ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

FISICA

Ciclo 36

Settore Concorsuale: 02/D1 - FISICA APPLICATA, DIDATTICA E STORIA DELLA FISICA

Settore Scientifico Disciplinare: FIS/07 - FISICA APPLICATA A BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

ESTABLISHMENT OF A RADIOMICS LABORATORY: DATA MINING AND QUANTITATIVE IMAGING TECHNIQUES IN CLINICAL RADIATION ONCOLOGY

Presentata da: Enrico Menghi

Coordinatore Dottorato

Alessandro Gabrielli

Supervisore

Gastone Castellani

Co-supervisore

Anna Sarnelli

Esame finale anno 2024

Establishment of a radiomics laboratory: data mining and quantitative imaging techniques in clinical radiation oncology.

Contents

CHAPTER 1

General introduction and outline of the thesis

Enrico Menghi

INTRODUCTION

In 2020, World Cancer Report published by the International Agency for Research on Cancer, states that cancer ranked as the second most prevalent cause of mortality worldwide, responsible for approximately 9.6 million deaths in 2018 [1].

There is inevitably a radical change in the approach to patient care which is increasingly directed towards a personalization of medicine based on imaging. The quantification of the data present in the images has in fact a great impact in the treatment of certain pathologies such as the increase in the accuracy of the diagnosis and the differentiation of the treatment.

In recent years, there has been a great deal of technological effort by manufacturers in the field of imaging diagnostics and radiant therapies to provide imaging tools such as ultrasound, computed tomography (CT), positron emission tomography (PET) and resonances nuclear magnetic (NMR), increasingly performing and at the same time capable of producing a high quality and quantity of images while minimizing the dose to the patient.

In this context, the physicist and in particular the medical physicist possesses the skills of acquisition, analysis and modeling of the data necessary to support the clinician in identifying, measuring, and quantifying possible mathematical descriptors of a given pathology. These descriptors extracted from the images (features) aim to become real specific biomarkers for that type of imaging and pathology (imaging biomarkers).

This science, called radiomics, cannot be implemented except through a new and dedicated multidisciplinary approach of the various professionals involved, and must make use of a dedicated infrastructure that is able to make complex patient databases containing large quantities ordered, interrogable and manageable of multimodal images (imaging biobanks).

DATA MINING AND QUANTITATIVE IMAGING IN ONCOLOGY

The scientific justification of radiomics in oncology comes from genomics and as you can see in **Figure 1** the biopsy of a tumor is shown with the gene expression profiling sites indicated (left) and the relative results (right). It is evident from the figure that a tumor does not present itself as a single genetic organism but is characterized by a set of heterogeneous genetic organisms [2]. The microenvironments that induce this genetic heterogeneity (main cause of resistance to modern molecular therapies) can be visualized through clinical imaging techniques (CT, PET, NMR).

The extraction of quantitative information from tumor images therefore allows to identify these microenvironments and thus quantify their genetic heterogeneity. With this we expect to find a connection between the mathematical descriptors extracted from the images, the patient's prognosis and more specifically with the tumor phenotype taken into consideration [3].

In the patient care process, radiomics data integrate with genomics and clinical data as shown in **Figure 2**. From the figure it is evident that the cycle is interrupted if it is not supported by an appropriate technology (data warehouse) that manages in structured the different sets of images and go to extract useful information for the construction of predictive and personalized models [4].

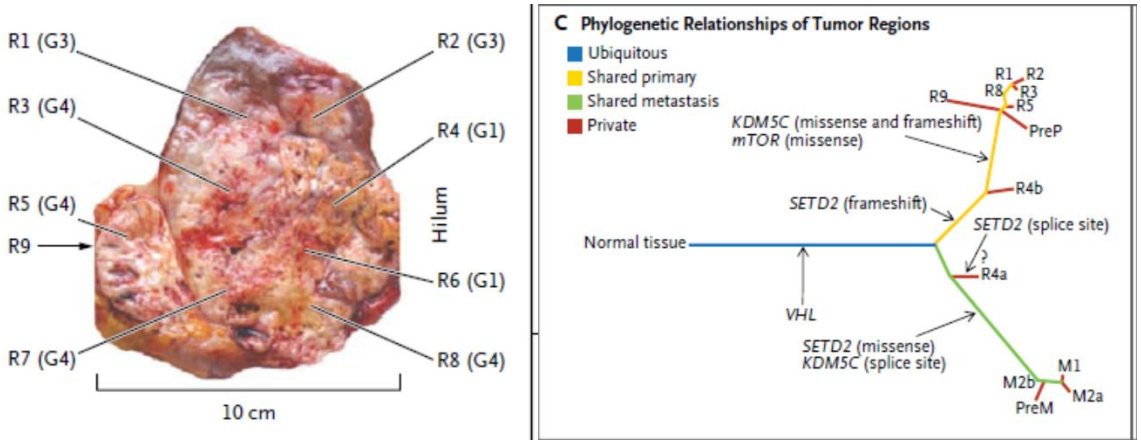


Figure 1 Biopsy of a kidney tumor with the gene expression profiling sites indicated (left) and related results where the genetic mutations present are evident compared to normal kidney tissue (right) from [1].

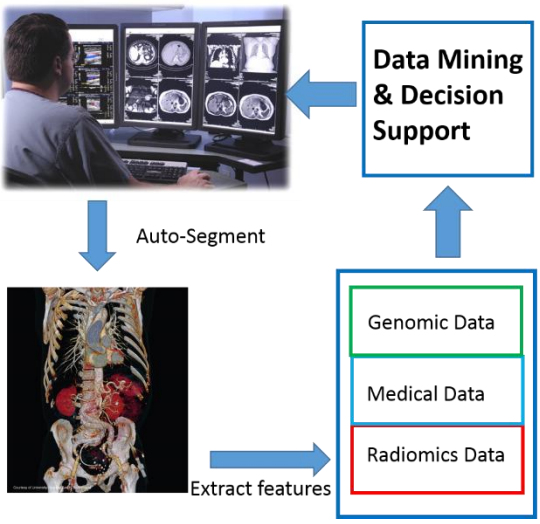


Figura 2 Integration of different kind of data for Data Mining and Decision Support (From a slide of plenary session of Radiological Society of North America 2015)

AIM OF THE THESIS

The objective of this PhD thesis is the establishment and the development of a so-called “Romagna Imaging Biobank” as a data mining and radiomics laboratory with the consequent study and application of radiomic descriptors to some clinical-oncological pathologies of primary interest in the IRCCS, Istituto Romagnolo per lo Studio dei Tumori “Dino Amadori” (IRST), (i.e. testicular, prostate and adrenal cancer). In particular, it is expected to:

- demonstrate the effective link between tumor phenotype and quantitative descriptors extracted from the imaging of the tumor for the pathologies considered.

- improve the stratification of patients with the identified descriptors and obtain feedback on the personalization of anti-cancer therapies

To acquire further development capabilities for the establishment of the laboratory we are collaborating with the working group of the Italian Association of Medical Physics (AIFM) called "Big Data and Artificial Intelligence": analysis of a survey (FM4AI) of AI and Machine Learning methods and areas used at medical physics services and in activities clinics and research [5], [6].

OUTLINE OF THE THESIS

This **Chapter 1** presents an introduction and a general overview.

The feasibility study and the development of a “Romagna Imaging Biobank” for the optimization of procedures and the implementation of the application of quantitative imaging presented in this thesis, is divided into the following two points:

1 - Development of a software platform capable of extracting, segmenting and analyzing radiomics descriptors from images from any clinical imaging instrument interfaced with IRCCS, IRST, starting from open-source and commercial software (i.e. Sophia DDM© Radiomics) and comparison of themselves through the use of digital and dedicated phantom as suggested by the Image Biomarkers Standardization Initiative (IBSI), is presented in **Chapter 2**. A multicentre study is conducted with digital and dedicated phantoms to test the effectiveness of the descriptors in controlled conditions (preprocessing parameters defined by set of images) to assess the agreement among different radiomic tools. In **Chapter 3** is presented a recent extension of the concept of data mining and extraction from classic structural and functional imaging: a multicentre evaluation of “dosiomics” features extracted from radiotherapy dose distribution converted in gray-scale level. This study derives from a first validation of predictive radiomic models with "real world data" in locally advanced rectal cancer through the Working Group Radiomics of the Alliance Against Cancer [7,8], then the WG Radiomics subgroup called "Dosiomics" aims to identify and standardize the extractors linked to the dose distributions delivered in clinical radiation oncological to predict the patient's clinical outcome (Dose Marker Standardization Initiative, DoMSI) [9].

2- Retrospective applications of radiomics to patients imaging with different pathologies and consequent identification of descriptors useful for improving the accuracy of the diagnosis, are presented in the next chapters of the thesis. In **Chapter 4** we present the potential role of MR-based radiomic biomarkers in the characterization of focal testicular lesions to investigate signatures for the preoperative prediction of testicular neoplasm histology. **Chapter 5** presents preliminary results of the Biopstage Trial, evaluating the ability of MRI-ADC and [68Ga]Ga-PSMA-11-based quantitative analysis to help differentiate low-risk prostate cancer patients (ISUP 1) from higher risk patient classes (ISUP>1) and aim to evaluate the benefits of the two imaging techniques combined. In **Chapter 6**, Radiomics in the characterization of lipid-poor adrenal adenomas at unenhanced CT is a retrospective observational study with a relevant clinical impact. Including more radiomic features in the identification of adenomas may improve the accuracy of not-enhanced CT and reduce the need for additional imaging procedures and clinical workup, according to this and other recent radiomics studies that have clear points of contact with current clinical practice.

A work under preparation [10] will have the aim of investigating the current state of adrenal tumors in order to further improve the study done in this chapter.

GENERAL DISCUSSION AND OUTLOOK OF THE RADIOMICS LABORATORY

Finally, **Chapter 7** is dedicated to a general discussion and conclusion, presenting a summary of the principal results of each chapter and a potential outlook for the future of the laboratory.

REFERENCES

[1] Wild CP, Weiderpass E, Stewart BW. World Cancer Report: Cancer research for cancer prevention 2020. IARC

[2] M. Gerlinger, D. Endesfelder, A. Stewart, and C. Santos, "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing," N. Engl. J. Med., vol. 366, pp. 883–892, 2012.

[3] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, B. Haibe-kains, P. Grossmann, S. Carvalho, J. Bussink, A. Dekker, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. Rene, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," Nat. Commun., vol. 5:4006, pp. 1–8, 2014.

[4] Tânia F. G. G. Cova , Daniel J. Bento and Sandra C. C. Nunes , "Computational Approaches in Theranostics: Mining and Predicting Cancer Data", Pharmaceutics 2019, 11, 119; doi:10.3390/pharmaceutics11030119 www.mdpi.com/journal/pharmaceutics

[5] Kortesniemi M, Tsapaki V, Trianni A, Russo P, Maas A, Källman HE, Brambilla M, Damilakis J.The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession. Phys Med. 2018 Dec;56:90-93.

[6] Neri E, Regge D. Imaging biobanks in oncology: European perspective. Future Oncol. 2017 Feb;13(5):433-441.

[7] Alex Zwanenburg, PhD, Martin Vallières, PhD, Mahmoud A. Abdalah, PhD, Hugo J. W. L. Aerts, PhD, Vincent Andrearczyk, PhD, Aditya Apte, PhD, Saeed Ashrafinia, PhD, Spyridon Bakas, PhD, Roelof J. Beukinga, PhD, Ronald Boellaard, PhD, Marta Bogowicz, PhD, Luca Boldrini, PhD, Irène Buvat, PhD Gary J. R. Cook, PhD, Christos Davatzikos, PhD, Adrien Depeursinge, PhD, Marie-Charlotte Desseroit, PhD, Nicola Dinapoli, PhD, Cuong Viet Dinh, PhD, Sebastian Echegaray, PhD, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping", Radiology · March 2020

[8] Boldrini, L., Cusumano, D., Chiloiro, G., Casà, C., Masciocchi, C., Lenkowicz, J., Cellini, F., Dinapoli, N., Azario, L., Teodoli, S., Gambacorta, M.A., De Spirito, M., Valentini, V., "Delta radiomics for rectal cancer response prediction with hybrid 0.35 T magnetic resonance-guided radiotherapy (MRgRT): a hypothesis-generating study for an innovative personalized medicine approach. Radiol Med. https://doi.org/10.1007/s11547-018-0951-y

[9] Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T et al. Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis, Front Oncol 2019 Apr 12;9:269. Doi: 10.3389/fonc.2019.00269. eCollection 2019

[10] Menghi et al. A review on radiomic studies for lipid-poor adrenal adenomas. Work in preparation

CHAPTER 2

A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools

Published in: **Radiology** 2022 Jun; 303(3): 533-541

A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools

Andrea Bettinelli, Francesca Marturano, Michele Avanzo, Emiliano Loi,
Enrico Menghi, Emilio Mezzenga, Giovanni Pirrone, Anna Sarnelli, Lidia Strigari,
Silvia Strolin, Marta Paiusco

ABSTRACT

Background: The translation of radiomic models into clinical practice is hindered by the limited reproducibility of features across software and studies. Standardization is needed to accelerate this process and to bring radiomics closer to clinical deployment.

Purpose: The aim of the study was twofold: 1) to assess the standardization level of seven radiomic software programs and 2) to investigate software agreement as a function of in-built image pre-processing (e.g. interpolation and discretization), feature aggregation methods and the morphology (i.e., volume and shape) of the region of interest (ROI).

Material and Methods: The study was organized into two phases: in Phase I, the two Image Biomarker Standardization Initiative's (IBSI) phantoms were used to evaluate the IBSI-compliance of 7 software programs. In Phase II, the reproducibility of all IBSI-standardized radiomic features across tools was assessed on two custom ImSURE digital phantoms that allowed, in conjunction with a systematic feature extraction, to observe whether and how feature matches between program pairs varied depending on the pre-processing steps, aggregation methods and ROI characteristics.

Results: In Phase I, we found that software programs presented different levels of completeness (i.e., the number of computable IBSI benchmark values). However, the IBSI-compliance assessment revealed that they were all standardized in terms of feature implementation. When considering additional pre-processing steps, for each individual program, match percentages fell by up to 30%. In Phase II, we found on the ImSURE phantoms that software agreement was dependent on discretization and aggregation as well as on ROI shape and volume factors.

Conclusion: The agreement of radiomic software varied in relation to factors that had already been standardized (e.g. interpolation and discretization methods) and to factors that need standardization. Both dependences must be resolved to ensure the reproducibility of radiomic features and to pave the way towards the clinical adoption of radiomic models.

ARTICLE TYPE: Original Research

ABBREVIATIONS

- ROI: Region of interest
- IBSI: Image Biomarker Standardization Initiative
- RFs: Radiomic features
- FBS: Fixed bin size
- FBN: Fixed bin number
- ImSURE: Italian multicenter Shared Understanding of Radiomic Extractors

KEY RESULTS

- On IBSI phantoms, radiomic software programs were able to compute different percentages (21-100%) of the IBSI benchmark values, however, they were all highly standardized regarding feature definition. When considering pre-processing, 'matching' values with the IBSI benchmark fell by up to 30% for the individual program.
- On ImSURE phantoms, software agreement was significantly dependent, α=0.05 (Bonferroni-adjusted α=9e-5), on discretization and aggregation methods, as well as on newly investigated factors (i.e., ROI shape and volume).

SUMMARY STATEMENT

We employed a novel approach to assess radiomic software agreement based on the ImSURE phantoms and a systematic feature extraction, finding that discrepancies were still present among standardized radiomic programs.

1. INTRODUCTION

Radiomics is the high-throughput extraction and analysis of quantitative imaging features with the goal of supporting medical decision-making by developing predictive and prognostic models (1–4). Radiomics has increasingly gained interest in radiology and oncology in recent years for cancer diagnosis, and for the prediction of prognosis or response to treatment (5–7). The lack of standardization in the definition and calculation of radiomic features (RFs), their ambiguous nomenclature and the limited reproducibility of radiomic studies (8–11) have all impeded the adoption of radiomics within clinical practice.

Some of these concerns were recently addressed by the Image Biomarker Standardization Initiative (IBSI) (12) which published a reference manual (13) comprising the definition of 169 standardized RFs and reporting guidelines on how to perform image pre-processing. Moreover, the IBSI shared two digital phantoms (14) with their respective benchmark feature values to assess the accuracy of software tools for radiomic analysis. Because of the raised awareness of the need for standardization, several radiomic tools have begun to conform to the IBSI-guidelines (15–18). However, a number of radiomic studies are based on in-house or public software with an unclear level of standardization concerning at least one of the several aspects involved for feature extraction (e.g. pre-processing methods, availability of tuning parameters).

In this context, the present study stems from a collaboration between four Italian clinical research institutes, the Italian multicenter Shared Understanding of Radiomic Extractors (ImSURE) group (i.e. Azienda Ospedaliero-Universitaria di Bologna, Centro di Riferimento Oncologico di Aviano, Veneto Institute of Oncology and Isstituto Romagnolo per lo Studio dei Tumori “Dino Amadori”), and aims 1) to assess IBSI-compliance of commonly available software tools for radiomics (Phase I of the study) and 2) to investigate the causes of possible discrepancy for a representative set of radiomic tools by designing a systematic workflow of RFs extraction performed on two custom digital phantoms comprising multiple regions of interest (ROIs) with various volumes and shapes (Phase II of the study). For completeness, we considered the entire set of IBSI-standardized RFs and we explored all possible combinations of pre-processing steps/aggregation methods. Finally, we exhaustively investigated the causes of discrepancy among the programs under consideration and discussed, for each software, whether differences were attributable to the limited or non-IBSI-compliant implementation of one or more aspects. Limited flexibility in parameter setting, as well

as software discrepancies, are non-negligible factors for the reproducibility of the RFs and for the general validity of the models proposed in the radiomic literature.

2. MATERIALS AND METHODS

2.1 Radiomic tools and radiomics features

Seven radiomic tools were evaluated, of which six were open-source (MIRP, S-IBEX, RaCaT, SERA, PyRadiomics, and RadiomiCRO) and one was commercial (SOPHiA DDM for Radiomics, SOPHIA GENETICS). The inclusion criteria were: 1) that the software was self-declared IBSI-compliant and/or IBSI participant and 2) that there was a consolidated experience in the software tuning by at least one of the four centers participating in the study. **Table 1** reports the salient characteristics of these tools. To ensure correctness of their use, the seven software programs were picked from the tools readily available at the project's participating centers and, when possible, were assigned to two centers for a consensus extraction of features.

Software	*IBSI-compliant*	*Version*	*Language*	*Data format*	*Characteristics*	*Documentation*	*Assigned centers*
A **(28)**	IBSI-participant	v1.0.2	Python	DICOM	Open source	✗	CRO, IOV
B **(16)**	self-declared	v2	Matlab	DICOM	Open source	✓	IOV, IRST
C **(29)**	self-declared	v2.2.0	-	DICOM	Commercial	✓	IRST, IOV
D **(18)**	IBSI-participant	v1.18	C++	DICOM, NRRD, NIfTI	Open source	✓	IOV, BO
E **(15)**	IBSI-participant	v2.1	Matlab	DICOM	Open source	✓	CRO, IRST
F **(17,30)**	IBSI-participant	v3.0.1	Python	NRRD, NIfTI	Open source	✓	BO, IRST
G **(31)**	self-declared	-	Matlab	DICOM	In-house	✗	CRO

Table 1**. Software packages included in the study**.

***A**=MIRP; **B** = S-IBEX; **C**=SOPHiA DDM; **D**=RaCaT; **E**=SERA; **F**=PyRadiomics; **G**=RadiomiCRO. IBSI-participant: tools that participated in IBSI for the standardization of pre-processing and feature calculation; self-declared: software for which there are independent works that state their IBSI-compliance.*

All one hundred and sixty-nine RFs that were standardized by IBSI and grouped into 11 feature families were considered (13). The corresponding processing requirements for each family are summarized in **Table 2**.

Feature family	Feature Count	Discretization	Aggregation
MORPH	25	none	none
LI	2		
IS	18		
IH	23	FBN or FBS	rotation dependent (2D:avg, 2D:mrg, 2.5D:dmrg, 2.5D:vmrg, 3D:avg, 3D:mrg)
IVH	6		
GLCM	25		
GLRLM	16		
GLSZM	16		rotation independent (2D, 2.5D, 3D)
GLDZM	16		
NGTDM	5		
NGLDM	17		

Table 2*. Standardized feature families and required settings*

*(**MORPH** = morphological features; **LI** = local intensity; **IS** = Intensity-based statistics; **IH** = Intensity histogram; **IVH** = Intensity-volume histogram; **GLCM** = Grey-level co-occurrence matrix; **GLRLM** = Grey-level run-length matrix; **GLSZM** = Grey-level size-zone matrix; **GLDZM** = Grey-level distance-zone matrix; **NGTDM** = Neighborhood grey tone difference matrix; **NGLDM** = Neighboring grey level dependence matrix; **FBN** = fixed bin number; **FBS** = fixed bin size; **2D:avg** = averaged over slices and directions; **2D:mrg** = merged directions per slice and averaged; **2.5D:dmrg** = merged per direction and averaged; **2.5D:vmrg** = merged over all slices; **3D:avg** = averaged over 3D directions; **3D:mrg** = merged 3D directions; **2D** = averaged over slices; **2.5D** = merged over all slices;**3D** = calculated from single 3D matrix).*

2.2 Phase I: Assessment of IBSI compliance

To quantify the level of IBSI-compliance of the seven software packages, RFs were extracted from two digital phantoms proposed by IBSI (13,14). In this context, a digital phantom is typically made up of an image containing intensity values and one or more regions of interest (ROIs), which enclose the image voxels to be used for feature calculation. The IBSI digital phantom consists of 5x4x4 isotropic voxels with one ROI, while the IBSI radiomic phantom is a CT image

from a patient with lung carcinoma where the gross tumor volume was used as the region of interest (19). RFs were extracted using the configuration settings proposed in the IBSI reference manual (13). We considered all five IBSI parameter configurations for the radiomic phantom, labeled A to E, and characterized by either 2D or 3D feature aggregation and either fixed bin size - FBS or fixed bin number - FBN discretization. In total, 482 feature values were extracted for the digital phantom applying neither interpolation nor discretization, whereas 1322 radiomic feature values were computed using all five configurations for the radiomic phantom. The calculated RFs were compared with the corresponding IBSI benchmark values and classified into 'matching' (differences ≤ IBSI-reported tolerance), 'partial matching' (differences ≤ three times the IBSI-reported tolerance) or 'no matching' (otherwise), accordingly to the evaluation criteria proposed by the initiative (12). Features that were not implemented within a tool were labeled as 'missing'.

2.3 Phase II: Software comparison on the ImSURE digital phantoms

The reproducibility of features across software programs was assessed for different pre-processing choices (e.g. interpolation and discretization) (20), feature aggregation methods (e.g. 2D, 2.5D, or 3D) and ROI characteristics (e.g. volume and shape). With this aim in mind, we designed two digital phantoms and a systematic feature extraction that included all possible combinations of the factors under investigation.

2.3.1 ImSURE phantoms

The ImSURE digital phantoms were designed with ROIs containing the texture of a medical image in order to mimic the content of a clinical ROI, and with geometrically defined morphologies, to control for ROI shape and volume.

A computed tomography (CT) image was selected from a pool of individuals who signed informed consent at one of our institutes and whose data had been previously used in studies approved by the Institutional Review Board. The image, that was artifact-free, was retrieved from the PACS and anonymized.

The CT scan was acquired from the skull base to mid-thigh with anisotropic voxel dimensions of 0.98x0.98x3.00 mm. The original image was used to create the ImSURE 'anisotropic phantom'. An IBSI-compliant trilinear interpolation was then applied to the CT to generate a second image with voxel dimensions of 1.00x1.00x1.00 mm, to be used for the ImSURE 'isotropic phantom'.

Nine different ROIs were obtained by combining three possible volumes (i.e., small, medium, and large) with three different shapes (i.e., bean, cube, and sphere). Ten different instances for each shape-volume combination were positioned in the space of each CT image, obtaining a total of 90 ROIs. **Figure 1** depicts the spatial arrangement of the ROIs, whereas Supplemental Table E1 summarizes each phantom's characteristics.

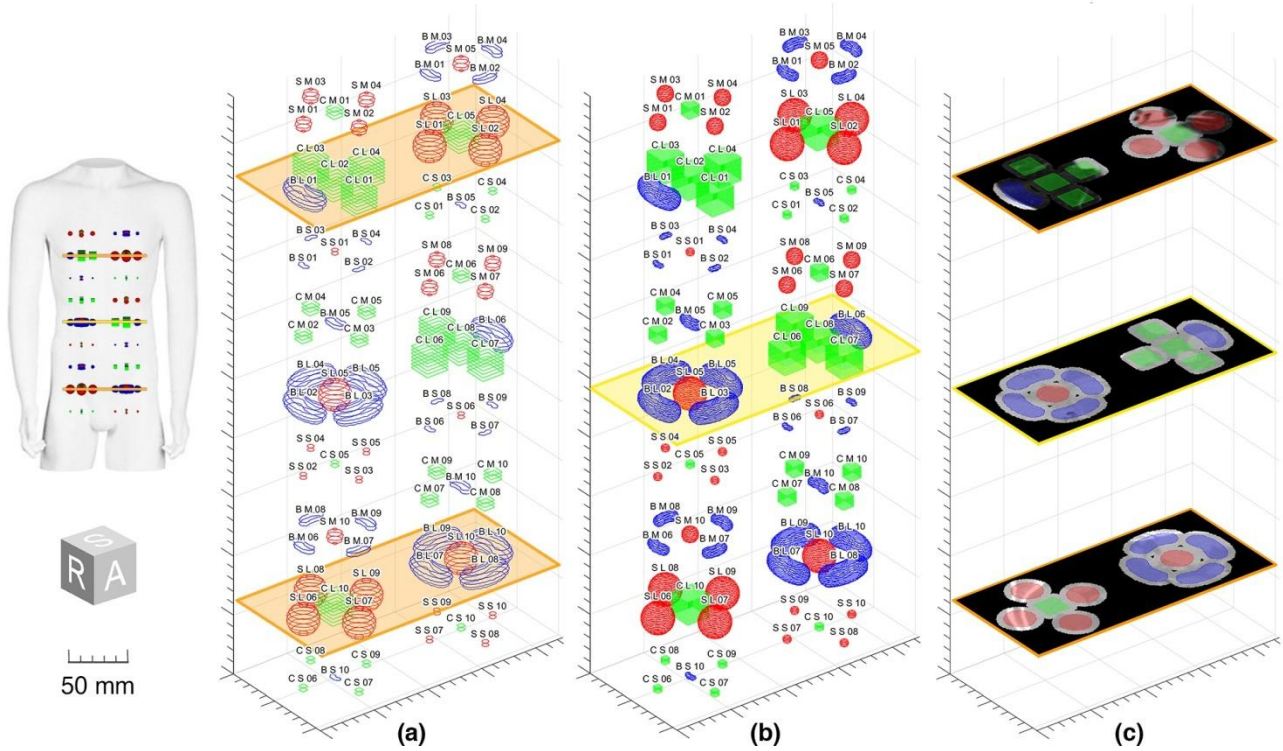


Figure 1. **(a-b)** Spatial arrangement of the 90 ROIs for the ImSURE anisotropic and isotropic phantoms, respectively. The 9 different shape-volume ROI configurations are visible: the spheres are in red, the beans in light blue, and the cubes in green. The ROIs were axially positioned from the apex of the lung upper lobe to the femoral head. The first letter of each ROI label indicates the shape (i.e., C = cube, S = sphere, and B = bean), the second indicates the size (i.e., S = small, M = medium, L = large), while the number indicates the specific instance (from 1 to 10); **(c)** Three representative slices of the phantoms. The texture of the underlying CT image was maintained within and around the ROIs, while the surrounding voxels were censored by setting their intensity to -1024 Hounsfield Units (black area). The ImSURE phantoms are available from the corresponding author upon request.

2.3.2 Extraction of radiomic features

The IBSI standardized RFs were extracted for each phantom by varying the configuration parameters reported in **Table 2**. All possible combinations were considered for the feature families that needed both grey-level discretization and feature aggregation. A total of 919 feature values were obtained from each phantom (Supplemental Figure S1, online). **Table 3** shows the image processing parameters used to harmonize feature extraction across the software.

Finally, for each tool, 'no matching' features on the IBSI digital phantom were excluded from the analysis, and the remaining feature values were rounded up to the third significant digit.

Pre-processing step	***Isotropic phantom***	***Anisotropic phantom***
Trilinear Interpolation		
resampled voxel spacing [mm]	none	1.00x1.00x1.00
Re-segmentation		
range [HU]	[-1000 400]	[-1000 400]
Discretization		
texture and IH	FBS: 25 HU; FBN: 32 bins	FBS: 25 HU, FBN: 32 bins
IVH	FBS: 2.5 HU; FBN: 1000 bins	FBS: 2.5 HU; FBN: 1000 bins

Table 3. *Pre-processing settings used for the isotropic and anisotropic phantom**.***

*(**FBN** = fixed bin number; **FBS** = fixed bin size; **HU** = Hounsfield units; **IH** = Intensity histogram; **IVH** = Intensity-volume histogram)*

2.3.3 Performance metrics and statistical analysis

We used the percentage of matching features between pairs of programs and their level of agreement to assess and compare software performances. For a pair of software s_i, s_j, the percentage of matching features, P, was calculated as:

$$P(s_i, s_j) = P(s_j, s_i) = \frac{\#\ matching\ features\ between\ s_i, s_j}{\#\ of\ comparable\ features\ between\ s_i, s_j}$$

where $i, j = 1, \ldots$, number of programs and $i \neq j$. For each feature, f, agreement across all software tools, A, was defined as:

$$A = \frac{1}{\#S}\sum_{S}[f_{s_i} = f_{s_j}]$$

where S is the set of unordered program pairs (s_i, s_j, with $i \neq j$) that are able to calculate the feature f, and $\#S$ is the dimension of S. Squared brackets represent the Iverson brackets, that is:

$$[f_{s_i} = f_{s_j}] = \begin{cases} 1 \ \ if \ f_{s_i} = f_{s_j} \\ 0 \ otherwise \end{cases}$$

The non-parametric Kruskal-Wallis test (21) was used to investigate whether A was significantly influenced by the factors being considered, under the null hypothesis that all groups came from populations with the same median. The significance level, $\alpha = 0.05$, was corrected with Bonferroni's method (adjusted α of 9e-5). The statistical analysis was performed in MATLAB (version 2018b, The MathWorks, Natick, 2018).

3. RESULTS

3.1 Phase I

For each software program, the resulting percentage of 'matching', 'partial matching', 'no matching' and 'missing' features, both for the IBSI digital and radiomic phantoms, are shown in **Figure 2a** and **2b**, respectively. For the latter, the results have been aggregated over the five parameter configurations proposed in the reference manual, while the outcomes stratified by configuration are reported in Supplemental Figure S2 (online). On the digital phantom MIRP, S-IBEX, RaCaT, SERA and SOPHiA all achieved percentages of matches above 94%, while PyRadiomics and RadiomiCRO had 52% and 25% of 'matching' features, respectively, due to 'missing' feature values. RaCaT, SERA and RadiomiCRO all exhibited a slight decrease in the percentage of matching features (90%, 85%, and 21%, respectively) on the radiomic phantom, while PyRadiomics showed a marked increase in partial matches and no matches. SOPHiA presented 16% missing features as config. E is currently not obtainable.

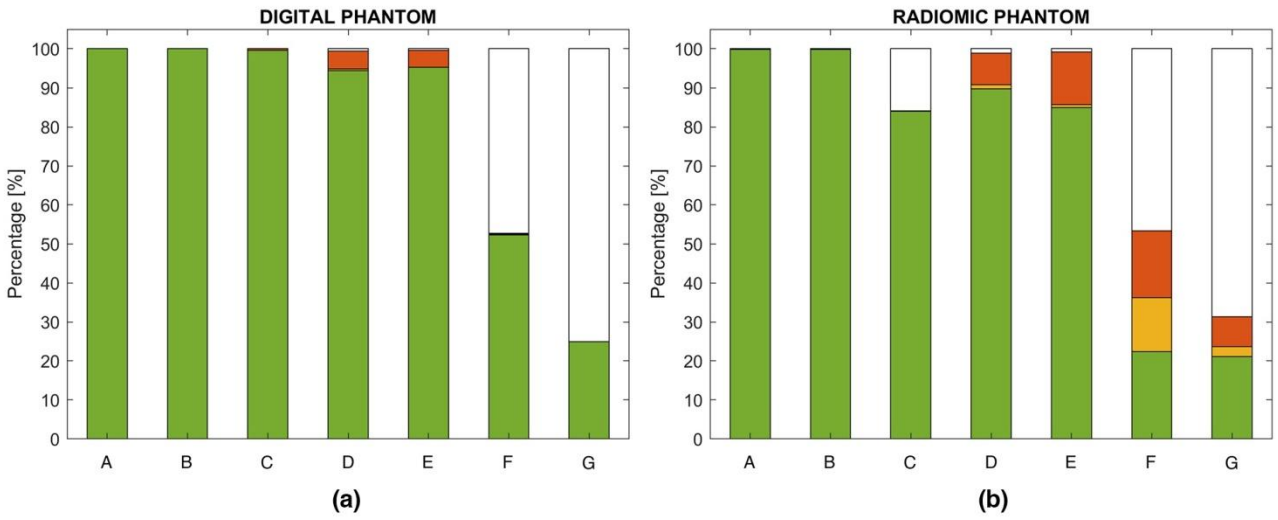


Figure 2. Percentages of "matching" (differences below the IBSI-reported tolerance, in green), "partial matching" (differences below three times the IBSI-reported tolerance, in yellow), "no matching" (otherwise, in red) feature values obtained for each software package on the *IBSI digital* **(a)** and *radiomic* **(b)** *phantoms*. The feature values that could not be calculated within a tool were labeled as 'missing' (white). The percentages for the *radiomic phantom* were averaged across the five IBSI configurations. A = MIRP; B = S-IBEX; C = SOPHiA DDM for radiomics; D = RaCaT; E = SERA; F = Pyradiomics; G = RadiomiCRO.

3.2 Phase II

The 'no matching' features that were excluded from Phase II were grouped by feature family and by aggregation method for each program (Supplemental Figure S3, online). No features were excluded for MIRP, S-IBEX and RadiomiCRO. One and two 'no matching' features were found for Pyradiomics and SOPHiA, respectively. A total of 21 and 22 features were eliminated for SERA and RaCaT, respectively, mostly belonging to the NGLDM 2.5D and LI families for the former and MORPH and LI for the latter.

Figure 3a shows the percentage of comparable features for each pair of programs out of a total of 919 possible values, while **Figures 3b** and **3c** compare the percentages of matches of the 'isotropic phantom' (no program-specific interpolation required) with those of the 'anisotropic phantom' (interpolated within each program before feature calculation). It should be noted that the reported match percentages were computed with respect to the total number of comparable features shared by each program pair. By comparing Figures 3b and 3c, we observed that program-based interpolation had an impact on the overall percentage of matching features. When interpolation was applied, the match percentages of PyRadiomics fell below 2.5%, suggesting that

the interpolation method used in this program may not be compliant with the IBSI guidelines, while those of SERA presented a marked decrease. SERA behavior could be ascribed to an erroneous interaction between interpolation and 2D/2.5D aggregation methods for FBS discretization, rather than to a non-compliant interpolation. Instead, program-specific interpolation had no effect on the MIRP, S-IBEX, SOPHiA, RaCaT, and RadiomiCRO values.

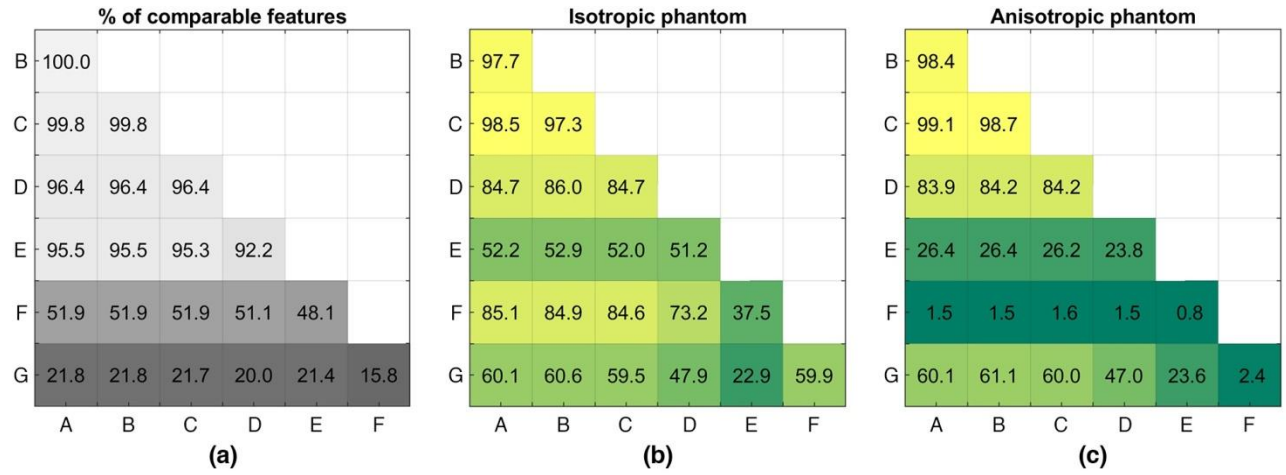


Figure 3. Analysis of interpolation effect. **(a)** Percentage of comparable features between program pairs out of the total of 919 features. **(b)** Percentage of matches between program pairs for the isotropic phantom (no interpolation required). **(c)** Percentage of matches for the anisotropic phantom (requiring program-based interpolation). A = MIRP; B = S-IBEX; C = SOPHiA DDM for radiomics; D = RaCaT; E = SERA; F = Pyradiomics; G = RadiomiCRO.

Following these results, in the subsequent analyses, we only focused on the isotropic phantom to rule out the discrepancies observed between programs caused by software-based interpolation. Consequently, we investigated the effect of the discretization approach on the percentage of matches between pairs of programs. **Figure 4** shows the results for the cases of no discretization **(a-b)**, FBN **(c-d)** and FBS **(e-f)**. **Figure 4a-b** only includes the MORPH, LI, and IS feature families, which do not require intensity discretization (see **Table 2**). In this case, all programs achieved match percentages higher than 80%. **Figures 4c-d** and **4e-f** aggregate the remaining families calculated using the FBN and FBS approaches, respectively, and highlight that SERA FBN discretization and RadiomiCRO FBS discretization are not concordant with the other programs.

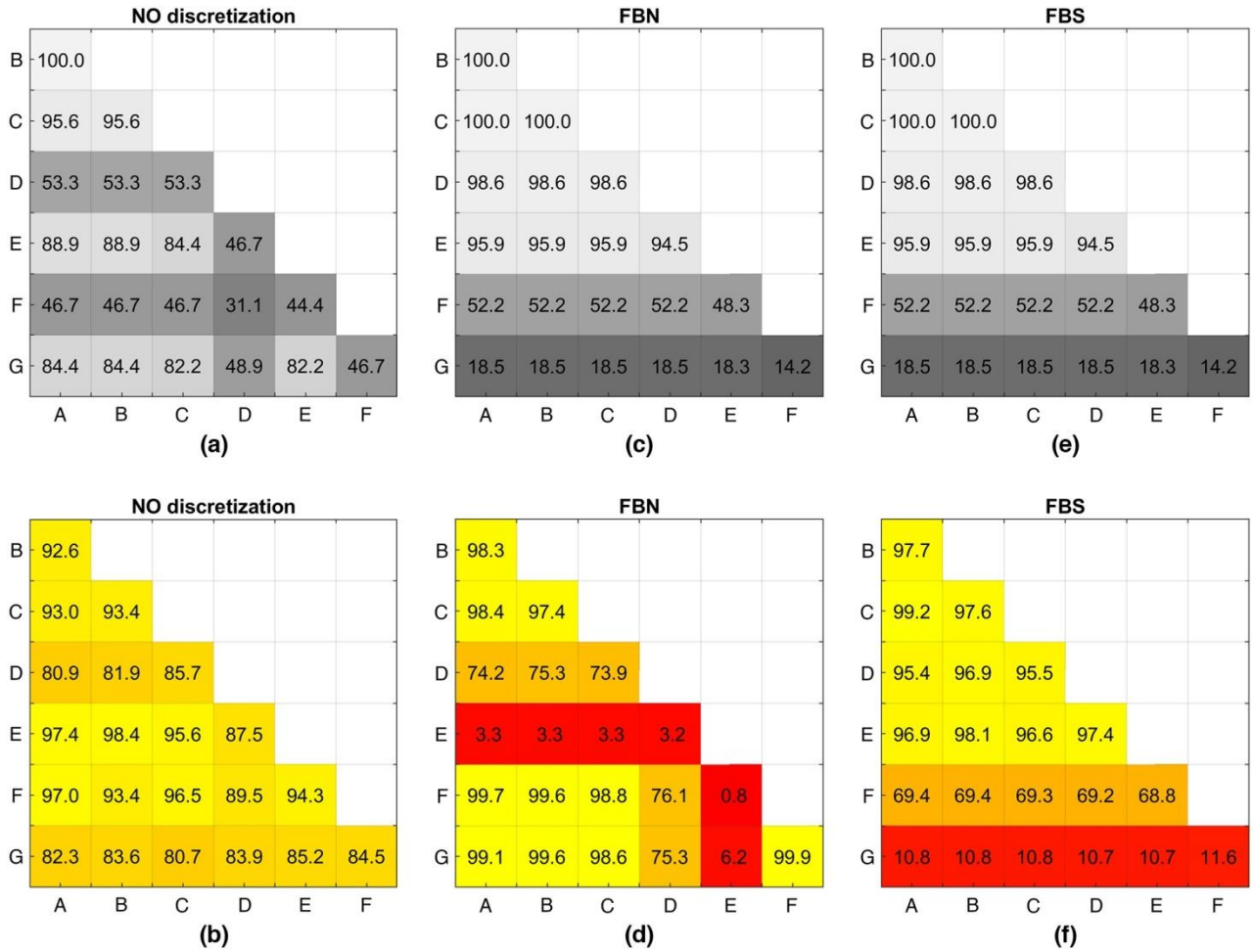


Figure 4. Analysis of the discretization effect. (**Top row**) Percentages of comparable features. (**Bottom row**) Percentages of matches between program pairs considering feature families without discretization, with FBN or with FBS discretization, respectively. FBN = fixed bin number; FBS = fixed bin size; A = MIRP; B = S-IBEX; C = SOPHiA DDM for radiomics; D = RaCaT; E = SERA; F = Pyradiomics; G = RadiomiCRO.

This suggests that their implementation is not IBSI-compliant for these programs. Notably, the RadiomiCRO discrepancy confirmed the results obtained in Phase I while the SERA discrepancy was not visible on the IBSI radiomic phantom. Regardless of the discretization method used, MIRP, S-IBEX, and SOPHiA achieved the highest match percentage. PyRadiomics showed greater match percentages for FBN discretization than for FBS, while the RaCaT results were complementary, with higher percentages for the FBS method.

The effect of the aggregation method on the percentage of matching features across program pairs, stratified by FBN and FBS approach, was also evaluated. Supplemental Figures S4 and S5 (online) illustrate the results in greater detail. PyRadiomics could not calculate the feature values associated with 2D aggregation, while RadiomiCRO was only designed to calculate 3D:mrg aggregation. The match percentages for MIRP were lower in 2D aggregation than in other aggregation methods. This result was observed for some ROI conformations that produced undefined results for the 2D aggregation method in the intermediate steps of feature calculation (further details can be found in the supplemental material, online).

Multiple ROIs with varied volumes and shapes were included in the two phantoms designed for this study, allowing us to also investigate the differences in program performance due to ROI characteristics. The data were stratified by ROI shape and ROI volume, and match percentages between software pairs were calculated in the two cases. The results are presented in Supplemental Figure S6 and S7, respectively (online). Unlike the other factors, this analysis showed no relevant differences between programs due to ROI shape or ROI volume, meaning that ROI morphology had no discernible impact on match percentages at the whole-feature level.

Finally, the non-parametric Kruskal-Wallis test (21) was applied to agreement values for insights at the single-feature level, distinguishing four main factors: discretization, aggregation methods, both rotation dependent and independent (see **Table 2**), ROI shape, and ROI volume. The test results are shown in **Figure 5** for each factor and feature under examination. This analysis showed that discretization was significant for almost every feature family requiring intensity discretization. The aggregation factor was significant for most of the features belonging to the GLCM and GLRLM classes, as well as for some GLSZM and NGLDM features. The ROI shape was only significant for the features belonging to the MORPH and LI families, while the ROI volume was significant for almost all the GLCM features, as well as for a portion of the MORPH, IS, IH, and NGTDM features.

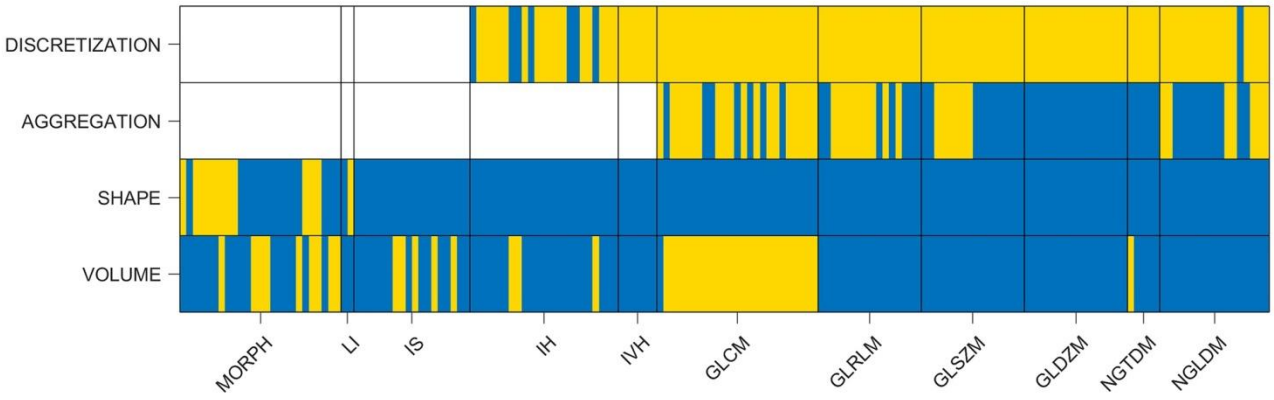


Figure 5. The results of the Kruskal-Wallis test applied to the agreement among programs. The test results are presented for each feature family and for four different factors, i.e., discretization, aggregation, ROI shape, and ROI volume. In the figure, the yellow color indicates significant differences after Bonferroni correction ($p \leq 9e\text{-}5$), the blue color denotes non-significant results, and the white cells correspond to non-existing combinations of feature families and factors.

4. DISCUSSION

We analyzed the performance of seven self-declared IBSI-compliant software packages. Phase I analysis on the IBSI digital phantom revealed that all programs achieved high percentages of 'matching' features, indicating a high standardization level in terms of RF implementation. However, programs showed different degrees of feature completeness, with PyRadiomics and RadiomiCRO having the highest number of non-computable feature values. The IBSI radiomic phantom analysis allowed us to consider the effects of multiple factors, such as image interpolation and intensity discretization, and highlighted the limited flexibility in the parameter settings of some tools.

By comparing our Phase I results with the ones of the IBSI study (12), we found them in accordance for MIRP and RaCaT, whereas SERA showed a higher percentage of 'matching' features on both the IBSI digital and radiomic phantoms in configurations A-D, but only a partial improvement in configuration E. Instead, PyRadiomics presented a lower 'matching' percentage and higher percentages of 'no match' and 'partial match' on the radiomic phantom. These differences could be the result of a missing update of either the software documentation or the version used by the IBSI.

In Phase II, we systematically investigated the effect of factors related to parameter setting (i.e., interpolation, discretization, and aggregation) as well as to ROI characteristics (i.e. volume and shape) on software agreement by employing two custom digital phantoms and a systematic

feature extraction. For the calculation of the percentage of matching features, P, and of software agreement, A, we considered all pairwise comparisons among tools instead of comparing them to a reference one, as we were unable to justify choosing one tool over the others: even IBSI-compliance assessed in Phase I was not a reasonable criterion, as in Phase II we explored aspects that were not analyzable on the IBSI phantoms.

The interpolation effect was analyzed by comparing the match percentages between isotropic and anisotropic phantoms. The results revealed that for SERA and Pyradiomics, the performances were influenced by program-specific interpolation. Notably, interpolation is one of the initial steps in the image processing scheme and has an impact on downstream processes as well as on final feature values. Thus, it should be a priority of standardization for all programs.

We subsequently evaluated the effect of intensity discretization, focusing on the isotropic phantom. We found that it was only among MIRP, S-IBEX, and SOPHiA that percentages of matching features were not impacted by the discretization method. This pre-processing step is typically applied to the ROI before the calculation of IH, IVH, and textural features. Therefore, correct implementation is also crucial for the reproducibility of these feature values across tools.

The analysis of the aggregation method allowed the identification of an aspect that still needs to be addressed by the IBSI, which caused the programs to calculate different RF values as they are currently not aligned in the implementation strategy.

In contrast to previous studies (22–24), we analyzed the entire set of IBSI-standardized RFs rather than only those that were common to all tools. Secondly, we disregarded program-default settings and only considered harmonized extraction (i.e. user defined parameter settings) because, in practice, users tweak the software to match a desired parameter configuration.

In literature, digital phantoms range from being purely synthetic (e.g. the IBSI digital phantom with artificial texture and arbitrarily-defined ROI) to image-based (e.g. the IBSI radiomic phantom with CT-derived pattern and GTV ROI). The ImSURE phantoms were designed with intermediate characteristics (textures derived from a CT image and geometrical ROIs) to allow the assessment, in a single investigation framework, of the impact on the software agreement of factors related to both image pre-processing and ROI morphology. Moreover, by placing multiple ROIs over a patient's image, different texture patterns were sampled, hence augmenting the

casuistry and heterogeneity (different anatomical regions were tested in the same run) of the ROIs that were used in the analysis.

Regarding the limitations of this study, the ImSURE phantoms used for the analysis were made of simplified morphologies arbitrarily positioned on a single image modality (i.e. CT). Nevertheless, the choice of the modality does not affect the overall outcome of the work, which was designed to assess and compare basic aspects of image processing among radiomic tools. In future, phantoms constructed with other modalities will allow for further investigations on modality-specific aspects (25,26). Concerning the morphologies, the chosen ROI shapes are less complex than clinical ROI, however, this simplification was necessary to systematically study the impact of ROI characteristics. Eventually, ROIs may intersect anatomical structures differently with respect to a ROI defined for clinical studies. However, multiple textural patterns were derived from different anatomical districts, which ensured covering of the feature range obtainable from clinical ROIs imaged with CT. In these terms, we are reasonably confident that the software concordance tested on our phantom could be translated into software concordance calculated for clinical targets in several districts.

The results we obtained are relative to a selected number of radiomic software programs and future studies might include additional packages to strengthen the present findings. However, we are reasonably confident that the considered packages are a representative set of the high-standardized radiomic tools available in the literature. Moreover, some of our findings are software-independent and have general validity.

It is important to note that the differences observed in extracted feature values might limit radiomic model reproducibility (9–11). Therefore, when building a model, it is recommended that the stability of selected features is checked by comparing the values obtained with at least two different tools. However, future studies are needed to assess the impact of software differences on clinical endpoint prediction (22,27).

In conclusion, we designed a new investigation scenario in which we demonstrated that, despite the ongoing efforts of both IBSI and software developers to standardize radiomic tools, additional efforts are needed to achieve full concordance. This would hasten the use of radiomic models in clinical practice and their application to improve cancer prognosis.

ACKNOWLEDGMENTS

We thank the SOPHiA Genetics data science team, namely Dr. Olivier Gallinato and Dr. Antoine Huc, for their collaboration and for providing us with the support needed for this work.

REFERENCES

1. Huang Y, Liu Z, He L, et al. Radiomics signature: A potential biomarker for the prediction of disease-free survival in early-stage (I or II) non-small cell lung cancer. Radiology. Radiological Society of North America Inc.; 2016;281(3):947–957. doi: 10.1148/radiol.2016152234.

2. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. Nat Rev Clin Oncol; 2017;14(12):749–762. doi: 10.1038/NRCLINONC.2017.141.

3. Coppola F, Mottola M, Lo Monaco S, et al. The Heterogeneity of Skewness in T2W-Based Radiomics Predicts the Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. Diagnostics. MDPI AG; 2021;11(5):795. doi: 10.3390/diagnostics11050795.

4. Guerrisi A, Loi E, Ungania S, et al. Novel cancer therapies for advanced cutaneous melanoma: The added value of radiomics in the decision making process–A systematic review. Cancer Med. Blackwell Publishing Ltd; 2020. p. 1603–1612. doi: 10.1002/cam4.2709.

5. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: The promise of radiomics. Phys. Medica. Associazione Italiana di Fisica Medica; 2017. p. 122–139. doi: 10.1016/j.ejmp.2017.05.071.

6. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology. Radiological Society of North America Inc.; 2016;278(2):563–577. doi: 10.1148/radiol.2015151169.

7. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to radiomics. J Nucl Med. Society of Nuclear Medicine Inc.; 2020;61(4):488–495. doi: 10.2967/JNUMED.118.222893.

8. Bogowicz M, Vuong D, Huellner MW, et al. CT radiomics and PET radiomics: Ready for clinical implementation? Q. J. Nucl. Med. Mol. Imaging. Edizioni Minerva Medica; 2019. p. 355–370. doi: 10.23736/S1824-4785.19.03192-3.

9. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys. Elsevier Inc.; 2018;102(4):1143–1158. doi: 10.1016/j.ijrobp.2018.05.053.

10. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys. Med. Biol. Institute of Physics Publishing; 2016. p. R150–R166. doi: 10.1088/0031-9155/61/13/R150.

11. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiother Oncol. Elsevier Ireland Ltd; 2018;127(3):370–373. doi: 10.1016/j.radonc.2018.03.004.

12. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. Radiological Society of North America Inc.; 2020;295(2):328–338. doi: 10.1148/radiol.2020191145.

13. Zwanenburg A, Leger S, Vallières M, Löck S. The image biomarker standardisation initiative — IBSI 0.0.1dev documentation. 2019. https://ibsi.readthedocs.io/en/latest/#. Accessed April 26, 2021.

14. Zwanenburg A. GitHub - theibsi/data_sets: Data sets used by the IBSI for benchmarking and standardisation. . https://github.com/theibsi/data_sets. Accessed April 23, 2021.

15. Ashrafinia S. Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics. Baltimore, Maryland; 2019. https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/61551/ASHRAFINIA-DISSERTATION-2019.pdf?sequence=1. Accessed April 26, 2021.

16. Bettinelli A, Branchini M, De Monte F, Scaggion A, Paiusco M. Technical Note: An IBEX adaption toward image biomarker standardization. Med Phys. John Wiley and Sons Ltd; 2020;47(3):1167–1173. doi: 10.1002/mp.13956.

17. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. American Association for Cancer Research Inc.; 2017;77(21):e104–e107. doi: 10.1158/0008-5472.CAN-17-0339.

18. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. Wang Y, editor. PLoS One. Public Library of Science; 2019;14(2):e0212223. doi: 10.1371/journal.pone.0212223.

19. Lambin P. Radiomics Digital Phantom | CancerData.org. 2016. doi: 10.17195/candat.2016.08.1.

20. Loi S, Mori M, Benedetti G, et al. Robustness of CT radiomic features against image discretization and interpolation in characterizing pancreatic neuroendocrine neoplasms. Phys Medica. Associazione Italiana di Fisica Medica; 2020;76:125–133. doi: 10.1016/j.ejmp.2020.06.025.

21. Bewick V, Cheek L, Ball J. Statistics review 10: Further nonparametric methods. Crit. Care. BioMed Central; 2004. p. 196–199. doi: 10.1186/cc2857.

22. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur Radiol. Springer Science and Business Media Deutschland GmbH; 2020;30(11):6241–6250. doi: 10.1007/s00330-020-06957-9.

23. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG. Variation in algorithm implementation across radiomics software. J Med Imaging. SPIE-Intl Soc Optical Eng; 2018;5(04):044505. doi: 10.1117/1.jmi.5.4.044505.

24. McNitt-Gray M, Napel S, Jaggi A, et al. Standardization in quantitative imaging: A multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets. Tomography. Grapho Publications LLC; 2020;6(2):118–128. doi: 10.18383/j.tom.2019.00031.

25. Lei M, Varghese B, Hwang D, et al. Benchmarking features from different radiomics toolkits / toolboxes using Image Biomarkers Standardization Initiative. arXiv. 2020. https://arxiv.org/pdf/2006.12761.pdf. Accessed April 14, 2021.

26. Baeßler B, Weiss K, Santos DP Dos. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Invest Radiol. Lippincott Williams and Wilkins; 2019;54(4):221–228. doi: 10.1097/RLI.0000000000000530.

27. Liang ZG, Tan HQ, Zhang F, et al. Comparison of radiomics tools for image analyses and clinical prediction in nasopharyngeal carcinoma. Br J Radiol. British Institute of Radiology; 2019;92(1102). doi: 10.1259/bjr.20190271.

28. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. Nature Publishing Group; 2019;9(1):1–10. doi: 10.1038/s41598-018-36938-4.

29. Genetics Soph. SOPHiA AI Makes Data-Driven Medicine More Valuable by Combining Genomics and Radiomics to Fight Cancer. Mountain View, California; 2018. https://www.prnewswire.com/news-releases/sophia-ai-makes-data-driven-medicine-more-valuable-by-combining-genomics-and-radiomics-to-fight-cancer-670669663.html. Accessed April 26, 2021.

30. PyRadiomics from the Computational Imaging & Bioinformatics Lab - Harvard Medical School. (Online). 2017. https://www.radiomics.io/pyradiomics.html. Accessed November 22, 2021.

31. Avanzo M, Pirrone G, Vinante L, et al. Electron Density and Biologically Effective Dose (BED) Radiomics-Based Machine Learning Models to Predict Late Radiation-Induced Subcutaneous Fibrosis. Front Oncol. Frontiers Media S.A.; 2020;10:490. doi: 10.3389/fonc.2020.00490.

SUPPLEMENTAL TABLE

Feature family	Isotropic phantom	Anisotropic phantom
Total Number of ROI	90	90
Pixel spacing [mm]	1.00x1.00	0.977x0.977
Slice Thickness [mm]	1.00	3.00
Volumes [mm^3]		
Bean - Small	120	125.89
Bean - Medium	998	881.20
Bean - Large	7964	7896.43
Sphere - Small	123	120.16
Sphere - Medium	1020	975.61
Sphere - Large	8025	7990.85
Cube - Small	125	143.05
Cube - Medium	1000	858.31
Cube - Large	8000	8010.87

Table E1*. Characteristics of the 9 different ROI types defined for both the isotropic and the anisotropic phantom. The volumes reported in the Table correspond to the number of ROI voxels multiplied by voxel dimension.*

SUPPLEMENTAL MATERIAL

2D:avg – 2D:smrg aggregation discrepancies

When dealing with 'bean small' and 'sphere small' ROIs, match percentages for 2D:avg and 2D:smrg aggregations for GLCM features produced discordant results among the software applications (Figure S8).

2D GLCM features are calculated by aggregating information from four different directional matrices calculated over each slice, that is, along the (x, y) directions (1, 0), (1, 1), (-1, 1), and (0, 1).

It is worth recalling that: 1) for 2D:avg aggregation, GLCM features are computed from each 2D directional matrix and averaged over directions and slices; 2) for 2D:smrg, features are computed from a single matrix after merging the four 2D directional matrices per slice and then averaged over slices.

To explain the discrepancies obtained in the computation of 2D aggregation, we need to look at the top and bottom slices of the ROIs reported in Figure S8 (i.e., slices B1, B5 for 'bean small', and slices S1, S7 for 'sphere small'). It is not possible to calculate all four directional matrices on these slices. In the case of 'bean small' for 3 directions over 4, GLCM matrices cannot be defined as there are no adjacent voxels in the mask along those directions (Figure S9a), while for 'sphere small', no directional matrices can be calculated as there are no adjacent voxels in the mask for all four directions (Figure S9b).

The observed discrepancies across software for 2D aggregation methods are due to the different handling of these undefined matrices in the calculation of GLCM features.

SUPPLEMENTAL FIGURE LEGENDS (ONLINE)

Figure S1. Flowchart of the feature extraction performed in Phase II on the isotropic phantom. The Figure also shows the 919 radiomic features extracted with each program.

Figure S2. Percentages of "matching" (differences below the IBSI-reported tolerance, in green), "partial matching" (differences below three times the IBSI-reported tolerance, in yellow), "no matching" (otherwise, in red), and "missing" (cannot be calculated within a tool, in white) feature values obtained for each software package on the *IBSI radiomic phantom* in the five different parameter configurations (A, B, C, D, and E).

Figure S3. Cases of matching (green), partial matching (yellow), no matching (red), and missing (white) features for each software tool, feature family, and type of aggregation method on the *digital phantom*. For each program, only matching and partial matching features were maintained for the analysis on the isotropic and anisotropic phantom.

Figure S4. Effect of the aggregation method for the FBN discretization approach (FBN = fixed bin number).

Figure S5. Effect of the aggregation method for the FBS discretization approach (FBS = fixed bin size).

Figure S6. Effect of the ROI shape.

Figure S7. Effect of the ROI volume.

Figure S8. Slice per slice visualization of the masks for (a) 'bean small' and (b) 'sphere small' ROIs.

Figure S9. (a) '*Bean small*' top/bottom slices (B1, B5). Only one direction results in a definite directional matrix for the top/bottom slice of the 'bean small' mask (in green). For the other directions, no adjacent voxels are available to calculate the GLCM matrices. (b) '*Sphere small*' top/bottom slices (S1, S7). No voxels are available in all four directions to calculate the GLCM matrices for the top/bottom slice of the 'sphere small' mask.

CHAPTER 3

A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity

Published in: **Cancers** 2021 Jul; 13(15), 3835

A Multicentre Evaluation of Dosiomics Features Reproducibility, Stability and Sensitivity

Lorenzo Placidi, Eliana Gioscio, Cristina Garibaldi, Tiziana Rancati, Annarita Fanizzi, Davide Maestri, Raffaella Massafra, Enrico Menghi, Alfredo Mirandola, Giacomo Reggiori , Roberto Sghedoni, Pasquale Tamborra, Stefania Comi, Jacopo Lenkowicz, Luca Boldrini and Michele Avanzo

Simple Summary: Dosiomics is born directly as an extension of radiomics: it entails extracting features from the patients' three-dimensional (3D) radiotherapy dose distribution rather than from conventional medical images to obtain specific spatial and statistical information. Dosiomic studies, in a multicentre setting, require assessing the features' stability to dose calculation settings and the features' capability in distinguishing different dose distributions. This study provides the first multicentre evaluation of the dosiomic features in terms of reproducibility, stability and sensitivity across various dose distributions obtained from multiple technologies and techniques and considering different dose calculation algorithms of TPS and two different resolutions of the dose grid. Harmonisation strategies to account for a possible variation in the dose distribution due to these confounding factors should be adopted when investigating a correlation between dosiomic features and clinical outcomes in multicentre studies.

Abstract: Dosiomics is a texture analysis method to produce dose features that encode the spatial 3D distribution of radiotherapy dose. Dosiomic studies, in a multicentre setting, require assessing the features' stability to dose calculation settings and the features' capability in distinguishing different dose distributions. Dose distributions were generated by eight Italian centres on a shared image dataset acquired on a dedicated phantom. Treatment planning protocols, in terms of planning target volume coverage and dose–volume constraints to the organs at risk, were shared among the centres to produce comparable dose distributions for measuring reproducibility/stability and sensitivity of dosiomic features. In addition, coefficient of variation (CV) was employed to evaluate the dosiomic features' variation. We extracted 38,160 features from 30 different dose distributions from six regions of interest, grouped by four features' families. A selected group of features (CV < 3 for the reproducibility/stability studies, CV > 1 for the sensitivity studies) were identified to support future multicentre studies, assuring both stable features when dose distributions variation is minimal and sensitive features when dose distribution variations need to be clearly identified. Dosiomic is a promising tool that could support multicentre studies, especially for predictive models, and encode the spatial and statistical characteristics of the 3D dose distribution.

Keywords: dosiomics; dose distribution texture analysis; multicentric study; reproducibility; stability; sensitivity; radiation dosimetry; radiotherapy

1. INTRODUCTION

In the era of personalised medicine and targeted therapy, one of the most promising methods introduced in clinical practice is radiomics [1]. The key idea behind radiomics is that we can mine images by extracting image descriptors, called radiomic features, which can provide rich information about the tumour or healthy tissue and can be used to build predictive or prognostic models. This method allows quantitative analysis of different

Image modalities and identification of patterns and correlations among voxels that can be of interest for improving diagnosis, prognosis and prediction of treatment outcomes [2–4]. Clinical outcomes can be therefore predicted employing radiomics features, potentially changing the treatment paradigm. Nevertheless, several studies highlight the importance of providing robust and unbiased descriptors. Objective quantification of reproducibility, stability and redundancy of features is a prerequisite for radiomics. This kind of process has been performed widely in radiomics [5–12], and it is even more meaningful when performed in a multicentric setting [13,14]. Dosiomics is born directly as an extension of radiomics; it entails extracting features from the patients' three-dimensional (3D) radiotherapy dose distribution rather than from conventional medical images [15] [16] to obtain specific spatial and statistical information. Furthermore, it can parameterise the dose distribution in particular regions of interest (ROIs) by intensity, textural and shape-based features allowing the description of the dose distribution at a high complexity level, distinct from those obtained from dose–volume histograms (DVHs) [17]. Indeed, 3D dose distribution optimisation and evaluation are still mostly based on DVH endpoints, dose distribution visual inspection and DVH-based metrics. Nevertheless, the well-known drawback of DVH is to collapse the 3D dose information in 2D metric, losing the information on its spatial and statistical distribution. The integration of dosiomics with the DVH could constitute an advanced tool to evaluate the radiotherapy plan quality [18] by identifying new dose distribution metrics based on dosiomic features. A second appealing development is introducing dosiomic features into Tumour Control Probability (TCP) and Normal Tissue Complication Probability (NTCP) models, thus overcoming the current limitation of these models [19]. Some authors recently employed dosiomics to improve the prediction of side effects [20–22] or local control after radiotherapy [23], including preliminary multicentre experiences [24]. The proposed dosiomic signatures must be highly stable and reproducible and need validation before being used in clinical practice. Developing robust models requires ample training and validation datasets with radiotherapy data from many patients for any specific cancer site. These needs settle dosiomics in the framework of

"big data" and push towards multicentre studies. Possible sources of variation for radiomic features include different radiotherapy techniques, treatment planning systems (TPSs), dose calculation algorithms and dose grid resolutions. The variability due to these sources may hide any potential variability associated with the dose–response, making at least some of the dosiomic models unreliable and preventing the generalization of results. In this frame, Placidi et al. evaluated the robustness of dosiomic signatures across grid resolution and algorithm for dose calculation [25] in a monocentric setting. The results of that study highlighted the not negligible variation in dosiomic features, especially for target region and for dosiomic textural features; therefore, dosiomic studies should always provide a reporting of grid resolution and algorithm dose calculation. We here propose to investigate the stability of dosiomic features in a multicentre setting with two main aims: (a) to provide an assessment of the stability of dosiomic features to dose calculation settings and (b) to assess the dosiomic features capability in discriminating dose distributions that are generated with different radiation therapy devices. This study provides the first multicentre evaluation of the dosiomic features in terms of reproducibility, stability and sensitivity across various dose distributions obtained from multiple technologies and techniques and considering different dose calculation algorithms of TPS and two different resolutions of the dose grid.

2. MATERIALS AND METHODS

The evaluation of dosiomic features' extraction from different dose distributions was performed by several centres, which participate in the Dosiomics Team of the Radiomics Working Group of "Alliance Against Cancer" (Alleanza Contro il Cancro, ACC), a national oncology network founded in 2002 by the Italian Ministry of Health. Specifically, nine centres have contributed to this analysis. Dose distributions were generated by eight out of the nine centres involved in the study on an image dataset acquired on a dedicated phantom (see the specific section below) shared among the centres. Each of these centres could provide more than one dose distribution based on the availability of technologies and delivery techniques.

2.1. Phantom

A computed tomography (CT) scan of a cylindrical heterogeneous phantom was acquired for treatment planning and dose calculation. In particular, the ArcCheck [26] PMMA insert (ArcCHECK MR Sun Nuclear Melbourne Florida, US) was modified by substituting 4 PMMA rectangular sub-inserts with the following equivalent densities: lung, bone, muscle and adipose. The planning CT (GE. Optima CT580 W HiSpeed DX/I Spiral) had a slice thickness of 1.25 mm, 140 kV, pixel size of 1.269 mm2, as shown in **Figure 1**.

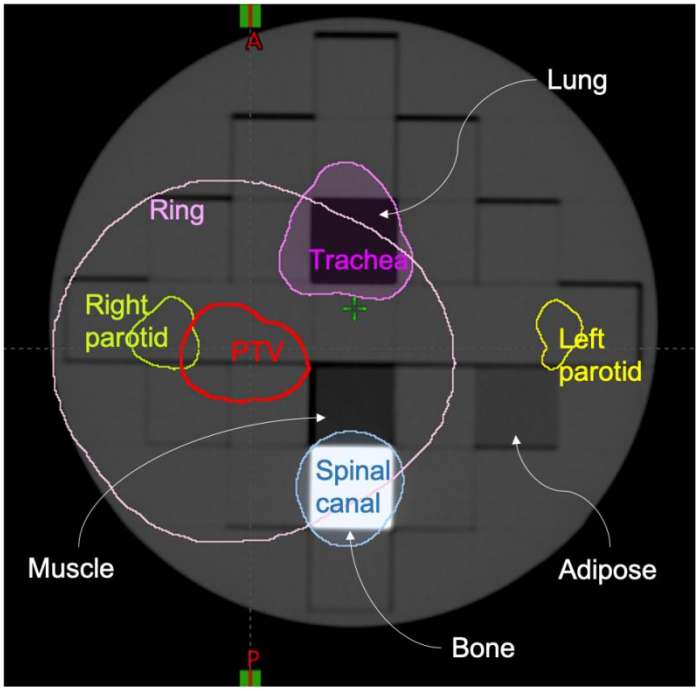


Figure 1. Planning CT of the phantom employed in the study and the countered Regions of Interest (ROIs). Six different ROIs were contoured: planning target volume (PTV), left parotid, right parotid, spinal canal, trachea and RING. Ring structure is the expansion of 3 cm from the PTV and cropped of 0.0 cm from the PTV edge.

On the acquired planning CT, the regions of interest (ROIs) simulating the tumour and organs at risk (OARs) of a head and neck radiotherapy treatment were contoured. These included a planning target volume (PTV), left parotid, right parotid, spinal canal, planning organ-at-risk volume for the spinal canal (PRV, i.e., isotropic 4 mm expansion of spinal canal) and trachea. Moreover, we added a RING structure defined as an expansion of 3 cm from the PTV and cropped of 0.0 cm from the PTV edge to ensure a high dose gradient outside the PTV. The planning CT and its ROIs were exported in DICOM format and shared among the centres. In terms of PTV coverage and dose–volume constraints to the OARs, two different planning protocols, including minimum

and maximum dose to the PTV and dose–volume constraints to the OARs, were followed by participants to produce comparable dose distributions, which were used to evaluate dosiomic features in terms of reproducibility, stability and sensitivity. Dose distribution computation was performed from eight different centres, named A, B, C, D, E, F, G, H.

2.2. Plan and Dose Prescription: Same Techniques, Technology and TPS

To evaluate the reproducibility and stability of dosiomic features, we planned a series of Intensity Modulated Radiation Therapy (IMRT) treatments with dose distributions as equivalent as possible, employing the same delivery technique, a unique dose distribution optimisation protocol and identical or similar LINACs. With reproducibility, we mean that a result obtained by an experiment should be achieved again with a high degree of agreement when the study is replicated with the same methodology by different researchers. A stable measure, on the other hand, is one in which the sources of variation are consistent over different inputs and conditions; here it is TPS and Technologies. This means that the process does not exhibit unpredictable variation for this purpose. We chose almost similar LINACs and the same photon energy, gantry angles, TPS and planning objectives. We considered only IMRT 6MV FF Varian machines (Trilogy, TrueBeam, TrueBeam Edge and Clinac) with the Eclipse-Aria TPS for this study phase. The normal tissue objective (NTO) tool was employed with the default setting, and dose grid resolution (optimisation and calculation) was set to 1 mm. **Table 1** reports the details of the IMRT protocol.

Table 1. Details of the IMRT protocol used to study the stability of the features when derived from almost equivalent plans: beam setup, dose prescription and constraints for organs at risk.

Beam settings	**Gantry Angles**	**Energy**	**Dose Rate**	**Collimator Angles**	**Dose Calculation Algorithm**	**Iteration**
	0°, 40°, 80°, 120°, 160°, 200°, 240°, 280°, 320°	6 MV FF	300 MU/min	15 for all the fields	AAA	At least 700
Planning objectives	Upper: Vol(%) = 0, Dose(Gy) = 68, Priority = 140			Lower: Vol(%) = 100, Dose(Gy) = 66, Priority = 140		
OARs constraints	**Trachea**	**Parotid L**	**Parotid R**	**PRV SC**	**Spinal canal**	**RING**
	D_{mean} = 49.5 Gy Priority = 80	D_{mean} = 5.0 Gy Priority = 50	D_{mean} = 23.0 Gy Priority = 100	D_{max} = 40.0 Gy Priority = 90	D_{max} = 62.0 Gy Priority = 110	D_{max} = 39.96 Gy Priority = 90

FF = flattening filter; **OARs** = Organs at Risk; **MU** = monitor units; **Dmean** = mean dose; **Dmax** = max dose; **AAA** = anisotropic analytical algorithm; **Vol** = volume; **Parotid L** = left parotid; **Parotid R** = right parotid; **PRV SC** = planning organ-at-risk volume for the spinal canal.

Eight IMRT dose distributions provided by different centres were included in the "stability" dataset. **Table 2** summarises the Varian (Varian Medical Systems) Linacs used, while all the other plan parameters, equal for all the centres, are shown in **Table 1**.

Table 2. List of centres that computed the IMRT dose distribution for the reproducibility and stability studies.

Centre_Plan	LINAC
G_1	CLinac
E_1	TrueBeam
E_2	Edge
B_1	TrueBeam
B_2	Edge
D_1	TrueBeam
D_2	Trilogy
A_1	TrueBeam

Figure 2 shows the eight dose distributions included in the studies. A reproducibility study (smaller green rectangle) was conducted on the dose distribution obtained by plans E_1, B_1, D_1 and A_1 (all TrueBeam Linac), while stability study (red rectangle) includes all eight dose distributions listed in **Table 2**.

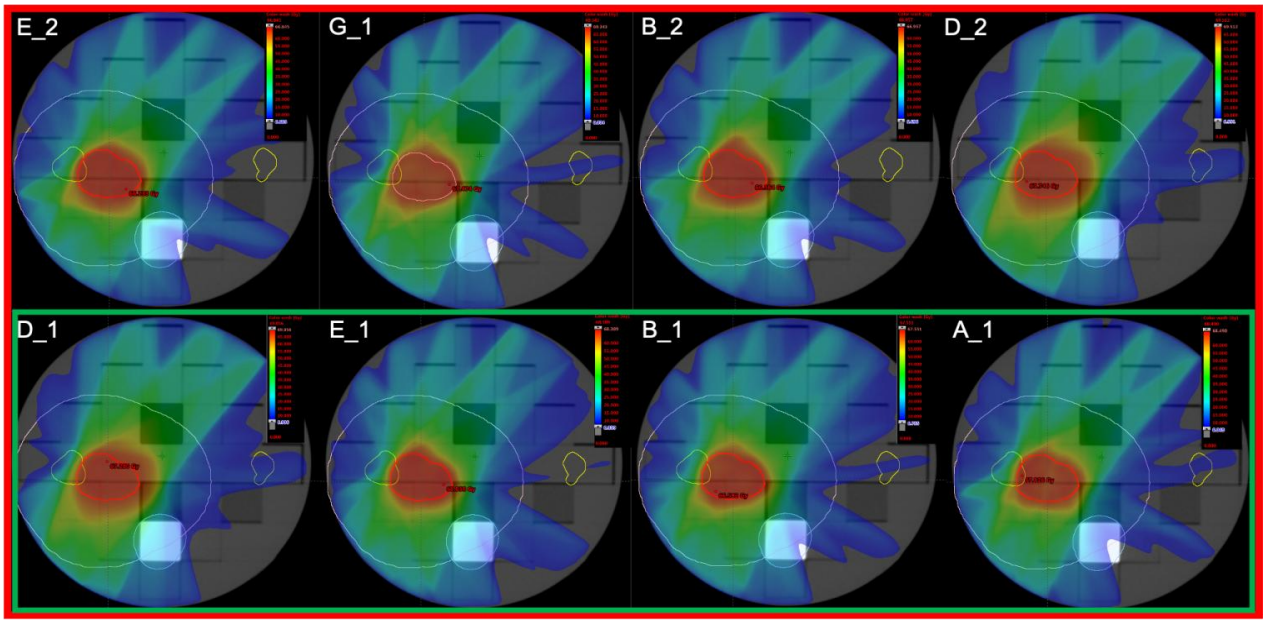


Figure 2. The four IMRT dose distributions included in the reproducibility study (within the green rectangle), and the eight IMRT dose distributions included in the stability study (within the red rectangle).

2.3. Plan and Dose Prescription: Different Techniques, Technologies and TPSs

To evaluate the sensitivity of the dosiomic features extraction to different techniques, technologies and TPSs, each centre planned one or more treatments using a range of different technologies among those available to the centres involved. Sensitivity is defined as the smallest absolute amount of change that can be detected by a measurement. The different delivery techniques, accelerators, TPS and dose calculation algorithms considered in this study are reported in **Table 3**.

Table 3. List of the eleven plans generated by the centres involved in the study, type of particle, beam energy, delivery technique, kind of Linac, treatment planning system and dose calculation algorithm

Centres_Plan	Particle	Energy (MV)	Technique	Accelerator Devices	TPS	Dose Calculation Algorithm
A_S1	photon	6 FF	VMAT	TrueBeam. Varian	Eclipse	AAA
B_S2	photon	6 FF	VMAT	TrueBeam. Varian	Eclipse	AAA
C_S3	proton	62.3–226.9 MeV/u	IMPT	Synchrotron (CNAO) [27]	RayStation	MC
D_S4	photon	6 FF	VMAT	TrueBeam. Varian	Eclipse	AAA
E_S5	photon	6 FF	VMAT	TrueBeam. Varian	Eclipse 15.6	Acuros
F_S6	photon	6 FF	VMAT	Synergy. Elekta	Pinnacle	CC
F_S7	photon	6 FF	VMAT	Synergy, Elekta	RayStation	CC
G_S8	photon	6 FF	VMAT	Clinac, Varian	Eclipse	AAA
H_S9	photon	6 FFF	TOMO	Tomotherapy, Accuray	Tomotherapy HT 2.1.6	CC
H_S10	photon	6 FF	DWA	Vero, Brainlab-Mitsubishi	Raystation 9B SP2	MC, CC
H_S11	photon	6 FF	VMAT	Trilogy, Varian	Eclipse 15.6	AAA

DWA = dynamic wave arc; **TPS** = treatment planning system. **FF** = flattening filter; **VMAT** = Volumetric Modulated Arc Therapy; **IMPT** = Intensity Modulated Proton Therapy; **TOMO** = Tomotherapy; **AAA** = anisotropic analytical algorithm; **MC** = MonteCarlo; **CC** = collapsed cone.

Eleven dose distributions provided by different centres were included in the dataset. Each dose distribution was calculated and optimised with two different dose grid resolutions: 1 mm and 2 mm, always keeping the dose to PTV and the OARs within prescription and constraints. No limitation was imposed in terms of beam setup and geometry. The dose prescription simulated a theoretical head and neck mono-lateral treatment plan with a prescribed dose to the PTV of 66 Gy and dose per fraction of 2.2 Gy. Dose prescription and OARs constraints are summarised in **Table 4**. **Figure 3** shows the resulting eleven dose distributions with 1 mm dose grid resolution.

Table 4. Dose prescription and constraints to organs at risk employed for the generation of the dose distributions for the sensitivity study.

ROIs	**Dose Prescription and Constraints**	
PTV	D98% > 95%	V105% < 10%
Spinal canal	D_{max} < 45 Gy	
PRV spinal canal	D_{max} < 45 Gy	
Trachea	D_{mean} < 50 Gy	
Parotids	D_{mean} < 25 Gy	
RING	D_{max} < 95% = 62.7 Gy	

PTV = Planning Target Volume; **PRV SC** = planning organ-at-risk volume for the spinal canal; **Dmean** = mean dose; **Dmax** = maximum dose; **D98%** = minimum dose to the 98% of the volume; **V105%** = percent of volume receiving at least 105% of the prescribed dose.

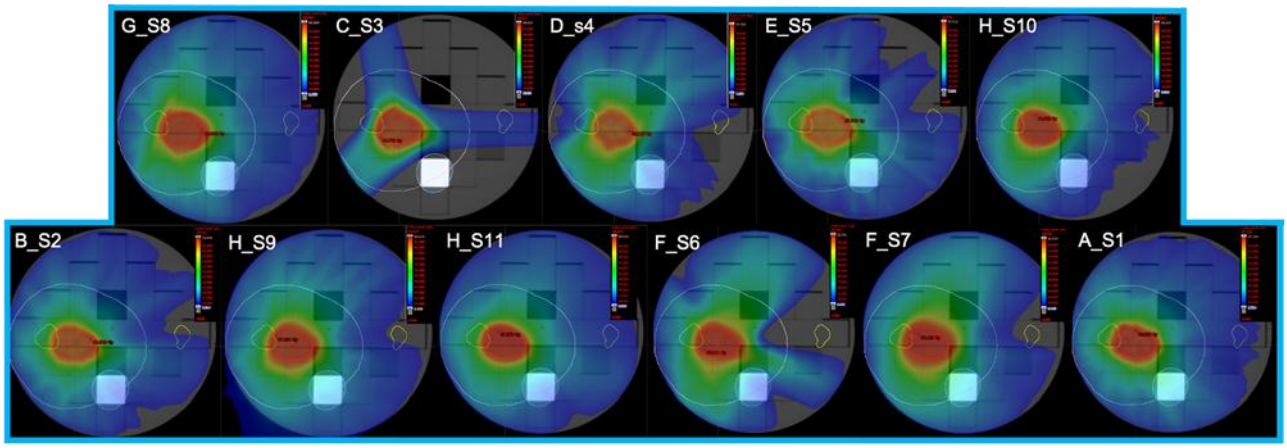


Figure 3. The eleven dose distributions obtained by different techniques, technologies and treatment planning systems, with 1 mm dose grid resolutions

2.4. Extraction of Dosiomic Features

The extraction of dosiomic features was centralised and carried out by a specific routine in the MODDICOM library, a free software package developed in R language optimised for automatic loading of DICOM images and radiomic analysis [28]. A specific routine for dose distribution texture analysis was realised for the purpose of this study, loading and processing the required DICOM dataset (planning CT, RT-Structure and RT-Dose). The features definition, nomenclature and extraction methodology following the one used for radiomic studies based on medical images, as accurately described by Zwanenburg et al. [29]. In dosiomics, the "image" is constituted by voxels with their grey level corresponding to the absolute dose in Gy. The absolute dose levels were binned in 100 discrete levels from zero to max dose before performing feature extraction. A total of 212 dosiomics features defined in the Image Biomarker Standardisation Initiative (IBSI) [29] were extracted from the selected ROIs (PTV, left parotid, right parotid, spinal canal, trachea and RING) belonging to the following families: 17 intensity-based statistics (STAT), 100 features from grey level co-occurrence matrix (GLCM), 63 from grey level run length matrix (GLRLM) and 32 from grey level size zone matrix (GLSZM). Morphological features were not included in this study since not considered relevant for dosiomic analysis.

2.5. Data Analysis

The analysis mirrored the two main goals of the study: to assess (a) the stability of dosiomic features to dose calculation settings and (b) the sensitivity to a change in dose distribution, that is, the ability of dosiomic features in distinguishing dose distributions generated with different radiation therapy devices.

As a preliminary test, we evaluated the software reproducibility for the computation of dosiomic features by extracting them in two different centres, both employing the MODDICOM library. We considered the complete set of 212 dosiomic features extracted from the same dose distribution (from centres A) computed with 1 mm and 2 mm calculation grid resolution and from all the contoured ROIs for this check. Differences in values for single features were then analysed. The expected differences are zero since dosiomic features extraction should not depend on the same software employed in different centres.

We used the coefficient of variation (CV) to evaluate the stability of the dosiomic features to dose calculation settings, i.e., when dosiomic features are derived from equivalent plans ("IMRT-Linac reproducibility"). The CV is a standardised measure of the dispersion of a distribution leading to the degree of intra-features variability. It is defined as the ratio of the standard deviation σ concerning the mean value μ (or to its absolute value |μ|). CV, to respect to standard deviation, is recommended when datasets with different units or widely different means were considered.

For assessing reproducibility, we computed the CV for the dosiomic features extracted by four IMRT dose distributions with the same technique, technology, TPS and Linac version (Varian, TrueBeam, Eclipse, AAA). This analysis investigates reproducibility as the features are extracted after the experiment (here, dose calculation) was replicated with the same methodology (here, same plan, same TPS, same LINAC) by different researchers (here, different centres). All ROIs were considered individually.

In the stability analysis, we still computed CV among the dosiomic features extracted by the entire set of eight IMRT dose distributions derived from the same technique, technology and TPS version. In this case, we considered different Linac Technologies (see Table 4). This analysis investigates stability as features are considered for their possible variation over different inputs and conditions (here, different TPSs and LINACs) to prove that the process does not exhibit unpredictable variation. Awareness and quantification of these variations should always be taken into account to avoid misinterpretation of results from studies, including dosiomic features.

The sensitivity of dosiomic features to dose distributions generated with different radiation therapy devices, TPSs and algorithms was also evaluated in terms of CV (see **Table 5**). In this case, CV describes how a dosiomic feature can change due to different dose distributions due to different techniques, technologies, TPSs, dose calculation algorithms, energies and beam quality.

Table 5. List of the percentage of common dosiomic features among different studies and different CV thresholds.

ROIs	Repr. (CV_{TH} <0.3) ∩ Stab. (CV_{TH} <0.3)	Sens. 1 mm (CV_{TH} >1) ∩ Sens. 2 mm (CV_{TH} >1)	Stab. (TH < 0.3) ∩ Sens. 1 mm (TH >1)	Stab. (CV_{TH} < 0.3) ∩ Sens. 2 mm (CV_{TH} >1)
PTV	63.2%	5.7%	9.4%	1.9%
Left parotid	49.5%	14.2%	2.4%	1.4%
Right parotid	46.2%	5.7%	9.0%	2.8%
Spinal canal	68.9%	12.7%	3.3%	2.8%
Trachea	58.5%	16.5%	2.8%	2.8%
RING	82.5%	1.4%	0.0%	0.5%

Repr. = Reproducibility (green), **Stab.** = Stability (red), **Sens. 1 mm**= Sensitivity 1 mm (light blue), **Sens. 2 mm**= Sensitivity 2 mm (dark blue).

We adopted a common guideline of thresholding CV value as a strategy to select stable, reproducible and sensitive features. With a view to future multicentric studies concerning tumour control and/or OARs toxicity, it is crucial to select dosiomic features with a small CV, for example, with a CV < 0.3, when dose distributions are expected to be stable and reproducible. Simultaneously, it would also be desirable to identify dosiomic features able to recognise true differences in the dose distributions, so with a large CV, e.g., with a CV > 1, which classifies the sensitivity of the dosiomic features to the dose distribution variation. Since the proposed threshold values are a completely arbitrary choice, the CV > 0.8 threshold was also employed to investigate further and evaluate the variation in the dosiomic feature's sensitivity on the selected threshold. Dosiomic features that are both stable (CV < 0.3) and sensitive (CV > 1 or CV > 0.8), i.e., that constitute an optimum set for modelling purposes, were described through Venn diagrams.

3. RESULTS

We extracted a total amount of 38,160 dosiomic features from 30 different dose distributions from six ROIs, grouped by four features' families. In terms of reproducibility of dosiomic features extraction using the same software, two centres extracted 212 dosiomic features for each calculation grid resolution size (1 mm and 2 mm), resulting in a comparison of 424 features between the centres. Single feature value differences, both for 1 mm and 2 mm

calculation grid resolutions, were found to be equal to zero for all the considered dosiomic features. This result confirms the reproducibility of the dosiomic features when extracted using the MODDICOM software package (version 0.52).

We evaluated 5088 and 10,176 dosiomic features to assess the reproducibility and stability of the extracted dosiomic features, respectively. Tables S1 and S2 show the CV values grouped by ROIs and features' family for reproducibility and stability studies, respectively.

Concerning the sensitivity study, we extracted 27,984 dosiomic features from the entire set of 11 dose distributions and six ROIs. Tables S3 and S4 depict the CV values grouped by ROIs and feature' family, respectively, for the 1 mm and 2 mm dose grid calculation sensitivity studies.

Results are also summarised in **Figures 4** and **5** in terms of box plots for the left parotid and PTV, respectively. The black horizontal line within the box display for each box the median CV value. All the other ROIs (right parotid, spinal canal, trachea and RING) are reported in Figures S1–S4 in the Supplementary Materials. Additionally, Table S5 lists the mean CV values for all the studies, ROIs and family's features.

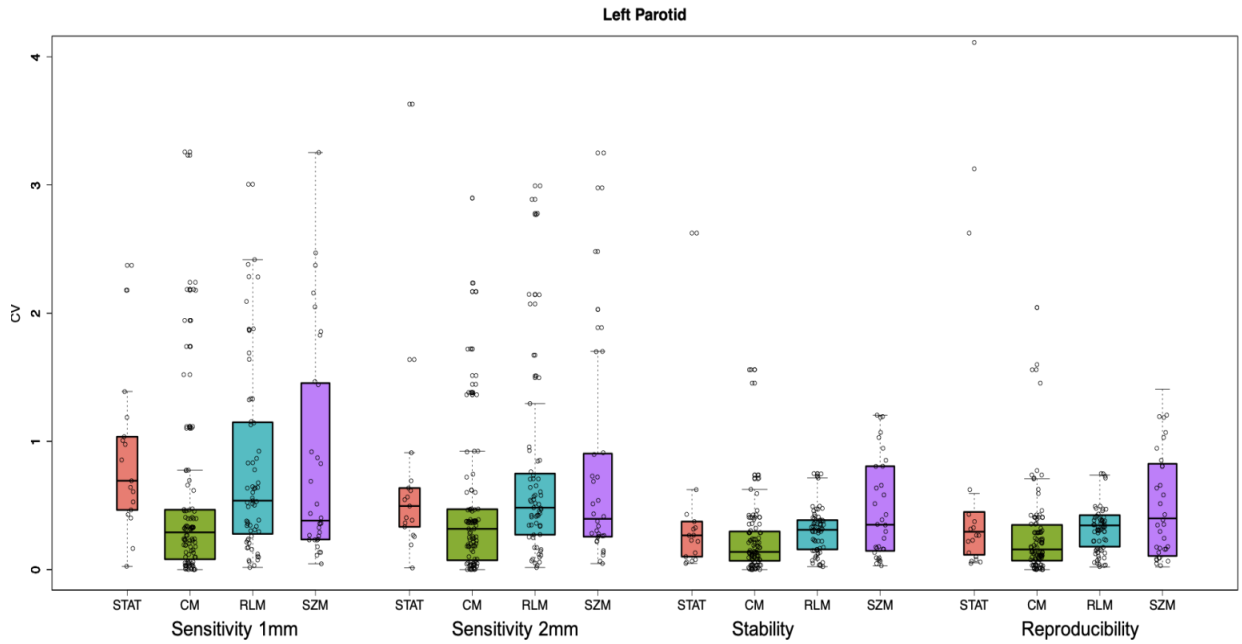


Figure 4. Box plots of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies, grouped for the four different features' families (STAT, CM, RLM and GSZ) for the left parotid.

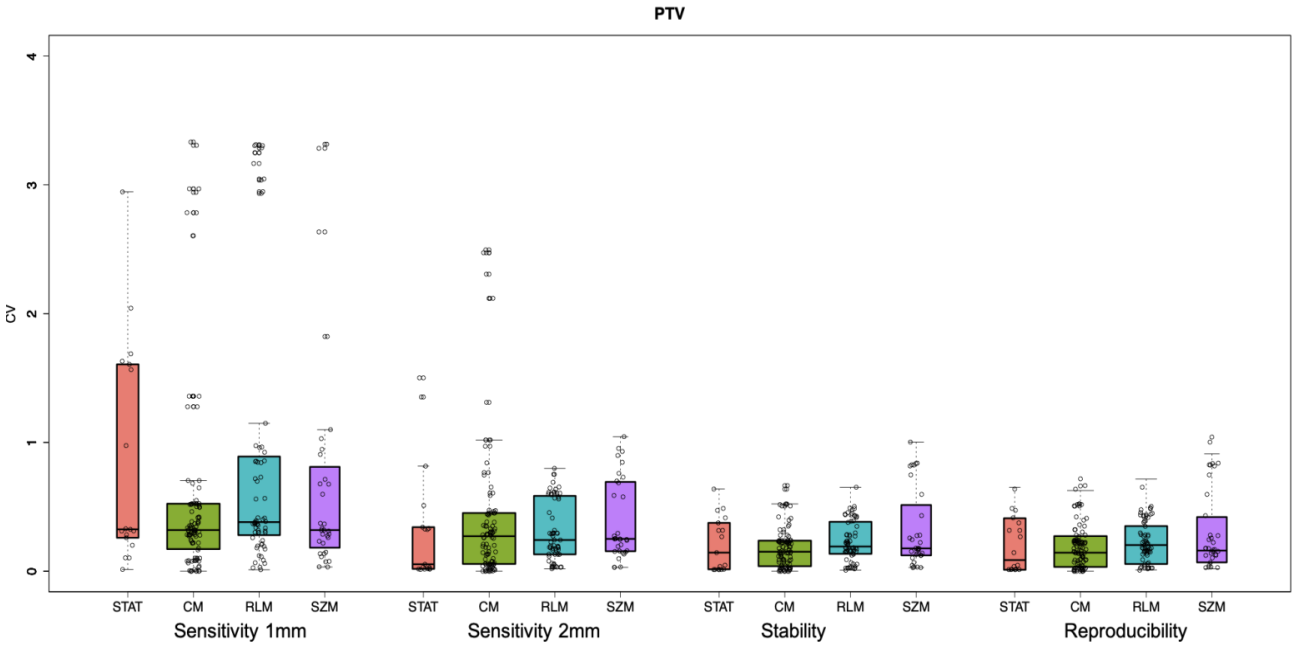


Figure 5. Box plots of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies grouped for the four different features' families (STAT, CM, RLM and GSZ) for the PTV.

The Venn diagrams in **Figure 6** highlight the features that are both stable and sensitive after the choice of specific CV threshold (CVTH) values. For example, results for the set of stable AND sensitive features for the ROIs PTV and left parotid are given in Figure 3 for two different CVTH values for the sensitivity, CV > 1 and CV > 0.8, while the CV threshold for stability is kept to 0.3.

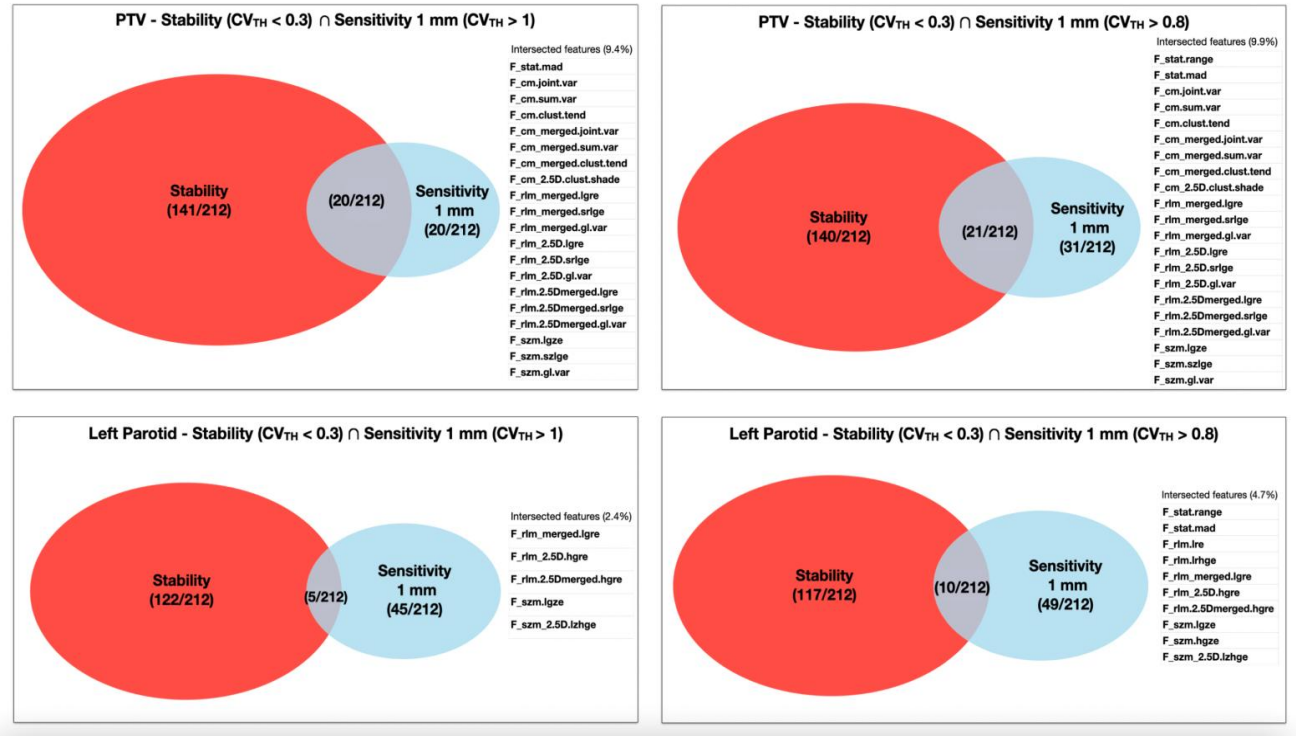


Figure 6. Venn diagrams showing the dosiomic features that are both stable and sensitive between the stability and sensitivity study for the PTV and left parotid. The sensitivity CV threshold was set to two different values: CVTH > 1 and CVTH > 0.8 for the PTV.

The Supplementary Materials (Tables S6–S10) report Venn diagrams highlighting dosiomic features that are both stable and sensitive for the other ROIs considered in this analysis. Table 5 summarises the percentage of features that are both stable and sensitive (CV threshold = 1 and different resolution of the grid for dose calculation, 1 mm vs. 2 mm) across all the ROIs, and the details are shown in Tables S6–S11 in the Supplementary Materials.

4. DISCUSSION

Dosiomic is increasingly used in clinical studies aiming to improve the prediction of clinical outcomes, e.g., locoregional recurrence after IMRT for head and neck cancer [23] or local control after carbon-ion radiotherapy in skull-base chordoma [21]. Dosiomic features were analysed by machine learning for the prediction of acute-phase weight loss in lung cancer patients treated with radiotherapy [30]. Among preliminary multicentre experiences, Adachi et al. [24] aimed at predicting radiation pneumonitis after lung stereotactic body radiation therapy using dosiomics. In both single and, especially, multicentre studies, consistent reporting of dose distribution to provide a robust setting for the study is a key point to ensure stronger validation of the use of dosiomic features in the clinical routine.

The presented study provided the first assessment of the variation in dosiomic features to dose calculation environments in a multicentre setting and the sensitivity of dosiomic features in distinguishing dose distributions generated with different radiation therapy devices.

If considered reproducibility and stability studies, dosiomic features' families with higher mean CV are always SZM apart from the RING and left parotid ROIs where STAT family shows the higher CV mean value. In terms of capability in describing and evaluating 3D dose distribution to a higher level than DVHs and employing dosiomic features in predictive modelling, CV mean values could not provide any useful information. Nevertheless, this study can describe a peculiar behaviour of the single dosiomic features (listed in Tables S1–S4) and families that could be representative for further studies.

The box plots in **Figures 1, 2** and Figures S1–S4, highlight how dosiomic features depend on dose distribution, ROIs and feature families. As expected, larger CV variations were observed in the sensitivity studies (with a dose calculation grid of 1 mm and 2 mm). It is difficult to generalise these results due to the dose distribution dependency, but, in terms of ROIs, it is visible that ROIs

that lay in the gradient region (RING and right parotid) show lower CV values on average (Table S5). Concerning the dosiomic features' families, GLCM has almost always the lowest value for all the studies and all the ROIs except for the right parotid in the sensitivity study using a dose calculation grid of 1 mm, PTV and spinal canal in the sensitivity study using a dose calculation grid of 2 mm. Even though the mean value of a single feature is considered, these results highlight how the GLCM features families show the lowest variation in terms of CV. Dosiomic features' families with higher mean CVs in the sensitivity studies (both with1 dose calculation grid of 1 mm and 2 mm) are RLM and SZM in the 87.5% of the cases for the RING, left parotid, right parotid and PTV ROIs, while STAT for spinal canal and trachea.

Of note, reproducibility and stability of features can also be evaluated in terms of intraclass correlation coefficient (ICC), as is customary in most studies in which radiomic and dosiomic characteristics are evaluated. ICC is defined as the ratio of the subject variance by the sum of the subject variance, the rater variance and the residual, where a lower rater variance implies a reliable scale. ICC expresses how strongly the components in the same group resemble each other [10]. The peculiar analysis presented in this study forces the TPSs/Linacs/RT techniques/RT Technologies to be the "raters" of the dose distributions, while the different ROIs would be the "subjects". Nevertheless, the ROIs present with very different dose distributions (high doses vs. low doses, almost uniform dose distribution vs. high gradient dose distribution), which means high variation between-subjects (possibly larger than variation among raters) that could lead to biases in ICC calculations. For these reasons, we chose to stick to the coefficient of variation that does not require an evaluation across different "subjects" and only needs evaluation across different "raters". Nevertheless, an example of ICC evaluation is reported in the Supplementary Materials (Figure S5) considering the STAT dosiomic features' family and the three possible ROIs groups: all ROIs, high dose region ROIs and low dose region ROIs.

The main aim of this study was to provide suggestions on a set of dosiomic features that are at the same time reproducible, stable and sensitive, i.e., robust across variations that are not related to true differences in the dose distribution and able to pick up even subtle true differences in the dose distributions. To achieve this result, a specific guideline of thresholding CV values was defined to filter out reproducible, stable and sensitive features.

The choice of CV thresholds, i.e., CV < 0.3 to define reproducibility and stability and CV > 1, or CV > 0.8, to define sensitivity, is somehow arbitrary. To date, there are no specific and shared reference

threshold values; our choice was driven by some statistical considerations. A CV < 0.3 means that the standard deviation of the value distribution for the single feature is less than 30% of the mean value of the same distribution, which means a reasonably low variation across the distribution, with 68% of values in the interval "mean value ±30%". A CV > 1 (or > 0.8) means that the standard deviation of the value distribution for the single feature is (almost) the same "size" as the mean value, which entails a high possibility that values sampled from such a distribution can be identified as significantly different after a statistical test, which would be desired in outcome modelling.

This approach allowed identification of dosiomic features that are both stable and sensitive, as depicted in **Figure 3** for PTV and left parotid, summarised in **Table 1** for all the ROIs and dosiomic features' families and detailed in Tables S6–S11 for all the dosiomic features. As an example, identifying features that overlap between sensitivity study with 1 mm and 2 mm dose calculation grid is potential information that could be useful retrospective multicentre studies where different dose grid resolutions were employed, or for prospective studies to evaluate the possible need of guidelines on the dose calculation grid settings.

We would like to emphasise once more that the CV threshold values selected to filter out and define reproducible, stable and sensitive dosiomic features are not absolute suggested values to take as completely "a priori" reference in future dosiomics studies. Our choice was grounded on some statistical considerations, and results are possibly associated with the peculiar nature of our study, i.e., a phantom study, fixed centralised contouring of ROIs, common dose calculation protocol including fixed-dose prescription, planning objectives and OARs constraints. Other thresholds on CVs could be selected for other studies considering different features distributions, e.g., a possible clear bimodal distribution which derives from "merging" of two separated distributions for patients with/without a selected clinical outcome.

The employment of dosiomic in clinical practice could represent a powerful tool to handle better the 3D dose spatial and statistical information if compared with conventional tools, such as DVH and DVH metrics. Potentially, the granularity and quantity of the information provided by the dosiomic features, and above all the usability of such information, could better support the clinical decision than standard parameters, such as DVH, DVH metrics and visually assessment of the 3D dose distribution. Obviously, what is still needed is a clinical translation of the meaning of each dosiomic feature both within the use of full 3D dose distribution as a new metric to better assess

plan quality during the optimisation phase as well as during the plan evaluation, but also to finally include dosiomics in the predictive models. Dosiomic features could represent additional parameters to be employed in the predictive models: this could lead to identifying some disomic features that, both during the plan optimisation and evaluation, could be considered to prevent or limit acute toxicities, as well as to improve local control.

Additionally, to exploit the full benefits of big data, machine and deep learning, multicentre trials are needed [31]. Multicentre studies (both retrospective and prospective) are strongly based on the quality of the selected parameters. How do we best use the parameters we have been using so far? Are there any other parameters that could support future studies? Dosiomics features could be one of these, being potentially much more sensitive to dose distribution variation. Moreover, in a multicentre trial, a priori selection of the optimal dosiomic features to be employed in the study would lead to more robust and unbiased studies. According to the present analysis, it is essential to highlight how even just the variation in different dosiomic features underlines a possible use of the dosiomics to select dose distribution within multicentric studies to avoid bias during the further clinical outcome correlation analysis.

The first limitation of the present study is related to the pool of the considered radiotherapy techniques and technologies. They are pretty diverse and representative but do not describe all the possible techniques and technologies available in clinical practice. Despite this, we believe that the employed number of radiotherapy techniques and technologies used by the eight centres are enough to support the message that a substantial number of dosiomic features are stable, and at the same time, they can distinguish or recognise dose distributions generated with different radiation therapy devices.

A second possible limitation is related to the number of considered features. We considered 212 dosiomic features, other dosiomic features of the second-order could indeed have been considered. Nevertheless, the selected features represent a robust dataset, internationally validated in the radiomics setting [29], available to proceed in the clinical implementation of the dosiomics, both in clinical outcome predictive models and in the 3D dose distribution description, optimisation and evaluation processes.

As a further study, dosiomic feature extraction analysis on different software [32] should also be considered to evaluate the possibility of employing different software to extract dosiomic features in multi-institutional studies.

5. CONCLUSIONS

The present study has assessed the stability of dosiomic features and their capability in distinguishing dose distributions generated with different radiation therapy devices in a multicentre setting. These results suggest that being dosiomic features sensitive to changes in dose calculation parameters, a consistent reporting of the TPS, dose calculation algorithms and pixel spacing used to calculate dose distributions is required. Harmonisation strategies to account for a possible variation in the dose distribution due to these confounding factors should be adopted when investigating a correlation between dosiomic features and clinical outcomes in multicentre studies.

REFERENCES

1. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur. J. Cancer* **2012**, *48*, 441–446, doi:10.1016/j.ejca.2011.11.036.

2. Cusumano, D.; Boldrini, L.; Yadav, P.; Yu, G.; Musurunu, B.; Chiloiro, G.; Piras, A.; Lenkowicz, J.; Placidi, L.; Romano, A.; et al. Delta Radiomics for Rectal Cancer Response Prediction Using Low Field Magnetic Resonance Guided Radiotherapy: An External Validation. *Phys. Med.* **2021**, *84*, 186–191, doi:10.1016/j.ejmp.2021.03.038.

3. Fave, X.; Zhang, L.; Yang, J.; Mackin, D.; Balter, P.; Gomez, D.; Followill, D.; Jones, A.K.; Stingo, F.; Liao, Z.; et al. Delta-Radiomics Features for the Prediction of Patient Outcomes in Non-Small Cell Lung Cancer. *Sci. Rep.* **2017**, *7*, 588, doi:10.1038/s41598-017-00665-z.

4. Comes, M.C.; La Forgia, D.; Didonna, V.; Fanizzi, A.; Giotta, F.; Latorre, A.; Martinelli, E.; Mencattini, A.; Paradiso, A.V.; Tamborra, P.; et al. Early Prediction of Breast Cancer Recurrence for Patients Treated with Neoadjuvant Chemotherapy: A Transfer Learning Approach on DCE-MRIs. *Cancers* **2021**, *13*, 2298, doi:10.3390/cancers13102298.

5. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajewski, B.J.; Hróbjartsson, A.; Roberts, C.; Shoukri, M.; Streiner, D.L. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) Were Proposed. *J. Clin. Epidemiol.* **2011**, *64*, 96–106, doi:10.1016/j.jclinepi.2010.03.002.

6. Bailly, C.; Bodet-Milin, C.; Couespel, S.; Necib, H.; Kraeber-Bodéré, F.; Ansquer, C.; Carlier, T. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. *PLoS ONE* **2016**, *11*, e0159984, doi:10.1371/journal.pone.0159984.
7. Zwanenburg, A. Radiomics in Nuclear Medicine: Robustness, Reproducibility, Standardization, and How to Avoid Data Analysis Traps and Replication Crisis. *Eur. J. Nucl. Med. Mol. Imaging.* **2019**, *46*, 2638–2655, doi:10.1007/s00259-019-04391-8.

8. Parmar, C.; Rios Velazquez, E.; Leijenaar, R.; Jermoumi, M.; Carvalho, S.; Mak, R.H.; Mitra, S.; Shankar, B.U.; Kikinis, R.; Haibe-Kains, B.; et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. *PLoS ONE* **2014**, *9*, e102107, doi:10.1371/journal.pone.0102107.

9. Zhao, B.; Tan, Y.; Tsai, W.-Y.; Qi, J.; Xie, C.; Lu, L.; Schwartz, L.H. Reproducibility of Radiomics for Deciphering Tumor Phenotype with Imaging. *Sci. Rep.* **2016**, *6*, 23428, doi:10.1038/srep23428.

10. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *102*, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.

11. Massafra, R.; Bove, S.; Lorusso, V.; Biafora, A.; Comes, M.C.; Didonna, V.; Diotaiuti, S.; Fanizzi, A.; Nardone, A.; Nolasco, A.; et al. Radiomic Feature Reduction Approach to Predict Breast Cancer by Contrast-Enhanced Spectral Mammography Images. *Diagnostics* **2021**, *11*, 684, doi:10.3390/diagnostics11040684.

12. La Forgia, D.; Fanizzi, A.; Campobasso, F.; Bellotti, R.; Didonna, V.; Lorusso, V.; Moschetta, M.; Massafra, R.; Tamborra, P.; Tangaro, S.; et al. Radiomic Analysis in Contrast-Enhanced Spectral Mammography for Predicting Breast Cancer Histological Outcome. *Diagnostics* **2020**, *10*, 708, doi:10.3390/diagnostics10090708.

13. Defeudis, A.; De Mattia, C.; Rizzetto, F.; Calderoni, F.; Mazzetti, S.; Torresin, A.; Vanzulli, A.; Regge, D.; Giannini, V. Standardization of CT Radiomics Features for Multi-Center Analysis: Impact of Software Settings and Parameters. *Phys. Med. Biol.* **2020**, *65*, 195012, doi:10.1088/1361-6560/ab9f61.

14. Kalendralis, P.; Traverso, A.; Shi, Z.; Zhovannik, I.; Monshouwer, R.; Starmans, M.P.A.; Klein, S.; Pfaehler, E.; Boellaard, R.; Dekker, A.; et al. Multicenter CT Phantoms Public Dataset for Radiomics Reproducibility Tests. *Med. Phys.* **2019**, *46*, 1512–1518, doi:10.1002/mp.13385.

15. Avanzo, M.; Pirrone, G.; Vinante, L.; Caroli, A.; Stancanello, J.; Drigo, A.; Massarut, S.; Mileto, M.; Urbani, M.; Trovo, M.; et al. Electron Density and Biologically Effective Dose (BED) Radiomics-Based Machine Learning Models to Predict Late Radiation-Induced Subcutaneous Fibrosis. *Front. Oncol.* **2020**, *10*, 490, doi:10.3389/fonc.2020.00490.

16. Avanzo, M.; Stancanello, J.; Pirrone, G.; Sartor, G. Radiomics and Deep Learning in Lung Cancer. *Strahlenther. Onkol.* **2020**, *196*, 879–887, doi:10.1007/s00066-020-01625-9.

17. Drzymala, R.E.; Mohan, R.; Brewster, L.; Chu, J.; Goitein, M.; Harms, W.; Urie, M. Dose-Volume Histograms. *Int. J. Radiat. Oncol. Biol. Phys.* **1991**, *21*, 71–78, doi:10.1016/0360-3016(91)90168-4.

18. Hernandez, V.; Hansen, C.R.; Widesott, L.; Bäck, A.; Canters, R.; Fusella, M.; Götstedt, J.; Jurado-Bruggeman, D.; Mukumoto, N.; Kaplan, L.P.; et al. What Is Plan Quality in Radiotherapy? The Importance of Evaluating Dose Metrics, Complexity, and Robustness of Treatment Plans. *Radiother. Oncol*. **2020**, *153*, 26–33, doi:10.1016/j.radonc.2020.09.038.

19. Gabryś, H.S.; Buettner, F.; Sterzing, F.; Hauswald, H.; Bangert, M. Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. *Front. Oncol.* **2018**, *8*, 35, doi:10.3389/fonc.2018.00035.

20. Rossi, L.; Bijman, R.; Schillemans, W.; Aluwini, S.; Cavedon, C.; Witte, M.; Incrocci, L.; Heijmen, B. Texture Analysis of 3D Dose Distributions for Predictive Modelling of Toxicity Rates in Radiotherapy. *Radiother. Oncol.* **2018**, *129*, 548–553, doi:10.1016/j.radonc.2018.07.027.

21. Buizza, G.; Paganelli, C.; D'Ippolito, E.; Fontana, G.; Molinelli, S.; Preda, L.; Riva, G.; Iannalfi, A.; Valvo, F.; Orlandi, E.; et al. Radiomics and Dosiomics for Predicting Local Control after Carbon-Ion Radiotherapy in Skull-Base Chordoma. *Cancers* **2021**, *13*, 339, doi:10.3390/cancers13020339.

22. Liang, B.; Tian, Y.; Chen, X.; Yan, H.; Yan, L.; Zhang, T.; Zhou, Z.; Wang, L.; Dai, J. Prediction of Radiation Pneumonitis With Dose Distribution: A Convolutional Neural Network (CNN) Based Model. *Front. Oncol.* **2019**, *9*, 1500, doi:10.3389/fonc.2019.01500.

23. Wu, A.; Li, Y.; Qi, M.; Lu, X.; Jia, Q.; Guo, F.; Dai, Z.; Liu, Y.; Chen, C.; Zhou, L.; et al. Dosiomics Improves Prediction of Locoregional Recurrence for Intensity Modulated Radiotherapy Treated Head and Neck Cancer Cases. *Oral. Oncol.* **2020**, *104*, 104625, doi:10.1016/j.oraloncology.2020.104625.

24. Adachi, T.; Nakamura, M.; Shintani, T.; Mitsuyoshi, T.; Kakino, R.; Ogata, T.; Ono, T.; Tanabe, H.; Kokubo, M.; Sakamoto, T.; et al. Multi-Institutional Dose-Segmented Dosiomic Analysis for Predicting Radiation Pneumonitis after Lung Stereotactic Body Radiation Therapy. *Med. Phys.* **2021**, *48*, 1781–1791, doi:10.1002/mp.14769.

25. Placidi, L.; Lenkowicz, J.; Cusumano, D.; Boldrini, L.; Dinapoli, N.; Valentini, V. Stability of Dosomics Features Extraction on Grid Resolution and Algorithm for Radiotherapy Dose Calculation. *Phys. Med.* **2020**, *77*, 30–35, doi:10.1016/j.ejmp.2020.07.022.

26. Li, G.; Zhang, Y.; Jiang, X.; Bai, S.; Peng, G.; Wu, K.; Jiang, Q. Evaluation of the ArcCHECK QA System for IMRT and VMAT Verification. *Phys. Med.* **2013**, *29*, 295–303, doi:10.1016/j.ejmp.2012.04.005.

27. Rossi, S. The National Centre for Oncological Hadrontherapy (CNAO): Status and Perspectives. *Phys. Med.* **2015**, *31*, 333–351, doi:10.1016/j.ejmp.2015.03.001.

28. Gatta, R.; Vallati, M.; Dinapoli, N.; Masciocchi, C.; Lenkowicz, J.; Cusumano, D.; Casá, C.; Farchione, A.; Damiani, A.; van Soest, J.; et al. Towards a Modular Decision Support System for Radiomics: A Case Study on Rectal Cancer. *Artif. Intell. Med.* **2018**, *96*, 145–153, doi:10.1016/j.artmed.2018.09.003.

29. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* **2020**, *295*, 328–338, doi:10.1148/radiol.2020191145.

30. Lee, S.H.; Han, P.; Hales, R.K.; Voong, K.R.; Noro, K.; Sugiyama, S.; Haller, J.W.; McNutt, T.R.; Lee, J. Multi-View Radiomics and Dosiomics Analysis with Machine Learning for Predicting Acute-Phase Weight Loss in Lung Cancer Patients Treated with Radiotherapy. *Phys. Med. Biol.* **2020**, *65*, 195015, doi:10.1088/1361-6560/ab8531.

31. Dong, D.; Fang, M.-J.; Tang, L.; Shan, X.-H.; Gao, J.-B.; Giganti, F.; Wang, R.-P.; Chen, X.; Wang, X.-X.; Palumbo, D.; et al. Deep Learning Radiomic Nomogram Can Predict the Number of Lymph Node Metastasis in Locally Advanced Gastric Cancer: An International Multicenter Study. *Ann. Oncol.* **2020**, *31*, 912–920, doi:10.1016/j.annonc.2020.04.003.

32. Foy, J.J.; Armato, S.G.; Al-Hallaq, H.A. Effects of Variability in Radiomics Software Packages on Classifying Patients with Radiation Pneumonitis. *J. Med. Imaging.* **2020**, *7*, 014504, doi:10.1117/1.JMI.7.1.014504.

Supplementary Materials: The following are available online at www.mdpi.com/article/10.3390/cancers13153835/s1. Figure S1: Box plot of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies for the ROI right parotid. Figure S2: Box plot of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies for the ROI spinal canal. Figure S3: Box plot of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies for the ROI trachea. Figure S4: Box plot of the CV values for the sensitivity (1 mm and 2 mm), stability and reproducibility studies for the ROI RING. Figure S5: Example of ICC results for the STAT dosiomic features' family and considering the three possible ROIs groups. Table S1: Reproducibility CV values for all the dosiomic features employed in the study and for all the six ROIs. Table S2: Stability CV values for all the dosiomic features employed in the study and for all the six ROIs. Table S3: Sensitivity (1 mm dose calculation grid) CV values for all the dosiomic features employed in the study and for all the six ROIs. Table S4: Sensitivity (2 mm dose calculation grid) CV values for all the dosiomic features employed in the study and for all the six ROIs. Table S5: mean values of the different CV for all the studies, ROIs and family's features. Table S6: common dosiomic features among the studies and different thresholds for the ROI RING. Table S7: common dosiomic features among the studies and different thresholds for the ROI left parotid. Table S8: common dosiomic features among the studies and different thresholds for the ROI right parotid. Table S9: common dosiomic features among the studies and different thresholds for the ROI PTV. Table S10: common dosiomic features among the studies and different thresholds for the ROI spinal canal. Table S11: common dosiomic features among the studies and different thresholds for the ROI Trachea.

Author Contributions: Conceptualisation and methodology: L.P., E.G., C.G., T.R. and M.A.; software: L.P., E.G. and J.L.; validation and formal analysis: L.P. and E.G.; investigation and data curation: L.P., E.G., C.G., T.R., A.F., S.C.; J.L., D.M., R.M., E.M., A.M., G.R., R.S., P.T., L.B. and M.A.; writing—original draft preparation: L.P. and E.G.; writing—review and editing: L.P., E.G., C.G., T.R., A.F., J.L., D.M., R.M., E.M., A.M., G.R., R.S., P.T., S.C., L.B. and M.A; funding acquisition: L.B.; project administration: M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Italian Ministry of Health through the network project RCR-2019-23669120_001 of the "Alleanza Contro il Cancro (ACC)" network. EG is funded by "ERA-NET ERA PerMed/FRRB grant agreement No ERAPERMED2018-244".

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reasons, according to GDPR.

Acknowledgments: The authors acknowledge the support from the Radiomics Group of “Alleanza Contro il Cancro” and the Italian Ministry of Health.

Conflicts of Interest: The authors declare no conflict of interest.

CHAPTER 4

The Potential role of MR-based radiomic biomarkers in the characterization of focal testicular lesions

Published in: **Scientific Reports** 2021 Feb 10; 11(1): 3456

The Potential role of MR-based radiomic biomarkers in the characterization of focal testicular lesions

Giacomo Feliciani, Lorenzo Mellini, Aldo Carnevale, Anna Sarnelli, Enrico Menghi, Filippo Piccinini, Emanuela Scarpi, Emiliano Loi, Roberto Galeotti, Melchiore Giganti and Gian Carlo Parenti

ABSTRACT

How to diferentiate with MRI-based techniques testicular germ (TGCTs) and testicular non-germ cell tumors (TNGCTs) is still under debate and Radiomics may be the turning key. Our purpose is to investigate the performance of MRI-based Radiomics signatures for the preoperative prediction of testicular neoplasm histology. The aim is twofold: (i), diferentiating TGCTs and TNGCTs status and (ii) diferentiating seminomas (SGCTs) from non-seminomatous (NSGCTs). Forty-two patients with pathology-proven testicular neoplasms and referred for pre-treatment MRI, were retrospectively enrolled. Thirty-two out of 44 lesions were TGCTs. Twelve out of 44 were TNGCTs or other histologies. Two radiologists segmented the volume of interest on T2-weighted images. Approximately 500 imaging features were extracted. Least Absolute Shrinkage and Selection Operator (LASSO) was applied as method for variable selection. A linear model and a linear support vector machine (SVM) were trained with selected features to assess discrimination scores for the two endpoints. LASSO identifed 3 features that were employed to build fvefold validated linear discriminant and linear SVM classifers for the TGCT-TNGCT endpoint giving an overall accuracy of 89%. Four features were employed to build another SVM for the SGCT-SNGCT endpoint with an overall accuracy of 86%. The data obtained proved that T2-weighted-based Radiomics is a promising tool in the diagnostic workup of testicular neoplasms by discriminating germ cell from non-gem cell tumors, and seminomas from non-seminomas.

ARTICLE TYPE: Original Research

ABBREVIATIONS

AJCC - American Joint Committee on Cancer

ADC - Apparent Diffusion Coefficient

AUC - Area Under the Curve

DWI - Diffusion-Weighted Imaging

DCE - Dynamic Contrast-Enanced imaging

MRI - Magnetic Resonance Imaging

NSGCT – Non-seminomatous Germ Cell Tumor

ROI - Region Of Interest

SGCT - Seminoma Germ Cell Tumors

SVM - Support Vector Machine

TGCT - Testicular Germ Cell Tumor

TNGCT - Testicular Non Germ Cell Tumor

US - Ultrasonography

VOI - Volume Of Interest

ZP - Zone Percentage

DECLARATIONS

Ethics approval and consent to participate

Written informed consent was waived by the Institutional Review Board.

Availability of supporting data

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

1. INTRODUCTION

Over the last decades, there has been a steady worldwide increase in the incidence of testicular cancer (1). The majority of these tumors are the germ cell tumors (TGCTs), which are then divided into two broad classes: seminomatous germ cell tumors (SGCTs) and nonseminomatous (NSGCTs). On this categorisation depend both the treatment and the prognosis (2). For istance, SGCT is more sensitive to radio and chemotherapy and thus a better prognosis. Although ultrasonography (US), including conventional grey-scale and color-Doppler US, still maintains the primary role in the diagnostic workup of scrotal pathology, magnetic resonance imaging (MRI) (3) has emerged as a supplemental imaging modality, which is mainly recommended as a problem-solving tool in challenging cases (4). Hence, MRI may provide additional information and help to clarify inconclusive or equivocal US findings in order to reduce the incidence of unnecessary surgery (4, 5). Albeit MRI may facilitate the differentiation between benign and malignant tumors (6), imaging alone is sometimes insufficient in making a clear distinction among testicular lesions. Previous studies have underlined the role of qualitative radiological assessment based on T1- and T2-weighted MR images that helps to differentiate between seminomas and non-seminomatous tumors (7). These studies have been further supported by quantitative investigation on diffusion weighted imaging (DWI) which have reported similar accuracy in discriminating SGCT – NSGCT status (8, 9); however, current existing data do not unequivocally support the role of DWI in being able to differentiate TGCT from non-germ cell tumors (TNGCT) (10). Given the rarity of these tumors, these results were obtained from small cohorts and still require validation. However, in the past decade, the breakthroughs in artificial intelligence and high-throughput computing have accelerated the application of radiomic analysis to medical imaging with the aim of guiding clinical decision-making.

The drive behind the spread of Radiomics is the attempt to derive quantitative features from digital images in order to provide information which is not obvious to human interpretation alone (11, 12). Radiomics appears to supply diverse imaging biomarkers in different medical fields, although medical oncology represents the main area of research, since such image analysis may be of help in tumor detection, diagnosis, prognostication and prediction of response to treatment (11, 13). Indeed, in most recent publication, Zhang et al (14), developed a radiomic signature to quantitatively discriminate seminomas from non-seminomatous tumors obtaining higher classification rate compared to the other standard MRI-based techniques (e.g. visual inspection, ADC and DWI value).

Therefore, this study extends and improves the work of Zhang et al by investigating the diagnostic performance of internally validated radiomic models in characterizing testicular neoplasms and more specifically differentiating between TGCTs and TNGTCs where classification is still under debate. Our findings show that in this field, MRI and Radiomics together allow an accurate characterization of testicular lesions, successfully guiding clinical decision-making.

2. METHODS

Patient selection

In this observational retrospective study, approved by our institutional review board of the Azienda USL della Romagna (informed consent is published in integral part on the website of Azienda USL della Romagna prot. N. 1683), a dataset of MR images of 42 patients who were referred for pathology has been analyzed. All research was performed in accordance with relevant regulations and informed consent was waived by the IRB.

After biopsy or orchiectomy, the Pathology Department of our hospital provided us with confirmation of histological diagnosis, in all testicular tumors which had undergone surgery from January 2006 to February 2019. All patients who had a scrotal MRI available in our imaging archive system (Carestream VuePACS, Carestream Health, Rochester, NY, USA) were consequently selected. Exclusion criteria were the following: (*a*) patients who underwent MRI after surgical or radiotherapy and/or chemotherapy treatment; (*b*) poor quality of the MR images due to movement artifacts; (*c*) no visible lesion on MRI; (*d*) not primary testicular tumor (**Figure 1**). MRI was perfomed in clinical practice as a second-level problem-solving tool when sonographic findings were equivocal or inconclusive, or following a request from the urology department to obtain a detailed local staging of a testis mass previously identified with US. The patient cohort was aged from 7 to 79 years old (average 39,3 ± 14,3 yrs). One of the patients had a bilateral classic seminoma and one had two different neoplasms years apart. We excluded 2 patients with testicular lymphoma and 1 with testicular localization of myeloma because of the uncertain metastatic origin; 1 more patient with classic seminoma was discarded due to bad image quality, so the final dataset consisted of 42 patients. Therefore, we analyzed MR studies of 44 testicular lesions (patient and lesion features are summarized in **Table 1**). Time difference between MRI and histologic final diagnosis was 25 ± 15 days. Thirty-two out of 44 were histologically classified as TGCTs,

including 23 classic seminomas and 9 NSGCTs (7 mixed germ cell tumors and 2 embryonal cell carcinomas). Twelve lesions out of 44 were TNGCTs or other histological types: 7 Leydig cell tumors, 2 Sertoli cells tumors, 2 adenomatoid tumors and 1 epidermoid tumor. For each lesion, laterality (left/right) and size have been considered; germ cell tumors were staged according to the 8th Edition of the American Joint Committee on Cancer (AJCC) Staging Manual. For a more detailed description of the lesions, also several visual features were analyzed in Supplementary Table S1, created by following indications found in (7). These features included signal intensity of the lesion compared to normal parenchyma, presence of necrotic or hemorrhagic areas, presence of tumor capsule. Furthermore, bandlike structures on T2w images were considered fibrovascular septa and the contrast of these septa was also analyzed.

Table 1: Patient demographics and lesion features

Germ cell tumors

AGE (yr)	Average ± standard deviation	36,8 ± 9
LATERALITY	Right/Left	19/13
SIZE (maximum diameter -cm)	Average ± standard deviation	3,2 ± 2,4
STAGING (T)	pT1/pT2/pT3/pT4	17/13/2/0

Non germ cell tumors

AGE (yr)	Average ± standard deviation	39,1 ± 18,6
LATERALITY	Right/Left	4/8
SIZE (maximum diameter - cm)	Average ± standard deviation	0,94 ± 0,46

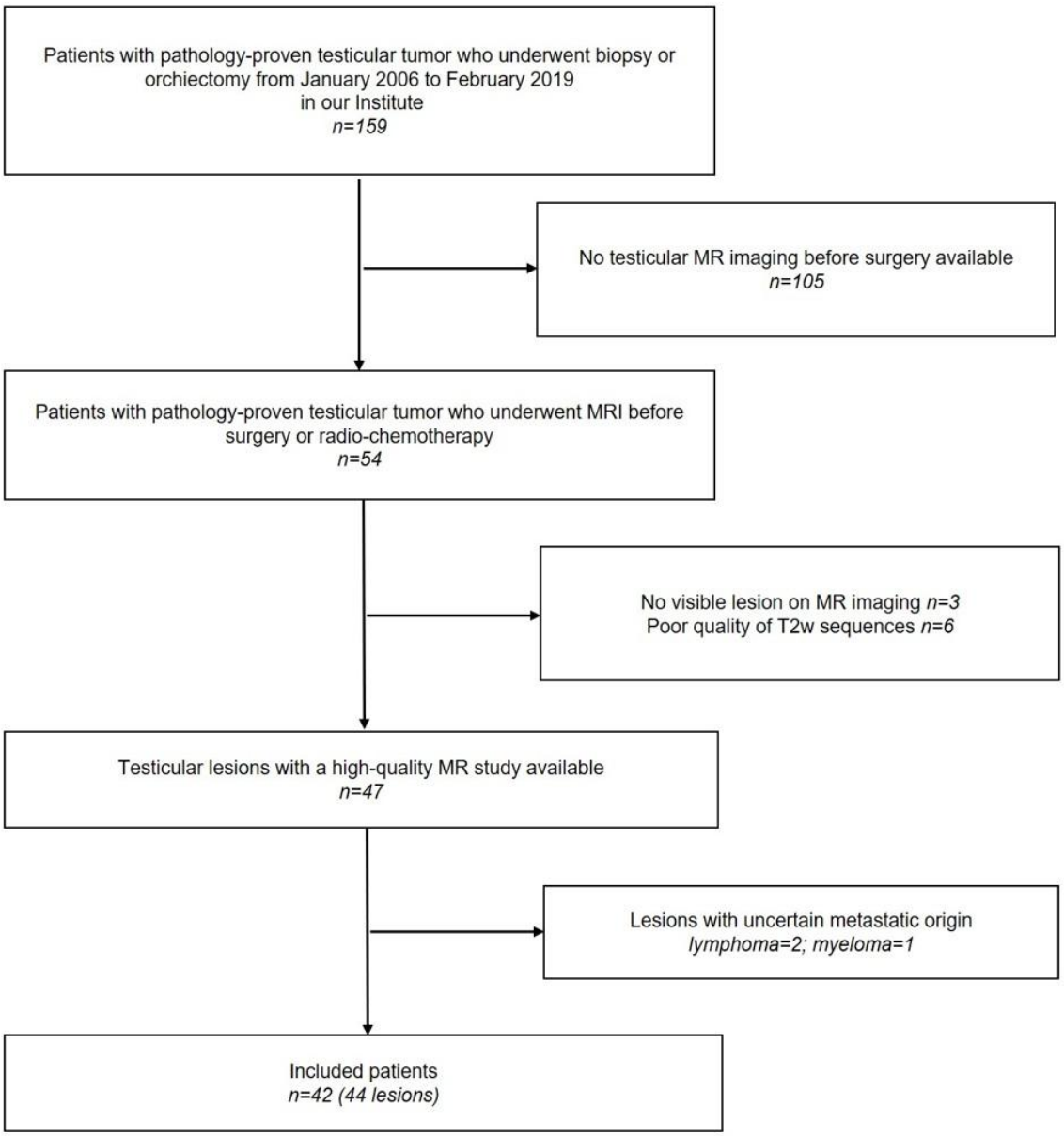


Figure 1. Flowchart summarizing patient accrual.

MR imaging protocol and radiomic analysis

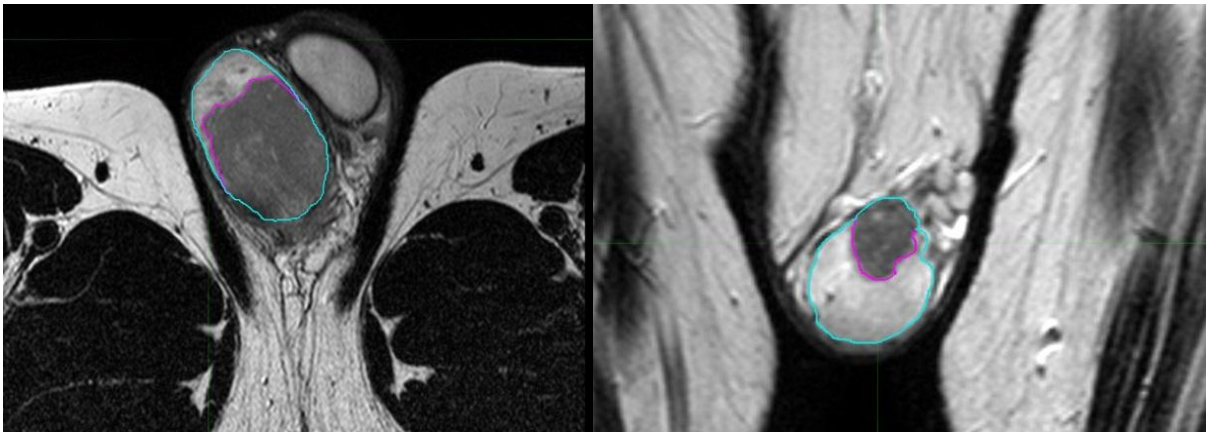
MR studies were acquired in our department on the same 1,5 T MR Scanner (Achieva Philips, Philips Healthcare, Best, Netherlands) by using a surface coil (Philips Sense Flex Medium coil).

The patient was placed in the scanner in the supine position, feet first. After adequate support and positioning of the scrotum, elevated by placing a towel between the thighs with the penis raised and fixed to the lower abdominal wall, the surface coil was placed over a

second towel covering the scrotum. A peripheral venous access (19-gauge) was obtained in an antecubital fossa vein. All the MRI study protocols included T1-weighted (T1w) sequences before and after paramagnetic contrast agent administration and T2w sequences in the axial, coronal and sagittal plane; some of the examinations also included DWI sequences and derived Apparent Diffusion Coefficient (ADC) maps. T2w sequences were selected for radiomic analysis since they are the most complete imaging set for each patient and are the best for lesion detection, localization and characterization, providing essential information on neoplastic tissue and anatomic detail (4). MRI parameters for T2w at our institution are summarized in **Table 2**. Spatial resolution varied from 0.3 to 0.7 mm in the axial direction and from 3 to 4 mm in the z direction. Resampling of the images was performed prior to contouring and radiomic analysis in order to uniform dataset to an average resolution of 0.5/0.5/3.5 mm. Contouring of the patient lesions was performed on the T2w sequences (**Figure 2A and 2B**) through consensus between two expert radiologists. First, second and higher order features were extracted with the open source MATLAB (The MathWorks, Inc., MA, USA) based software CGITA version 1.3. A quick guide on setting up and run CGITA for feature extraction is present in Supplemental Material S2. First order features were derived from the histogram of voxel intensities. Second and higher order features were calculated from Intensity size-zone, co-occurrence and run-length based matrices. Detailed description of the 72 imaging features extracted can be found in (15). Grey level quantization was fixed to 64 bins between minimum and maximum value inside the Region Of Interest (ROI).

Table 2: Turbo Spin-Echo T2-weighted image acquisition parameters

Acquisition Parameter	**Value**
Slice thickness	3.5 mm 3-4 mm
Min. slice gap	0
Repetition time	5899 ms
Echo time	120 ms
Flip angle	90°
Field of View	
Right/Left dimension	160 mm
Anterior/Posterior dimension	90 mm
Foot/Head dimension	160 mm



(**A**) (**B**)

Figure 2. MR images showing the segmentation process in a 27-year-old man with testicular seminoma (**A**) and a 31-year-old man with Leydig cell tumor (**B**), axial and coronal T2-weighted images, respectively. Testicles are contoured in blue, whereas neoplasms are contoured in violet.

Statistical analysis

The endpoint of this study was to investigate the diagnostic performance of textural features against two different biopsy responses. The first response was to discriminate between germinal and non-germinal lesions, whereas the second was to assess whether the tumor was a seminoma or not. Mann-Whitney test was used for the germinal-non germinal (TGCTs-TNGCTs) test, with TGCTs labelled as 0 and TNGCTs labelled as 1. Features that showed a p-value < 0.01 were further analyzed applying again Mann-Whitney test for SGCTs (labelled as 0) vs NSGCTs (labelled as 1) endpoint alone leaving benign tumors outside from the dataset and labelled as 2. Features that had a p-value < 0.01 in every test were further investigated. A correlation test was performed among significant features to remove redundancy through Spearman-Rho correlation coefficient. Features which correlated with each other were discarded and the features with the lowest p were kept in order to build a diagnostic model. Logistic regression was performed by employing R and the open source software RStudio (16) to assess imaging biomarkers prediction significance together with patient age and lesion volume for both the endpoints in order to unveil potential confounders. MATLAB R2018a statistical toolbox (17) was employed to generate a validated classifier and evaluate its performance. All the designed scripts are provided on request. In order to reduce overfitting of the classifiers 5- fold cross validation has been performed. A linear model and a Support Vector Machine (SVM) were trained to assess discrimination scores of statistical models.

3. RESULTS

From the 44 lesions fnally identifed, a total of 487 features were extracted. LASSO algorithm was independently applied for the two endpoints of this study. In the pool of features identifed by LASSO and afer evaluating the correlations with spearman ρ we fnally identifed 3 features for the association with TGCT-TNGCT discrimination endpoint and 4 features for the SGCT-NSGCT status. Zone percentage (ZP) calculated from the Gray Level Size Zone Matrix was the strongest predictor with a p-value < 0.001 for TGCT-TNGCT and p-value < 0.01 for SGCT-NSGCT discrimination endpoint employing Mann-Whitney U in both cases. Detailed description of ZP calculation is available in Supplemental Material S3. .

Figure 3A and **3C** show the box plot of ZP against the two endpoints, whereas **Figure 3B** and **3D** present the same for lesion volume.

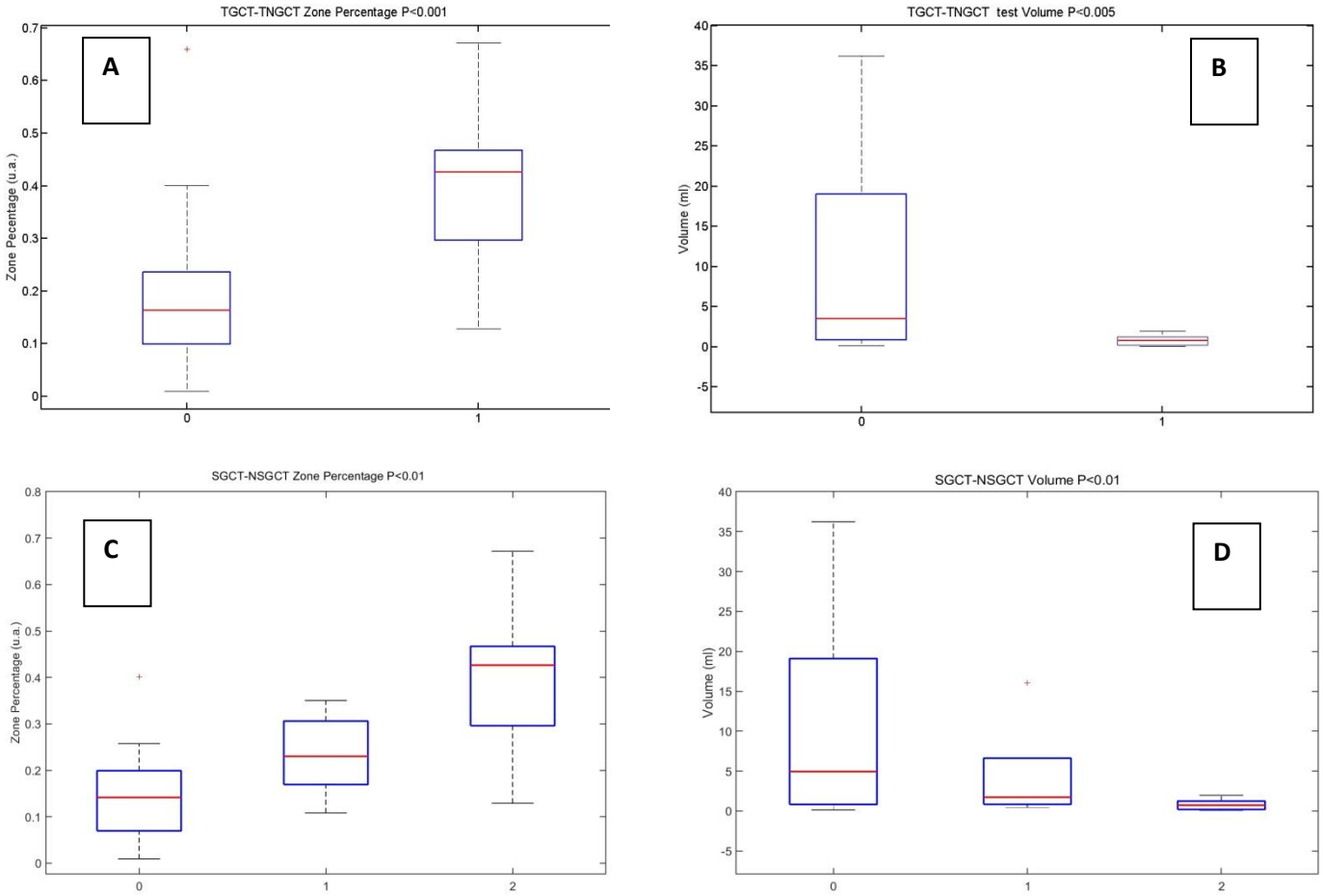


Figure 3. In **A** we show Zone Percentage (ZP) calculated values for germinoma(TGCT) and non germinoma (TNGCT) tumor cell cancers. TGCT is labeled as 0 whereas TNGCT is labeled as 1. The same applies for tumor volume in **B**. In **C** we show Seminoma (SGCT) vs non-seminoma (NSGCT). SGCT is labelled as 0, NSGCT as 1 and other histology as 2. The same labelling is used in **D** for tumor Volume

Figure 4 illustrates the ROC curves of ZP and volume in discriminating TGCT and TNGCT showing an AUC of 88% and 78%, respectively. The AUCs for the SGCT vs NSGCT endpoints were 83 % and 73% respectively. Furthermore, a logistic regression together with age as demographic data was performed. Results of linear regression for TGCT-TNGCT endpoint are shown in **Table 3**. ZP is the only statistically significant index associated to TGCT-TNGCT with a Hazard Ratio of 1.608. Volume is correlated with TGCT-TNGCT endpoint but with no statistical significance. The same results hold for SGCT-NSGCT status prediction as shown in Table 3. ZP was able to discriminate between both TGCT-TNGCT and SGCT-NSGCT.

We trained and tested through 5-fold cross validation one linear discriminator and one linear SVM employing ZP and volume as separated predictors.

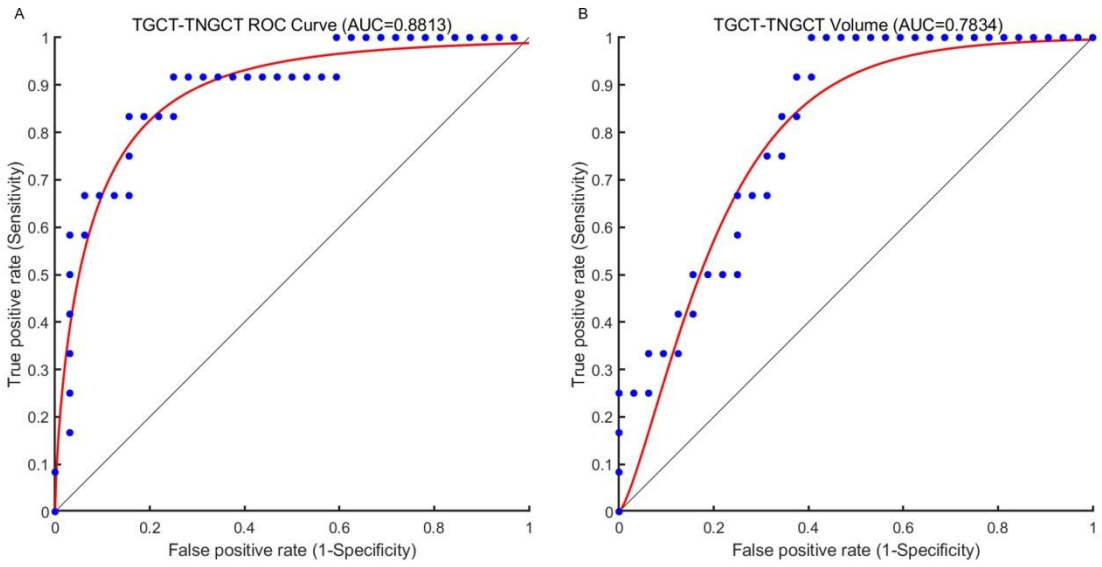


Figure 4. Zone percentage and Volume Receiver Operating Curve for TGCT-TNGCT endpoint showing an Area Under the Curve (AUC) of 0.88 and 0.78, respectively

Table 3. Logistic Regression of clinical and imaging variables for discriminating germ cell from non-germ cell tumors and seminomas from nonseminomatous ones excluding benign tumors.

TGCT-TNGCT	Coef.	*p*-value	95.0% CI for Coef. Lower	Upper	SGCT-NSGCT	Coef.	*p*-value.	95.0% CI for Coef. Lower	Upper
AGE	0.001	0.814	-0.008	0.011	AGE	-0.005	0.580	-0.023	0.013
ZP	**1.608**	**0.000**	**0.814**	**2.402**	**ZP**	**3.121**	**0.000**	**1.612**	**4.631**
Volume	0.000	0.916	-0.004	0.005	Volume	0.000	0.980	-0.009	0.009

TGCT/TNGCT – Testicular Germ Cell / Non Germ Cell Tumor

SGCT/NSGCT – Seminoma Germ Cell Tumor / Non Seminomatous Germ Cell Tumor

Coef. – Coefficient of logistic regression

CI – Confidence Interval

ZP – Zone Percentage

For TGCT-TNGCT endpoint the two models gave a final accuracy of 84% and 86%, respectively. Confusion matrix and the ROC curve of the best model (linear SVM) to discriminate TGCT-TNGCT are shown in **Figure 5A**. SVM accuracy in predicting SGCT-NSGCT status was 81% for both models, confusion matrix and ROC curve are similarly shown in **Figure 5B**. Training the same models with volume we obtained a 72.9% and 66% accuracy for TGCT-TNGCT status, and 63% and 61.4% for SGCT-NSGCT.

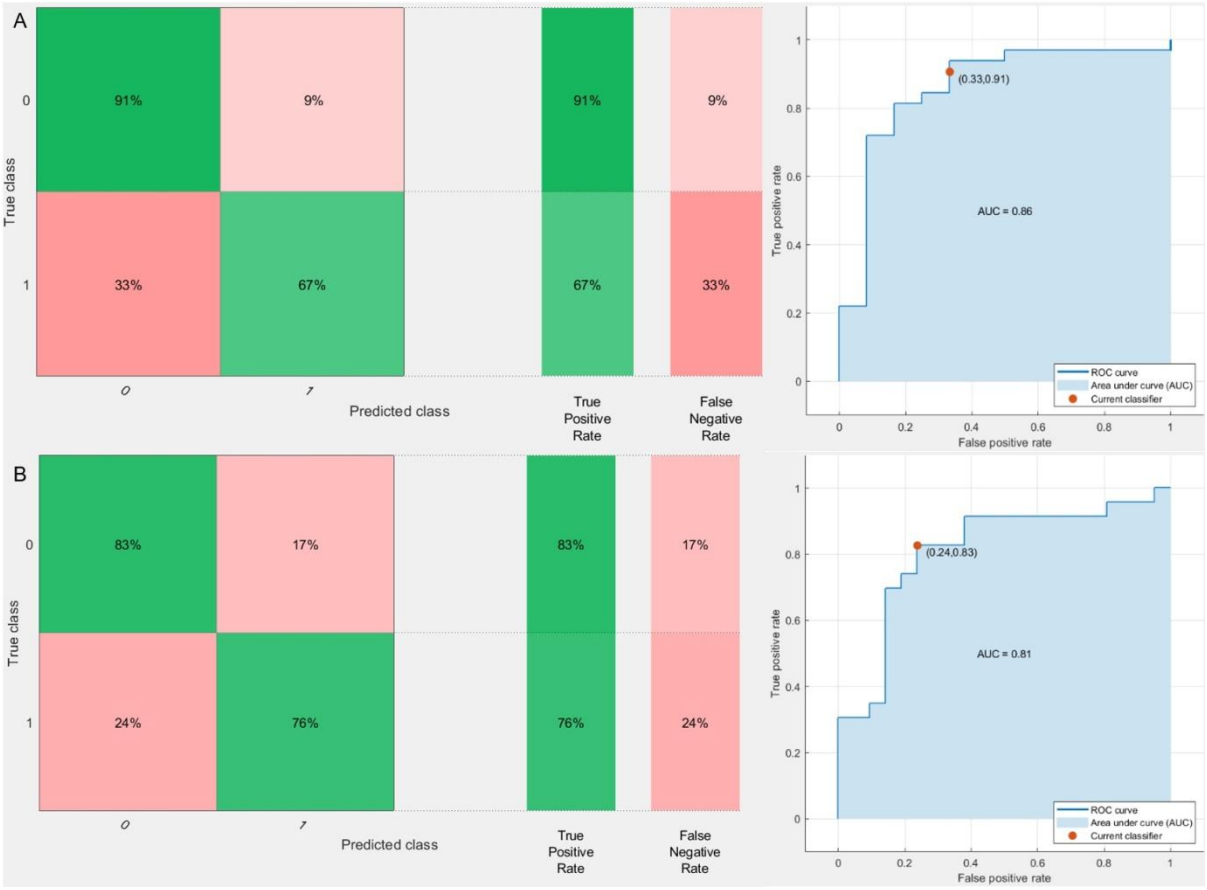


Figure 5. At the top (**A**), Confusion matrix and ROC curve of ZP based Support Vector Machine (SVM) trained and cross-validated for TCGT-TNGCT endpoint. Below (**B**), we show the ZP based SVM model performance for SGCT-NSGCT endpoint

4. DISCUSSION

This study evaluated the ability of T2w MR-based quantitative analysis to help differentiate germinal from non-germinal tumors and seminomas from non-seminomas. In the United States, testicular cancer represents the most common malignancy among men aged 15–44, with almost 9600 new cases estimated in 2019 (19); In young men, germ cell-derived tumors constitute by far the vast majority of testis neoplasms (almost 95%), with benign sex cord–stromal tumors representing approximately the remaining 5% (20); moreover, germ cell tumors are almost equally composed of seminomas and non-seminomas (21), with differences in treatment strategies and prognosis (22). Advances in multimodality treatments, including surgery, chemotherapy and radiation, have yielded a noticeable decline in mortality rates of testis cancer, particularly when the diagnosis is made early in the clinical course; The preoperative diagnosis with US has been shown to have a 92-98% sensitivity and a 95-99,8% specificity (21)) but cannot be use to accurately predict tumor histology and to differentiate benign from malignant types. MRI for scrotal pathology has proved to be a valuable second-level imaging modality that could help to elucidate diagnostic dilemmas found at US. Indeed, characterization of scrotal lesions at US may sometimes be difficult as a result of several limitations of this technique compared with MRI, which include the small field of view, operator dependence, and limited tissue characterization (23). In selected cases MR could represent a useful adjunct for patients with inconclusive clinical and US findings, since it could modify and direct treatment strategies towards more conservative approaches, including biopsy, tumor enucleation and testicular-sparing surgery, or even clinical and imaging follow-up when deemed possible (6, 24). Nevertheless, a confident characterization of the nature of scrotal masses is not always achievable even with MRI. Not surprisingly, Radiomics represents a rapidly-growing translational field of research that has been applied to cancer care in an effort to find imaging biomarkers as decision support tools for clinical practice, given the increased number and availability of imaging data in oncology. Lung, breast, colorectal, renal cell, pancreatic, brain cancer and sarcoma have all been previously investigated through medical image processing and analysis (25, 26), whereas only one study (27) has applied radiomic analysis to retroperitoneal nodal masses from germ cell testis cancer after chemotherapy.

In the literature, a previous study has tested the ability to discriminate between seminomas and non-seminomas through qualitative observation by the radiologist examining on MRI images morphologic features, including tumor volume, infiltrative margins, fibrovascular septa, necrosis (7). This study reported high inter-radiologist agreement and an accuracy of MRI findings in predicting histologic diagnosis of 91%. However, the number of patients was limited to 21 cases and the interpretation of MRI findings will always be dependent on radiologist expertise. Other studies have focused on quantitative MRI imaging, such as DWI with Apparent Diffusion Coefficient (ADC) values giving promising results in discriminating SGCT-NSGCT status with an AUC of 0.906 (9, 10, 28, 29). The robustness of these results was also proven against different ROI definitions (28). Furthermore Dynamic Contrast Enhanced (DCE) – MRI has been also proven to be a valuable semi-quantitative method to discriminate TGCT and TNGCT lesions with a maximum AUC of 0.89; however, these methods do not provide numerical data for a standardized assessment (10). Recently Zhang et al proposed a radiomic signature based on multiple features able to discriminate quantitatively SGCT-NSGCT status with high reproducibility scoring an AUC of 0.979. Unfortunately, a radiomic signature comparison is hard to assess due to the complexity and high number of features employed for its development and it is beyond the scope of this study. Here we propose a simplified classifier based on a single feature, namely ZP, which is able to discriminate germinoma from non-germinoma cancers and seminomas from non-seminomas. ZP quantitatively depicts the coarseness of the texture of the lesion and high values indicate a large portion of the ROI having a fine texture and thus higher homogeneity. In our results benign lesion have the highest value of ZP whereas seminoma have the lowest as can be seen in Figure.3A and 3C. This seems to be informative of the more heterogeneous nature of malignant lesions. Volume informative contribution was also investigated, but ZP has proven to be a far stronger predictor of histopathological status at logistic regression. Internal 5-fold cross-validation was employed to build a stable SVM model and avoid overfitting; this represents a limitation, as an external validation will be required to confirm the results. However, the data supporting the conclusions of this manuscript may be available under request for additional. 44 lesions were included in the models and SVM gave a final accuracy of 86% in discriminating TGCT-TNGCT status. The model correctly identified 29 malignant lesions out of 32 (91%). Another SVM ZP- based model had a final accuracy of 81% in differentiating SGCT from other histologies, with 19 seminomas correctly classified out of 23 (83%). Following these promising results, we strongly believe that radiomics can be integrated with other quantitative techniques such as ADC and DCE to improve testicular

mass classification accuracy. We acknowledge that another limitation of this study lies in its retrospective nature and in the relatively low number of patients. Furthermore, dependency of ZP on contouring method and scanner vendors was not explored in this study.

In conclusion, our preliminary study shows that the radiomic measures obtained by scrotal MR image analysis may be useful in the diagnostic workup of testicular lesions, since they could add valuable information and help to discriminate among testicular neoplasms by differentiating germ cell from non-gem cell tumors, and seminomas from other histologies. Further independent validation is required to assess whether quantitative imaging features, possibly in conjunction with standard clinical markers and other quantitative techniques, may allow more accurate characterization of testicular lesions.

DATA AVAILABILITY

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

REFERENCES

1. Rosen A, Jayram G, Drazer M, Eggener SE: Global Trends in Testicular Cancer Incidence and Mortality. *Eur Urol* 2011; 60:374–379.

2. Honecker F, Aparicio J, Berney D, et al.: ESMO Consensus Conference on testicular germ cell cancer : diagnosis , treatment and follow-up Special article. *Ann Oncol* 2018(August):1658–1686.

3. Bertolotto M, Muça M, Currò F, Bucci S, Rocher L, Cova MA: Multiparametric US for scrotal diseases. *Abdom Radiol* 2018; 43:899–917.

4. Tsili AC, Bertolotto M, Turgut AT, et al.: MRI of the scrotum: Recommendations of the ESUR Scrotal and Penile Imaging Working Group. *Eur Radiol* 2018; 28:31–43.

5. Parenti GC, Feletti F, Carnevale A, Uccelli L, Giganti M: Imaging of the scrotum: beyond sonography. *Insights Imaging* 2018; 9:137–148.

6. Tsili AC, Bertolotto M, Rocher L, et al.: Sonographically indeterminate scrotal masses: how MRI helps in characterization. *Diagn Interv Radiol* 2018; 24:225–236.

7. Tsili AC, Tsampoulas C, Giannakopoulos X, et al.: MRI in the histologic characterization of testicular neoplasms. *Am J Roentgenol* 2007; 189:1473.

8. Tsili AC, Argyropoulou MI, Giannakis D, Tsampalas S, Sofikitis N, Tsampoulas K: Diffusion-weighted MR imaging of normal and abnormal scrotum: Preliminary results. *Asian J Androl* 2012; 14:649–654.

9. Algebally AM, Tantawy HI, Yousef RRH, Szmigielski W, Darweesh A: Advantage of adding diffusion weighted imaging to routine MRI examinations in the diagnostics of scrotal lesions. *Polish J Radiol* 2015; 80:442–449.

10. Manganaro L, Saldari M, Pozza C, et al.: Dynamic contrast-enhanced and diffusion-weighted MR imaging in the characterisation of small, non-palpable solid testicular tumours. *Eur Radiol* 2018; 28:554–564.

11. Gillies RJ, Kinahan PE, Hricak H: Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016; 278:563–577.

12. Rizzo S, Botta F, Raimondi S, et al.: Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp* 2018; 2:36.

13. de Leon AD, Kapur P, Pedrosa I: Radiomics in Kidney Cancer: MR Imaging. *Magn Reson Imaging Clin N Am* 2019; 27:1–13.

14. Zhang P, Feng Z, Cai W, You H, Fan C, Lv W: T2-Weighted Image-Based Radiomics Signature for Discriminating Between Seminomas and Nonseminoma. 2019; 9(November):1–9.

15. Fang YHD, Lin CY, Shih MJ, et al.: Development and evaluation of an open-source software package "cGITA" for quantifying tumor heterogeneity with molecular images. *Biomed Res Int* 2014; 2014.

16. RStudio Team: RStudio: Integrated Development Environment for R. 2015.

17. 2018a M: The MathWorks, Inc. .

18. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM: The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs—Part A: Renal, Penile, and Testicular Tumours. *Eur Urol* 2016; 70:93–105.

19. American Cancer Society: Cancer facts and Figuers 2019. *Am Cancer Soc Web site* 2019.

20. Ulbright TM: Germ cell tumors of the gonads: a selective review emphasizing problems in differential diagnosis, newly appreciated, and controversial issues. *Mod Pathol* 2005; 18:S61–S79.

21. Coursey Moreno C, Small WC, Camacho JC, et al.: Testicular Tumors: What Radiologists Need to Know—Differential Diagnosis, Staging, and Management. *RadioGraphics* 2015; 35:400–415.

22. Algaba F, Bokemeyer C, Cohn-Cedermark G, et al.: *EAU Guidelines on Testicular Cancer*. 2018.

23. Mittal PK, Abdalla AS, Chatterjee A, et al.: Spectrum of Extratesticular and Testicular Pathologic Conditions at Scrotal MR Imaging. *RadioGraphics* 2018; 38:806–830.

24. Cassidy FH, Ishioka KM, McMahon CJ, et al.: MR Imaging of Scrotal Tumors and Pseudotumors. *RadioGraphics* 2010; 30:665–683.

25. Limkin EJ, Sun R, Dercle L, et al.: Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol* 2017; 28:1191–1206.

26. Cheng S-H, Cheng Y-J, Jin Z-Y, Xue H-D: Unresectable pancreatic ductal adenocarcinoma: Role of CT quantitative imaging biomarkers for predicting outcomes of patients treated with chemotherapy. *Eur J Radiol* 2019; 113:188–197.

27. Lewin JH, Dufort P, Halankar J, et al.: Applying radiomics to predict pathology of post chemotherapy retroperitoneal nodal masses in germ cell tumors (GCT). *J Clin Oncol* 2017; 35(15_suppl):4559–4559.

28. Tsili AC, Ntorkou A, Astrakas L, et al.: Diffusion-weighted magnetic resonance imaging in the characterization of testicular germ cell neoplasms: Effect of ROI methods on apparent diffusion coefficient values and interobserver variability. *Eur J Radiol* 2017; 89:1–6.

29. Tsili AC, Sylakos A, Ntorkou A, et al.: Apparent diffusion coefficient values and dynamic contrast enhancement patterns in differentiating seminomas from nonseminomatous testicular neoplasms. *Eur J Radiol* 2015; 84:1219–1226.

ACKNOWLEDGEMENTS

We would like to thank Alessandro Vagheggini for statistical support; Giorgio Mazzotti and Tiziana Licciardello for technical assistance and important discussions on the topic.

CHAPTER 5

Radiomics Analysis on [^{68}Ga]Ga-PSMA-11 PET and MRI-ADC for the Prediction of Prostate Cancer ISUP Grades: Preliminary Results of the BIOPSTAGE Trial

Published in: **Cancers** 2022 Apr; 14(12), 1888

Radiomics Analysis on [^{68}Ga]Ga-PSMA-11 PET and MRI-ADC for the Prediction of Prostate Cancer ISUP Grades: Preliminary Results of the BIOPSTAGE Trial

Giacomo Feliciani, Monica Celli, Fabio Ferroni, Enrico Menghi, Irene Azzali, Paola Caroli, Federica Matteucci, Domenico Barone, Giovanni Paganelli and Anna Sarnelli

Simple Summary: Radiomics analysis is used on MRI-ADC maps and [68Ga]Ga-PSMA-11 PET uptake maps to assess unique tumor traits not visible to the naked eye and to predict histology-proven ISUP grades in a cohort of 28 patients. Our study's main goal is to report imaging features that can distinguish low ISUP grades patients from higher grades (ISUP 1+) employing logistic regression statistical models based on MRI-ADC and 68Ga-PSMA data, as well as to assess the features' stability under small contouring variations. Our findings reveal that MRI-ADC and [68Ga]Ga-PSMA-11 PET imaging features based models are equivalent and complementary to predict low ISUP grade patients. These models can be employed in broader studies to confirm their ISUP grade prediction ability and eventually impact clinical workflow in reducing overdiagnosis of indolent, early-stage PCa.

Abstract: Prostate Cancer (PCa) risk categorization based on clinical/PSA results in a substantial number of men being overdiagnosed with indolent, early-stage PCa.Clinically non-significant PCa is characterized by the presence of ISUP grade 1 where PCa is found in no more than two prostate biopsy cores. Mp-MRI and [68Ga]Ga-PSMA-11 have been proposed as tools to predict ISUP 1 grade patients and consequently reduce overdiagnosis. In this study radiomics analysis is applied on MRI-ADC and [68Ga]Ga-PSMA-11 PET maps to quantify tumor characteristics to predict histology-proven ISUP grades. ICC was applied with a threshold of 0.6 to assess features' stability for variations in contouring. Logistic regression predictive models based on imaging features were trained on 31 lesions to differentiate ISUP 1 from ISUP 2+ patients. The best model based on [68Ga]Ga-PSMA-11 PET returned a prediction efficiency of 95% in the training phase and 100% in test phase whereas the MRI-ADC best model had an efficiency of 100% in both phases. Employing both imaging modalities, prediction efficiency was 100% in the training phase and 93% in test phase. Although our patient cohort was small, it was possible to assess that both imaging modalities add information to build the prediction models and show promising results for further investigations.

Keywords: prostate cancer; retrospective studies; MRI-ADC Scans; [68Ga]Ga-PSMA-11 PET; radiomics

1. INTRODUCTION

Prostate cancer (PCa) is the second most frequent cancer diagnosis made in men and the fifth leading cause of death worldwide with an ever-increasing incidence [1].

Current clinical-/PSA-based risk stratification for PCa still leads to a large number of men being overdiagnosed with indolent, early-stage PCa that may only require active surveillance rather than immediate treatment with unjustified comorbidities. According to pertinent societal guidelines clinically non-significant PCa (cns-PCa) is characterized by the presence of ISUP grade group 1 where PCa is found in no more than two prostate biopsy cores, each affected by less than 50% of its length, with a total PSA inferior to 10ng/ml[2], [3]. At the same time systematic trans-rectal ultrasound-guided 12-core biopsies may fail to detect the most aggressive components of PCa and their real size underestimating clinically-significant PCa (cs-PCa) in up to 30% of cases, delaying active treatments. Noninvasive determination of the real ISUP grade group would be of great help in informing biopsy targeting and treatment decision [4]–[6]. In this scenario, there is an emerging need of non-invasive methods that better correlate with histology proven ISUP grade.

Multi parametric magnetic resonance (mp-MRI) combining T1, T2 weighted sequences with Diffusion Weighted MRI and [68Ga]Ga-PSMA-11 PET have proven to be good candidates to bridge this gap [7]–[10]. For instance, the PROMIS trial demonstrated that mp-MRI triage might avoid unnecessary biopsies in 27% of cases and allowing for 18% increased detection of clinically significant cancer for TRUS biopsies guided by mp-MRI compared to standard TRUS biopsies[7]. In subsequent studies also Kasivisvanathan et al [11] and Ahdoot et al [8] found out that MRI targeted biopses are superior to standard transrectal ultrasonography–guided biopsy in men at clinical risk of prostate cancer. Furthermore, quantitative parameter extracted from Apparent Diffusion Coefficient maps (calculated from DWI sequences of mp-MRI) showed a negative correlation with histology proven ISUP grade (former Gleason score) [12]–[14]. However, the positive predictive value of mp-MRI is still poor, ranging from 20% to 68% [12] resulting in needless biopsies [11] and a need for improvement, particularly for individuals classified as intermediate risk. Despite these new findings, the analysis of the MRI-ADC maps' histogram alone leaves many grey areas in the discrimination of low ISUP grade patients (1 vs 2+), which is critical in treatment guidance, such as deciding between active surveillance, surgery, or radiotherapy according to NICE guidelines [15].

Prostate-specific membrane antigen (PSMA) is a type II transmembrane glycoprotein overexpressed on PCa epithelial cells surface. PSMA degree of overexpression is associated with higher aggressive biology (Gleason Score / ISUP grade group), luminal subtype, high androgen receptor activity and with higher serum PSA and it is related to tumor progression and disease recurrence[16]–[22]. Pioneering studies evaluating the potential of [68Ga]Ga-PSMA-11 PET to detect intraprostatic tumour foci have documented proportionality between the intensity of PSMA tumour uptake and pathology ISUP grade group, the size of tumor foci, tumor growth pattern (infiltrative versus expansive), serum PSA and higher D'Amico score[23]–[32]. A recent meta-analysis carried out on 389 patients with clinical/biochemical suspicion of PCa documented for [68Ga]Ga-PSMA-11 PET an overall sensitivity and specificity of 97% and 66%, respectively. Despite[68Ga]Ga-PSMA-11 PET returned a poor specificity similar to that of mpMRI, its negative likelihood ratio was found to be 0.05 leading to a 20-fold decrease in the odds of PCa being present in patients with negative PET findings. Also, [68Ga]Ga-PSMA-11 PET diagnostic accuracy for detecting clinically significant PCa returned pooled sensitivity and negative likelihood ratio of 0.99 and 0.02, respectively, potentially implying a role as a non-invasive risk stratifier[33].

Thus, PSMA-targeted PET imaging has been proposed in recent years to increase the mpMRI diagnostic accuracy in defining the malignant potential of lesions detected and scored according to PIRADS version 2.1. Studies evaluating the added value of [68Ga]Ga-PSMA-11 PET and mpMRI to detection of clinically-significant PCa documented a significantly increased diagnostic accuracy of the multimodality approach compared to individual modalities. PSMA uptake (SUVmax) and DWI MRI (ADCmax and ADCmin) were found to be distinct biomarkers able to differentiate between clinically significant PCa and and normal prostatic tissue in naïve prostate cancer patients with Gleason Score ≥ 7 [34], [35]. In this study texture analysis, which applies advanced mathematical functions to medical images, will be employed both in MRI-ADC maps and [68Ga]Ga-PSMA-11 uptake maps to quantify peculiar tumor characteristics, not visible to naked eye, in order to predict histology proven ISUP grade. Therefore their application to MRI-ADC maps has been reported to be helpful in reducing grey areas in ISUP grade prediction [36] and employed together with [68Ga]Ga-PSMA-11 PET they may show even more promising results. Despite this evidence, radiomics features are strongly affected both by acquisition parameters and contouring methods [37]–[39].

The primary objective of our study is to report features able to discriminate low ISUP grade patients from higher grades (ISUP 1+) employing both MRI-ADC and [68Ga]Ga-PSMA-11 data and to test the stability of the features under small contouring variations.

2. MATERIALS AND METHODS

2.1. Patient Selection

We retrospectively analyzed a dataset of mp-MRI and [68Ga]Ga-PSMA-11 PET images from 28 patients with biopsy-proven prostate adenocarcinoma enrolled in our institutional prospective multi-cohort study BIOPSTAGE (EudraCT number: 2017-002651-28) in the time span between May 2018 and May 2020. In this prospective study, patients with high-risk prostate cancer are staged by pelvic mp-MRI and [68Ga]Ga-PSMA-11 PET prior to radical prostatectomy and pelvic lymph node dissection to rule out metastases and for correlation of pelvic imaging findings with axial step section histopathology analysis. Both mp-MRI and [68Ga]Ga-PSMA-11 PET scans were performed in patients fulfilling the following cohort-specific inclusion criteria:

(a) patients 18 years of age or older, able to express informed consent for study participation and compliant with BIOPSTAGE on-protocol imaging;

(b) biopsy evidence of prostate cancer with any of the following high-risk characteristics:

(1) clinical T stage ≥ T2c;

(2) clinical stage N1;

(3) ISUP grade group ≥ 4;

(4) serum PSA > 20 ng/mL;

(c) biopsies performed at least 4 weeks prior to mp-MRI and [68Ga]Ga-PSMA-11 PET;

(d) patients opting for radical prostatectomy and pelvic lymph node dissection.

Exclusion Criteria included:

(a) ongoing hormone therapy at the time of screening and within the previous six months;

(b) previous pelvic radiation therapy;

(c) any medical condition incompatible with MRI scanning or with the administration of MRI contrast medium or any condition that impairs the quality of pelvic MRI imaging;

(d) history of allergic reactions attributed to compounds of similar chemical or biological composition to [68Ga]Ga-PSMA-11;

(e) other known malignant neoplastic disease in the patient's medical history with a disease-free interval of less than 5 years; chemotherapy or radiation therapy in the 4 weeks prior to study entry;

(f) a history of other malignant neoplastic disease in the patient's medical history with a disease-free interval of less than 5 years.

An outline of the workflow prior to statistical analysis employed to obtain the results described below is given in **Figure 1**.

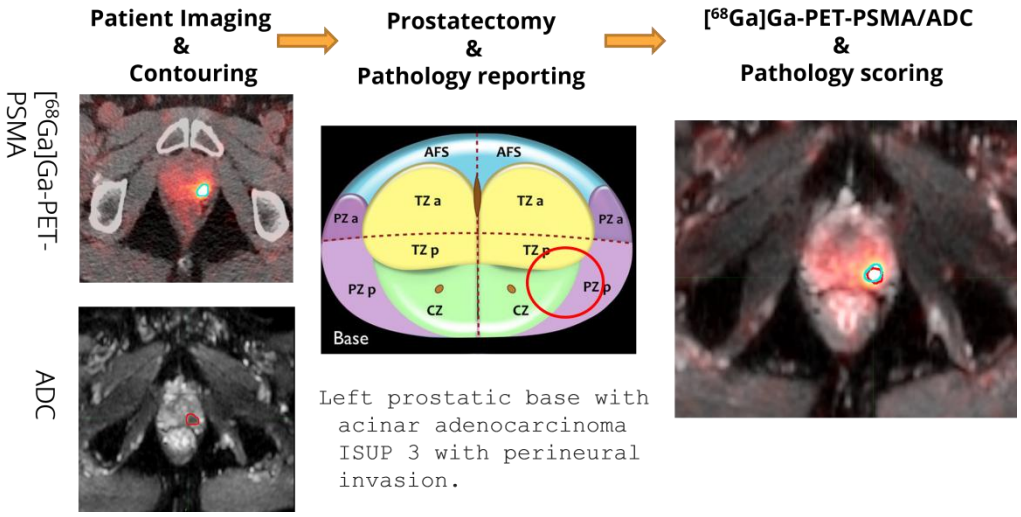


Figure 1. Detail of prostate contouring for the two imaging modalities performed by Nuclear Physician and Radiologist respectively (Left). In the center it is represented an example of anatomo-pathology reporting with details about ISUP grading. In the right side [68Ga]Ga-PSMA-11 PET and MRI-ADC are being fused with MIM maestro software with respective contouring superimposed.

2.2. MR Imaging Protocol and Lesion Contouring

Mp-MRI studies were acquired at our department on a 3 Tesla MR Scanner (Philips Ingenia 3.0T, Philips Healthcare, Best, Netherlands) by using a Philips Sense Flex Medium surface coil.

The patient was placed in the scanner in the supine position, feet first. T1- weighted (T1w), T2-weighted (T2w) and ADC maps generated by AXIAL Diffusion-Weighted Imaging (DWI) sequences for prostate / small pelvis were acquired.

Four b-values were used (b100, b800, b1000, b2000) to provide more accurate ADC calculations. Echo time (TE) and repetition time (TR) were ≤90 msec and ≥3000 msec respectively. The field of view (FOV) was 16-22 cm with an in-plane dimension of 2.5mm . Slice thickness was set to 3 mm without gap.

ADC maps were selected for radiomic analysis since they are the most informative for lesion detection, localization, and characterization, providing essential information on neoplastic tissue and anatomic detail (25). Contouring of the 28 patient lesions was performed on mp-MRI through Watson Elementary software (Watson Medical, Nijmegen, Netherlands) by an expert radiologist employing T1w, T2w, and ADC maps for a total of 37 lesions contoured.

2.3. [68Ga]Ga-PSMA-11 PET/CT Imaging Protocol and Lesion Contouring

[68Ga]Ga-PSMA-11 was prepared according to national regulations, good radiopharmaceutical practices (GRP) as outlined in specific EANM guidelines (26). All patients were intravenously injected with a mean activity of 159MBq of [68Ga]Ga-PSMA-11 (activity range: 112 - 202 MBq) via an indwelling catheter in an antecubital vein according to patient weight. A whole-body PET/CT scan was performed 60-80 min after i.v. administration of [68Ga]Ga-PSMA-11 covering a volume from the skull vertex through the mid-thigh in 3D flow motion. Whole-body PET acquisitions were corrected for attenuation and scatter and adjusted for system sensitivity and providing parametric images in terms of Standardized Uptake Values (SUVbw: KBq found / gm tissue / KBq injected / gm body mass). PET reconstruction matrix was 400x400 (Hi-REZ processing), achieving an axial resolution of 2.5 mm and a slice thickness of 4 mm. The CT component of the studies was performed using the CARE Dose4D protocol for CT dose adaptation (mAs weighed on z-axis, patient's dimensions, and x-y axis) HDFOV 512 x 512 matrix, slice thickness 3mm for PET attenuation correction and co-registration. [68Ga]Ga-PSMA-11 PET images were contoured by an

expert Nuclear Physician on MIMmaestro software employing as minimum positivity threshold an arbitrary maximum SUVbw of 3 g/ml and outlining 62 positive lesions.

2.4. Histopatological Reporting and ISUP Grade Assignment

The post-surgical histopathology results are considered the standard of truth for ISUP grade determination of the lesions. The anatomo-pathology specimens were sectioned serially from apex to base and submitted as 12 whole-mount sections for examination. After detailed microscopic revision, the ISUP grade pattern present in each section was determined. Then each lesion detected with mp-MRI or [68Ga]Ga-PSMA-11 PET was compared with histopathology results and consequently, an ISUP grade was associated. In case an imaging-detected lesion had negative correspondence on histopathology this same lesion was classified as false positive and consequently discarded from the analysis. All the lesions with tumour correspondence on histopathology were classified as true positives.

2.5. Images Pre-Processing and Radiomic Analysis

[68Ga]Ga-PSMA-11 PET and MRI-ADC images were resampled to a resolution of 1/1/1 mm to uniform the dataset.

For radiomics feature stability assessment, physicians' contours were isotropically expanded by 1 and 2 mm and contracted by 1 mm. Contraction of 2 mm was not considered in the analysis due to the small size of many lesions that may cause failure in the subsequent radiomics analysis

Lesions subset visible both in MRI-ADC and [68Ga]Ga-PSMA-11 PET were further contoured according to the following rules: a) if [68Ga]Ga-PSMA-11 PET Lesion contour is included in MRI-ADC lesion contour – [68Ga]Ga-PSMA-11 PET contour is chosen and vice versa b) In case of partial overlapping (> 80% of the volume) intersection between lesion contours was performed c) in other scenarios association between lesions was not considered.

First, second and higher order features were extracted with the Image Biomarker Standardisation Initiative (IBSI) [40] compliant tool, SOPHiA DDM™ For Radiomics (2021 SOPHIA GENETICS s.p.a., Boston, MA 02116, USA), for MRI-ADC and [68Ga]Ga-PSMA-11 PET images. First order features were derived from the histogram of voxel intensities. Second and higher order features were calculated from Intensity size-zone, co-occurrence and run-length based matrices.

Detailed description of the 218 imaging features extracted can be found in the IBSI Reference manual [40]. Grey level quantization was fixed to 32 bins between the minimum and maximum value inside the Region Of Interest (ROI). Features extracted from physicians' contours were compared with isotropically expanded and contracted ROIs (+1 and +2 mm -1 mm) through Intra Class Correlation coefficient (ICC) to select stable features under small variations in contouring with ICC>0.6.

2.6. Statistical Analysis

The endpoint of this study was to investigate the diagnostic performance of radiomics features extracted from multimodality imaging (MRI-ADC and [68Ga]Ga-PSMA-11 PET) against ISUP grade obtained from histology evaluation. In particular the ability of radiomic features to discriminate ISUP 1 from higher grades in order to help treatment stratification. In **Figure 2** it is summarized the entire process of statistical analysis from feature extraction to final model evaluations.

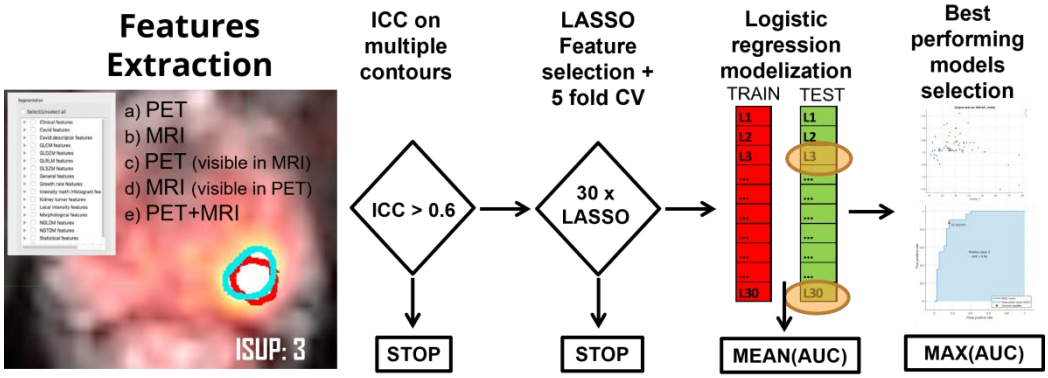


Figure 2. Detail of the workflow employed from the features extraction to the selection of the final statistical models.

Five independent ISUP Grade predictive logistic models were developed based on:

a) lesions visible only through [68Ga]Ga-PSMA-11 PET

b) lesions visible only with MRI-ADC

c) lesions visible with [68Ga]Ga-PSMA-11 PET and MRI-ADC but only employing 68-[68Ga]Ga-PSMA-11 PET imaging features

d) lesions visible with [68Ga]Ga-PSMA-11 PET and MRI-ADC but only employing mp-MRI imaging features

e)lesions visible both with [68Ga]Ga-PSMA-11 PET and MRI-ADC with features extracted from both imaging modalities

The models were built through a stochastic cross-validation process to evaluate their performance.

The modeling process followed this procedure:

Lesion feature datasets were divided into a training (2/3) and test (1/3) set. Subsequently, a logistic regression model was trained on the training set, employing features selected by a least absolute shrinkage and selection operator (LASSO) algorithm with internal 3-fold cross validation. The predictive ability of the model was then calculated on the test set. This operation was repeated 30 times and subsequently receiver operating curves (ROC) and their area under the curve (AUC) of each iteration were recorded both for the training and test set.

Models' quality was reported by averaging AUC across iterations. ROC and AUC were reported for the best-performing iteration to evaluate the model's prediction power and to compare the performances of mixed imaging features model e) with standalone imaging models c) and d). The most frequently selected features across iterations were reported as the most informative features for ISUP Grade prediction.

All statistical analyses were carried out with R and the open-source software RStudio[41]. The raw data of this study ([68Ga]Ga-PSMA-11 PET, MRI-ADC and Pathology records) are available as supplementary material.

3. RESULTS

Patients were aged between 44 and 72 years (mean age: 62 years). The median total PSA at time of prostate cancer diagnosis was 6.8ng/ml (IQR: 4.4 - 8,7). Eleven patients had ISUP 1 prostate cancer on post-prostatectomy pathology, eight patients had ISUP2, three patients had ISUP3, five patients had ISUP4, and 1 one patient had ISUP5 prostate cancer. The median time between [68Ga]Ga-PSMA-11 PET and mp-MRI was 8 days whereas the median time between advanced imaging and prostatectomy was 45 days. On post-prostatectomy pathology, organ-confined disease (pT2a to pT2c) was documented in 21 patients; seven patients were found with locally advanced disease (pTa to pT3b). **Table 1** provides an overview of the patient's features, while Supplementary Table S1 provides a more detailed description of the patient's characteristics.

Table 1. Summary of patients' characteristics.

Patients Characteristics	Value
mean age (years), age range	62.0 [44 - 72]
median age (years), IQR	63.0 [58.5 - 66.5]
median total PSA (ng/ml), IQR	6.8 [4.4 - 8.7]
median PSA density (ng/ml/g), IQR	0.15 [0.11 - 0.23]
median prostate volume (ml), IQR	48 [37.3 - 59.3]
overall ISUP grade group (post-prostatectomy pathology)	
1	n = 11
2	n = 8
3	n = 3
4	n = 5
5	n = 1
pathology T stage	
T2a - T2b	n = 6
T2c	n = 15
T3a	n = 4
T3b	n = 3
median time between [^{68}Ga]Ga-PSMA-11 PET and mpMRI (days), IQR	8 [4 - 13]
median time between imaging and surgery (days), IQR	45 [24 - 86]

In this cohort of high-risk prostate cancer patients candidates for surgery, MRI-ADC and [68Ga]Ga-PSMA-11 PET yielded similar sensitivity (71.5% and 72.3%, respectively) and a specificity of 99.5% and 90.5%, respectively, in detecting prostate cancer foci.

For the purpose of this study, we analyzed only true positive lesions on MRI-ADC imaging (n= 37) and on [68Ga]Ga-PSMA-11 PET imaging (n = 49 lesions) that is all those lesions that had positive correspondence on histopathology and that were used to build model a) and b). The small unbalance in the number of discovered lesions between imaging modalities is due to the fact that in 4 patients multiple PET lesion had correspondence with only one big lesion in MRI-ADC maps and for 3 patients MRI-ADC was low quality or unreadable. Among these lesions, 31 were topographically paired at fusion and employed to build models c), d) and e).

We extracted 218 imaging features with the Radiomics software Sofia from MRI-ADC and [68Ga]Ga-PSMA-11 PET imaging. The extraction was performed on the original images and on expanded lesion contours. Subsequently, ICC was applied with a threshold of 0.6 to assess features stability for small variation in contouring. Twenty-nine and 87 features successfully passed the ICC test for [68Ga]Ga-PSMA-11 PET and MRI-ADC imaging respectively. These features were further investigated and employed to build the 5 logistic models described in "Materials and Methods" section. **Table 2** summarizes the performances of the models in train and test phase and the overall best performing model for each cathegory whereas details are described below. On [68Ga]Ga-PSMA-11 PET features a) the average models performance in terms of area under the curve (AUC) on training and test sets was 0.58 and 0.53, respectively. One iteration out of the 30 showed a very good predictive power with an AUC of 0.90 on training set and of 1.00 on test set. MRI-ADC based models b) exhibited higher performance with an average AUC of 0.91 in the train phase and 0.67 in the test set. Furthermore, 8 of 30 iterations showed a high predictive performance both on training and test set with an AUC higher than 0.80. The average performance of model c) based on [68Ga]Ga-PSMA-11 PET features but trained on lesions visible also for mp-MRI, was 0.80 and 0.60 on training and test set, respectively. One iteration returned an AUC of 0.95 on the training set and an AUC of 1.00 on the test set. The most frequently selected features for models' development were area density, inverse elongation, zone size non uniformity, flatness and volume fraction difference between intensity fractions.

Table 2. Summary of trained and tested imaging biomarkers based models.

MODEL Type.	# of Lesions	TRAIN mean AUC	TEST mean AUC	TRAIN best AUC	TEST best AUC
a) PET	49	0.58	0.53	0.9	1
b) MRI	37	0.91	0.67	0.92	1
c) PET (MRI-visible)	31	0.8	0.6	0.95	1
d) MRI (PET-visible)	31	0.74	0.45	1	1
e) MRI+PET	31	0.75	0.49	1	0.93

The average performance of models d) based only on MRI-ADC features and trained on commonly detected lesions was 0.74 and 0.45 on the training and test set, respectively. Two iterations scored an AUC higher than 0.80 and the most selected features were joint maximum, zone distance non-uniformity, 90th discretized intensity percentile, compactness, information correlation, and skewness. Models e) based on both [68Ga]Ga-PSMA-11 PET and MP-MRI features,

showed a mean performance of 0.75 on the training set and of 0.49 on the test set. Two iterations had AUC higher than 0.80 and the most informative features were normalized inverse difference ([68Ga]Ga-PSMA-11 PET), zone distance non uniformity (MRI-ADC), joint maximum (MRI-ADC), large zone low grey level emphasis ([68Ga]Ga-PSMA-11 PET), 90th discretized intensity percentile(MRI-ADC), area density ([68Ga]Ga-PSMA-11 PET).

In **Figure 3** we report the ROC curves of the best performing iterations of models c) d) and e), both on training and test set.

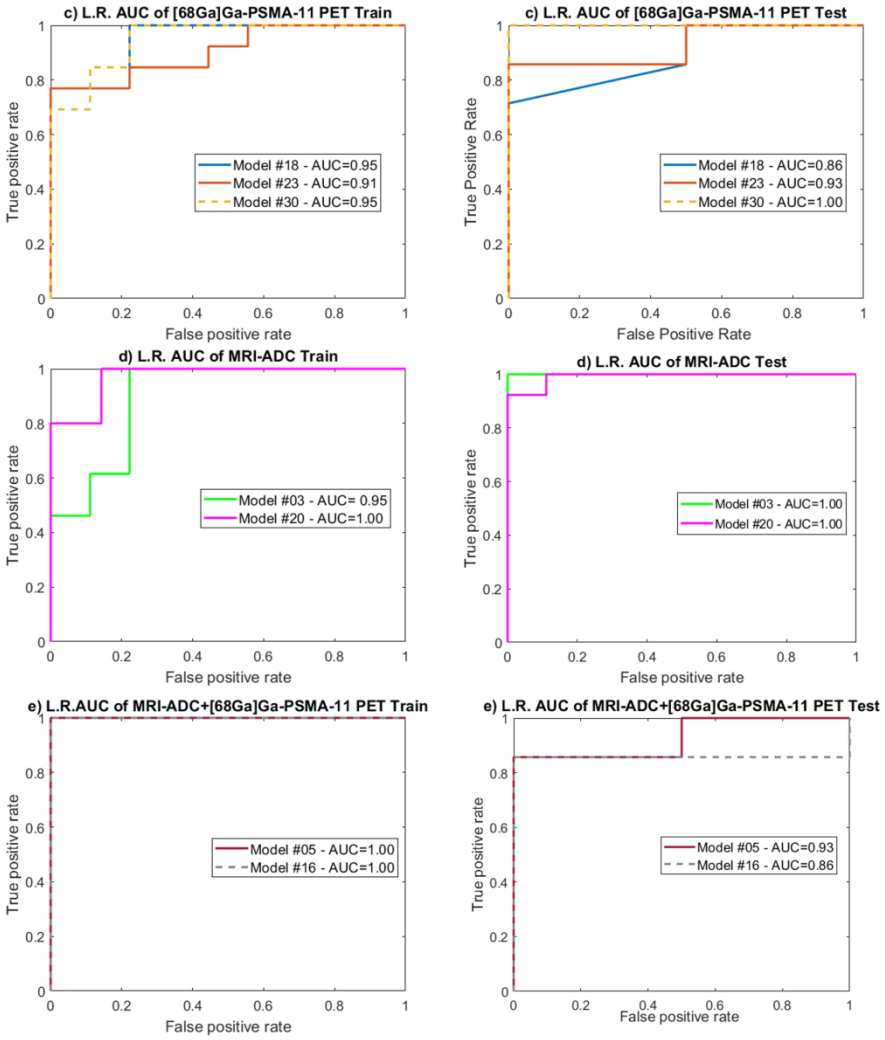


Figure 3. Logistic regression (L.R.) models performance in the training and test phase in terms of area under the curve (AUC).

4. DISCUSSION

Biopsy ISUP grade differs from the final ISUP determined after surgery in around one-third of patients, with biopsies tending to underestimate cancer aggressiveness. The differences between the two ISUPs can have a big impact on how patients are managed. As a result, incorporating pre-therapeutic imaging characteristics to accurately determine PCa aggressiveness is of great clinical importance.

This study evaluated the ability of MRI-ADC and [68Ga]Ga-PSMA-11-based quantitative analysis to help differentiate low-risk prostate cancer patients (ISUP 1) from higher risk patient classes (ISUP>1) and aimed to evaluate the benefits of the two imaging techniques combined. However, the results of this paper can be only intended as proof of concept as the number of concordant lesions on MRI-ADC and [68Ga]Ga-PSMA-11 PET is low, and this represents the major limitation of this study. To overcome this limitation, we employed a stochastic cross-validation approach and run LASSO-logistic modelling process on 30 partitions of the datasets into training and test set. Another limitation is represented by the laborious and time-consuming process required to contour, fuse and evaluate lesions on different imaging modalities. Furthermore, Radiomics feature variability due to imaging acquisition and reconstruction is another disadvantage that to date limits the widespread in clinical practice of this approach. The average predictive power in terms of AUC for the training phase is very variable across models a-e) and reaches a maximum of 0.91 for model b). In the test phase, the performances are quite low ranging from 0.45 of model d) to 0.65 of model b). From these average AUC, it is difficult to speculate about the benefits of employing both [68Ga]Ga-PSMA-11 PET and MRI-ADC for ISUP predictions and these low performances can be justified by the small datasets and mild class imbalance involved in the analysis that may compromise the training of the majority of the models. For model e) we had a performance drop in the test phase probably caused by the augmented number of features involved in the analysis together with a reduction of the number of lesions. Following these results, we are convinced that models' predictive power is strongly influenced by the data repartition in the train and test phase and thus it is our opinion only in higher AUC models the datasets were correctly balanced to give an idea of the real benefit of imaging features. For these reasons, we should take a closer look at every single model to give further details about the contribution of the two imaging modalities. Best performing [68Ga]Ga-PSMA-11 PET models "a-c" have very high accuracies (>90%) both in training and test phase and outperforms similar models models reported by Solari et al [42] for PCa ISUP grade prediction.

It's interesting to point out that in "b" models the MRI-ADC mean value that is the imaging predictor currently employed in clinical practice to assess patient risk was not selected by LASSO. This evidence suggests that Radiomics approach can provide a significant improvement to patient classification for MRI-ADC sequences. Furthermore, the performances of the best trained models "b" are in line with previously reported performances of mp-MRI Radiomics based analysis [43] in particular in the work of Fehr et al [44] combining ADC and T2w mean values with textural features they achieve an accuracy higher than 90% in differentiating low Gleason (6) prostate lesions from higher scores(>7).

Building the predictive models "c" and "d", including the lesions visible both in MRI-ADC and [68Ga]Ga-PSMA-11 PET, we can assess that the two imaging modalities are equivalent in discriminating low risk patients from higher risk ones with an AUC of the best performing iteration of 1.00 for test phase as visible from Figure. 3a,b. Finally combining the 29 features of [68Ga]Ga-PSMA-11 PET and the 87 of the MRI-ADC imaging we obtained model "e" where performances are slightly lower and a maximum AUC of 0.93 (Figure 3c). It is important to point out that in this model, LASSO algorithm always chooses as most informative features both MRI-ADC- and [68Ga]Ga-PSMA-11-based features to build the logistic regression prediction model indicating that the two modalities contribute to adding unique information for lesion classification. However, with our dataset it is difficult to observe statistically significant improvements in the performances given by the integration of the two modalities due to the restricted number of lesions and further investigations will be required to confirm our hypothesis.

5. CONCLUSIONS

Among developed models each imaging modality seems to provide similar results in ISUP grade prediction.. Preliminary results suggests that aside of MRI-ADC average value, currently employed in clinical practice to assess lesion severity, other imaging biomarkers may provide complementary information for ISUP grade prediction but further broader studies are necessary to confirm these findings.

Both [68Ga]Ga-PSMA-11 PET and MRI-ADC imaging biomarkers showed to be complementary about ISUP grade assessment when employed together to build prediction models.

REFERENCES

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA. Cancer J. Clin., vol. 71, no. 3, pp. 209–249, 2021.

2. National Comprehensive Cancer Network, "Prostate Cancer," 2020. [Online]. Available: https://www.nccn.org/professionals/physician_gls/pdf/prostate.pdf.

3. European Association of Urology, "EAU guidelines for prostate cancer," 2021. [Online]. Available: https://uroweb.org/guideline/prostate-cancer/#6.

4. V. Narain, F. J. Bianco, D. J. Grignon, W. A. Sakr, J. E. Pontes, and D. P. Wood, "How accurately does prostate biopsy Gleason score predict pathologic findings and disease free survival?," Prostate, vol. 49, no. 3, pp. 185–90, 2001.

5. R. Kvoele, B. Møller, R. Wahlqvist, S.D. Fosså, A. Berner, C. Busch, A.E. Kyrdalen, A. Svindland, T. Viset and Ole J. Halvorsen "Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: A population-based study," BJU Int., vol. 103, no. 12, pp. 1647–1654, 2009.

6. A. Rajinikanth, M. Manoharan, C. T. Soloway, F. J. Civantos, and M. S. Soloway, "Trends in Gleason Score: Concordance Between Biopsy and Prostatectomy over 15 Years," Urology, vol. 72, no. 1, pp. 177–182, Jan. 2018.

7. H. U. Ahmed, A.E. Bosaily, L.C. Brown, R. Gabe, R. Kaplan, M. K. Parmar, Y. Collaco-Moraes, K. Ward, R.G. Hindley, A. Freeman et al, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study," Lancet, vol. 389, no. 10071, pp. 815–822, 2017.

8. M. Ahdoot, A.R. Wilbur, S.E. Reese, A.H. Lebastchi, S. Mehralivand, P.T. Gomella, J. Bloom, S. Gurram, M. Siddiqui, P. Pinsky et al, "MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis," N. Engl. J. Med., vol. 382, no. 10, pp. 917–928, 2020.

9. M. Chen, Q. Zhang, C. Zhang, X. Zhao, G. Marra, J. Gao, X. Lv, B. Zhang, Y. Fu, F. Wang et al, "Combination of 68Ga-PSMA PET/CT and multiparametric MRI improves the detection of clinically significant prostate cancer: A lesion-by-lesion analysis," J. Nucl. Med., vol. 60, no. 7, pp. 944–949, 2019.

10. H. Rhee P. Thomas, B. Shepherd, S. Gustafson, I. Vela, P. J. Russell., C. Nelson, E. Chung, G. Wood, G. Malone et al, "Prostate Specific Membrane Antigen Positron Emission Tomography May Improve the Diagnostic Accuracy of Multiparametric Magnetic Resonance Imaging in Localized Prostate Cancer," J. Urol., vol. 196, no. 4, pp. 1261–1267, 2016.

11. V. Kasivisvanathan, A.S. Rannikko, M. Borghi, V. Panebianco, L.A. Mynderse, M.H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R.G. Hindley et al, "MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis," N. Engl. J. Med., pp. 1767–1777, 2018.

12. Y. Mazaheri, A. Shukla-Dave, H. Hricak, S.W. Fine, J. Zhang, G. Inurrigarro, C.S. Moskowitz, N.M. Ishill, V.E. Reuter, K. Touijer, et al, “Prostate Cancer: Identification with Combined Diffusion-weighted MR Imaging and 3D 1H MR Spectroscopic Imaging—Correlation with Pathologic Findings,” Radiology, vol. 246, no. 2, pp. 480–488, Feb. 2008.

13. S. Il Jung, O. F. Donati, H. a Vargas, D. Goldman, H. Hricak, and O. Akin, “Transition Zone Prostate Cancer: Incremental Value of Diffusion-weighted Endorectal MR Imaging in Tumor Detection and Assessment of Aggressiveness,” Radiology, vol. 269, no. 2, pp. 493–503, 2013.

14. O. F. Donati , Y. Mazaheri, A. Afaq, H.A. Vargas, J. Zheng, C. S. Moskowitz, H. Hricak and O. Akin “Prostate cancer aggressiveness: assessment with whole-lesion histogram analysis of the apparent diffusion coefficient.,” Radiology, vol. 271, no. 1, pp. 143–52, 2014.

15. NICE, “Overview | Prostate cancer: diagnosis and management | Guidance | NICE,” Nice, no. May 2019, p. 6, 2019.

16. G. L. Wright, C. Haley, M. Lou Beckett, and P. F. Schellhammer, “Expression of prostate-specific membrane antigen in normal, benign, and malignant prostate tissues,” Urol. Oncol. Semin. Orig. Investig., vol. 1, no. 1, pp. 18–28, 1995.

17. S. D. Sweat, A. Pacelli, G. P. Murphy, and D. G. Bostwick, “Prostate-specific membrane antigen expression is greatest in prostate adenocarcinoma and lymph node metastases,” Urology, vol. 52, no. 4, pp. 637–640, 1998.

18. S. Perner M.D. Hofer, R. Kim, R.B. Shah, H. Li, P. Möller, R.E. Hautmann, J.E. Gschwend, R. Kuefer and M.A. Rubin “Prostate-specific membrane antigen expression as a predictor of prostate cancer progression,” Hum. Pathol., vol. 38, no. 5, pp. 696–701, 2007.

19. S. Minner, C. Wittmer, M. Graefen, G. Salomon, T. Steuber, A. Haese, H. Huland, C. Bokemeyer, E. Yekebas, J. Dierlamm,, “High level PSMA expression is associated with early psa recurrence in surgically treated prostate cancer,” Prostate, vol. 71, no. 3, pp. 281–288, Feb. 2011.

20. M. C. Hupe, C. Philippi, D. Roth, C. Kümpers, J. Ribbat-Idel, F. Becker, V. Joerg, S. Duensing, V. H. Lubczyk, J. Kirfel et al, “Expression of Prostate-Specific Membrane Antigen (PSMA) on Biopsies Is an Independent Risk Stratifier of Prostate Cancer Patients at Time of Initial Diagnosis ,” Frontiers in Oncology , vol. 8. p. 623, 2018.

21. S. Bravaccini, M. Puccetti, M. Bocchini, S. Ravaioli, M. Celli, E. Scarpi, U. De Giorgi, M. M. Tumedei, G. Raulli, L. Cardinale, and G. Paganelli, “PSMA expression: a potential ally for the pathologist in prostate cancer diagnosis,” Sci. Rep., vol. 8, no. 1, p. 4254, 2018.

22. C. D. Bahler, M. Green, G. D. Hutchins, L. Cheng, M. J. Magers, J. Fletcher and M. O. Koch, “Prostate Specific Membrane Antigen Targeted Positron Emission Tomography of Primary Prostate Cancer: Assessing Accuracy with Whole Mount Pathology,” J. Urol., vol. 203, no. 1, pp. 92–99, Jan. 2020.

23. K. Rahbar, M. Weckesser, S. Huss, A. Semjonow, H. Breyholz, A. J. Schrader, M. Schäfers and M. Bögemann, “Correlation of Intraprostatic Tumor Extent with & 68-Ga-PSMA Distribution in Patients with Prostate Cancer,” J. Nucl. Med., vol. 57, no. 4, pp. 563 LP – 567, Apr. 2016.

24. W. P. Fendler , D. F. Schmidt, V. Wenter, K. M. Thierfelder, C. Zach, C. Stief, P. Bartenstein, T. Kirchner, F. J. Gildehaus, C. Gratzke and Claudius Faber 6., 68-PSMA PET/CT Detects the Location and Extent of Primary Prostate Cancer," J. Nucl. Med., vol. 57, no. 11, pp. 1720 LP – 1725, Nov. 2016.

25. C. Uprimny, A. S. Kroiss, C. Decristoforo, J. Fritz, E. von Guggenberg, D. Kendler, L. Scarpa, G. di Santo, L. G. Roig, J. Maffey-Steffan et al, "68Ga-PSMA-11 PET/CT in primary staging of prostate cancer: PSA and Gleason score predict the intensity of tracer accumulation in the primary tumour," Eur. J. Nucl. Med. Mol. Imaging, vol. 44, no. 6, pp. 941–949, 2017.

26. S. A. Koerber, M. T. Utzinger, C. Kratochwil, C. Kesch, M. F. Haefner, S. Katayama, W. Mier, A. H. Iagaru, K. Herfarth, U. Haberkorn et al, "68Ga-PSMA-11 PET/CT in Newly Diagnosed Carcinoma of the Prostate: Correlation of Intraprostatic PSMA Uptake with Several Clinical Parameters," J. Nucl. Med., vol. 58, no. 12, pp. 1943 LP – 1948, Dec. 2017.

27. N. Woythal, R. Arsenic, C. Kempkensteffen, K. Miller, J. Janssen, K. Huang, M. R. Makowski, W. Brenner and V. Prasad "Immunohistochemical Validation of PSMA Expression Measured by 68Ga-PSMA PET/CT in Primary Prostate Cancer," J. Nucl. Med., vol. 59, no. 2, pp. 238 LP – 243, Feb. 2018.

28. I. Berger, C. Annabattula, J. Lewis, D. V. Shetty, J. Kam, F. Maclean, M. Arianayagam, B. Canagasingham, R. Ferguson, M. Khadra, R. Ko et al, "68Ga-PSMA PET/CT vs. mpMRI for locoregional prostate cancer staging: correlation with final histopathology," Prostate Cancer Prostatic Dis., vol. 21, no. 2, pp. 204–211, 2018.

29. S. Dekalo, J. Kuten, N. J. Mabjeesh, A. Beri, E. Even-Sapir, and O. Yossepowitch, "68Ga-PSMA PET/CT: Does it predict adverse pathology findings at radical prostatectomy?," Urol. Oncol. Semin. Orig. Investig., vol. 37, no. 9, pp. 574.e19-574.e24, 2019.

30. E. Lopci, A. Saita, M. Lazzeri, G. Lughezzani, P. Colombo, N. M. Buffi, R. Hurle, K. Marzo, R. Peschechera and A Benetti, "68Ga-PSMA Positron Emission Tomography/Computerized Tomography for Primary Diagnosis of Prostate Cancer in Men with Contraindications to or Negative Multiparametric Magnetic Resonance Imaging: A Prospective Observational Study," J. Urol., vol. 200, no. 1, pp. 95–103, Jul. 2018.

31. E. Demirci, L. Kabasakal, O. E. Şahin, E. Akgün, M. H. Gültekin, T. Doğanca, M. B. Tuna, C. Öbek, M. Kiliç and T. Esen, "Can SUVmax values of Ga-68-PSMA PET/CT scan predict the clinically significant prostate cancer?," Nucl. Med. Commun., vol. 40, no. 1, pp. 86–91, Jan. 2019.

32. J. H. Rüschoff, D. A. Ferraro, U. J. Muehlematter, R. Laudicella, T. Hermanns, A. Rodewald, H. Moch, D. Eberli, I. A. Burger, and N. J. Rupp "What's behind 68Ga-PSMA-11 uptake in primary prostate cancer PET? Investigation of histopathological parameters and immunohistochemical PSMA expression patterns," Eur. J. Nucl. Med. Mol. Imaging, vol. 48, no. 12, pp. 4042–4053, 2021.

33. S. Satapathy, H. Singh, R. Kumar, and B. R. Mittal, "Diagnostic Accuracy of 68Ga-PSMA PET/CT for Initial Detection in Patients With Suspected Prostate Cancer: A Systematic Review and Meta-Analysis," Am. J. Roentgenol., vol. 216, no. 3, pp. 599–607, Jan. 2021.

34. D. Margel, H. Bernstine, D. Groshar, Y. Ber, O. Nezrit, N. Segal, M. Yakimov, J. Baniel and L. Domachevsky “Diagnostic Performance of 68Ga Prostate-specific Membrane Antigen PET/MRI Compared with Multiparametric MRI for Detecting Clinically Significant Prostate Cancer,” Radiology, vol. 301, no. 2, pp. 379–386, Aug. 2021.

35. S. Y. Park, C. Zacharias, C. Harrison, R. E. Fan, C. Kunder, N. Hatami, F. Giesel, P. Ghanouni, B. Daniel and A. M. Loening., “Gallium 68 PSMA-11 PET/MR Imaging in Patients with Intermediate- or High-Risk Prostate Cancer,” Radiology, vol. 288, no. 2, pp. 495–505, May 2018.

36. A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, V. E. Reuter, “Haralick texture analysis of prostate MRI : utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores,” vol. l, pp. 2840–2850, 2015.

37. D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones and L. Court “Measuring Computed Tomography Scanner Variability of Radiomics Features.,” Invest. Radiol., vol. 50, no. 8, pp. 1–9, 2015.

38. R. T. H. Leijenaar, S. Carvalho, E. R. Velazquez, W. J. C. van Elmpt, C. Parmar, O. S. Hoekstra, C. J. Hoekstra, R. Boellaard, A. L. A. J. Dekker, R. J. Gillies, H. J. W. L. Aerts,, “Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability.,” Acta Oncol., vol. 52, no. 7, pp. 1391–7, Oct. 2013.

39. R. T. H. M. Larue, J. E. van Timmeren, E. E. C. de Jong, G. Feliciani, R. T. H. Leijenaar, W. M. J. Schreurs, M. N. Sosef, F. H. P. J. Raat, F. H. R. van der Zande, M. Das 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. (Madr)., vol. 56, no. 11, pp. 1544–1553, 2017.

40. A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, “Image biomarker standardisation initiative,” 2016.

41. RStudio Team, “RStudio: integrated development environment for R.” Boston, MA, 2015

42. E. L. Solari, A. Gafita, S. Schachoff, B. Bogdanović, A.V. Asiares, T. Amiel, W. Hui, I. Rauscher, D. Visvikis, T. Maurer et al The added value of PSMA PET/MR radiomics for prostate cancer staging. Eur. J. Nucl. Med. Mol. Imaging 49, 527–538 (2022).

43. F. Midiri, F. Vernuccio, P. Purpura, P. Alongi and T. V. Bartolotta, “Multiparametric MRI and Radiomics in Prostate Cancer : A Review of the Current Literature” Diagnostics vol. 11 (2021).

44. D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, H. Vargas, E. Sala, H. Hricak, J.O. Deasy et al “Automaticclassification of prostate cancer Gleason scores from multiparametric magnetic resonance images.” Proc. Natl. Acad. Sci. USA 2015, 112, E6265–E6273

Funding: This research was partly funded by AIRC Investigator Grant, grant number IG 20476, Ministry of Health Ricerca Finalizzata, grant number RF-2016-02364230 and Fondazione Cassa di Risparmio di Cesena

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board CEIIAV (protocol code IRST185.05 and date of approval: 15/11/2017).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All raw data employed for the development of logistic regression models are available upon reasonable request to the corresponding author.

Conflicts of Interest: "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

CHAPTER 6

Radiomics in the characterization of lipid-poor adrenal adenomas at unenhanced CT: time to look beyond usual density metrics

Published in: **European Radiology** 2023 Aug 11; (11)

Radiomics in the characterization of lipid-poor adrenal adenomas at unenhanced CT: time to look beyond usual density metrics

Giacomo Feliciani, Francesco Serra, Enrico Menghi,
Fabio Ferroni, Anna Sarnelli, Carlo Feo, Maria Chiara Zatelli, Maria Rosaria Ambrosio,
Melchiore Giganti, Aldo Carnevale

ABSTRACT

Objectives: In this study, we developed a radiomic signature for the classification of benign lipid-poor adenomas, which may potentially help clinicians limit the number of unnecessary investigations in clinical practice. Indeterminate adrenal lesions of benign and malignant nature may exhibit different values of key radiomic features.

Methods: Patients who had available histopatology reports and a non-contrast enhanced CT scan were included in the study. Radiomic feature extraction was done after the adrenal lesions were contoured. The primary feature selection and prediction performance scores were calculated using the Least absolute shrinkage and selection operator (LASSO). To eliminate redundancy, the best performing features were further examined using the Pearson correlation coefficient, and new predictive models were created.

Results: This investigation covered 50 lesions in 48 patients. After LASSO-based radiomic feature selection, the test dataset's 30 iterations of logistic regression models produced an average performance of 0.72. The model with the best performance, made up of 13 radiomics features, had an AUC of 0.99 in the training phase and 1.00 in the test phase. The number of features was lowered to 5 after performing Pearson correlation to prevent overfitting. The final radiomic signature trained a number of machine learning classifiers, with an average AUC of 0.93.

Conclusions: Including more radiomic features in the identification of adenomas may improve the accuracy of NECT and reduce the need for additional imaging procedures and clinical workup, according to this and other recent Radiomics studies that have clear points of contact with current clinical practice.

Clinical relevance statement: The study developed a radiomic signature using unenhanced CT scans for classifying lipid-poor adenomas, potentially reducing unnecessary investigations. After feature selection and correlation, a final signature using 5 radiomic features had an average AUC of 0.93. The study suggests that incorporating more radiomic features may improve accuracy and reduce the need for additional imaging procedures.

Key points

- Radiomics has potential for differentiating lipid poor adenomas and avoiding unnecessary further investigations.
- Quadratic mean, strength, maximum 3D diameter, volume density and area density are promising predictors for adenomas.
- Radiomics models reach high performance with average AUC of 0.95 in training phase and 0.72 in test phase.

Keywords

Abdomen; Adrenocortical Adenoma; Adrenal incidentaloma; X-Ray Computed Tomography; Artificial Intelligence.

Abbreviations

AUC - area under the curve

IBSI - Image Biomarker Standardisation Initiative

LASSO - least absolute shrinkage and selection operator

NECT - non-enhanced CT

ROC - receiver operating curve

ROI - region of interest

VOI - volume of interest

1. INTRODUCTION

An adrenal incidentaloma is defined as an asymptomatic adrenal mass discovered on imaging that was not performed to investigate a suspected adrenal disease [1]. In most cases, adrenal incidentalomas represent benign non-functioning adenomas, but they may also correspond to different conditions requiring full clinical attention and therapeutic intervention (e.g. adrenocortical carcinoma, pheochromocytoma, hormone-producing adenoma or metastasis). As a consequence of the burgeoning use of advanced diagnostic imaging in daily medical practice, in the last decades, we have observed a constantly increasing incidence rate of incidentally discovered adrenal nodules. Indeed, adrenal incidentalomas are common, estimated to occur in approximately 3% to 7% of adults [2, 3].

Incidental adrenal masses represent diagnostic challenges for both radiologists and referring clinicians, particularly when the initial imaging features are equivocal or non-diagnostic. The main challenge is correctly identifying the infrequent unexpected malignant lesions (or hyperfunctioning adenomas), while sparing the vast majority of patients with clinically insignificant disease from unnecessary further examinations.

Diagnostic imaging is crucial in the classification of adrenal masses, since the precise etiology can be determined on both computed tomography (CT) and magnetic resonance imaging (MRI) for several entities without the need for further tests [1, 4].

In particular, CT could aid the diagnosis of adrenal adenomas in two ways, namely density measurement and contrast washout. A density lesser than 10 HU on non-enhanced CT (NECT) is almost always diagnostic of a lipid-rich adenoma, regardless of size [2]. By contrast, if there are not benign diagnostic imaging features (for instance, macroscopic fat, adrenal density <10 HU), a dedicated adrenal CT protocol including a 15-minutes delayed acquisition after contrast media administration is advisable, in order to assess the absolute - or relative - percentage washout. However, pheochromocytomas and adrenal metastases from hypervascular primary extra-adrenal malignancies could sometimes exhibit a washout pattern similar to that of adrenal adenomas [5–8].

Other imaging modalities may be useful to clarify the nature of the nodule, in particular MRI, in which a signal loss between in- and opposed-phase images at chemical-shift imaging is diagnostic of adenoma, or positron emission tomography (PET)-CT, in which most adenomas show FDG uptake less than 3.1 [9].

However, the need for additional tests puts patients at risk of anxiety and unnecessary harm from diagnostic procedures; additionally, the costs incurred can be significant.

Radiomics refers to a rapidly emerging discipline based on the extraction of mineable data from medical imaging. It has been used in oncology to support the diagnosis, prognostication, and clinical decision making, with the goal of delivering precision medicine [10–13].

In recent research, O'Shea A. et al. and Cao L. et al. [14, 15] demonstrated that early-stage metastases may be differentiated from lipid-poor adenomas using contrast-enhanced CT and NECT radiomic features-based models with high performances. In other research, radiomics was used to distinguish lipid-poor adenomas from paragangliomas, phrochromocytomas or carcinomas [16, 17].

To discriminate lipid-poor adenomas from other adrenal lesions, Zhang et al.[18] recently developed three prediction models using conventional, Radiomics, and combined feature nomograms. However, there was no significant difference in performance between the radiomic and traditional models.

In this study, we retrospectively assessed a dataset of adrenal masses with pathological confirmation that had been classified at NECT as indeterminate and that had not been distinguished by standard clinical demographic or radiological characteristics.

We hypothesized that indeterminate adrenal lesions of benign and malignant nature may exhibit different values of key radiomic features, and we developed a radiomic signature for the classification of benign lipid-poor adenomas, which may potentially help clinicians limit the number of unnecessary investigations in clinical practice.

2. MATERIALS AND METHODS

This retrospective study was conducted according to the Declaration of Helsinki; local Ethics Committee approval for data collection was obtained (Ethics Committee of Area Vasta Emilia Centrale (AVEC); protocol code: 146/2022/Oss/AOUFe, approved on 17/02/2022). All investigations were performed by routine clinical practice and retrospectively retrieved.

Population

Hospital discharge form (Scheda di Dimissione Ospedaliera – SDO) database of Sant'Anna University Hospital of Ferrara was searched to find all ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) coded interventions that included surgical

resection of unilateral or both adrenal glands, between January 2003 and December 2018, independently by clinical suspicion or diagnosis.

A total of 251 patients that underwent adrenalectomy were identified.

All histopathology reports were reviewed, aiming to exclude patients with large infiltrating lesions (adrenalectomy done as “en-bloc” resection with other near tissues and organs during large retroperitoneal tumour debulking) or adrenal cortical hypertrophy, thus including only focal adrenal lesions. Subjects who were missing a complete histopathologic electronic report in our Institutional Pathology Database were excluded from further evaluation. Lesions with a maximum diameter less than 1 cm were not included in this study.

Preoperative radiologic imaging data were retrieved by querying the institutional Radiological Information Systems - Picture Archiving and Communication system (RIS-PACS; Philips VuePACS, Philips Medical Systems), and only the patients for whom a non-enhanced CT (NECT) scan was available were included.

CT data

Each CT examination was reviewed independently by two abdominal radiologists, with 3 and 10 years of experience respectively, who were blinded to the patients’ pathological data, in order to exclude the lesions with gross fat component or showing median attenuation less than 10 HU, by applying a single region of interest (ROI) encompassing more than 50% of the target lesion in the axial plane demonstrating the maximal lesion extent. Disagreements between the readers were resolved through consensus.

The CT studies analysed after application of the inclusion and exclusion criteria were acquired on 4 different multidetector scanners: Philips Brilliance 64 (Philips Medical Systems), GE Lightspeed VCT (GE Healthcare), Philips iCT 256 (Philips Medical Systems), Siemens Biograph 64 (Siemens Medical solutions).

The CT examinations were acquired using standard acquisition parameters adjusted to patients’ biometrics and accordingly to the purpose of the investigation (10–400 effective mAs, 120 kVp, 1.375-1.75 pitch and 1.5-3mm slice reconstruction thickness). Images were reconstructed using a standard soft tissue kernel used in clinical practice (namely for GE – standard, Philips – B. Siemens – Qr40).

Imaging analysis and segmentation

A last-year radiology resident retrieved the CT images from the RIS-PACS database, fully anonymized and de-identified, in DICOM (Digital Imaging and Communications in Medicine) format. First-pass segmentations were manually performed by the same radiology fellow who contoured the adrenal lesions using 3DSlicer software with SlicerRT extension [19] on each axial image, finally obtaining a three-dimensional contoured volume of interest (VOI).

The contoured volumes had to contain the whole adrenal mass, including the edges but avoiding the peri-adrenal soft tissues (fat, vessels, and the parenchyma of adjacent organs) (**Figure 1**).

The appropriateness of the contouring process was determined by the same two experienced abdominal radiologists.

The segmented VOIs were subsequently exported as DICOM files with the RT option enabled from the SlicerRT extension.

Image Pre-Processing and Radiomic Analysis

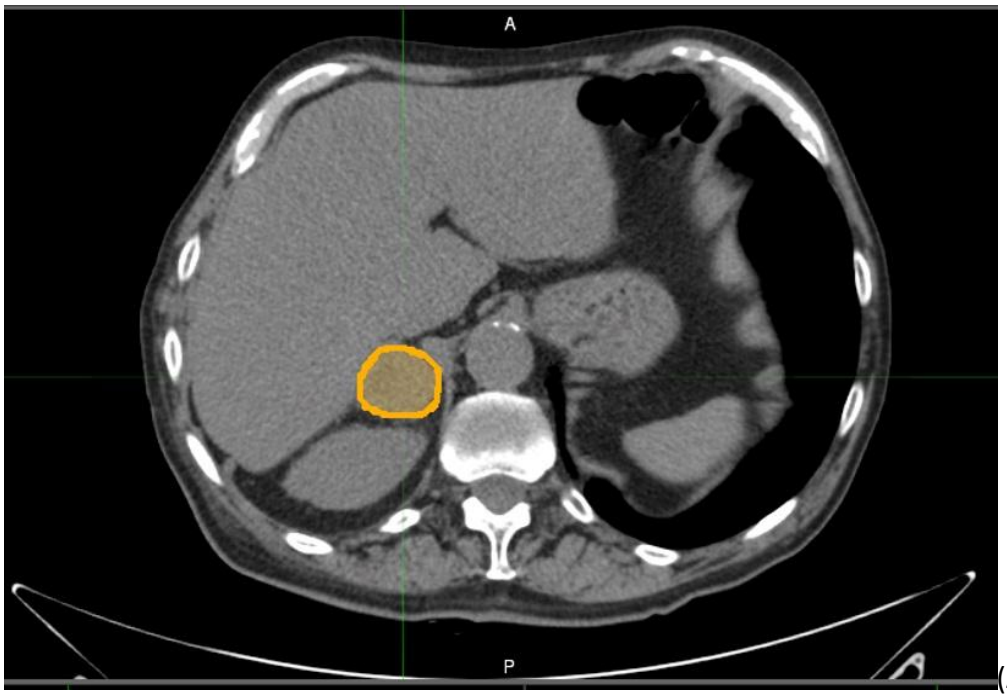
Following manual segmentation, images were exported to the Image Biomarker Standardisation Initiative (IBSI) [16] compliant software SOPHiA DDMTM Radiomics (Sophia Genetics). The patients' CT images were resampled to a resolution of 1/1/1 mm to standardize the dataset, and grey-level quantization was performed at 32 bins prior to radiomic analysis.

Radiomics analysis software extracted 209 imaging features for each segmented volume.

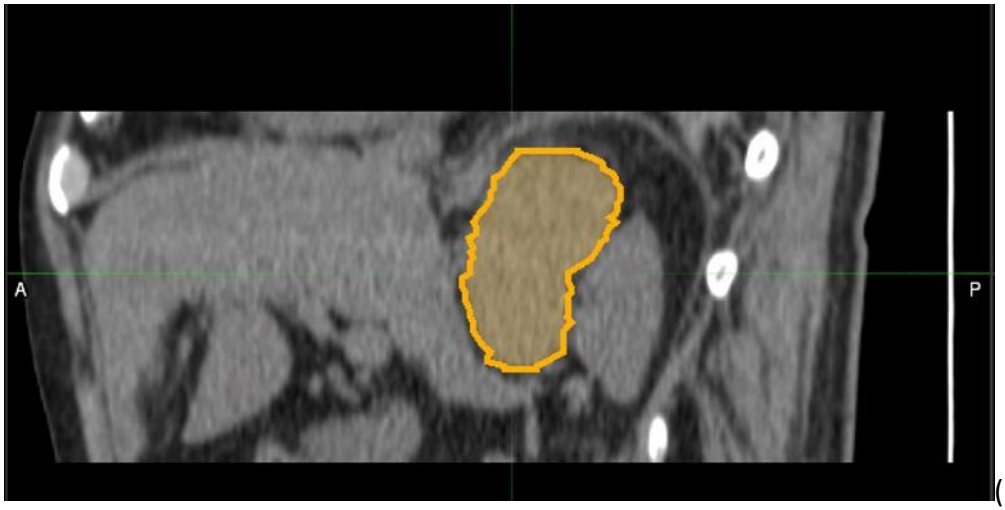
Features included first-, second- and higher-order features. The histogram of voxel intensities was employed to calculate first-order features. Intensity size-zone, co-occurrence, and run-length based matrices were used to calculate second- and higher-order features. The IBSI Reference Handbook contains a detailed description of the 209 imaging features extracted [20].

Statistical analysis

The endpoint of this study was to investigate the diagnostic performance of radiomic features extracted from patients' CT images to differentiate between pathologically proven adenomas (labelled from now on as 0) and other adrenal histotypes (labelled from now on as 1). Clinical and demographical characteristics of the cohort were analysed with a multivariable logistic regression for the endpoint of this study. We summarize the entire process of Radiomics analysis from feature extraction to statistical model evaluations in **Figure 2**.



(A)



(B)

(C)

Figure 1: Segmentation process performed on a nodule in the right adrenal gland showed respectively in the axial (**A**), sagittal (**B**) and coronal planes (**C**).

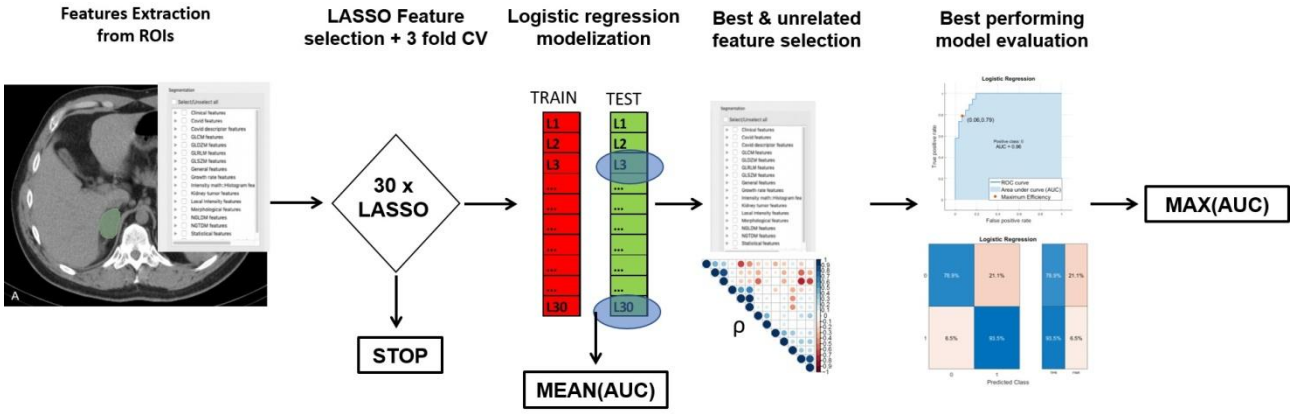


Figure 2: Radiomics and statistical workflow from features extraction to selection of the best performing machine learning models.

To assess the models' performance, we employed a stochastic cross-validation technique. The lesion feature datasets were separated into a training (2/3) and test (1/3) set during the modelling process. The training set was then used to train a logistic regression model using features picked using a least absolute shrinkage and selection operator (LASSO) technique with internal 3-fold cross validation with the objective of maximizing the distinction between adenomas and non-adenomas.

On the test set, the predictive ability of the model was calculated. This procedure was repeated 30 times, with each iteration's receiver operating curves (ROC) and area under the curve (AUC) being recorded for both the training and test sets. The average value of the latter ones was then used to assess the overall diagnostic performance of the model. The features that performed better in the best model of the test phase were further processed with Pearson ρ correlation coefficient to remove redundancy setting a threshold of 0.80. As final feature selection method, we employed the selection frequency of LASSO in the 30 repetition. In the end, we built four machine learning models (logistic regression, linear discriminant, support vector machine, and decision tree) with a fixed number of lesions per feature (10 lesions per feature included in the model) to avoid overfitting on the whole lesion dataset. The performance of the best model was evaluated with calibration and decision curve to assess the consistency of the classification and its clinical usefulness.

Data availability

Radiomics features extracted from the 50 lesions and corresponding status (adenoma/non-adenoma) that were used to develop the models are available in supplementary material 1. CT images of the patients are available upon reasonable request to the corresponding author.

3. RESULTS

Patients' selection workflow is shown in **Figure 3**.

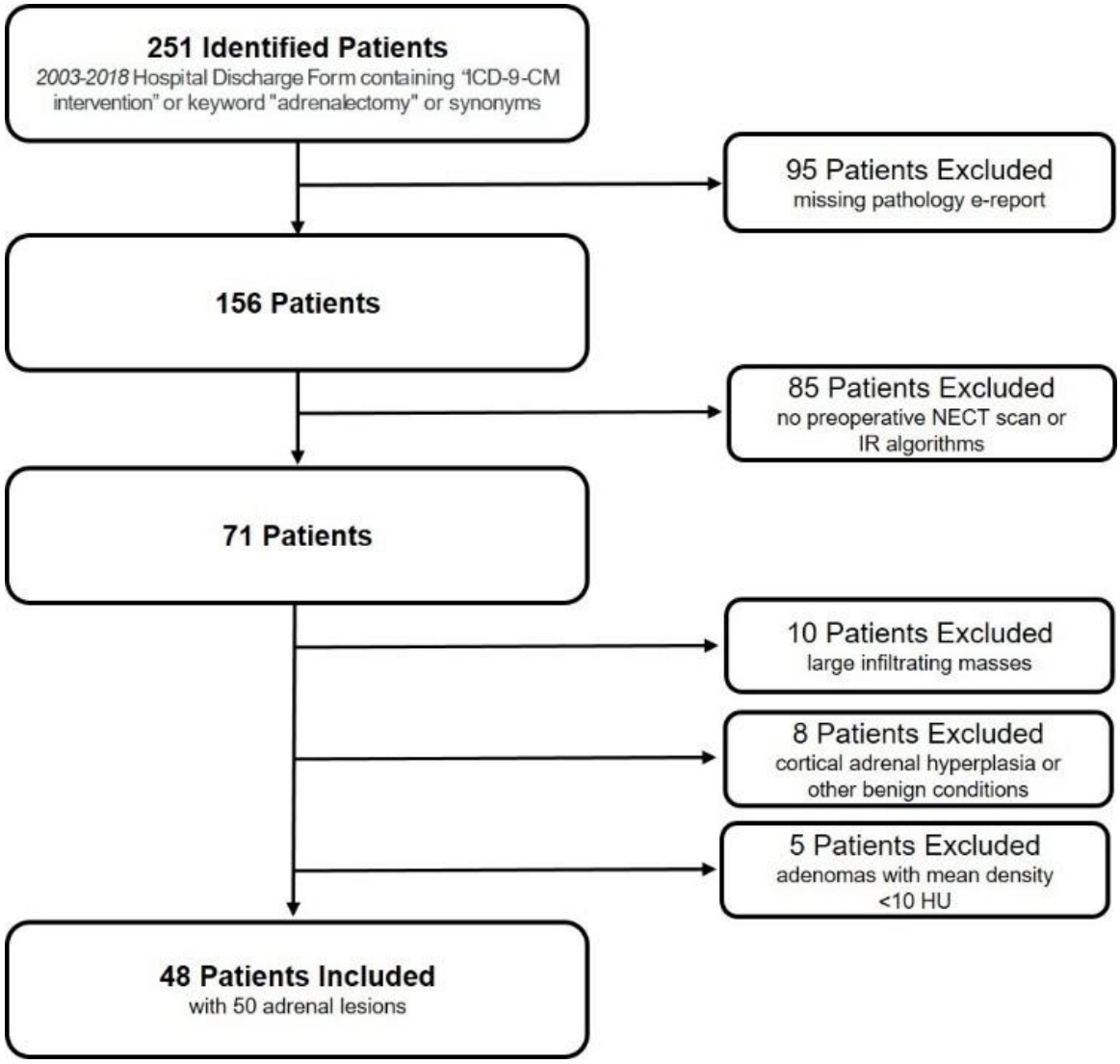


Figure 3: flowchart of patients' selection.

The final study population consisted of 48 patients (26 males, 22 females) accounting for 50 lesions (24 in the female population, 26 in male). The age of patients ranged between 27 and 86 years old, with an average age of 72 for woman and 70 for men.

In details, from the initial dataset of 251 patients who underwent adrenalectomy in our Institution we excluded from further analysis: 95 patients for missing histopathologic digital report; 85 patients due to missing NECT before surgery; 4 patients for cortical adrenal hyperplasia; 10 patients for infiltrating masses; 4 because of other benign conditions (i.e. haematomas, cysts);

only 5 patients were ineligible because their NECT scans could reliably diagnose adenomas with a mean density lower than 10 HU.

The histopathological classification of the lesions was the following: 19 adenomas (38%), 9 pheochromocytomas (18%), 5 adrenal carcinomas (10%), 7 myelolipomas (14%), 8 metastases (16%), 1 mesothelioma and 1 cavernous hemangioma (4%). Patients had an average lesion diameter of 5.5 cm with a minimum diameter of 1.5 cm and a maximum of 14.7. The characteristics of the patients and lesions included in the final analysis are summarized in **Table 1**.

Patients (n)	48
females; males (n)	22; 26
age (average; range, y)	61; 27 - 86
Lesions (n)	50
diameter (average; range, cm)	5.5; 1.5 - 14.7
laterality	
monolateral (n; %)	46; 92%
bilateral (n; %)	4; 8%
histology (n; %)	
adenoma	19; 38%
pheocromocytoma	9; 18%
metastasis	8; 16%
adrenal carcinomas	5; 10%
myelolipoma	7; 14%
other histology	2; 4%

Table 1: characteristics of patients and lesions included in the final analysis.

Multivariable logistic regression on clinical, demographical, and radiological characteristics of the patients is shown in **Table 2** demonstrating that none of these parameters are associated with the outcome.

	Multivariate Logistic regression			
	p-Value	HR	95% C.I. of HR	
			inf.	Sup.
Max3Ddiameter	0.189	1.230	0.903	1.673
MeanHU	0.900	1.002	0.976	1.028
Age	0.382	1.020	0.975	1.067
M/(F)	0.564	0.688	0.193	2.454

Table 2: Multivariable logistic regression of clinical (Age, Sex) and standard radiological characteristics (maximum 3D diameter and mean HU). HR – Hazard ratio.

The 30 logistic regression models trained by LASSO resulted in an average AUC of 0.95 (0.81-1.00) (excluding repetitions when the algorithm didn't reach convergence (5 times). On the test set, the models had an average AUC of 0.72 (0.48-1.00). The best performing logistic regression model had an AUC of 0.99 in the training phase and 1.00 in the test phase and was composed of 13 features.

In **Figure 4**, we show Pearson correlation coefficient results and features with a Rho > 0.80 were eliminated. To prevent overfitting, only the top 5 informative features from the 30 LASSO iterations (quadratic mean, strength, maximum 3D diameter, volume density, and area density) were retained for further analysis.

In the end, we trained the 4 final models on the entire lesion dataset employing logistic regression, linear discriminant, support vector machine, and decision tree as classifiers, obtaining an AUC of 0.95, 0.94, 0.91 and 0.96, as can be seen in **Figure 5**. True Positive Rates and False Negative Rate and other classification performances can be appreciated in the confusion matrices of the models reported in **Figure 6**. Calibration and decision curves of the logistic regression model with the comparison with standard clinical parameters are available as supplementary material 2.

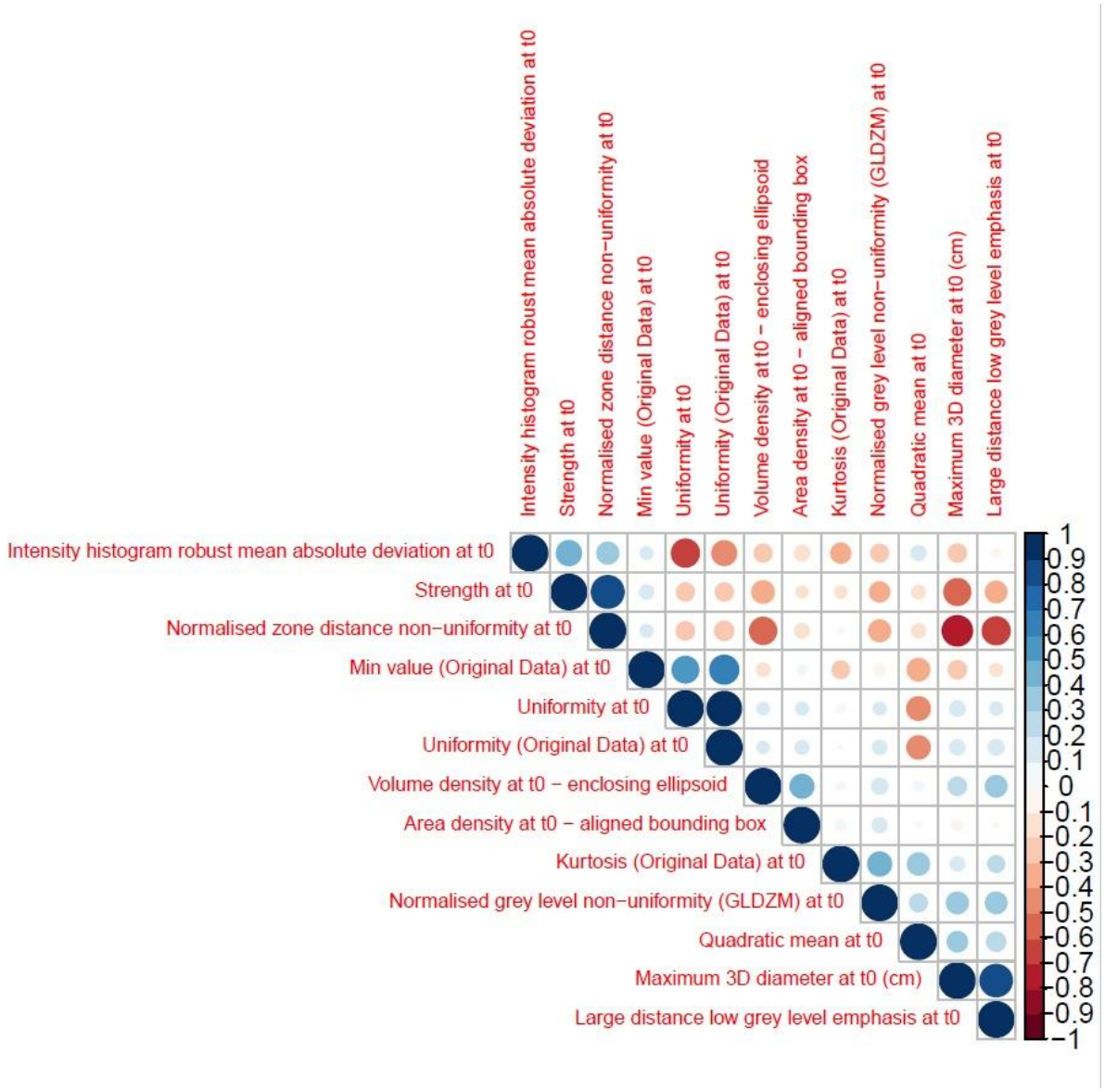


Figure 4: Pearson ρ correlation coefficient of the 13 features selected by LASSO in the best performing model among the 30 iteration.

4. DISCUSSION

Physicians' desire for diagnostic certainty and, on the other hand, discomfort with diagnostic uncertainty when faced with an unexpected or unexplained imaging finding can lead to an increase in test ordering. As a result, further imaging and clinical evaluation are often performed when an adrenal incidentaloma is discovered and when imaging findings are equivocal or inconclusive [2, 21, 22].

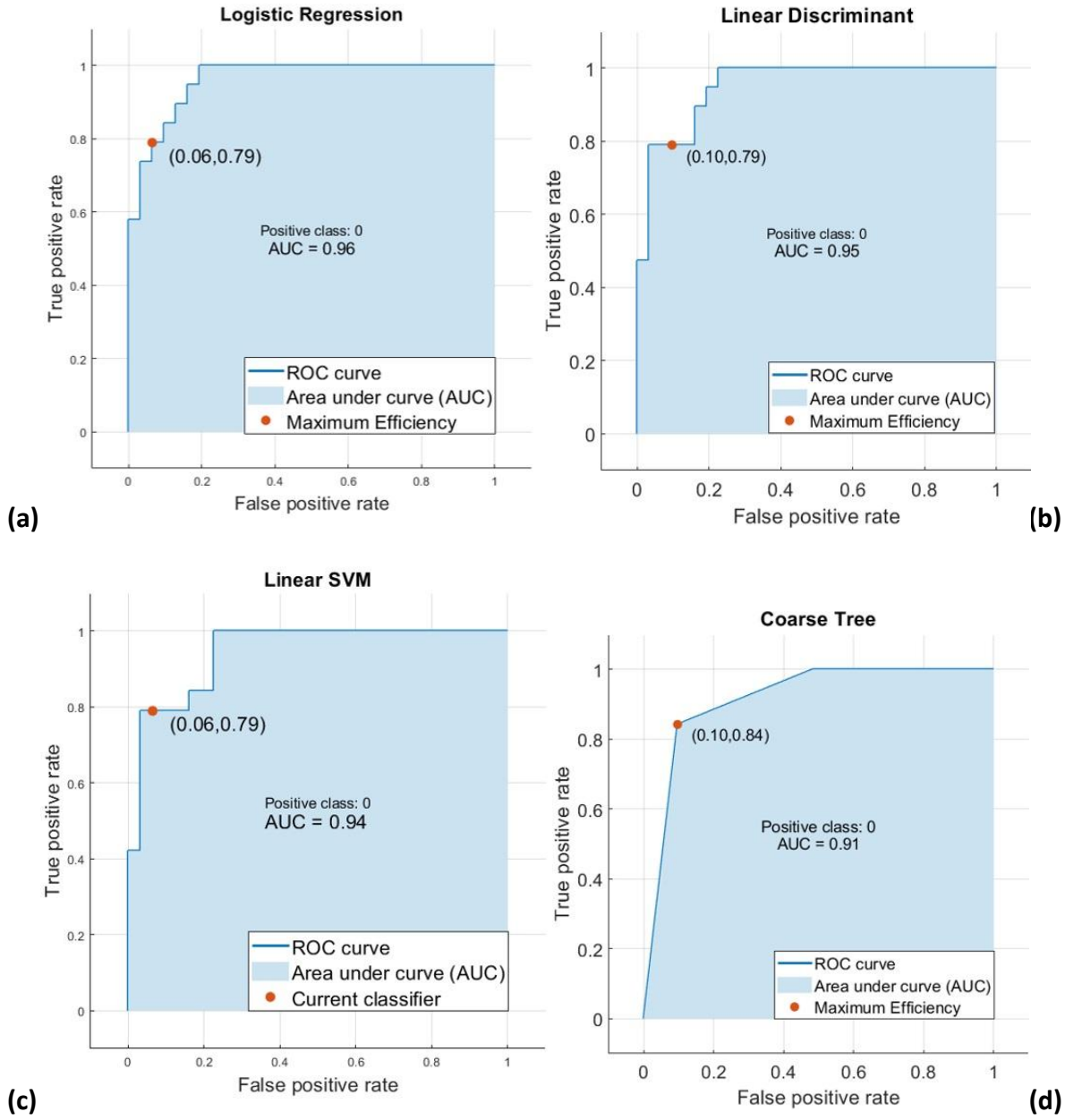


Figure 5: Model performances in terms of ROC curves and AUC for **(a)** Logistic regression, (**b**) Linear Discriminant, (**c**) Linear SVM, (**d**) Coarse Tree

In the present study, a radiomic signature composed of first- and higher-order features, namely quadratic mean, strength, maximum 3D diameter, volume density, and area density, showed a very good average performance with AUC = 0.94 (0.91-0.96) among the four final classifiers to discriminate adenomas from other adrenal lesions at NECT.

The performances of our machine learning models did not differ significantly, with logistic regression showing the best results with an AUC of 0.96 (**Figure 5a**). The True Positive Rate for adenomas, according to the model, was 79%, whereas it was higher (93.5%) for non-adenomas, as

shown in **Figure 6a**. The decision tree performed better for both classes, with a True Positive Rate of 84.2% and 90.3% (**Figure 6d**).

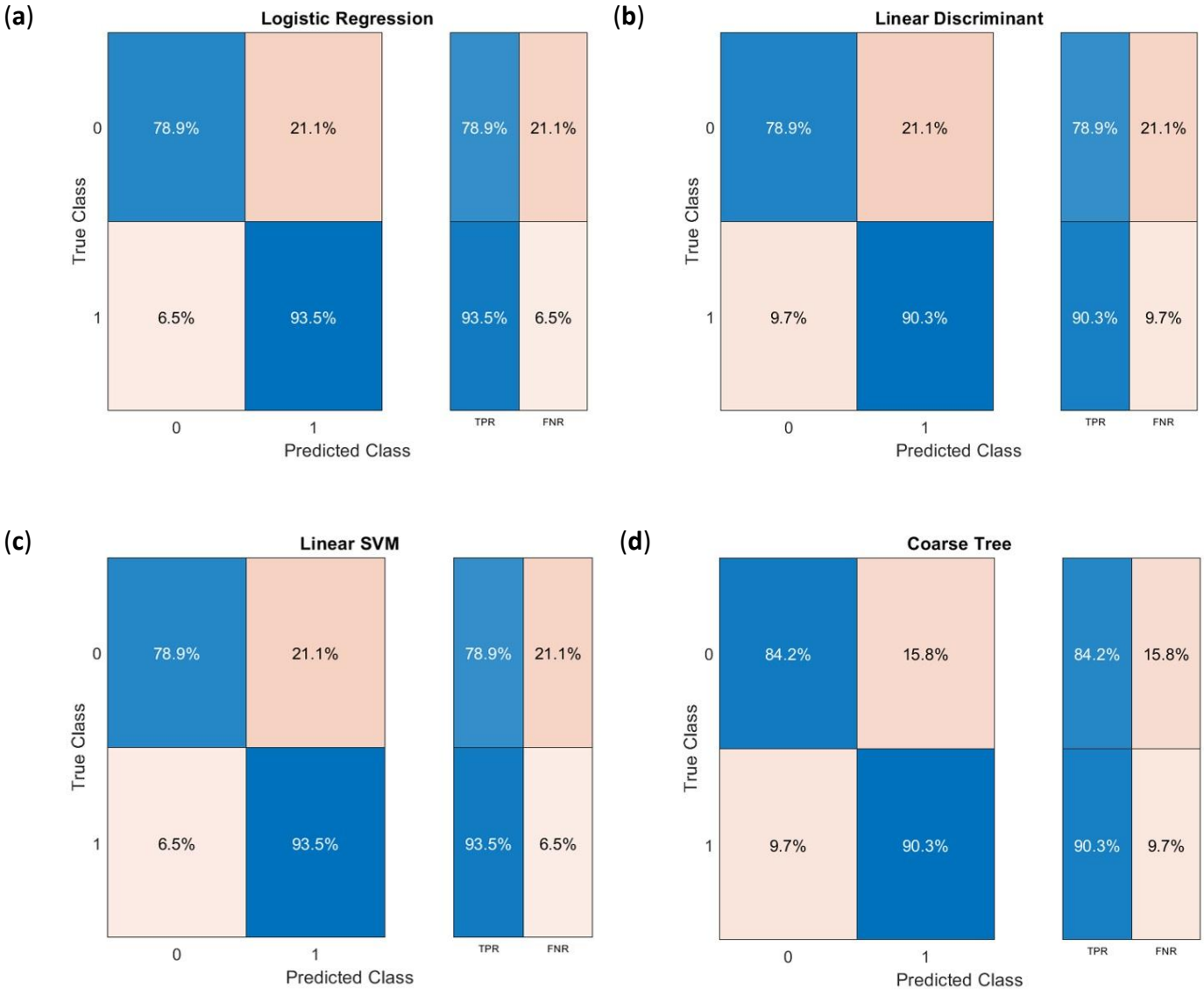


Figure 6: Model performances in terms of confusion matrices and TPR and FNR for (**a**) Logistic regression, (**b**) Linear Discriminant, (**c**) Linear SVM, (**d**) Coarse Tree.

Our signature's performance is comparable to that of the three models developed by Zhang et al. [18] for differentiating lipid-poor adenomas using conventional, radiomic, and integrated conventional-radiomic CT features. These models had an AUC of 0.94, 0.93, and 0.96, respectively. However, in their cohort, conventional parameters such as gender, age, mean HU, and tumor diameter were strong predictors of the outcome at both univariable and multivariable logistic regression. As a result, their radiomic signature did not significantly improve the

performance of the conventional model, i.e. employing standard radiological features and demographic data, thereby diminishing the scientific impact of their results.

Conversely, the net benefit of using our radiomic signature in comparison to standard parameters is clearly visible from the decision curves shown in supplementary figure 2. In our opinion, these differences are easily addressed by the different cohorts of patients used to train and test the models in the two studies, which may have led to some selection biases while considering broader inclusion criteria for eligible lesions in the aforementioned work. As shown in table 2, indeed, no standard parameters in our cohort were statistically significant predictors of the outcome using multivariable regression.

The composition of the radiomic signature reveals additional distinctions. Indeed, our signature is composed of quadratic mean, strength, maximum 3D diameter, volume density, and area density. Quadratic mean is a first-order feature derived from histogram that has fair correlation to the HU median (or mean) value, which is found in the work of Zhang et al., Cao et al., and O'Shea et al. for the differentiation of lipid-poor adenomas from other histotypes.

Strength is a more complex second order feature that is related to the texture of the image; in particular, it can be correlated to the concepts of coarseness, as specifically described by Amadasun and King [23]. In this context, a high strength means that the patterns that composes the texture of the tumor appears more large with broader areas of uniform pixel intensities whereas a low strength would correspond to a finer texture leading to higher variations in local pixel intensities. To our knowledge, this predictor has not been investigated in any other published study.

Maximum 3D diameter is a parameter already employed in clinical practice and reported in previous studies cited above. In the end, area and volume density are related to the shape and extent of the tumors and may provide additional information on their morphological appearance. In fact, adenomas present more frequently as well-demarcated round or oval lesions [9]. These two parameters likely reflect and quantify these visual characteristics of the tumor that were not quantified in previous studies or were only partially considered when the greatest or shortest diameters and tumor volume were used [15]. Our findings suggest that additional metrics, beyond the mere measurement of the mean density, should be considered for inclusion in routine radiological evaluation of adrenal lesions to reduce the number of incidentalomas regarded as indeterminate at NECT examination, thus avoiding unnecessary clinical workup and follow-up examinations.

It is known that imaging has a significant impact on the clinical management of patients and is crucial in determining whether an adrenal tumour is benign or malignant. The use of a radiomic signature in clinical practice, particularly in the case of an incidental adrenal nodule discovered in a non-dedicated CT examination, such as a chest high-resolution CT, may be extremely useful in reducing the number of unnecessary tests and, as a result, in containing health costs. Indeed, one of the biggest challenges in medicine is the development of accurate, cost–effective tools with the end goal of personalized patient management.

At NECT, adenomas present more frequently low attenuation (less than 10 HU) due to a microscopic fat component. This cut-off is highly specific (sensitivity 71%, specificity 98%) [4, 6], widely accepted in the scientific literature, and routinely employed in radiological practice [24, 25]. However, NECT alone is not always diagnostic, since 15–30% of adenomas are lipid-poor, namely contain insufficient intracytoplasmic lipid to conform to the non-contrast features previously described, thereby demonstrating higher attenuation values [26]. Previous works have shown that decreasing the HU threshold for the identification of adenomas could improve the specificity but reduce the sensitivity, whereas increasing such a threshold could result in improved sensitivity but reduced specificity [27, 28].

In a study by Yi and colleagues [17], aiming to differentiate histology-confirmed lipid-poor adenomas from pheochromocytomas, the authors built two radiomic nomograms using NECT and contrast-enhanced CT data respectively, and concluded that the additional contrast-enhanced adrenal CT may not be necessary. Indeed, the drawbacks of a second scan can include additional cost, radiation risks, and potential harms associated with contrast media administration, including allergy and potential renal injury. A dedicated adrenal CT protocol including a 15-minute delayed acquisition and considering a 60% threshold for contrast washout) has been shown to properly classify 96% of adrenal masses, with 98% sensitivity and 92% specificity for discriminating adenomas from non-adenomas [29]. However, it should be noted that the additional role of dedicated CT protocols in characterizing incidental adrenal masses, based on washout calculation, is still being debated in the literature, particularly in the case of suspected pheochromocytomas or metastases from hypervascular tumors which frequently demonstrate rapid contrast washout [22]. Hypervascular metastases from renal cell carcinoma and hepatocellular carcinoma are examples that may include intracellular lipid and have washout values similar to adenomas [30].

Furthermore, the patient is required to return for the dedicated adrenal imaging if the initial NECT, in which the lesion had been incidentally detected, was inconclusive; this will obviously lengthen the diagnostic process and cause psychological distress to the patient.

There are several limitations to this study that should be considered. One major limitation is the small sample size of the final cohort. This was inevitable because our efforts to find patients who had adrenal nodules that were indeterminate at NECT, with the necessity of histological confirmation, resulted in a relatively small number of lesions meeting the inclusion criteria. Another limitation is the retrospective nature of the image data acquisition: in this observational study, the type of scanner used for each patient was not controlled. When considering the robustness of radiomic applications in the clinical setting, the potential impact of variation in CT data acquired from different scanners should not be understated. However, our radiomic signature is based on 1 histogram-based feature, 1 second-order feature, and 3 shape features that have been shown to be robust in previous Radiomics studies [31, 32].

5. CONCLUSIONS

Including additional imaging indicators for the identification of lipid-poor adenomas can increase the accuracy of NECT and reduce the need for additional imaging and clinical workup, according to this and other recent studies focusing on Radiomics that have distinct points of contact with current clinical practice.

Our radiomic signature based on 1 histogram-based feature, 1 second-order feature, and 3 shape features could be considered for integration in routine radiological assessment of adrenal lesions, beyond the mere measurement of median density. This may serve as a method of enhancing the diagnostic power of NECT in order to substantially limit the number of adrenal incidentalomas initially regarded as indeterminate.

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at https:// doi. org/ 10. 1007/ s00330- 023- 10090-8.

ACKNOWLEDGMENTS

This work was partly supported thanks to the contribution of Ricerca Corrente by the Italian Ministry of Health within the research line innovative therapies, phase I-III clinical trials and therapeutic strategy trials based on preclinical models, onco-immunological mechanisms, and nanovectors.

ETHICAL APPROVAL

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Area Vasta Emilia Centrale (AVEC) (146/2022/Oss/AOUFe, 17/02/2022).

Study subjects or cohorts overlap

No study subjects or cohorts overlap.

Methodology

- retrospective
- observational
- performed at one institution

REFERENCES

1. Fassnacht M, Arlt W, Bancos I, et al (2016) Management of adrenal incidentalomas: European Society of Endocrinology Clinical Practice Guideline in collaboration with the European Network for the Study of Adrenal Tumors. Eur J Endocrinol 175:G1–G34. https://doi.org/10.1530/EJE-16-0467
2. Mayo-Smith WW, Song JH, Boland GL, et al (2017) Management of Incidental Adrenal Masses: A White Paper of the ACR Incidental Findings Committee. J Am Coll Radiol 14:1038–1044. https://doi.org/10.1016/j.jacr.2017.05.001
3. Willatt J, Chong S, Ruma JA, Kuriakose J (2015) Incidental adrenal nodules and masses: The imaging approach. Int J Endocrinol 2015:. https://doi.org/10.1155/2015/410185
4. Garrett RW, Nepute JC, El Hayek M, Albert SG (2016) Adrenal Incidentalomas: Clinical controversies and modified recommendations. Am J Roentgenol 206:1170–1178. https://doi.org/10.2214/AJR.15.15475
5. Barat M, Cottereau AS, Gaujoux S, et al (2022) Adrenal Mass Characterization in the Era of Quantitative Imaging: State of the Art. Cancers (Basel) 14:. https://doi.org/10.3390/CANCERS14030569
6. Nandra G, Duxbury O, Patel P, et al (2020) Technical and Interpretive Pitfalls in Adrenal Imaging. RadioGraphics 40:1041–1060. https://doi.org/10.1148/rg.2020190080
7. Schieda N, Alrashed A, Flood TA, et al (2016) Comparison of quantitative MRI and CT washout analysis for differentiation of adrenal pheochromocytoma from adrenal adenoma. Am J Roentgenol 206:1141–1148. https://doi.org/10.2214/AJR.15.15318
8. Sasaguri K, Takahashi N, Takeuchi M, et al (2016) Differentiation of benign from metastatic adrenal masses in patients with renal cell carcinoma on contrast-enhanced CT. Am J Roentgenol 207:1031–1038. https://doi.org/10.2214/AJR.16.16193
9. Dong A, Cui Y, Wang Y, et al (2014) (18)F-FDG PET/CT of adrenal lesions. AJR Am J Roentgenol 203:245–252. https://doi.org/10.2214/AJR.13.11793
10. Limkin EJ, Sun R, Dercle L, et al (2017) Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Ann Oncol 28:1191–1206. https://doi.org/10.1093/annonc/mdx034
11. Shur JD, Doran SJ, Kumar S, et al (2021) Radiomics in oncology: A practical guide. Radiographics 41:1717–1732. https://doi.org/10.1148/rg.2021210037
12. van Timmeren JE, Cester D, Tanadini-Lang S, et al (2020) Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11:91. https://doi.org/10.1186/s13244-020-00887-2
13. Feliciani G, Mellini L, Carnevale A, et al (2021) The potential role of MR based radiomic biomarkers in the characterization of focal testicular lesions. Sci Rep 11:1–9. https://doi.org/10.1038/s41598-021-83023-4
14. O'Shea A, Kilcoyne A, McDermott E, et al (2022) Can radiomic feature analysis differentiate adrenal metastases from lipid-poor adenomas on single-phase contrast-enhanced CT abdomen? Clin Radiol 77:e711–e718. https://doi.org/10.1016/j.crad.2022.06.015
15. Cao L, Xu W (2022) Radiomics approach based on biphasic CT images well differentiate "early stage" of adrenal metastases from lipid-poor adenomas: A STARD compliant article. Medicine (Baltimore) 101:e30856. https://doi.org/10.1097/MD.0000000000030856
16. Laderian B, Ahmed FS, Zhao B, et al (2019) Role of radiomics to differentiate benign from malignant pheochromocytomas and paragangliomas on contrast enhanced CT scans. J Clin Oncol 37:e14596–e14596.

https://doi.org/10.1200/JCO.2019.37.15_suppl.e14596

17. Yi X, Guan X, Zhang Y, et al (2018) Radiomics improves efficiency for differentiating subclinical pheochromocytoma from lipid-poor adenoma: a predictive, preventive and personalized medical approach in adrenal incidentalomas. EPMA J 9:421–429. https://doi.org/10.1007/S13167-018-0149-3/FIGURES/4
18. Zhang B, Zhang H, Li X, et al (2022) Can Radiomics Provide Additional Diagnostic Value for Identifying Adrenal Lipid-Poor Adenomas From Non-Adenomas on Unenhanced CT? Front Oncol 12:. https://doi.org/10.3389/fonc.2022.888778
19. 3D Slicer image computing platform (2023) available via https://www.slicer.org/ Accessed 17 May 2023
20. Zwanenburg A, Leger S, Vallières M, Löck S (2016) The image biomarker standardisation initiative. CoRR abs/1612.0: https://doi.org/10.1148/radiol.2020191145
21. Ierardi AM, Carnevale A, Angileri SA, et al (2020) Outcomes following minimally invasive imagine-guided percutaneous ablation of adrenal glands. Gland Surg. 9:859–866
22. Corwin MT, Remer EM (2021) Adrenal washout CT: Point-not useful for characterizing incidentally discovered adrenal nodules. Am. J. Roentgenol. 216:1166–1167
23. Amadasun M, King R (1989) Texural Features Corresponding to Texural Properties. IEEE Trans Syst Man Cybern 19:1264–1274. https://doi.org/10.1109/21.44046
24. Sherlock M, Scarsbrook A, Abbas A, et al (2020) Adrenal Incidentaloma. Endocr Rev 41:775–820. https://doi.org/10.1210/ENDREV/BNAA008
25. Boland GWL, Lee MJ, Gazelle GS, et al (1998) Characterization of adrenal masses using unenhanced CT: an analysis of the CT literature. AJR Am J Roentgenol 171:201–204. https://doi.org/10.2214/AJR.171.1.9648789
26. Albano D, Agnello F, Midiri F, et al (2019) Imaging features of adrenal masses. Insights Imaging 10:1–16. https://doi.org/10.1186/s13244-019-0688-8
27. Kirsch MJ, Kohli MW, Long KL, et al (2020) Utility of the 10 Hounsfield unit threshold for identifying adrenal adenomas: Can we improve? Am J Surg 220:920–924. https://doi.org/10.1016/J.AMJSURG.2020.04.021
28. Lattin GE, Sturgill ED, Tujo CA, et al (2014) From the Radiologic Pathology Archives: Adrenal Tumors and Tumor-like Conditions in the Adult: Radiologic-Pathologic Correlation. https://doi.org/101148/rg343130127 34:805–829. https://doi.org/10.1148/RG.343130127
29. Caoili EM, Korobkin M, Francis IR, et al (2002) Adrenal Masses: Characterization with Combined Unenhanced and Delayed Enhanced CT. Radiology 222:629–633. https://doi.org/10.1148/radiol.2223010766
30. Grajewski KG, Caoili EM (2020) Adrenal Washout CT: Counterpoint—Remains a Valuable Tool for Radiologists Characterizing Indeterminate Nodules. Am J Roentgenol 216:1168–1169. https://doi.org/10.2214/AJR.20.24490
31. Mackin D, Fave X, Zhang L, et al (2015) Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 50:1–9. https://doi.org/10.1097/RLI.0000000000000180
32. Larue RTHM, van Timmeren JE, de Jong EEC, et al (2017) Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol (Madr) 56:1544–1553. https://doi.org/10.1080/0284186X.2017.1351624

CHAPTER 7

General discussion and outlook of the radiomics laboratory

Enrico Menghi

GENERAL DISCUSSION

The establishment of the "Romagna Imaging Biobank" and the application of radiomics in clinical radiation oncology mark a significant step forward in the field of personalized medecine. The comprehensive journey outlined in this thesis presents several important points for consideration and discussion.

This thesis has highlighted the potential for quantitative imaging to transform cancer diagnosis and treatment, emphasizing the importance of multidisciplinary collaboration and robust data management. The "Romagna Imaging Biobank" and the research conducted herein have the potential to pave the way for more personalized and effective cancer care in the future.

IMAGING BIOMARKERS – PHANTOMS

In **Chapter 2** we analyzed the performance of seven self-declared IBSI-compliant software packages. Phase I analysis on the IBSI digital phantom revealed that all programs achieved high percentages of 'matching' features, indicating a high standardization level in terms of Radiomic Features implementation. In Phase II, we systematically investigated the effect of factors related to parameter setting (i.e., interpolation, discretization, and aggregation) as well as to ROI characteristics (i.e. volume and shape) on software agreement by employing two custom digital phantoms and a systematic feature extraction. The results we obtained are relative to a selected number of radiomic software programs and future studies might include additional packages to strengthen the present findings.

However, we are reasonably confident that the considered packages are a representative set of the high-standardized radiomic tools available in the literature. Moreover, some of our findings are software-independent and have general validity.

In conclusion, we designed a new investigation scenario in which we demonstrated that, despite the ongoing efforts of both IBSI and software developers to standardize radiomic tools, additional efforts are needed to achieve full concordance.

In **Chapter 3** we presents a study that provides the first multicentre evaluation of the dosiomics features in terms of reproducibility, stability and sensitivity across various dose distributions obtained from multiple technologies and techniques and considering different dose calculation algorithms of treatment planning systems and two different resolutions of the dose

grid. The study has assessed the stability of dosiomic features and their capability in distinguishing dose distributions generated with different radiation therapy devices in a multicentre setting.

A limitation of the present study is related to the pool of the considered radiotherapy techniques and technologies. They are pretty diverse and representative but do not describe all the possible techniques and technologies available in clinical practice. Despite this, we believe that the employed number of radiotherapy techniques and technologies used by the eight centres are enough to support the message that a substantial number of dosiomic features are stable, and at the same time, they can distinguish or recognise dose distributions generated with different radiation therapy devices.

IMAGING BIOMARKERS - PATIENTS

The study of **Chapter 4** evaluates the ability of T2w MR-based quantitative analysis to help differentiate germinal from non-germinal tumors and seminomas from non-seminomas. This preliminary study shows that the radiomic measures obtained by scrotal MR image analysis may be useful in the diagnostic workup of testicular lesions, since they could add valuable information and help to discriminate among testicular neoplasms by differentiating germ cell from non-gem cell tumors, and seminomas from other histologies. Further independent validation is required to assess whether quantitative imaging features, possibly in conjunction with standard clinical markers and other quantitative techniques, may allow more accurate characterization of testicular lesions.

Biopsy International Society of Urological Pathology (ISUP) grade differs from the final ISUP determined after surgery in around one-third of patients, with biopsies tending to underestimate cancer aggressiveness. The differences between the two ISUPs can have a big impact on how patients are managed. As a result, incorporating pre-therapeutic imaging characteristics to accurately determine Prostate Cancer (PCa) aggressiveness is of great clinical importance. This study in **Chapter 5** evaluates the ability of MRI-ADC and [68Ga]Ga-PSMA-11-based quantitative analysis to help differentiate low-risk prostate cancer patients (ISUP 1) from higher risk patient classes (ISUP>1) and aimed to evaluate the benefits of the two imaging techniques combined. Both [68Ga]Ga-PSMA-11 PET and MRI-ADC imaging biomarkers showed to be complementary about ISUP grade assessment when employed together to build prediction models.

In the study contained in **Chapter 6**, we developed a radiomic signature for the classification of benign lipid-poor adenomas, which may potentially help clinicians limit the number of unnecessary investigations in clinical practice. Indeterminate adrenal lesions of benign and malignant nature may exhibit different values of key radiomic features. Including additional imaging indicators for the identification of lipid-poor adenomas can increase the accuracy of not-enhanced CT (NECT) and reduce the need for additional imaging and clinical workup. Our radiomic signature could be considered for integration in routine radiological assessment of adrenal lesions, beyond the mere measurement of median density. This may serve as a method of enhancing the diagnostic power of NECT in order to substantially limit the number of adrenal incidentalomas initially regarded as indeterminate.

A review on radiomic studies for lipid-poor adrenal adenomas is in preparation to further investigate and stress the scenario to create a first multicentre study and to increase the small sample size of the final cohort, one of major limitation.

OUTLOOK OF THE RADIOMICS LABORATORY

In the coming years, radiomics will continue to evolve.

Modern medicine requires large amounts of data, particularly in the domain of cancer care. The future of personalized medicine lies with ‘‘genomic medicine’’, ‘‘precision medicine’’, but also with ‘‘data medicine’’ (DM) (big data, data mining). This requires far-reaching changes, to establish four essential elements connecting patients and doctors: biobanks, databases, bioinformatic platforms and genomic platforms as shown in **Figure.1**. Molecular tumor boards (MTB) are one response to these changes [1], an evolution from classical Tumor Board Model or Multidisciplinary Team (**Figure 2**).

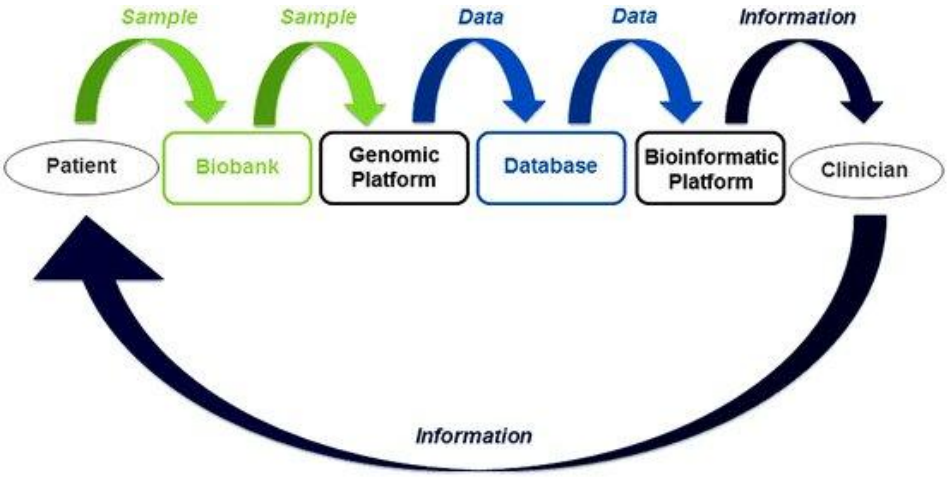


Figure 1 Model of the data medicine process. Biological sample flows and biobanks are shown in green. Data flows and databases are shown in light blue. Information flows are shown in dark blue [1]

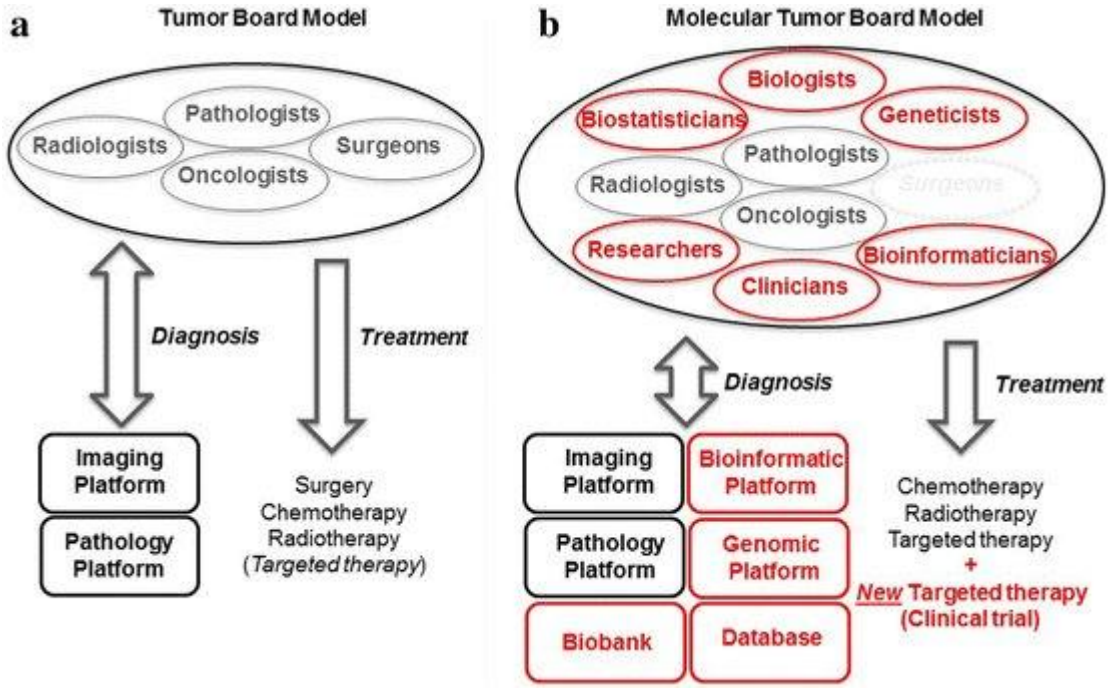


Figure 2 a The tumor board model, versus **b** the molecular tumor board model [1]

Collaborating with the working group of the Italian Association of Medical Physics (AIFM, WG FM4AI), Alliance Against Cancer (Alleanza Contro il Cancro, ACC, WG Radiomics, WG Dosiomics) who pursue clinical and translational research in order to bring state of the art diagnostics and advanced therapeutics to patient care [2] and working with a joint AIFM-INFN Italian initiative for a dedicated cloud-based computing infrastructure to enhancing the impact of Artificial Intelligence in Medicine [3], this thesis represents a critical experience to

the evolving landscape of clinical oncology, where the union of technology and clinical insight paves the way for a brighter data-driven future in cancer management.

Another potential outlook for the future of the laboratory will also be the “RIS-PACS Romagna”, that will be installed to standardize the flows of digital images within the different hospitals within the Intercompany Program of “COMPREHENSIVE CANCER CARE NETWORK” for the activation of the Onco-hematology Network of Romagna [4].

Finally, here is a step-by-step guide on how to set up a radiomics laboratory:

Define Your Goals and Research Focus:

- Clearly define the objectives of your radiomics laboratory. What specific clinical oncology problems do you intend to address? What are your research goals?

Secure Funding:

- Establish a budget for your laboratory, which includes costs for equipment, software, personnel, and ongoing operation. Seek funding from government grants, private foundations, or institutional sources.

Build a Team:

- Assemble a multidisciplinary team with expertise in radiology, oncology, data science, and computational biology. Your team should include radiologists, oncologists, data scientists, and IT specialists.

Infrastructure and Equipment:

- Acquire the necessary imaging equipment, such as CT and MRI scanners. Ensure they are up-to-date and capable of high-resolution imaging.

- Invest in high-performance computing infrastructure to process and analyze the large amount of data generated.

Data Management and Storage:

- Develop a secure data management and storage infrastructure. Ensure compliance with patient data privacy regulations..

Software and Tools:

- Identify and acquire radiomics software tools, such as 3D Slicer, PyRadiomics, or proprietary solutions, depending on your research needs.
- Consider using machine learning and deep learning libraries (e.g., TensorFlow, PyTorch) for data analysis.

Image Data Collection and Preprocessing:

- Establish protocols for image data acquisition, storage, and preprocessing.
- Develop standard operating procedures (SOPs) for image acquisition, including calibration and quality control.

Feature Extraction:

- Implement radiomics feature extraction algorithms to extract quantitative features from medical images.
- Standardize the feature extraction process to ensure consistency and reproducibility.

Data Mining and Analysis:

- Apply data mining and machine learning techniques to analyze radiomics data.
- Develop predictive models for cancer diagnosis, prognosis, and treatment response.

Quality Control and Validation:

- Implement quality control measures to ensure data accuracy and reliability.
- Validate your radiomics models using independent datasets and clinical studies.

Collaborate and Publish:

- Collaborate with clinical partners to apply radiomics in real patient cases.
- Publish your research findings in scientific journals and present at conferences to contribute to the field's knowledge.

Ethical and Regulatory Considerations:

- Ensure that your research adheres to ethical guidelines and regulatory requirements, especially those related to patient privacy and data handling.

Education and Training:

- Train your team in radiomics methodologies and stay updated with the latest advancements in the field.

Continual Improvement:

- Continually assess and improve your laboratory's processes, software, and research methodologies.

Outreach and Collaboration:

- Collaborate with other radiomics laboratories and research institutions to share knowledge and resources.

REFERENCES

[1] Molecular Tumor Boards: Ethical Issues in the New Era of Data Medicine. Stoeklé HC, Mamzer-Bruneel MF, Frouart CH, Le Tourneau C, Laurent-Puig P, Vogt G, Hervé C.Sci Eng Ethics. 2018 Feb;24(1):307-322. doi: 10.1007/s11948-017-9880-8. Epub 2017 Mar 9.PMID: 28281147

[2] https://www.alleanzacontroilcancro.it/en

[3] Enhancing the impact of Artificial Intelligence in Medicine: A joint AIFM-INFN Italian initiative for a dedicated cloud-based computing infrastructure. Retico A, Avanzo M, Boccali T, Bonacorsi D, Botta F, Cuttone G, Martelli B, Salomoni D, Spiga D, Trianni A, Stasi M, Iori M, Talamonti C.Phys Med. 2021 Nov;91:140-150. doi: 10.1016/j.ejmp.2021.10.005. Epub 2021 Nov 18.PMID: 34801873

[4] Bravi F., Gilibertoni D., Marcon A., Sicotte C., Minvielle E., Rucci P., Angelastro A., Carradori T., Fantini M.P.Hospital network performance: A survey of hospital stakeholders' perspectives, Health Policy, Vol. 109, Issue 2, 2013: 150-157

Abstract

This thesis focuses on the field of data mining and radiomics and its application in clinical radiation oncology from cancer diagnosis to therapies.

One of the thesis objective is to establish a "Romagna Imaging Biobank" and apply radiomics to specific oncological pathologies. The goals are to establish a link between tumor phenotype and quantitative image descriptors, enhance patient stratification, and personalize anti-cancer therapies.

First the thesis outlines a software platform development and a multicentre study for radiomic tools testing, investigating their reproducibility, sensitivity and stability in terms of features extraction.

We examine the use of imaging biomarkers, both through phantom-based testing and patient-focused studies:

In the first study, software packages compliant with the Image Biomarkers Standardization Initiative (IBSI) were assessed, revealing high standardization in feature implementation. However, the study also indicates the necessity for additional efforts to achieve full concordance among radiomic tools. Then we present a multicentre evaluation of dosiomics features, emphasizing their stability and effectiveness in distinguishing dose distributions across various radiation therapy technologies and techniques.

Patient-focused studies explore the potential of radiomic analysis in patient diagnosis. The former investigates the use of MR-based quantitative analysis to differentiate testicular tumors, while the latter assesses the complementary roles of MRI-ADC and [68Ga]Ga-PSMA-11-based quantitative analysis in distinguishing prostate cancer patients of varying risk levels. The last patient-focused study introduces a radiomic signature for the classification of benign lipid-poor adrenal adenomas, with potential applications in reducing unnecessary investigations and enhancing the accuracy of not-enhanced CT scans. A review on radiomic studies for lipid-poor adrenal adenomas is in preparation.

In conclusion, the outlook of the radiomics laboratory is rooted in the ever-evolving field of data-driven medicine. Collaboration with various medical physics associations and research initiatives, including artificial intelligence, promises to drive progress in clinical oncology. The integration of digital image flows within hospitals and the implementation of standardized practices further enhance a comprehensive cancer care network and strengthen the establishment of a “Romagna Imaging Biobank”.