Computed Tomography–based Radiomics for Risk Stratification in Prostate Cancer


Published in:
International Journal of Radiation Oncology Biology Physics

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2019 Elsevier. This manuscript is distributed under a Creative Commons Attribution-NonCommercial-NoDerivs License (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Accepted Manuscript

CT-based Radiomics for Risk Stratification in Prostate Cancer

Sarah OS. Osman, PhD, Ralph TH. Leijenaar, PhD, Aidan J. Cole, PhD FRCR, Ciara A. Lyons, MD, Alan R. Hounsell, PhD, Kevin M. Prise, PhD, Joe M. O'Sullivan, FRR(RCSI) PhD, Philippe Lambin, MD PhD, Conor K. McGarry, PhD, Suneil Jain, FRCR PhD

PII: S0360-3016(19)33392-9
DOI: https://doi.org/10.1016/j.ijrobp.2019.06.2504
Reference: ROB 25812

To appear in: International Journal of Radiation Oncology • Biology • Physics

Received Date: 15 January 2019
Revised Date: 14 May 2019
Accepted Date: 14 June 2019


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
CT-based Radiomics for Risk Stratification in Prostate Cancer

Sarah OS Osman, PhD$^{1,2,*}$, Ralph TH Leijenaar, PhD$^3$, Aidan J Cole, PhD FRCR$^{1,4}$, Ciara A Lyons, MD$^{1,4}$, Alan R Hounsell, PhD$^{1,2}$, Kevin M Prise, PhD$^1$, Joe M O’Sullivan, FRR(RCSI) PhD$^{3,4}$, Philippe Lambin, MD PhD$^3$, Conor K McGarry, PhD$^{1,2}$ and Suneil Jain, FRCR PhD$^{1,4}$

$^1$ Centre of Cancer Research and Cell Biology, Queen’s University Belfast, Belfast, UK
$^2$ Radiotherapy Physics, Northern Ireland Cancer Centre, Belfast Health and Social Care Trust, Belfast, UK
$^3$ The D-Lab: Decision Support for Precision Medicine, GROW - School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, the Netherlands
$^4$ Clinical Oncology, Northern Ireland Cancer Centre, Belfast Health and Social Care Trust, Belfast, UK

Authors email addresses
Sarah O S Osman; s.osman@qub.ac.uk
Ralph TH Leijenaar; ralph.leijenaar@maastrichtuniversity.nl
Aidan J Cole; a.Cole@qub.ac.uk
Ciara A Lyons; ciaralyons@gmail.com
Alan R Hounsell; alan.hounsell@belfasttrust.hscni.net
Kevin M Prise; k.prise@qub.ac.uk
Joe M O’Sullivan; joe.osullivan@qub.ac.uk
Philippe Lambin; philippe.lambin@maastrichtuniversity.nl
Conor K McGarry; conor.mcgarry@belfasttrust.hscni.net
Suneil Jain; s.jain@qub.ac.uk

* Corresponding author
Dr Sarah O S Osman
Centre for Cancer Research and Cell Biology
Queen’s University Belfast
Belfast, BT7 1NN
Northern Ireland
Tel +44 (0) 28 95043443
s.osman@qub.ac.uk

Key words:
Radiomics, Prostate Cancer, Classification, Gleason Score, Risk Group.

Running title: Prostate Cancer CT-based Radiomics

Authors responsible for statistical analysis: Dr Sarah Osman, Dr Ralph Leijenaar

Funding statement: This work was supported by grants from Prostate Cancer UK, the Belfast-Manchester Movember Centre of Excellence (CE013-2-004), and the Northern Ireland Health and Social Care Trust R&D Office grant (COM/4965/14).

Conflict of interest statement: Dr Lambin reports grants/sponsored research from Oncoradiomics, pTHERagnostic, advisor fee from Oncoradiomics. Dr Lambin is inventor of two patents on radiomics and one non patentable invention (software), licensed to Oncoradiomics and has (minority) shares in the company Oncoradiomics. Dr. Leijenaar reports personal fees and other from OncoRadiomics SA, outside the submitted work; In addition, Dr. Leijenaar has a patent EP3207521A1 pending. Dr. Prise reports personal fees from Bayer, outside the submitted work. Dr. O’Sullivan reports grants and personal fees from Bayer, personal fees from Janssen, personal fees from Astellas, and personal fees from Sanofi, outside the submitted work.
Title: CT-based Radiomics for Risk Stratification in Prostate Cancer

Abstract

Purpose

To explore the role of CT-based radiomics features in prostate cancer risk stratification.

Methods and Materials

The study population consisted of 506 prostate cancer patients from a clinically annotated database. After applying exclusion criteria, 342 patients were included in the final analysis. CT-based radiomics features were extracted from planning CT scans for prostate gland only (PO) structure, and machine learning was employed to train models for Gleason Score (GS) and Risk Group (RG) classifications. Repeat cross-validation was used. The discriminatory performance of the developed models was assessed using receiver operating characteristic (ROC) area under the curve (AUC) analysis.

Results

Classifiers employing CT-based radiomics features distinguished between GS≤6 vs. GS≥7 with (AUC=0.90) and GS7(3+4) vs. GS7(4+3) with (AUC=0.98). Developed classifiers also showed excellent performance in distinguishing low vs. high RG (AUC=0.96), low vs. intermediate RG (AUC=1.00), but poorer performance was observed for GS7 vs. GS>7 (AUC=0.69). An overall modest performance was observed for validation on holdout datasets with the highest AUC of 0.75 for classifiers of low vs. high RG and AUC of 0.70 for GS7 vs. GS>7.

Conclusion

Our results show that radiomics features from routinely acquired planning CT scans may provide insights into prostate cancer aggressiveness i.e. GS and RG in a non-invasive manner. Assessing models on training datasets, the classifiers were especially accurate in identifying high-risk from low risk patients, and in classifying GS7 vs. GS>7 and GS7(3+4) vs G7(4+3), however classifiers were less good in distinguishing high vs intermediate risk
groups. External validation and prospective studies are warranted to verify the presented findings. These findings could potentially guide targeted radiotherapy strategies in radical intent radiotherapy for prostate cancer.

**Introduction**

Prostate cancer is one of the most frequently diagnosed cancers globally; it is the most common cancer in males in the UK with around 130 new cases per day [1], and other than skin cancer, it is also the most common cancer in males in the United States of America with around 64,690 new cases per year [2]. Traditionally, prostate cancer is stratified into risk categories using prostate-specific antigen (PSA) levels, biopsy-based Gleason scores (GS), and TNM staging [3]. Although extremely useful in assessing prostate cancers, GS are typically determined from a limited number of trans-rectal ultrasound (TRUS) guided biopsies that are randomly sampled. This poses a challenge as prostate cancer is associated with significant intra-tumour heterogeneity [4]. The behaviour of prostate cancer is also highly variable with some tumours progressing rapidly and others remaining indolent for many years. These factors make early diagnosis, accurate prognosis and tumour risk stratification critical to increase the therapeutic ratio and avoid unnecessary intervention.

Medical scans are conventionally used for diagnosis and treatment planning, however, in recent years this role has expanded to also include the extraction and quantification of imaging features (i.e. radiomics analysis) to serve as imaging biomarkers for staging, and to enhance the performance of prognostic and predictive models [5]–[10]. The prognostic and predictive value of MRI-based radiomics features for prostate cancer has been demonstrated in a number of publications [8]–[11]. In recent years, there has been an increased focus on the development of multi-parametric (mp)-MRI as a tool for effective image characterization, and for treatment planning as well as response assessment [12], [13]. Despite this, there is no clear consensus on the most important imaging biomarkers for prostate or their clinical applicability. This could be due to the lack of interoperability (i.e. variable imaging protocols, scanners, software) or due to uncertainties in the manual definition of tumours [14], [15]. In contrast to MRI, planning Computed Tomography (CT) scans are typically highly standardized and they are available for almost all patients treated with external beam radiation therapy (EBRT) for outlining structures of interest, planning treatments, dose calculation, and as a reference for image guidance. Several studies have reported high repeatability and reproducibility of several CT based radiomics
features extracted from patient and from phantoms scans [16]. Only a limited number of studies have investigated reproducibility with MRI scans reporting that MRI-based radiomics features are less robust [16], [17]. In radiotherapy, standardized planning CT scans are frequently assessed for CT number/electron density calibration, uniformity and noise as this is critical for dose calculation, however as MRIs were traditionally used for visualization, less standardization has been typically employed. Now that the role of medical imaging and radiomics analysis has been realised, there are more international efforts to standardize and control differences in medical scans to allow their optimal use for prospective imaging biomarkers studies [18], [19].

To date, no studies have examined the role of non-contrast CT-based prostate radiomics in risk stratification or prognosis. This could be partially due to the poor visual discrimination of prostate cancer on CT scans. However, accumulating evidence of the value of imaging in diagnosis and prognosis of prostate cancer warrants more comprehensive investigation [5]–[11].

We hypothesize that the computerized interpretation of prostate intensity levels and their spatial distribution, as quantified from CT scans, will reveal additional information about prostate cancer and radiomics analysis could serve as virtual biopsies.

In this study, we investigate the value of CT-based radiomics features to classify prostate cancer GS and risk group (RG). To the best of our knowledge, this is the first investigation on non-contrast CT-based radiomics features for prostate cancer.

**Materials and Methods**

The use of CT scans and disease staging information of the study population in the present study was approved by XXXX (XXXX) with research governance permission number XXXX.

**Study Population**

All patients with localized prostate cancer treated at XXXX (XXXX), between 1 January 2005 and 31 December 2009 were considered in this study. Patients were treated with 70–74Gy external beam radiation therapy (EBRT) in 2Gy fractions with 3D conformal or intensity modulated radiotherapy techniques over 7–7.5 weeks. Patients had short (≤6 months) or long (6–36 months) course Androgen deprivation therapy (ADT) commenced at least 3 months before radiation with luteinizing hormone-releasing hormone (LHRH) agonists or
antiandrogens monotherapy (Bicalutamide 150mg OD). For patients in this cohort, GS was determined from TRUS biopsies. The median number of cores was 12 [range: 6–21] and the median volume of the tissue involved was 20% [range: 1%–90%] (statistics from available from 53% of the patients). The median volume of tissue involved in different GS subgroups were: GS6 = 7.5%, GS7 (3+4) = 18.8%, GS 7(4+3)=20% and GS >7= 20%. Patients were assigned to low-, intermediate- and high-risk groups based on GS, iPSA and TNM stage as per the United Kingdom National Institute for Health and Care Excellence (NICE) guidelines [20].

CT Image Acquisition and Image Segmentation

506 patients were initially identified for this study, of which 447 had complete data sets (pCT scans, and DICOM structures) to qualify their inclusion in this study. All patients were scanned at XXXX using our local pelvis protocol on one of two CT scanners; Emotion 6 (Siemens Healthcare Siemens Healthcare GmbH, Erlangen, Germany) and LightSpeed RT (GE Medical Systems, Chicago, Illinois, United States). Patients were scanned in helical mode using the following acquisition parameters; tube current, 370–437mAs for the GE scanner and 130–194mAs for Siemens; tube voltage, 120kV(p) for GE and 130kV(p) for Siemens; slice thickness 2.5mm; convolution kernel was standard for GE and B41s for Siemens; 512x512 pixels; pixel spacing 0.98–1.17mm. All DICOM images and structure sets were exported into CERR software (Computational Environment for Radiotherapy Research) and converted into MATLAB data file format [21]. To reduce contouring variability, all prostate structures were reviewed and adjusted when necessary to construct a prostate gland only structure (PO) by a single observer and reviewed by an experienced consultant clinical oncologist. The PO structure, that includes the whole prostate, was used for further analysis. Only CT scans with 2.5mm slice thickness and scans without artefacts were included in the analysis (e.g. patients with prosthetic hips were excluded). Using these selection criteria, a total of 342 scans were then identified as suitable for further analysis. Risk groupings are detailed in Table 1 along with the number of patients assigned to each group.

Radiomics Features Extraction

RadiomicTM software (Oncoradiomics, Liège, Belgium), was used to extract 1618 radiomic features from PO structures for each patient. After an initial investigation on optimal radiomics extraction settings, (supplementary material S(I)), CT scans were re-sampled to 2.5mm isotropic voxels and the range of
Hounsfield (HU) discretised to 10HU bins prior to features extraction. Radiomics features extracted were the first order gray-level statistics from the intensity level histograms, and features based on; gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), gray-level distance zone matrix (GLDZM), neighbourhood gray tone difference matrix (NGTDM) and neighbouring gray-level dependence matrix (NGLDM). The textural and statistics features were also calculated after applying 8 sub-bands of three-dimensional wavelet transforms and 16 sub-bands of Laplace of Gaussian (LoG) transform to the scans. Detailed explanation of the extracted features and their mathematical definitions were described previously [6], [22]. In this analysis, geometrical features (except for prostate volume) were excluded as the entire prostate gland was contoured and analysed rather than tumour specific regions.

Test-retest Cohort

Twenty prostate cancer patients with available pre-treatment repeat pCT scans were also included in this study. These patients underwent repeat scans due to either sub-optimal bladder/rectal filling, or due to temporary presence of rectal gas pocket on the first pCT scans. The average time between scans and re-scans was 11 days ranging between [same day scans–28 day gap]. Radiomics analysis of PO structures were conducted on the scans and the re-scans. The intra-class correlation coefficient (ICC) [23], [24] was calculated to provide an indication of the test-retest repeatability and reproducibility of the features extracted. The ICC ranged between 0 and 1, with 0 indicating no reliability and 1 indicating perfect reliability. A cut-off value of 0.8 was used to arrive at a set of stable and robust features set for further analysis.

Feature Selection and Classification

For each developed model, stratified partitioning was used to randomly assign 80% of the data for model training and 20% for validation. To reduce the risk of models over-fitting the following feature reduction methods were employed. The function `findCorrelation` in the R package “Caret” was used to remove (redundant) highly correlated features to reduce multi-co-linearity [25]–[27]. This function searches through a correlation matrix of all the features and returns a vector of integers corresponding to columns to remove pair-wise correlations. Generalized linear models, via penalized maximum likelihood, were fitted using the R package “glmnet” and the default parameter tuning with “Caret” interface was employed. Two methods of variable selection were implemented; regularization through 1) the least absolute shrinkage and selection
operator (LASSO) and 2) ElasticNet. Prior to the analysis, radiomics features were scaled and centred using Caret’s function PreProc. This function scales feature values by subtracting the mean value and centering the values by dividing by the standard deviation [25]. A grid search was employed to choose optimal model hyperparameters that minimizes cross-validation errors and hyperparameters one standard error away from optimal values were then used. The average predictive performance of each model on the training dataset was estimated using 10-fold cross-validation repeated 100 times. Sample augmentation using the Syntactic Minority Oversampling Technique (SMOTE) for class balancing in training data was also investigated [28].

Statistical Analysis and Model Assessment

Following the guidelines of the transparent reporting of multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement for type 2a studies (i.e. study using a single data set, data is randomly split into two groups: one to develop the prediction model and one to evaluate its predictive performance) [29], Mann Whitney U tests (Wilcoxon Rank Sum Test) were performed to test the null hypothesis that feature values have equal medians in different patient risk groups, against the alternative that they are not.

In this work, all models were trained with and without sample augmentation and model performances were assessed on training and on testing datasets using area under the ROC curve (AUC), accuracy (Acc), sensitivity and specificity and for comparison with previous work [30], Youden index [31] was also reported (YI=specificity+sensitivity-1). For all models, the AUC 95% confidence intervals (CI) was computed with 2000 stratified bootstrap replicates.

Results

A flow diagram of the study cohort, exclusion criteria and pre-modelling feature filtering is shown in Figure 1. After applying our exclusion criteria, 342 patients’ CT scans were eligible for analysis. Forty PO structures from 20 patients were included in the test-retest analysis (Figure1 and Figure2 (a)).

The performance of the trained models for classifying patients into different GS and risk groups was assessed in multiple steps (Table 2 and 3). Sample augmentation was employed to balance training data when required (detailed analysis of performance of models developed on non-augmented and augmented data is provided in supplementary Table ST1 and ST2). For example, the first radiomics classifier was trained to distinguishing low...
from high GS (GS≤6 vs. GS≥7). Figure 2(b) shows a circular bar-plot of a subset of robust radiomics features (n=114, unfiltered) in the two patients groups GS≤6 and GS≥7. The length of the bar represents the median value of each feature (scaled and centred) in each patient group represented with different colours. The stars in the circular bar-plots indicate features that had significantly different median values in different groups. A comprehensive list of significant features is provided in supplementary material S(II).

Assessing the apparent performance of the trained models on training data, excellent performance was observed in classifying patients with low GS (GS 6) from high GS (GS > 6) and low-risk from high-risk patients with (AUC=0.90 [95% CI (0.86-0.95)] and (AUC= 0.96 [95% CI (0.93-0.98)]), respectively; (Figures 3 and Tables 2 and 3). Excellent performance was observed for classifiers of GS 7(3+4) vs. 7(4+3) with an AUC of 0.98 [95% CI (0.96-1.00) with high specificity and sensitivity. Developed classifiers had lower performance in distinguishing GS7 from GS>7 (AUC = 0.69 [95% CI (0.61-0.76)]. The most important variables (features) in distinguishing GS and risk groups were found to be the wavelet transforms of the Neighbouring Gray-Tone Difference Matrix (NGTDM) coarseness (HLL, HLH, HHH) and contrast (HHH), the Low Gray-Level Small Dependence Emphasis of the Neighbouring Gray-Level Dependence Matrix NGLDM LGSDE (HHL, LLL, LHL, HLL) and the (LLL) Gray-Level Size Zone (GLSZM) Intensity Non-uniformity Normalized (INN). All features above capture different aspects of texture heterogeneity.

All models were subsequently validated on unseen test data. The resulting AUC are presented in Table 2 and 3 and Figure 3. The best observed performance was for classifiers of low risk vs. high risk cancers with an AUC of 0.75 [95% CI (0.48-1.00)] and for GS7 vs >7 (AUC = 0.70 [95% CI (0.48-0.81)]). Lower performance was observed for models of GS 7(3+4) vs GS (4+3) and low risk vs. Intermediate risk (AUC = 0.65 and 0.49, respectively).

In Figure 3, examples the area under the ROC curve (AUC) for different radiomics-based GS and RG classification are presented for models developed using ElasticNet on the training data as well as on testing datasets.

Multiclass models for GS and RG classification were also explored and results are presented in the supplementary Table ST3. Both GS and RG multiclass classifiers (built on SMOTEd data) had reasonable apparent performance on training data, however only the RG multiclass classifier showed reasonable accuracy on test dataset (Acc = 0.70 [95% CI (0.47-0.71)]). Applying sample augmentation through SMOTE, all the
classifiers outperformed those built on non-augmented samples with all models showing high sensitivity and specificity for all tested cases (supplementary material S(III)).

None of the developed classifiers had prostate volume in the final model. Pairwise t-test were conducted to test differences in prostate volume different risk groups. No significant differences in prostate volume were found between different GS groups, or between different ADT types, however prostate volume was significantly larger in intermediate vs high risk group (p = 0.02), supplementary Figures S1.

Pearson and Spearman’s correlation coefficients were also calculated for the 20 most important variables, picked by each classifier, and are presented in supplementary Figure S2. The cross-correlation of patients’ (initial) PSA, GS, risk group (RG) and prostate volumes (PV) with radiomics features revealed only weak correlations between individual radiomics features and clinical factors. It is also observed that several radiomics features exhibit a monotonic relationship indicated by a high value for their Spearman’s correlation coefficients. A Venn diagram showing the intersections of the top features selected by LASSO for different classification problems is presented in Figure S3.

Discussion

In this study, the value of CT-based radiomics analysis in prostate GS and risk group classification is demonstrated. This work describes the extraction of quantitative imaging features and the application of machine learning methods for features selection and reduction and model training. A test-retest analysis was conducted to investigate the robustness and stability of extracted features and to eliminate non-reliable features. Given the retrospective nature of the database used in this study and limitations associated with the variable scan re-scan data (scanners, scan parameters and time gap between scans), this test-retest analysis is likely to eliminate more features than truly required. However, this method was adopted to provide a conservative measure to account for the variations in the dataset used in this analysis.

Previous studies on prostate cancer radiomics mainly focussed on the use of multi parametric-MRI (mpMRI) [8], [9], [37], [38], [10], [11], [30], [32]–[36]. Several studies have demonstrated the value of mpMRI-based
features for PCa detection [8], [32]–[34], PCa aggressiveness and GS classification [30], [35], biochemical failure risk prediction [36], [39] and in targeted radiation therapy planning [38].

Fehr et al. [30] investigated the use of MRI-based texture features in cancer detection and GS assessment in prostate cancer patients. They showed that accurate models for GS classification could be achieved by employing machine learning methods to combine several apparent diffusion coefficients ADC and T2 measures even in highly imbalanced datasets. Similar to our work, they also showed that employing sample augmentation methods always resulted in better performing classification models in all the cases.

Only one study was found for perfusion-CT based radiomics for prostate cancer [7]. Motivated by previous studies showing the correlation of perfusion-CT parameters with PSA values, GS, and TNM stage [40], [41]), Tanasini-Lang et al. investigated perfusion-CT-based radiomics for GS classifications [7]. In their study, prior to radical prostatectomy, all patients (n=41) had PCT and perfusion parameters were then calculated from PO contours. Three perfusion maps were calculated, i.e. blood volume (BV), blood flow (BF), mean transit time (MTT) and 1701 radiomics features from these maps were calculated for each patient. Using principle component analysis (PCA), 10 feature groups were identified. From each group, a single radiomics feature (that correlated best with the group) was chosen to represent the group in a multivariate regression analysis to predict the post-surgical GS. They found one radiomics feature, the HHL wavelet transform of BF joint average, to be prognostic for GS (7 vs. >7). Comparing patients with GS7(3+4) vs. GS7(4+3), both BF HHH fractal dimension and BV HLH root mean square were found different between the two groups. Interestingly, they also reported that for 28% of patients there was a difference in grading between the biopsy and post-surgical histopathology [7].

Although it is not possible to directly correlate the CT-based radiomics features in this study with previous work on MRI or perfusion-CT, a general comparison of the performance of final classification models is presented for non-augmented data. CT-based radiomics ElasticNet-based classifiers of GS6 vs. GS>6 (n=80 vs. 194) achieved an accuracy of 0.72 with YI=0.04 compared to 0.83 accuracy and YI=0.03 reported for the MRI based classifier in reference [30] (n=34 vs. 159). Excellent performance was observed in this work for classifiers of GS 7(3 + 4) vs. 7(4 + 3) with Acc=0.98 and YI=0.80 (n=70 vs. 40) compared to Acc=0.83 and YI=0.11 in reference [24] (n=114 vs. 26). CT-based classifier for GS (7 vs. > 7) showed lower performance (AUC=0.69; Acc=0.64) compared to (AUC=0.81) in perfusion-CT based models [7], however, the number of patients in
reference [7] was much smaller (n=32 vs. n=8) than the number of patients in the present study (n=109 vs. n=85).

Assigning prostate cancer patients to different risk categories depend on their GS, blood PSA level, and if the cancer has spread outside the prostate [14]. In this study the role of radiomics features in classifying PCa patients into different risk groups was also investigated. Despite the highly imbalanced data sets, excellent apparent performances were observed for classifiers of low- (n=26) and high-risk (n=187) groups (Acc=0.98, AUC=1.00 and YI =0.81) and low- vs. intermediate- (n=61) risk (Acc= 1.00, AUC=1.00 and YI=1.00). However, poor performance was found for classifiers of intermediate- vs. high-risk (Acc=0.75, AUC=0.68 and YI=0).

In all the studies discussed thus far, the data was used to develop the models and the performance of classifiers was assessed using re-sampling techniques. In this study, classifiers were also tested on unseen data showing overall modest performance. This may be influenced by the small number of patients in test datasets and by the class imbalances. Nonetheless, this study shows that CT-based radiomics features may provide insights into the nature of prostate cancer beyond the scope of visual assessment. Our models classified patients with low SG/RG vs. high GS/RG and with GS7(3+4) vs. 7(4+3) with high accuracy. The developed classifiers were especially accurate in identifying high-risk patients with high Gleason score’ disease. This study does not suggest to replace MRI based radiomics for prostate cancer risk assessment. This study demonstrates that CT-based radiomics carry information that may be utilized to improve current risk stratification methods. Combining CT and MRI based radiomics could give more accurate assessments of prostate cancer aggressiveness and is a subject that needs further investigations. It is acknowledged that this study has some limitations. Firstly, this is a retrospective study with no external validation data set. In an ongoing study, three datasets for external validation are being prepared; the first dataset is from our own institution sampled from a later period of time (2010-2015), the two other cohorts will come from other external institutions. Moreover, radiomics features in this study were extracted from the PO structures and not PCa explicitly (albeit this is a more accurate representation of current PCa RT treatments). In several investigations on MRI based PCa, radiomics features were found to be different in cancers arising in the transitional zone from those in the peripheral zone of the prostate [30], [32]. This limitation can be addressed in future prospective studies by contouring suspicious areas of the prostate on CT scans aided by MRI fusion [11]. This technique is employed in clinical practice for defining boost volumes of dominant intra-prostatic lesions. Cancer specific contouring
may also allow better understanding of the shape features. Moreover, the end point in this study was GS based on TRUS biopsies. It is expected that mp-MRI guided biopsies will provide more accurate grading of prostate cancers and consequently allow the training of more accurate image based classifiers. Another observation in this study was the correlation between prostate volume and some of the radiomics features included in the final models. As discussed by Welch at al. on their work on radiomics safeguards [42], this may have implications on models’ performance and generalizability. Therefore, these correlations between radiomics features and prostate volume should be thoroughly investigated in future studies.

Despite these limitations, the current study is proof of concept that standardly obtained pCT scans can reproducibly characterise risk stratification in prostate cancer and may guide future studies investigating bespoke RT based on CT radiomics. External validation and prospective studies are warranted to verify our findings.

**Figure legend**

Figure 1: Flow diagram of study cohort, radiomics workflow and radiomics features generated. Patient numbers denoted by N and features by n.

Figure 2: (a) a histogram of the inter-class correlation coefficients for radiomics features (n= 1618) from the test re-test data. The black dotted line indicates the cut-off value of 0.8 for reproducible features (n=522). (b) A circular stacked bar plot of stable radiomics features (excluding filtered features) - features were scaled and centred for this presentation. Yellow and purple colours in the bar indicate median value of feature in patients with low and high GS cohorts respectively indicates a feature is significantly different between the two cohorts with Mann Whitney U test p-values≤0.05,**≤0.01,***≤0.001.

Figure 3: The receiver operating characteristics for Gleason score (left) and risk group (right) classification with ElasticNet. Classifiers performance on training (top) and testing data (bottom) is presented.
References:


[22] H. J. W. L. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative


### Table 1(a): Patient grouping based on Gleason Score (GS)

<table>
<thead>
<tr>
<th>Gleason Score</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>6(3+3)</td>
<td>100</td>
</tr>
<tr>
<td>7(3+4)</td>
<td>87</td>
</tr>
<tr>
<td>7(4+3)</td>
<td>49</td>
</tr>
<tr>
<td>&gt;7</td>
<td>106 (GS 8 = 55, GS 9 = 51)</td>
</tr>
</tbody>
</table>

### Table 1(b): Patient grouping based on Risk Group (RG)

<table>
<thead>
<tr>
<th>Risk Group</th>
<th>UK NICE Criteria</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td>T1 or T2a and</td>
<td>n = 33</td>
</tr>
<tr>
<td>Intermediate risk</td>
<td>Gleason score 6 and PSA &lt;10 ng/ml</td>
<td>n = 76</td>
</tr>
<tr>
<td></td>
<td>T2b or T2c or</td>
<td></td>
</tr>
<tr>
<td>High Risk</td>
<td>Gleason score 8-10 or PSA &gt; 20ng/ml</td>
<td>n = 233</td>
</tr>
<tr>
<td></td>
<td>&gt;T2c</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Accuracy results for Gleason Score classification using Lasso (a), and elastic net regularization (b).

<table>
<thead>
<tr>
<th>(a) LASSO</th>
<th>training</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gleason Score</strong></td>
<td><strong>AUC</strong></td>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>GS 6 vs. &gt;6</td>
<td>0.77</td>
<td>0.65</td>
</tr>
<tr>
<td>n = (80/194)*</td>
<td>(0.73-0.84)</td>
<td>(0.59-0.71)</td>
</tr>
<tr>
<td>GS 7(3+4) vs. 7(4+3)</td>
<td>0.97</td>
<td>0.86</td>
</tr>
<tr>
<td>n=70/40</td>
<td>(0.95-1)</td>
<td>(0.79-0.92)</td>
</tr>
<tr>
<td>GS 7 vs. &gt;7</td>
<td>0.67</td>
<td>0.62</td>
</tr>
<tr>
<td>(109/85)</td>
<td>(0.59-0.74)</td>
<td>(0.55-0.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) ElasticNet</th>
<th>training</th>
<th>testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gleason Score</strong></td>
<td><strong>AUC</strong></td>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>GS 6 vs. &gt;6</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>n = (80/194)*</td>
<td>(0.86-0.95)</td>
<td>(0.81-0.90)</td>
</tr>
<tr>
<td>GS 7(3+4) vs. 7(4+3)</td>
<td>0.98</td>
<td>0.92</td>
</tr>
<tr>
<td>n=70/40</td>
<td>(0.96-1.00)</td>
<td>(0.85-0.96)</td>
</tr>
<tr>
<td>GS 7 vs. &gt;7</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td>(109/85)</td>
<td>(0.61-0.76)</td>
<td>(0.57-0.71)</td>
</tr>
</tbody>
</table>

Youden Index (YI = sensitivity + specificity -1). n is the number of patients in each cohort. * indicates that training data was augmented prior to model training.
Table 3: Accuracy results for prostate cancer risk group classification using Lasso (a), and elastic net regularization (b).

<table>
<thead>
<tr>
<th></th>
<th>(a) LASSO</th>
<th></th>
<th>(b) ElasticNet</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>training</td>
<td>testing</td>
<td>training</td>
<td>testing</td>
</tr>
<tr>
<td><strong>Risk Group</strong></td>
<td>AUC</td>
<td>Accuracy  YI</td>
<td>AUC</td>
<td>Accuracy  YI</td>
</tr>
<tr>
<td>Low vs. High</td>
<td>0.96</td>
<td>0.89  0.87</td>
<td>0.96</td>
<td>0.80  0.87</td>
</tr>
<tr>
<td>(26/187)*</td>
<td>(0.94-0.99)</td>
<td>(0.84-0.93)</td>
<td>(0.93-0.98)</td>
<td>(0.65-0.89)</td>
</tr>
<tr>
<td>Low vs. Inter</td>
<td>1.00</td>
<td>1.00  1.00</td>
<td>0.56</td>
<td>0.71  0.30</td>
</tr>
<tr>
<td>(33/76)</td>
<td>(1.00-1.00)</td>
<td>(0.96-1.00)</td>
<td>(0.21-0.90)</td>
<td>(0.48-0.89)</td>
</tr>
<tr>
<td>Inter vs. High</td>
<td>1.00</td>
<td>0.99  0.97</td>
<td>0.49</td>
<td>0.54  -0.10</td>
</tr>
<tr>
<td>(61/187)*</td>
<td>(1.00-1.00)</td>
<td>(0.99-0.98)</td>
<td>(0.33-0.63)</td>
<td>(0.41-0.67)</td>
</tr>
</tbody>
</table>

Youden Index (YI = sensitivity + specificity -1). n is the number of patients in each cohort. * indicates that training data was augmented prior to model training.
Patient population N= 506

- 59 patients excluded no pCT scans available

Patients with pCT N= 447

- 105 patients excluded slice thickness ≠2.5mm or/and presence of artefacts

Patients suitable for final analysis N= 342

CT scans and GS and risk group N= 342

Segmenting PO structure N= 342

Test-retest dataset N = 20

Segmenting PO structure N= 40

Radiomics analysis n = 1618 features

Yes

ICC > 0.8

Filter features

No

Unreliable features no further analysis

Final set of features to use in analysis n = 522
\( n = 1096 \)

\( n = 522 \)

\( n = 114 \)

<table>
<thead>
<tr>
<th>Stats</th>
<th>First order gray-level statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLCM</td>
<td>Gray-Level Co-occurrence Matrix</td>
</tr>
<tr>
<td>GLDZM</td>
<td>Gray-Level Distance-Zone Matrix</td>
</tr>
<tr>
<td>GLRLM</td>
<td>Gray-Level Run-Length Matrix</td>
</tr>
<tr>
<td>GLSZM</td>
<td>Gray-Level Size-Zone Matrix</td>
</tr>
<tr>
<td>NGLDM</td>
<td>Neighbourhood Gray-Level Dependence Matrix</td>
</tr>
<tr>
<td>NGTDM</td>
<td>Neighbourhood Gray-Level Difference Matrix</td>
</tr>
</tbody>
</table>
Gleason Score

<table>
<thead>
<tr>
<th>Condition</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 vs. &gt; 6</td>
<td>0.78</td>
</tr>
<tr>
<td>7(3+4) vs. 7(4+3)</td>
<td>0.98</td>
</tr>
<tr>
<td>7 vs. &gt;7</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Risk Group

<table>
<thead>
<tr>
<th>Condition</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low vs. High</td>
<td>0.96</td>
</tr>
<tr>
<td>Low vs. Inter</td>
<td>1.00</td>
</tr>
<tr>
<td>Inter vs. High</td>
<td>1.00</td>
</tr>
<tr>
<td>Low vs. High</td>
<td>0.75</td>
</tr>
<tr>
<td>Low vs. Inter</td>
<td>0.56</td>
</tr>
<tr>
<td>Inter vs. High</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Summary

The clinical behaviour of localised prostate cancer (PCa) is highly variable, with some cancers remaining indolent for many years and others progressing to metastatic disease in a short time. In this work, the value of CT–based radiomics for risk stratification in PCa was investigated. Our results show that radiomics features from routinely acquired planning CT scans can provide insights into PCa aggressiveness in a non-invasive manner.