior net benefit compared to the clinical and clinical-radiomic models for almost all threshold probabilities, using both the outcomes of OS and DFS.

5.7. Feature harmonization

Hierarchal clustering of radiomic features was performed, stratified by CT technical parameters, however no clear batch effect was apparent (figure 14). Therefore, empirical batch groupings were created: manufacturer, slice thickness, reconstruction kernel, manufacturer+slice thickness and kernel+slice thickness. No batching combination provided an improvement in results. For example, using the 'manufacturer' batch, 6 features and 5 interaction terms were selected for Rad-score. Discrimination performance (table 8) demonstrated c-index for the clinical-radiomic model of 0.622 (95% CI: 0.585-0.659) in the training cohort and 0.56 (0.559-0.561) in the external cohost, which was no better than the performance achieved without Combat feature harmonization.

Figure 14. Hierarchal clustering of radiomic features, stratified by CT technical parameters.

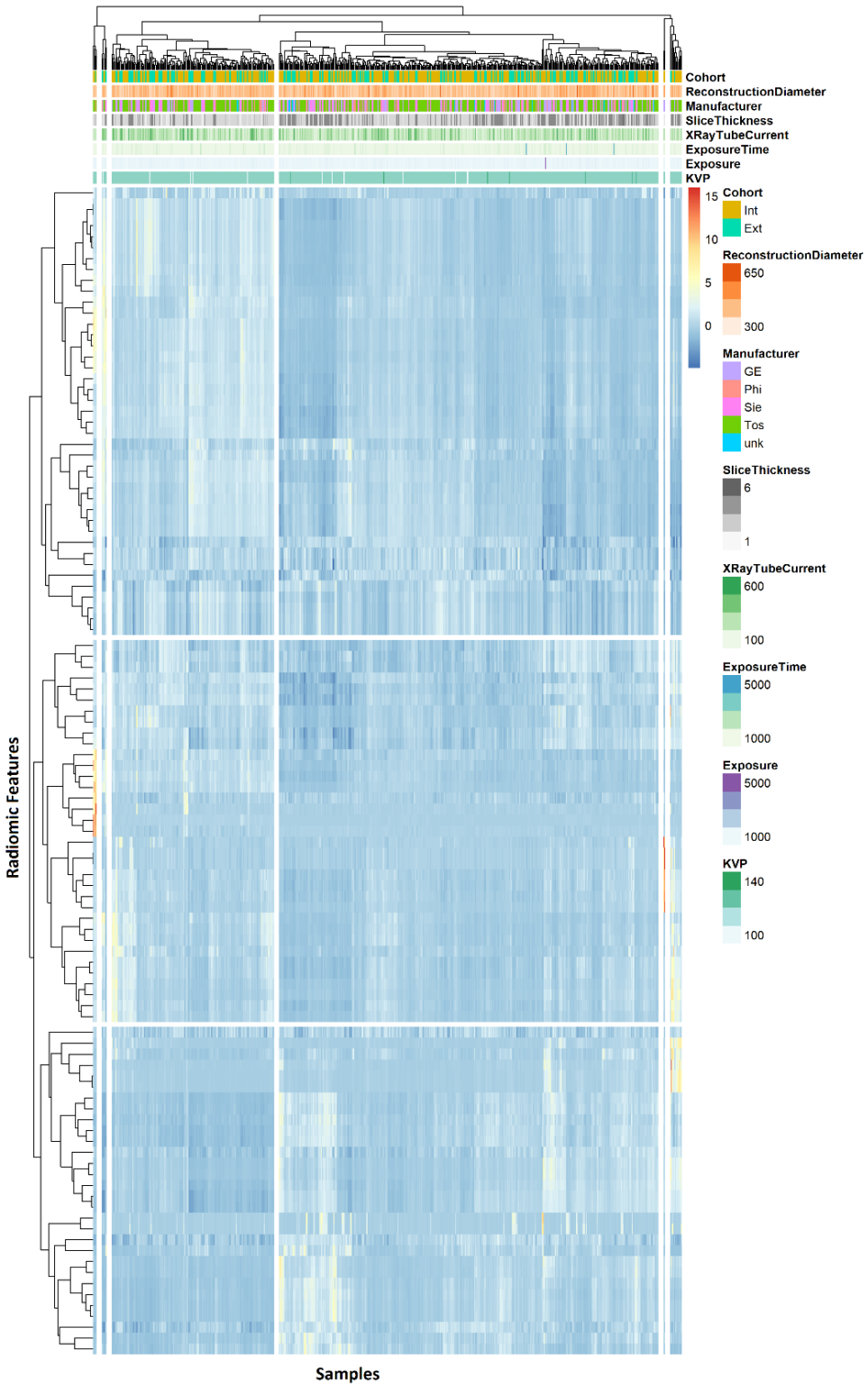


Table 9. Discrimination performance of the model overall survival in the training and external cohorts created using Combat (with manufacturer batching).

	Training C-Index (95% CI)	External C-Index (95% CI)
Clinical model	0.548 (0.509-0.587)	0.495 (0.494-0.497)
Rad-score	0.613 (0.576-0.650)	0.560 (0.559-0.561)
Rad-score + clinical	0.622 (0.585-0.659)	0.560 (0.559-0.561)

5.8. Analysis with filtered features

As described in section 4.5.3, all available PyRadiomics filters were applied in order to expand the total number of radiomics features to 1037. The feature selection pipeline described in section 4.5 was repeated using this combination of original and filtered. 48 features with zero variance were excluded. The remaining features were entered into univariable Cox proportional hazard model for OS and 257 were selected. This was reduced to 26 after selection using the Spearman correlation step. Interaction terms were then identified and all were entered into LASSO, resulting in a Rad-score containing 13 terms. The performance results (Table 9) demonstrated a large performance drop in the external cohort, with performance that was inferior using only original features for the analysis. This is likely due to overfitting in the internal cohort.

Table 10. Discrimination performance of the model for overall survival in the training and external cohorts using original + filtered features (n=1037).

	Training C-Index	External C-Index
Rad-score	0.636	0.488
Rad-score + clinical	0.646	0.49

5.9 Analysis of manufacturer influence on Radiomic features.

To assess the influence of CT scanner manufacturer on the radiomic results, we compared the radiomic features values from the four features included in the Rad-score between manufacturers. We first performed analysis of variance (ANOVA) to compare the features across the manufacturer groups. Features which demonstrated significant difference between groups were retained and further assessed by Tukey honest significance test (HSD) test for pairwise comparisons between different manufacturers. A separate analysis is performed for internal (Toronto) and external (Irish) cohorts. For most pairwise comparisons between manufacturers the value of the radiomic features did not change significantly. In 9 cases, there was a significant difference in the value of radiomic features based on the manufacturer, however, the results are not consistent both across cohort and radiomic features (table 10).

Table 11. Comparison of Radscore radiomic features vs manufacturer.

Feature	**Cohort**	**Pair**	**p.adj**	**Cohort**	**Pair**	**p.adj**
original_firstorder_Median	Ext	Phi-GE	0.9993	Int	Phi-GE	0.9691
original_firstorder_Median	Ext	Sie-GE	0.7898	Int	Sie-GE	0.9773
original_firstorder_Median	Ext	Tos-GE	0.7917	Int	Tos-GE	0.0024
original_firstorder_Median	Ext	Sie-Phi	0.9752	Int	Sie-Phi	1.0000
original_firstorder_Median	Ext	Tos-Phi	0.7615	Int	Tos-Phi	0.0152*
original_firstorder_Median	Ext	Tos-Sie	0.3150	Int	Tos-Sie	0.0140*
original_firstorder_Median	Ext	unk-Phi	0.8617			
original_firstorder_Median	Ext	unk-GE	0.6957			
original_firstorder_Median	Ext	unk-Sie	0.9604			

original_firstorder_Median	Ext	unk-Tos	0.2981			
original_glcm_JointEntropy	Ext	Phi-GE	0.2859	Int	Phi-GE	0.9835
original_glcm_JointEntropy	Ext	Sie-GE	0.9899	Int	Sie-GE	0.9811
original_glcm_JointEntropy	Ext	Tos-GE	0.7367	Int	Tos-GE	0.0001*
original_glcm_JointEntropy	Ext	Sie-Phi	0.0984	Int	Sie-Phi	0.9194
original_glcm_JointEntropy	Ext	Tos-Phi	0.0770	Int	Tos-Phi	0.0360*
original_glcm_JointEntropy	Ext	Tos-Sie	0.8563	Int	Tos-Sie	0.0015*
original_glcm_JointEntropy	Ext	unk-Phi	0.0091*			
original_glcm_JointEntropy	Ext	unk-GE	0.2220			
original_glcm_JointEntropy	Ext	unk-Sie	0.3042			
original_glcm_JointEntropy	Ext	unk-Tos	0.9391			
original_ngtdm_Coarseness	Ext	Phi-GE	0.6513	Int	Phi-GE	1.0000
original_ngtdm_Coarseness	Ext	Sie-GE	0.9725	Int	Sie-GE	0.7178
original_ngtdm_Coarseness	Ext	Tos-GE	0.9889	Int	Tos-GE	0.8021
original_ngtdm_Coarseness	Ext	Sie-Phi	0.2993	Int	Sie-Phi	0.7906
original_ngtdm_Coarseness	Ext	Tos-Phi	0.5853	Int	Tos-Phi	0.9071
original_ngtdm_Coarseness	Ext	Tos-Sie	1.0000	Int	Tos-Sie	0.9313
original_ngtdm_Coarseness	Ext	unk-Phi	0.0578			
original_ngtdm_Coarseness	Ext	unk-GE	0.3240			
original_ngtdm_Coarseness	Ext	unk-Sie	0.4848			
original_ngtdm_Coarseness	Ext	unk-Tos	0.7648			
original_firstorder_Minimum	Ext	Phi-GE	0.9609	Int	Phi-GE	0.9993
original_firstorder_Minimum	Ext	Sie-GE	0.3395	Int	Sie-GE	0.8902
original_firstorder_Minimum	Ext	Tos-GE	0.9929	Int	Tos-GE	0.9683

original_firstorder_Minimum	Ext	Sie-Phi	0.2053	Int	Sie-Phi	0.8916
original_firstorder_Minimum	Ext	Tos-Phi	0.9999	Int	Tos-Phi	0.9981
original_firstorder_Minimum	Ext	Tos-Sie	0.4756	Int	Tos-Sie	0.6413
original_firstorder_Minimum	Ext	unk-Phi	0.0020*			
original_firstorder_Minimum	Ext	unk-GE	0.0031*			
original_firstorder_Minimum	Ext	unk-Sie	0.0559			
original_firstorder_Minimum	Ext	unk-Tos	0.0092*			
*indicates statistical significance at the <0.05 level.						

5.10. Radiomics Quality score

The Radiomics quality score for our study is 18/36 (figure 15), which compared to a score of 14/36 in the two highest scoring studies to date in this field (see section 3.3). The areas where our study gains points are: the use of median value for cut-off analysis when determining risk groups, the use of bootstrap resampling when measuring discrimination and validation using cohorts from two distinct institutions. Our study will also gain an additional point for 'detection and discussion of biological correlates' when our future work in this area is completed (see section 6.7), increasing the score to 19/36.

Figure 15. Radiomics Quality Score calculated on www.radiomics.world.

Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability	☑ protocols well documented ☐ public protocol used ☐ none
Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities	◉ yes ○ no
Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability	○ yes ◉ no
Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)	○ yes ◉ no
Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	◉ Either measure is implemented ○ Neither measure is implemented
Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features	◉ yes ○ no
Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology	○ yes ◉ no
Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	◉ yes ○ no
Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	☑ a discrimination statistic and its statistical significance are reported ☑ a resampling method technique is also applied ☐ none

Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	☑ a calibration statistic and its statistical significance are reported ☐ a resampling method technique is applied ☐ none
Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	○ yes ◉ no
Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	☐ No validation ☐ validation is based on a dataset from the same institute ☐ validation is based on a dataset from another institute ☑ validation is based on two datasets from two distinct institutes ☐ the study validates a previously published signature ☐ validation is based on three or more datasets from distinct institutes
Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics	◉ yes ○ no
Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).	◉ yes ○ no
Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)	○ yes ◉ no
Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	☐ scans are open source ☐ region of interest segmentations are open source ☐ the code is open sourced ☐ radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source
Total score	**18** (50.00%)

5.11. Results summary.

We developed a clinical-radiomic model for prognostication in PDAC, based on pre-operative CT imaging in a multi-institutional cohort of patients from Canada and externally assessed this model in a cohort of patients from Ireland. Our study is the highest quality evidence in this field to date and the only study to incorporate robust external validation. The model identified a signal capable of stratifying patient prognosis and it generalized to the external dataset, but despite superior performance compared to a clinical model alone, overall performance was

suboptimal. The reference post-operative TNM model demonstrated superior net benefit compared to both the clinical and clinical-radiomic models. It is likely that heterogeneity of CT technical factors within and between the cohorts contributed to poor model performance, however there was no clear 'batch effect' apparent to indicate one CT parameter in particular which was the primary influence on the results. Dedicated analysis of the influence of manufacturer on radiomic values did not reveal any consistent relationship and attempts to improve the model performance using Combat feature harmonization (to reduce CT parameter heterogeneity) and the addition of filtered radiomic features, did not prove useful.

5.12. Chapter 5 references.

1. Shi, H., et al., *Survival prediction after upfront surgery in patients with pancreatic ductal adenocarcinoma: Radiomic, clinic-pathologic and body composition analysis.* Pancreatology, 2021. **21**(4): p. 731-737.
2. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.
3. Xu, D., et al., *Prognostic Nomogram for Resected Pancreatic Adenocarcinoma: A TRIPOD-Compliant Retrospective Long-Term Survival Analysis.* World J Surg, 2020. **44**(4): p. 1260-1269.
4. *NCCN Cinical Practice Guidelines in Oncology : Pancreatic Adenocarcinoma*. 2021.
5. Zhang, Z., et al., *Decision curve analysis: a technical note.* Annals of Translational Medicine, 2018. **6**(15): p. 19.

Chapter 6 – Discussion and Conclusion

6.1. Summary of results and study rationale.

We have developed and externally validated a pre-operative clinical-radiomics prognostic model for PDAC, which significantly outperformed a pre-operative clinical model, for both outcomes of disease-free and overall survival. The Rad-score was the only pre-operative prognostic variable in the training cohort and the only prognostic factor, other than age, in the external cohort. The addition of Rad-score to the pre-operative clinical model significantly improved prognostication and added approximately 0.05 to the value of the c-Index, whereas it is suggested that a biomarker should be deemed relevant if it adds more than 0.005 to a model [1].

Compared to the reference standard TNM model, our clinical-radiomic model performed slightly better in terms of discrimination (clinical-radiomics model: C-index 0.545 for OS in external cohort, 0.554 for DFS; TNM C-index 0.525 for OS external cohort, 0.485 for DFS) however TNM demonstrated higher net benefit (Figure 13). The TNM system is based upon post-operative pathological data, which is only available after surgical resection is complete. While it is possible to estimate an 'imaging' TNM pre-operatively, we have shown that estimating these variables on imaging is not accurate (for example, the CT-N-stage for the internal and external cohorts was 39% and 40% positive respectively vs 76.4% and 63.7% positive on pathology, which the gold standard. Table 3). Thus, TNM is not useful for pre-operative decision making, which was the goal of this study.

The importance of this work resides in the pre-operative treatment decisions for patients with radiologically resectable PDAC. While surgical resection is the only potential for cure, this treatment carries a significant risk of morbidity and mortality and achieves cure in only a small percentage of patients. Hence, one potential strategy to improve long term outcomes focuses upon pre-operative (neoadjuvant)

chemotherapy, which has been shown to improve disease-free survival[2]. However, prospective trials have not found that treating resectable PDAC patients with neoadjuvant therapy has improved survival and reviews of the literature have concluded that, on the basis of current evidence, this strategy should only be used in research settings at present [3]. It has been proposed that better stratification of patients is needed in order to identify those who will benefit from neoadjuvant therapy [2] based upon the theory that there is a subgroup of resectable PDAC patients who may benefit most. It is clear that pre-operative clinical models alone are inadequate, therefore the ability to contribute additional prognostic signal to a pre-operative model is valuable.

6.2. Comparison to prior studies.

Based upon the Radiomics Quality Score (RQS), our study is now the highest ranked work in the field of CT radiomics for PDAC prognostication, with a score of 18/36, compare to 14/36 in the two highest ranking studies to date. As highlighted in section 5.10, the areas where our study gained points compared to these prior studies are: the use of median value for cut-off analysis when determining risk groups, the use of bootstrap resampling when measuring discrimination and validation using cohorts from two distinct institutions.

Our study is only the second in prognostic PDAC CT radiomics to utilize an external test cohort. It is recommended that such validation datasets should experience at least 100 events [4] and compared to current body of literature in PDAC radiomics (table 1, page 41), our large Irish cohort (n=215) is the only test cohort to have exceeded this threshold. Our study thus robustly assessed clinical transportability, which are not sufficiently captured by prior studies. Larger number of events results in less variance of the c statistic, translating to narrower confidence intervals and hence greater confidence in the result [4] and this is true of our data (C-index for OS in external cohort 0.545, 95% CI 0.543-0.546) compared to prior studies from Xie [5]

(0.726 [0.646–0.806]) and Shi [6] (0.73 [0.66-0.79]). The international, multi-institutional nature of our cohorts resulted in substantial heterogeneity, with a significantly higher proportion of pathological T4 disease (AJCC stage III) in the validation set compared to the training cohort and this is reflected in the performance decrease at validation.

Our radiomic model identified a signal capable of stratifying patient prognosis, which might be enhanced in the future by the additional of novel preoperative biomarkers [7, 8]. However, despite the fact that our model outperformed the pre-operative clinical model, the overall discriminatory performance falls short of requirements for clinical implementation. Our model c-indices of 0.545 and 0.554 for overall and disease-free survival in the validation cohort compare to 0.72-0.74 in prior clinical-radiomic models, who used internal validation and incorporated post-operative pathological data [5, 6, 9]. The group which employed external validation reported c-index of 0.651 in a test set of 30 patients, which is too small to draw conclusions [10]. Performance metrics such as c-index can be influenced by length of follow-up, with longer follow up associated with lower results [11]. Our median survival of 58.6 and 61 months are considerably longer than all previous publications (table 1), however our model performance falls short of prior studies in both training and validation and minimal clinical utility is observed in the decision curve analysis, therefore it is important to explore the potential reasons for this low discrimination.

Our training cohort (n=352) is the largest to date in this field (prior studies ranged 30-210 patients, table 1) and this was achieved by employing broad inclusion criteria. Some prior studies included only CTs with one specific reconstruction slice thickness [6, 12, 13], excluded any which deviated from a specific pancreatic CT protocol [12, 13], and/or excluded all patients with biliary stents [13]. We did not follow such restrictive approaches because we wanted to create and test our model using real-world populations which match the type of the patient data expected in routine clinical practice. This is recommended in guidelines on the use of radiomics

in clinical trials from the European Society of Radiology, who highlight that "Radiomic signature are best developed initially on datasets that represent diversity of acquisition protocols, rather than within clinical trials with standardized and optimised protocols, as this would risk the selection of radiomic features linked to the imaging process rather than the pathology" [14]. In three prior CT radiomic PDAC studies, all CTs were acquired with identical contrast and reconstruction protocols using one or two CT scanner models [9, 15, 16] whereas our training and external cohorts were scanned on 26 and 24 different models, respectively (table 4). Consequently, our data is considerably more heterogenous regarding CT parameters than prior publications. This is likely the major factor limiting the performance of our model, since radiomic models are known to be highly sensitive to variation in CT acquisition parameters [17]. Attempts to mitigate this effect using Combat for feature harmonization did not result in better discrimination performance (table 9). Our study design eliminate one potential source variability by using IBSI compliant software [18, 19], since there are at least 14 software packages available [18-20] and compliance with IBSI has been shown to improve feature reproducibility [18].

In our inclusion criteria, we allowed ISI up to 120 days, whereas this varied from 1-6 months in prior studies who reported this metric (Table 1). While the number of patients with a very long ISI was low (5 patients had ISI between 90-120 days, three from Ireland and two from Canada, p=0.306), there was a significant difference in the ISI between our two cohorts (table 3), with a median of 22 days (IQR 24) for the Canadian cohort compared to median 29 (IQR 31, p=0.0036) for the Irish cohort. This is why we included ISI as a variable within the pre-operative clinical model, in order to account for this difference between the cohorts in our analysis. As demonstrated in our regression analysis, the ISI did not have a significant influence on patient overall survival or progression free survival in either cohort (Multivariable analysis results, table 6). This is supported by prior research which has demonstrated that ISI does not influence overall survival/progression free survival, either in an intention-to-treat or as-treat analysis [21], suggesting that underlying

tumour biology is probably a larger determinate of prognosis. This also suggests that the decision to use surgery date as time zero for time-to-event analysis was reasonable from a survival statistical viewpoint, albeit there is no data as to the dynamic change in radiomic features values over time, during the pre-operative phase of PDAC.

There are other notable differences between our study cohorts and the populations in prior studies; three studies included patients who received neoadjuvant therapy [10, 15, 22], which may influence the appearance of the tumour and increase the ISI. In addition, the inclusion of patients with metastatic (M stage) disease in the Xie *et al* study was unusual [5] since this is a major determinate of survival and is a contraindication to resection in guidelines [23]. We excluded patients who died within 30 days of surgery (in order to exclude deaths related to surgical complications) whereas some groups extended this to 60 [12] or 90 days [15, 22], albeit there are no guidelines as to the best strategy in this regard.

6.3. Biological meaning of the Rad-score.

The decision to focus primarily on original (non-filtered) radiomic features in this study was made in an attempt to improve the interpretation of the biological meaning of the Rad-score and also because IBSI has not yet performed analysis and standardization of radiomic image filters [19]. Two themes emerged in the Rad-score: tumours with lower attenuation (lower values of firstorder_Median and firstorder_Minimum) and/or tumours with more heterogenous texture (higher value for glcm_JointEntropy and low value for ngtdm_Coarseness) demonstrate worse survival. In the most extreme cases in our cohort, these characteristics were visible to the human eye (Figure 10) however computer automated quantification is required to classify the majority of cases. The finding that tumour attenuation is of prognostic importance is in line with previous publications. Several studies have shown that PDAC lesions with less enhancement on CT (i.e., lower attenuation),

demonstrated inferior response to neoadjuvant chemoradiation and shorter survival [24, 25] and there is evidence that the degree of enhancement correlates with underlying tumour biology, particularly the volume of extracellular stroma within the lesion [24]. Similarly, lower attenuation is associated with higher grade lesions, higher risk of lymph node metastasis and lower DFS [12, 13].

The finding that more heterogenous PDAC tumour are associated with worse survival has been previously demonstrated [16, 22] and it is consistent with many other tumour, such as non-small cell lung cancer [26], oesophageal squamous cell cancer [27] and glioblastoma [28]. However, the literature is contradictory regarding the effects of homogeneity vs heterogeneity in PDAC, with two early radiomic studies reporting that more homogenous PDAC lesions are associated with worse survival [13, 15]. In one publication, features from the same Gray-level co-occurrence matrix (GLCM) family (a group of second order radiomic features) demonstrated opposing effects within the same model, when measured at different angles, offsets or contrast phases [6]. Therefore, the true biological meaning of specific heterogeneity features in a PDAC tumour are yet to be determined. The inclusion of interaction terms in our study was intended to better capture the complex relationship between features. Interaction effects exist when the effect of a predictor variable changes depending on the value of other predictor variables. In our analysis, the relationship between original_firstorder_minimum and survival probability reversed depending on the values of original_firstorder_median and glcm_JointEntropy, whereas the relationship between glcm_JointEntropy and original_firstorder_median was synergistic (Figure 9).

6.4. Rad-score vs clinical-radiomic score.

The Radiomics Quality Score (RQS) encourages the inclusion of clinical variables into a radiomics prediction model because this is ‘expected to provide a more holistic model’[29]. In particular, the authors of the RQS highlight that it is important to

include established prognostic clinical variables in the model in order to identify whether any of the radiomic features are highly correlated with the clinical data, which would mean that the radiomics would not add any additional predictive ability to the model. The is the reason why our study focused primarily upon the discriminatory performance of the combined clinical-radiomic model, even though the Rad-score model alone performed slightly better than the combined model for both overall and disease-free survival in the external cohort (table 8).

6.5. Study limitations

A potential limitation of our study is the use of 2D contours. Although this is the most common strategy used in this field to date (Table 1, page 41), it has been shown that 2D and 3D contours yield different results for most variables in PDAC, other than a small number of attenuation features [30]. There is evidence from colorectal cancer that 3D contours may be more accurate [31] although there is opposing evidence from a recent study in gastric cancer[32], hence it is unclear whether any performance gain from using 3D contours is worth the extra labor required to perform it (3D contouring requires every axial slice of the lesion to be contoured, which takes considerably longer than 2D contours) [33]. It has been shown that the impact of CT parameter heterogeneity outweighs variation in segmentation [34] and this is suspected to be the dominant factor limiting model performance in our study. Future work should focus upon standardization of CT protocols. Other limitations include missing ca19.9 and ISI data, managed with multiple imputations and missing DFS data, albeit <10% in both cohorts. Multiple imputation has been demonstrated to introduce less estimation bias than excluding cases with missing information [4, 35] and has been used in recent similar studies in PDAC prognostication [36]. Finally, the 7th edition of the AJCC TNM system was used in this study, because this reflects the time period from which the patients in our study were treated (the 8th edition was published in 2016).

6.6. Conclusion.

We have developed and externally validated a pre-operative clinical-radiomic prognostic model for patients with PDAC using multi-institutional cohorts. The model significantly improved risk stratification compared to established pre-operative clinical variables and demonstrated similar discrimination (but lower net benefit) compared to the gold standard TNM model, which uses post-operative pathological information. While the presented model may help with pre-operative treatment decisions, the low discriminative performance suggests that these decisions remain challenging. The comprehensive statistical methodology utilised in this study sets a new standard for studies in PDAC CT radiomics, however heterogeneity of CT acquisition parameters may have contributed to the low model performance and future work should focus upon standardization of these protocols.

6.7. Next steps: Building upon our study results.

In our study, we designed a robust test to assess the accuracy and clinical utility of state-of-the-art CT radiomics for PDAC prognostication and we can conclude that radiomics failed this test. The reasons for the poor performance have not been conclusively identified, however heterogeneity of CT acquisition parameters is the most likely explanation. Our results are an important stepping stone for researchers in this area and there are multiple potential pathways which the field may take from this point. These are the next steps which our group has chosen:

(1) We have established a new collaboration with a hospital in Munich, Germany which is affiliated with the Technical University of Munich. We have done this in order to expand the size of our PDAC database and to include a third country into the project. This will facilitate comparisons such as North America vs Europe, Ireland vs Germany etc.

(2) We are continuing our collaboration with the biostatistics group of Professor Wie Xu at the University of Toronto. They will take our data and attempt to improve upon our results using a novel high dimension feature selection technique which was recently developed at their lab [37]. We decided not to use this in our initial work, since we wanted to stay within the realm of classic statistics, because we felt that would be more acceptable to clinicians. In addition, the work we have completed thus far will serve as a baseline which can be compared to as they attempt to improve the results.

(3) Our lab will now embark on a biology-correlation study, using a sub-cohort from the Toronto patients who have additional biological information available, including transcriptomic patterns which are known to associate with patient survival, including the Collison, Moffit, and Bailey Squamous subtypes [38]. This biological validation work is being led by my colleague, Emmanuel Salinas-Miranda, a post-doctoral fellow in the Radiomics and Machine Learning Lab. The provisional results indicate that there is an association between Rad-Score and the Bailey squamous transcriptional subtype, but this work is ongoing and a manuscript is in draft currently (January 2022).

(4) Recent studies have identified additional pre-operative prognostic biomarkers in PDAC, which may contribute to our model. In particular, our lab is interested in body composition biomarkers derived from pre-operative CT, such as skeletal muscle index and visceral adiposity. Our lab has recently published on skeletal muscle index (SKI) in unresectable PDAC [39], so we plan to build upon this experience to amalgamate this biomarker into our prediction model. One group has already combined SKI and radiomics, where SKI was remained a significant predictor of prognosis in multivariable analysis [6]. Thus, we plan to incorporate this in the next iteration of our work. Other pre-operative biomarkers which we would like to incorporate include

neutrophil-lymphocyte ratio and patient symptoms at diagnosis, both of which have been shown to have prognostic significance [7].

(5) We have focused in this manuscript upon the use of radiomics for PDAC prognostication, however there are other forms of artificial intelligence which can be applied to medical imaging. There have been studies using deep learning for PDAC prognostication, but none with a cohort as large as ours, and none with external validation. Hence, the dataset we have generated is ideal for testing this more advanced method, and that is something which are lab intends to pursue.

6.8. Chapter 6 references.

1. Nguyen, C.T. and M.W. Kattan, *How to tell if a new marker improves prediction.* Eur Urol, 2011. **60**(2): p. 226-8; discussion 228-30.
2. Gaskill, C.E., et al., *History of preoperative therapy for pancreatic cancer and the MD Anderson experience.* J Surg Oncol, 2021. **123**(6): p. 1414-1422.
3. Müller, P.C., et al., *Neoadjuvant Chemotherapy in Pancreatic Cancer: An Appraisal of the Current High-Level Evidence.* Pharmacology, 2021. **106**(3-4): p. 143-153.
4. Steyerberg, E., *Clinical Prediction Models: A practical Approach to Development, Validation and Updating. Second Edition.* 2009: Springer.
5. Xie, T., et al., *Pancreatic ductal adenocarcinoma: a radiomics nomogram outperforms clinical model and TNM staging for survival estimation after curative resection.* European Radiology, 2020. **30**(5): p. 2513-2524.
6. Shi, H., et al., *Survival prediction after upfront surgery in patients with pancreatic ductal adenocarcinoma: Radiomic, clinic-pathologic and body composition analysis.* Pancreatology, 2021. **21**(4): p. 731-737.
7. Guo, S.W., et al., *A preoperative risk model for early recurrence after radical resection may facilitate initial treatment decisions concerning the use of neoadjuvant therapy for patients with pancreatic ductal adenocarcinoma.* Surgery, 2020. **168**(6): p. 1003-1014.
8. Bhatti, I., et al., *Preoperative hematologic markers as independent predictors of prognosis in resected pancreatic ductal adenocarcinoma: neutrophil-lymphocyte versus platelet-lymphocyte ratio.* Am J Surg, 2010. **200**(2): p. 197-203.
9. Attiyeh, M.A., et al., *Survival Prediction in Pancreatic Ductal Adenocarcinoma by Quantitative Computed Tomography Image Analysis.* Ann Surg Oncol, 2018. **25**(4): p. 1034-1042.
10. Zhang, Y., et al., *CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging.* BMC Med Imaging, 2020. **20**(1): p. 11.
11. Vickers, A.J. and A.M. Cronin, *Everything you always wanted to know about evaluating prediction models (but were too afraid to ask).* Urology, 2010. **76**(6): p. 1298-301.

12. Cassinotto, C., et al., *Resectable pancreatic adenocarcinoma: Role of CT quantitative imaging biomarkers for predicting pathology and patient outcomes.* Eur J Radiol, 2017. **90**: p. 152-158.
13. Yun, G., et al., *Tumor heterogeneity of pancreas head cancer assessed by CT texture analysis: association with survival outcomes after curative resection.* Sci Rep, 2018. **8**(1): p. 7226.
14. Fournier, L., et al., *Incorporating radiomics into clinical trials: expert consensus endorsed by the European Society of Radiology on considerations for data-driven compared to biologically driven quantitative biomarkers.* Eur Radiol, 2021. **31**(8): p. 6001-6012.
15. Eilaghi, A., et al., *CT texture features are associated with overall survival in pancreatic ductal adenocarcinoma - a quantitative analysis.* BMC Med Imaging, 2017. **17**(1): p. 38.
16. Kim, H.S., et al., *Preoperative CT texture features predict prognosis after curative resection in pancreatic cancer.* Sci Rep, 2019. **9**(1): p. 17389.
17. He, L., et al., *Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule.* Sci Rep, 2016. **6**: p. 34921.
18. Fornacon-Wood, I., et al., *Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform.* Eur Radiol, 2020. **30**: p. 6241-6250.
19. Zwanenburg, A., et al., *The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping.* Radiology, 2020. **295**(2): p. 328-338.
20. Chu, L.C., et al., *Diagnostic performance of commercially available vs. in-house radiomics software in classification of CT images from patients with pancreatic ductal adenocarcinoma vs. healthy controls.* Abdom Radiol (NY), 2020. **45**(8): p. 2469-2475.
21. Healy, G.M., et al., *Preoperative CT in patients with surgically resectable pancreatic adenocarcinoma: does the time interval between CT and surgery affect survival?* Abdom Radiol (NY), 2018. **43**(3): p. 620-628.
22. Khalvati, F., et al., *Prognostic Value of CT Radiomic Features in Resectable Pancreatic Ductal Adenocarcinoma.* Sci Rep, 2019. **9**(1): p. 5449.
23. *NCCN Cinical Practice Guidelines in Oncology : Pancreatic Adenocarcinoma*. 2021.
24. Koay, E.J., et al., *Transport properties of pancreatic cancer describe gemcitabine delivery and response.* J Clin Invest, 2014. **124**(4): p. 1525-36.
25. Fukukura, Y., et al., *Contrast-enhanced CT and diffusion-weighted MR imaging: performance as a prognostic factor in patients with pancreatic ductal adenocarcinoma.* Eur J Radiol, 2014. **83**(4): p. 612-9.
26. Ganeshan, B., et al., *Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival.* Eur Radiol, 2012. **22**(4): p. 796-802.
27. Larue, R., et al., *Pre-treatment CT radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer.* Acta Oncol, 2018. **57**(11): p. 1475-1481.
28. Shim, K.Y., et al., *Radiomics-based neural network predicts recurrence patterns in glioblastoma using dynamic susceptibility contrast-enhanced MRI.* Scientific reports, 2021. **11**(1): p. 9974-9974.

29. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nat Rev Clin Oncol, 2017. **14**(12): p. 749-762.
30. Kulkarni, A., et al., *Pancreas adenocarcinoma CT texture analysis: comparison of 3D and 2D tumor segmentation techniques.* Abdom Radiol (NY), 2020. **46**(3): p. 1027-1033.
31. Ng, F., et al., *Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis?* Eur J Radiol, 2013. **82**(2): p. 342-8.
32. Zhao, H., et al., *TCGA-TCIA-Based CT Radiomics Study for Noninvasively Predicting Epstein-Barr Virus Status in Gastric Cancer.* AJR Am J Roentgenol, 2021: p. 1-11.
33. Lubner, M.G., et al., *CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges.* Radiographics, 2017. **37**(5): p. 1483-1503.
34. Yamashita, R., et al., *Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation.* Eur Radiol, 2020. **30**(1): p. 195-205.
35. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.
36. Xu, D., et al., *Prognostic Nomogram for Resected Pancreatic Adenocarcinoma: A TRIPOD-Compliant Retrospective Long-Term Survival Analysis.* World J Surg, 2020. **44**(4): p. 1260-1269.
37. Jain, R. and W. Xu, *HDSI: High dimensional selection with interactions algorithm on feature selection and testing.* PLoS One, 2021. **16**(2): p. e0246159.
38. Espiau-Romera, P., et al., *Molecular and Metabolic Subtypes Correspondence for Pancreatic Ductal Adenocarcinoma Classification.* J Clin Med, 2020. **9**(12).
39. Salinas-Miranda, E., et al., *Prognostic value of early changes in CT-measured body composition in patients receiving chemotherapy for unresectable pancreatic cancer.* Eur Radiol, 2021. **31**(11): p. 8662-8670.