A European Cancer Image Platform Linked to Biological and Health Data for
Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D6.1:

Definition of metrics, criteria and procedures for testing performance and robustness

Reference	D6.1_ EuCanImage_UM_final
Lead Beneficiary	UM
Author(s)	Philippe Lambin, Zohaib Salahuddin, Shruti A. Mali, Henry Woodruff
Dissemination level	Confidential
Type	Radiomics Quality Score 2.0
Official Delivery Date	31-3-2022
Date of validation of the WP leader	31-3-2022
Date of validation by the Project Coordinator	31-3-2022
Project Coordinator Signature	

EuCanImage is funded by the European Union's H2020 Framework
Under Grant Agreement No 952103

Version log

Issue Date	Version	Involved	Comments
07/02/2022	v.1.1	Philippe Lambin, Zohaib Salahuddin, Shruti Mali, Henry Woodruff	Initial Draft
06/03/2022	v.1.2	Karim Lekadir	Feedback and Comments
17/03/2022	v.1.3	Maciej Bobowicz, Katrine Riklund	Feedback and Comments
22/03/2022	v.1.4	Philippe Lambin, Zohaib Salahuddin, Shruti Mali, Henry Woodruff	Final Document incorporating feedback and comments from AI WG, Clinical WG, WP6 Partners and UM D-Lab.
31/03/2022	Final	Karim Lekadir, Isabell Tributsch	Final and revised version.

Executive Summary

In the field of quantitative image analysis, the Radiomics Quality Score 1.0 is a popular tool that has been widely used to benchmark radiomics studies and encourages best scientific practices (https://www.nature.com/articles/nrclinonc.2017.141). Recent advancements and challenges impeding the clinical translation of radiomics have created the need for an improved benchmarking tool. We propose a new consensus-derived Radiomics Quality Score 2.0 as a new standard for the quality assessment and facilitation of planning of radiomics studies. This document was prepared within the EuCanImage consortium in collaboration across work packages and centres.

Contents

Acronyms

Name	Abbreviation
Area under the curve	AUC
Deep Learning	DL
Handcrafted Radiomics	HCR
Image biomarker standardization initiative	IBSI
Quality-adjusted life years	QALYs
Radiomics Quality Score	RQS
Receiver operating characteristics curve	ROC
SHapley Additive exPlanations	SHAP
Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis	TRIPOD
Transparent Reporting of Medical Image Acquisition	TRIAC

Radiomics Quality Score 2.0: Towards Trustworthy Clinical Translation

1 EuCanImage and RQS 2.0

1.1 Document Formulation

This document was first disseminated to the UM Precision Medicine Department and feedback was taken with an intention to fill the gaps in RQS 1.0. The department consists of individuals from multi-disciplinary backgrounds ranging from medicine to computer science. Feedback was taken and incorporated for every checkpoint present in the RQS tool and a new version was formed namely RQS 2.0. After incorporating the feedback from the UM Precision Medicine Department, the document was disseminated to multiple working groups (WG) in EUCanImage, spanning multiple work packages (WP), and institutes, and was further developed within the AI W, Clinical WG, and WP6 partners. Discussions were carried on regarding each checkpoint in the latest version of the document and feedback coming from these WGs were incorporated into the document. Figure 1 shows the procedure for the conception of the consensus document.

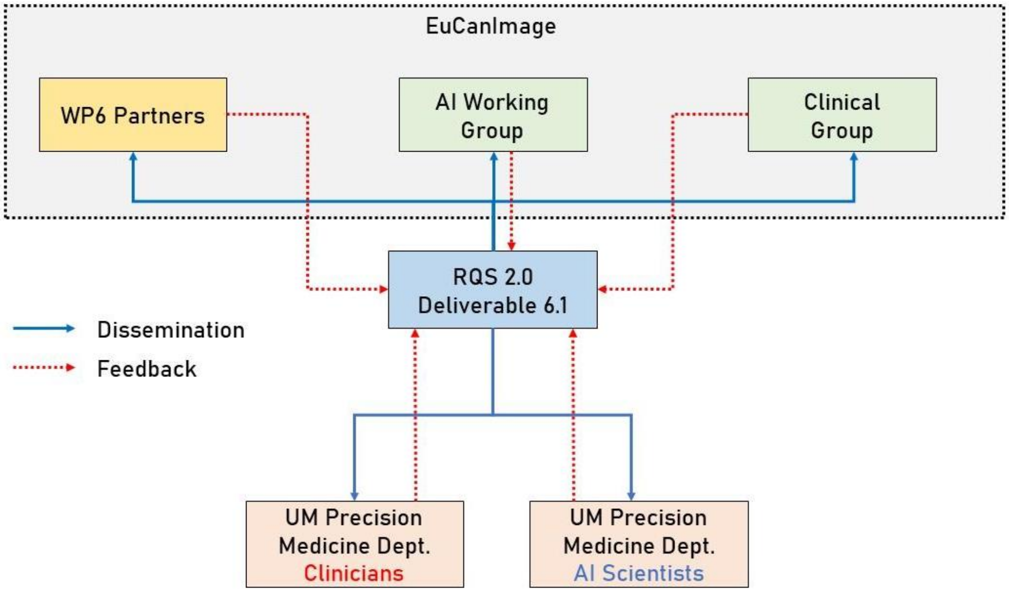

Figure 1: The workflow for the formulation of the consensus document outlining the metrics, criteria and procedures for testing performance and robustness.

2 Abstract

Radiomics, the quantitative analysis of medical images to extract image features and consequently incorporate them within decision support systems for clinical applications, is gaining research traction every year. After a decade of research, the clinical translation of radiomics is still sparse due to the insufficient quality of the radiomics studies that do not meet the clinical standards. Radiomics Quality Score 1.0 is a popular tool that has been widely used to benchmark radiomics studies and encourages best scientific practices. Due to recent

advancements and challenges such as harmonization and interpretability impeding the clinical translation of radiomics, there is a need for an improved benchmarking tool to assess the quality of the radiomics studies and for encouraging best scientific practice to translate radiomics into clinical practice. We propose Radiomics Quality Score 2.0 as a new standard for the quality assessment of radiomics that alleviates the problems of RQS 1.0.

3 Introduction

Medical imaging such as magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) is regularly used in clinical practice to aid the decision-making process for diagnostic, theragnostic, predictive, prognostic, follow-up, and treatment purposes [1,2]. Radiomics is a field where medical images are converted to mineable data. This information is achieved by extracting quantitative features which are further used for supporting clinical decisions [3,4]. The radiomics theory presumes that the quantitative analysis of medical images provides additional knowledge in a prompt and reproducible way to aid radiologists in reporting and clinicians in their decision-making process [5]. This acquired additional knowledge when combined with clinical data and associated with predicted data can foster the development of clinical decision support systems. Presently, there are two categories of radiomics studies (Figure 2): First, hand-crafted radiomics (HCR), which is dependent on a traditional workflow of extracting standardized hand-crafted features that are image biomarker standardization initiative (IBSI)-compliant. These features are either texture, shape, or intensity features that are extracted from a specific region of interest/s such as a tumor [5]. The modeling is usually done using machine learning techniques. The second category is a deep learning (DL) approach. It is a data-driven method that learns complex visual representative features by performing classification/segmentation tasks using neural networks.

The term radiomics was coined in 2012 and the research interest in radiomics is growing every year as indicated by Figure 3,4. The clinical translation of radiomics research is rare even after a decade of research because the quality of radiomics study is insufficient to satisfy the requirements for clinical use [6]. Radiomics quality score 1.0 (RQS)[1] was introduced to aid in the assessment of past and future radiomics studies and consequently increase the scientific rigor and quality of the radiomics studies[4]. A mean RQS score of 20.4%, 26.1%, and 27.4% was obtained after recent analyses of radiomics studies [7–9]. This shows that RQS is a stringent and demanding criterion [9–13] that aims to encourage the best scientific practice. RQS is a popular tool for the quality assessment of radiomics studies [9,12,14–18]. In the light of recent advancements and new challenges such as interpretability[19], harmonization [20], and reproducibility, and to alleviate the shortcomings of RQS[21], we introduce a newer version of RQS referred to as radiomics quality score 2.0 (RQS 2.0). RQS 2.0 also makes a distinction between handcrafted radiomics and deep learning radiomics.

[1] https://www.radiomics.world/rqs

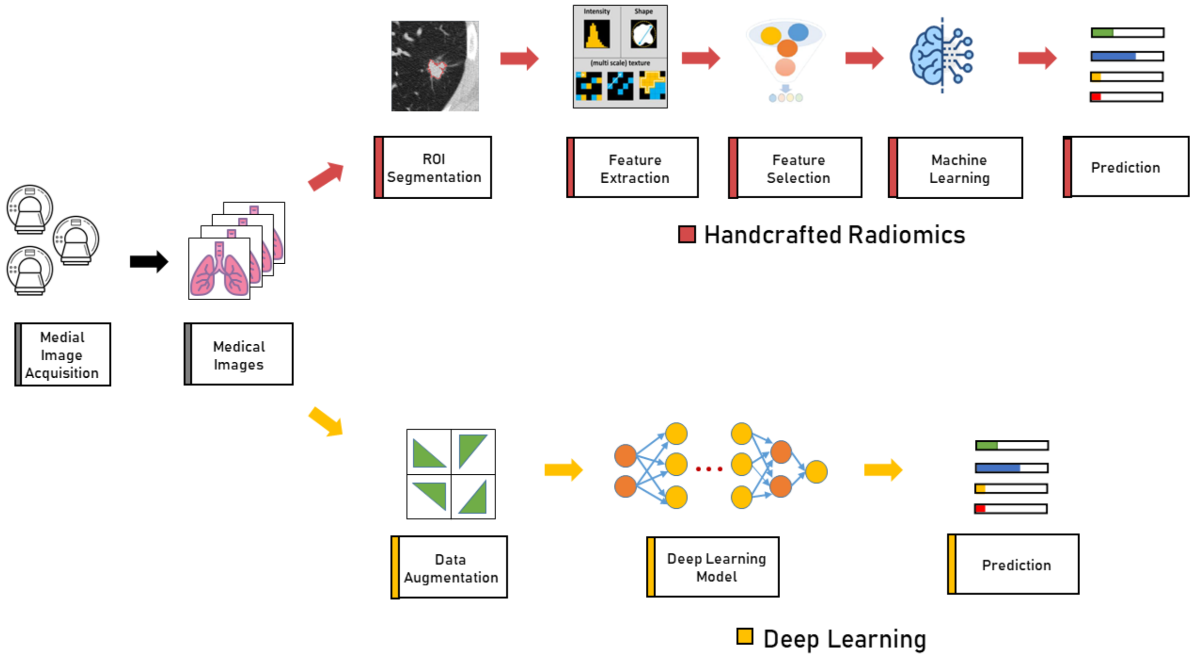

Figure 2: Radiomics involves quantitative analysis of imaging data and comprises handcrafted radiomics (HCR) and deep learning (DL). The figure shows a summary of the few steps involved for quantitative prediction using HCR and DL. (Full Resolution: *https://tinyurl.com/dlhcrpipeline*)

Quality assessment of radiomics studies is essential for clinical translation. QUADAS tool for systematic reviews was developed in 2003 for the quality assessment of diagnostic studies and later improved to QUADAS-2[22]. Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) initiative consists of a set of recommendations for reporting on developing, validating, and updating of a prediction model for prognosis or diagnosis thereby facilitating comparison of future studies [23]. Future-AI recommendations provide guidelines on implementing fairness, universality, traceability, usability, robustness, and explainability principles for trustworthy artificial intelligence in medical imaging. RQS 2.0 emulates the TRIPOD initiative and the guiding principles of FUTURE-AI are also reflected in the recommendations of RQS 2.0. By integrating the principles of FUTURE-AI, RQS 2.0 is intended to enhance the clinical safety, technical robustness, clinical acceptance, as well as ethical compliance of future radiomics tools, to make them more trustworthy and applicable in the real world.

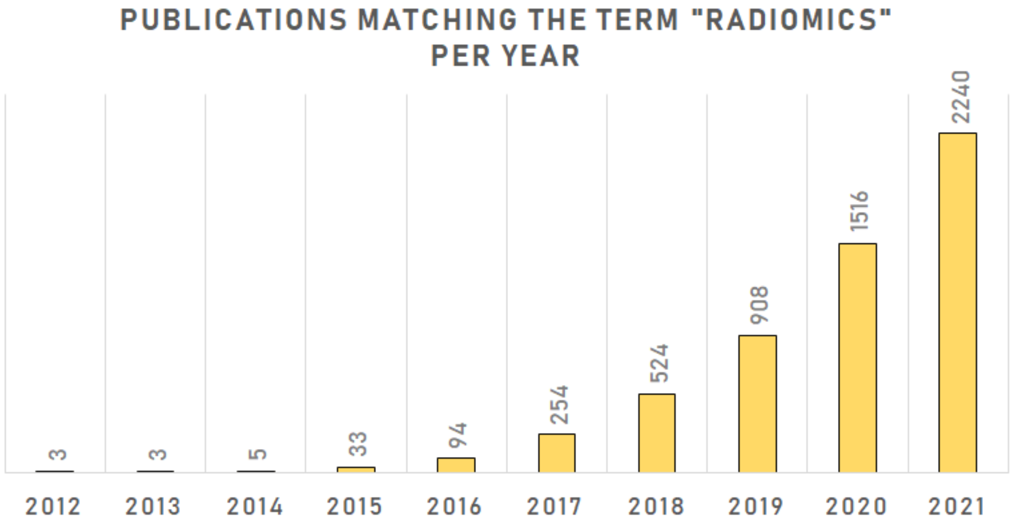

Figure 3: The number of articles published until 2022 matching the term *"radiomics"* on PubMed (*https://pubmed.ncbi.nlm.nih.gov/?term=radiomics&timeline=expanded*).

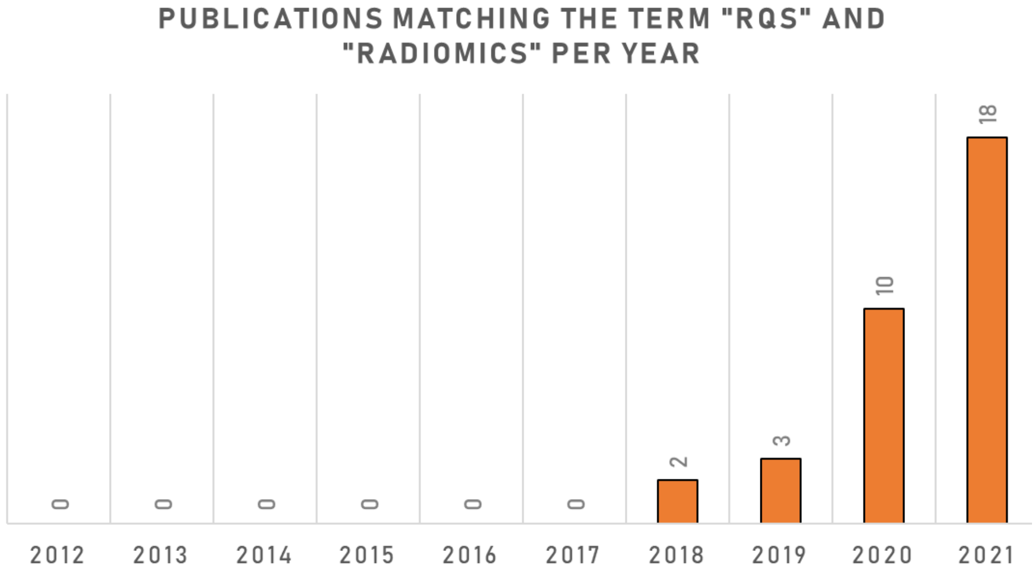

Figure 4: The number of articles published until 2022 matching the term *"radiomics" and "RQS"* on PubMed (*https://pubmed.ncbi.nlm.nih.gov/?term=radiomics&timeline=expanded*).

4 Radiomics Quality Score 2.0

Radiomics is the quantitative analysis of imaging data that further aids clinicians in their decision-making process. The information collected from radiomics studies can help foster the clinical decision support systems by establishing a relationship between radiomic features and clinical endpoints by developing diagnostic, prognostic, and predictive models. In the context of radiomics studies, the workflow can be divided into six checkpoints: Objectives and Discussion, Input Data, Method, Evaluation, Interpretability/Explainability, and Utility (Figure

5). To assess the quality of radiomics studies and to overcome the roadblocks of RQS 1.0, we propose RQS 2.0.

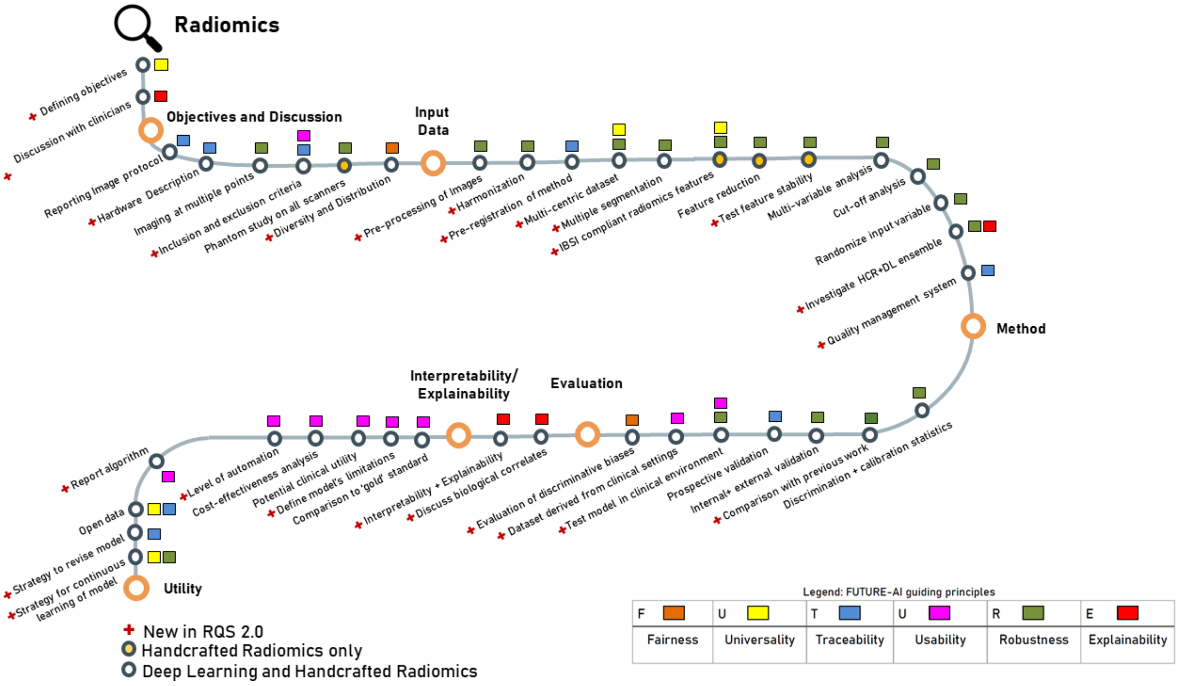
Radiomics
✱ Defining objectives
Discussion with clinicians
✱
Objectives and Discussion
Reporting Image protocol
✱ Hardware Description
Imaging at multiple points
✱ Inclusion and exclusion criteria
Phantom study on all scanners
✱ Diversity and Distribution
Input Data
✱ Pre-processing of Images
✱ Harmonization
✱ Pre-registration of method
✱ Multi-centric dataset
✱ Multiple segmentation
✱ IBSI compliant radiomics features
Feature reduction
✱ Test feature stability
Multi-variable analysis
Cut-off analysis
Randomize input variable
✱ Investigate HCR+DL ensemble
✱ Quality management system
Method
Discrimination + calibration statistics
✱ Comparison with previous work
Internal+ external validation
Prospective validation
✱ Test model in clinical environment
✱ Dataset derived from clinical settings
✱ Evaluation of discriminative biases
Evaluation
✱ Discuss biological correlates
✱ Interpretability + Explainability
Interpretability/ Explainability
Comparison to 'gold' standard
✱ Define model's limitations
Potential clinical utility
Cost-effectiveness analysis
✱ Level of automation
✱ Report algorithm
Open data
✱ Strategy to revise model
✱ Strategy for continuous learning of model
Utility
✚ New in RQS 2.0
Handcrafted Radiomics only
Deep Learning and Handcrafted Radiomics
Legend: FUTURE-AI guiding principles
F 🟧 Fairness
U 🟨 Universality
T 🟦 Traceability
U 🟪 Usability
R 🟩 Robustness
E 🟥 Explainability

Figure 5: The flowchart demonstrates the workflow of radiomics and the necessary steps that the radiomics quality score (RQS) 2.0 rewards or penalizes to encourage best scientific practice. The radiomics workflow takes into account handcrafted radiomics and deep learning radiomics. The new steps introduced in RQS 2.0 are shown by the plus (+) sign. (Full Resolution: https://tinyurl.com/rqsflow)

4.1 Objectives and Clinical Discussion

Before setting the criteria for data selection, the unmet clinical need must be clearly defined. The aims and objectives of the clinical question (e.g. classification or a segmentation task) at hand must be properly identified before moving on to the next steps. The requirements needed to carry out the experiments should be clearly defined (e.g. uni-centric or multi-centric data). Furthermore, discussions should be carried out with the clinicians before model development to come to a consensus for choosing an appropriate explainability method.
Input Data

Radiomics studies pipelines begin with the data selection procedure of selecting the appropriate image modality, imaging protocol, the region of interest, and the prediction target/event. Standardized imaging protocols are important to eliminate variabilities arising due to different scanners and their acquisition and reconstruction settings [24,25]. One such way is to ensure that the protocols are well documented e.g., protocol following Transparent Reporting of Medical Image Acquisition (TRIAC [26,27]) guidelines for future proof radiomics or if a public protocol is used. TRIAC guidelines describe five different levels of evidence for reporting imaging protocols. Level 0 indicates that the protocol has not been formally approved with a reference number; Level 1 indicates that the protocol has been approved with a reference number in the archive of the department; Level 2 indicates that the protocol has been approved with formal quality assurance (recommended minimum level for prospective trials); Level 3 indicates that the protocol is established internationally and has been published in guideline documents and peer-reviewed papers; Level 4 indicates that the protocol is Future proof i.e., the protocol follows TRIAC Level 3, FAIR principles and retains raw data.

Reporting the scanner hardware settings, image acquisition, and reconstruction methods are also critical in view of standardizing imaging protocols. Most of the radiomics studies include retrospective datasets that have already been imaged in the past with preset scanner/s and standardizing imaging protocols at this point might not be feasible. To tackle this either pre-processing of images could be done before image analysis or a phantom study could be performed to detect inter-scanner differences and vendor-dependent features to assess the feature robustness [28–33]. This is intended to increase feature reproducibility. Few pre-processing methods have been used in previous work including isotropic voxel resampling, bias field correction [34], normalizing intensity scales using histogram equalization [35,36], gray-level discretization [37], and processing of raw sensor-level image data [38,39]. Pre-processing step is crucial for standardizing heterogeneous datasets to increase the reproducibility of features [27]. Acquiring images from individuals at multiple time points also allows analyzing feature robustness across temporal variabilities (e.g., organ movement). Once the model has been decided, defining model constraints is crucial for its development. E.g., defining inclusion and exclusion criteria for model inputs; detecting and eliminating biases (e.g., sex, ethnicity, socio-economic factors, data imbalance) occurring due to diversity and distribution across diverse patient groups within the dataset/s.

4.2 Method

The next step is to build a generalizable model that fits the input data to predict outcome/s. If the method is already pre-registered on a public platform2, the model is already one step closer to being called a ‘generalizable’ model. To build a generalizable model that fits external data too, harmonization/normalization/correction methods could be implemented at the image level and/or feature level that produces robust images/features [20,40]. Many studies have shown that variabilities across the scanner protocol settings affect the reproducibility of radiomic features [41–49]. Hence, various harmonization methods are available that could be implemented to reduce multi-centric acquisition variability e.g., ComBat [50,51] for HCR and adversarial networks [52–59] for DL. Accounting for variabilities present within the dataset (uni-centric or multi-centric), across multiple segmentations (inter-observer delineations) and different scanner protocol settings, gives more insight into the nature of the dataset. This leads the workflow/study into producing more robust images/features for further analysis. Image biomarker standardization initiative (IBSI)-compliant radiomics features should be extracted. To reduce the high dimensionality of the extracted HCR features, feature reduction is needed to get rid of redundant features. Feature reduction is achieved using either test-retest data, correlation-based analysis, cluster analysis, harmonization methods, and/or machine learning algorithms. It is a plus point if the study implements multivariable analysis with non-radiomic (clinical) features to give a more holistic model (applicable for HCR only). After modeling the features, a cut-off analysis should be implemented and the cut-off values (e.g., log-rank tests) should be checked and compared with previous studies to assess the performance of the prediction model. This would reduce the risk of reporting overly optimistic

[2] www.osf.io

results. Even randomizing permutations within the data would help assess the risk of overfitting. To increase the robustness of the algorithm, investigations must be carried out for HCR and DL methods, or a combination thereof, in an ensemble. This way radiomics features might be helpful in interpreting the predicted outcomes from DL algorithms [60].

4.3 Evaluation

Validation of the radiomics model is critical to assess its robustness and benchmark the performance of the model [7,61–63]. The validation should at least be carried out internally. Preferably, external validation on multi-centric heterogeneous data should be performed. The composition of the external dataset should reflect the distribution of assessed classes in the real-world clinical setting which is critical in terms of translation to clinical usage. The validation should be performed without retraining or adaptation of the cut-off value. Qualitative and quantitative sources of bias should be identified due to diversity present in the patient group such as sex, ethnicity, data imbalance, and difference in breast density [64]. Model performance should also be evaluated with respect to the identified biases. Discriminatory statistics such as receiver operating characteristics curve (ROC), the area under the curve (AUC), sensitivity, specificity, Mathews correlation coefficient (MCC) [65] should be used to evaluate the performance of the classification problems. Statistical significance (p-value) of the results should also be reported. For regression problems, mean squared error or root mean squared error should be reported and for prognostic problems, statistics such as C-index [66] should be reported. Calibration of a prediction model shows how closely the predicted probabilities agree numerically with the actual probability [67,68]. The predictions are grouped to assess the calibration of the model. Calibration statistics such as calibration plots, brier score [69], and calibration-in-the-large/slope should be reported to assess the robustness of the predicted probabilities. Bootstrapping techniques can be utilized to report the confidence interval of the discriminatory and calibration statistics. The validation performance should be compared with previously published radiomics signatures and algorithms. A prospective clinical trial (real or in-silico) to confirm the clinical validity and usefulness of the radiomics biomarker should be pre-registered in a trial database. This prospective clinical trial will provide the highest level of evidence of the utility of the radiomics study. Finally, the radiomics pipeline should be tested in a clinical environment as a final step for the clinical translation.

4.4 Interpretability and Explainability

One of the hurdles in the clinical translation of radiomics studies is the lack of transparency concerning the decision-making process of the Machine Learning models [19,70,71]. Transparency of radiomics prediction models is a legal [72] and ethical requirement, and it is a necessity for troubleshooting purposes. Intrinsic or post-hoc interpretability methods should be used for HCR e.g. SHAP analysis [73,74] and also for DL e.g. attribution methods [75]. Radiomics prediction models, in particular deep learning models, can fail due to noisy and out-of-distribution data. Radiomics models should provide an uncertainty estimate to allow clinicians to refrain from trusting predictions that have a high uncertainty [76]. The link between radiomics features and tumor biology should be investigated by correlating the image features with ground truth pathology substrates [17]. This can help in determining the relationship between tumor-biology-related genomic, cellular and metabolic information and image features. Evaluation of explanations should be carried out quantitatively to determine the sanity of the explanations [77,78]. Moreover, evaluation of explanations in a real world setting with clinicians should be carried out to ensure explanation satisfaction and trust.

4.5 Utility

The radiomics workflow consists of a series of complex steps and each step needs to be carefully reported to allow researchers to reproduce and replicate results [79]. Reproducibility refers to the use of the same workflow and the same dataset to verify the results. Replication aims for a stronger affirmation of the workflow using a different dataset [80–82]. For reproducibility and replicability, the algorithm, source code, and model weights should be made publicly available. Furthermore, the medical image dataset along with segmentation, clinical data, and outcomes should also be made publicly available to allow future development. The added value of the radiomics should be highlighted by performing a comparison with the "gold standard" for performing the clinical task. The model's limitations and scenarios under which the model demonstrates lower performance should be highlighted so that the users are made aware of the shortcomings beforehand. The level of automation for the clinical task due to radiomics can be described using the analogy of the level of automation of the car [83]. At level 0 (No Automation), a clinician performs the clinical task without using the radiomics model. At level 1 (Clinical Assistance), the clinician uses the radiomics model's prediction for a part of the clinical task. At level 2 (Partial Automation), the clinician considers the radiomics model's prediction for the clinical task before making the final recommendation. At level 3 (Conditional Automation), the radiomics model provides the predictions for the clinical task under the supervision and the clinician can intervene at any time. At level 4 (High Automation), the radiomics model provides the predictions and the clinician's intervention is required for special (out-of-distribution) cases. At level 5 (Full Automation), the radiomics model provides predictions for the clinical task without human intervention. The current and potential clinical utility of the radiomics model in a clinical setting should be reported. For example, the radiomics models can be used in clinical decision support systems to predict the need for clinical intervention. Decision-curve analysis can help in visualizing the benefit of using the radiomics model to guide the decision [84]. Cost-effectiveness of the decision support system using the radiomics model should be reported e.g. using quality-adjusted life years (QALYs) [85,86]. To continuously improve the radiomics pipeline, a strategy should be defined on improving and re-training the model from the errors after deployment in the clinical environment. It is important to evaluate the model periodically to ensure that the performance of the AI tool remains consistent with data shifts.

5 Radiomics Quality Score 2.0 Table

Table 1: The radiomics quality score (RQS) 2.0 consists of 35 checkpoints that reward or penalize radiomics studies to encourage best scientific practice. Each checkpoint shows the FUTURE-AI principle it promotes (Fairness, Universality, Traceability, Usability, Robustness, Explainability). Each checkpoint either belongs to Handcrafted Radiomics (HCR), deep learning (DL), or both.

No.	Criteria	Points	HCR/DL
(A) Objectives and Clinical Discussion			
1. 🟨	Unmet clinical need (UCN) defined where: • Uni-centre means UCN is defined by one centre • Multi-centre means UCN is agreed upon and defined by more than one centres • International multi-centre means UCN is agreed upon across borders	-1 (uni-centre) +1 (multi-centre) +2 (international multi-centre)	Both
2. 🟨	Classification of the model: diagnostic, theragnostic, predictive, prognostic, follow-up	-1 (not clearly defined), +1 (defined)	Both
3. 🟥	Input from Radiologist/Imaging Specialists for interpretable pipeline development. Discussion regarding choosing appropriate explainability method	+1 (Clinical knowledge incorporated in the pipeline or explainability method decided and agreed with the clinician before model development)	Both
(B) Input data			
4. 🟦	Image protocol quality to be documented following the TRIAC level (Transparent Reporting of Medical Image Acquisition for a future proof radiomics). TRIAC guidelines describe five different levels of evidence for reporting imaging protocols. • **Level 0** indicates that the protocol has not been formally approved with a reference number • **Level 1** indicates that the protocol has been approved with a reference number in the archive of the department • **Level 2** indicates that the protocol has been approved with formal quality assurance (recommended minimum level for prospective trials) • **Level 3** indicates that the protocol is established internationally and has been published in guideline documents and peer-reviewed papers • **Level 4** indicates that the protocol is Future proof i.e., the protocol follows TRIAC **Level 3**, FAIR principles and retains raw data.	+1 (TRIAC Level 1 and 2) +2 (TRIAC Level 3 and 4)	Both
5. 🟦	Hardware's used described, image reconstruction method specified	+1 (description of the hardware used for image acquisition), +1 (information about image reconstruction method e.g. convolutional kernel)	Both

6. 🟩	Preprocessing of the images	+1 (well-motivated preprocessing steps that account across variation images)	Both
7. 🟩	Imaging at multiple time points - collect individuals' images at additional time points. Analyze feature robustness to temporal variabilities (e.g., organ movement, organ expansion/shrinkage)	+1	Both
8. 🟦🟪	Inclusion and exclusion criteria defined (e.g. a CT with 6 mm slice thickness cannot be analyzed)	+1 (inclusion criteria defined) +1 (exclusion criteria defined)	Both
9. 🟩	Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyze feature robustness.	+1	HCR
10 🟧	The diversity and distribution across diverse patient groups (e.g. according to sex/gender, ethnicity, age) in the datasets should be reported at training and testing to identify potential biases.	+1 (if diversity and distribution across diverse patient groups in the datasets have been reported and mitigation strategies need to be applied.)	Both
	(C) Method		
11 🟩	Use of post-processing harmonization to reduce multi-center acquisition variability e.g. Combat for HCR and CycleGANs for DL	+1	Both
12 🟦	Method and statistical plan pre-registered on a public platform (e.g. www.osf.io)	+1	Both
13 🟨🟩	The number of participating clinical sites in the training dataset	- 1 one centre +1 two centres +2 > three centres	Both
14 🟨🟩	Use of IBSI compliant radiomics features	+1	HCR
15 🟩	Multiple segmentations - possible actions are segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyze feature robustness to segmentation variabilities	+1 (multiple segmentation)	Both
16 🟩	Feature reduction based on the test-retest dataset, other method or adjustment for multiple testing - decreases the risk of overfitting. Consider feature robustness when selecting features	-3 (if neither measure is implemented), +3 (if either measure is implemented)	HCR
17 🟩	Multivariable analysis with non-radiomics features (e.g., age, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non-radiomics features	+1	HCR

18 🟩	Cut-off analyses - determine risk groups by either median, a previously published cut-off, or report a continuous risk variable or published method. Reduces the risk of reporting overly optimistic results	+1	Both
19 🟩	Random permutations to assess the risk of overfitting. Randomize the input variable to get ideally an AUC not different than 0.5 and therefore assess the risk of overfitting	+1	Both
20. 🟩 🟥	Investigate both handcrafted radiomics and deep learning, or a combination thereof, in an ensemble. Radiomics features may also help in the interpretability of deep learning	+1 for comparative analysis or ensemble of HCR and DL approaches.	Both
21 🟦	Quality Management System	+1 available online with internal audit, +3 iso certification or equivalent with external audit	Both
	(F) Evaluation		
22 🟩	Discrimination statistics - report discrimination statistics (e.g., C-statistic, ROC curve, AUC) and their statistical significance (e.g., p-values, confidence intervals). One can also apply a resampling method (for example, bootstrapping, cross-validation).	+1 (if a discrimination statistic and its statistical significance are reported), +1 (if also a resampling method technique is applied)	Both
23 🟩	Calibration statistics - report calibration statistics (e.g., Calibration-in-the-large/slope, calibration plots) and their statistical significance (e.g., p-values, confidence intervals). One can also apply a resampling method (for example, bootstrapping, cross-validation).	+1 (if a calibration statistic and its statistical significance are reported), +1 (if also a resampling method technique is applied)	Both
24 🟩	Comparison with previously published radiomics signatures and models by evaluation on a common dataset.	+1	Both
25 🟩	Validation - the validation is performed without retraining and adaptation of the cut-off value, providing crucial information about credible clinical performance.	-5 if validation is missing +2 if validation is based on a dataset from the same institute +3 if validation is based on a dataset from another institute +4 if validation is based on two datasets from two distinct institutes +5 if validation is based on three or more datasets from distinct institutes +6 if the validation is carried out on a third-party framework on an external dataset	Both
26 🟦	Prospective study registered in a trial database (real-world or In Silico), with sample size calculation - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+5 (for prospective validation), +1 (if the trial is pre-registered), +1 (if sample size calculation)	Both

27 🟩 🟪	Algorithm tested in a clinical environment e.g. a department of Radiology or Nuclear medicine	+2	Both
28 🟪	The composition of the external dataset should reflect the distribution of assessed classes in the real-world clinical setting.	+1	Both
29 🟧	Evaluation should also assess the performance due to discriminative biases (e.g. sex/gender, age, ethnicity) identified, and the level of fairness.	+1	Both
	(G) Interpretability and Explainability		
30 🟥	Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene-protein expression patterns) deepens understanding of radiomics and biology	+1	Both
31 🟥	Details on the intrinsic or post-hoc interpretability method or uncertainty estimation method utilized (e.g. attribution maps, SHAP analysis).	+1 (for details on interpretability methods or uncertainty estimation)	Both
32 🟥	Evaluation of the explanations using in-silico trials or by the clinicians (e.g. explanation satisfaction, trust score etc) .	+1 (for the sanity and/or evaluation of explanations)	Both
	(H) Utility		
33 🟪	Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (e.g. Dr. evaluation, TNM-staging for survival prediction, Dr. Assessment). This comparison shows the added value of radiomics	+2	Both
34 🟪	Potential clinical utility - report on the current and potential application of the model in a clinical setting (e.g., decision curve analysis)	+2	Both
35 🟪	Define the model's limitations and underperformance	+1	Both
36 🟪	Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (e.g., QALYs generated).	+1	Both
37 🟪	Level of automation for the clinical practice. 1. At level 0 (No Automation), a clinician performs the clinical task without using the radiomics model. 2. At level 1 (Clinical Assistance), the clinician uses the radiomics model's prediction for a part of the clinical task. 3. At level 2 (Partial Automation), the clinician considers the radiomics model's prediction for the clinical task before making the final recommendation. 4. At level 3 (Conditional Automation), the radiomics model provides the predictions for the clinical task under supervision	One point per level of automation of the software. Level 1 (Clinical Assistance) +1 Level 2 (Partial Automation) +2	Both

	and the clinician can intervene at any time. 5. At level 4 (High Automation), the radiomics model provides the predictions and the clinician's intervention is required for special (out-of-distribution) cases. 6. At level 5 (Full Automation), the radiomics model provides predictions for the clinical task without human intervention.	Level 3 (Conditional Automation) +3 Level 4 (High Automation) +4 Level 5 (Full Automation) +5	
38 🟦	The algorithm, source code, and coefficients are made publicly available. Add a table detailing the different versions of software & packages used.	+1	Both
39 🟦🟨	Open data - make data publicly available. Open data facilitates knowledge transfer and reproducibility of the study	+1 if scans are open source, +1 if the ROI/segmentations are open source, +1 if clinical, non-DICOM data, and outcomes are open source.	Both
40 🟦🟩	Define strategy for continuous learning to learn and improve over time from errors.	+1	Both
41 🟦	Define strategy to evaluate the model performance periodically due to data shifts	+1	Both
	Total Points (HCR = 70, DL = 64) = 100 %		

6 Discussion

A consensus was reached with all the participating partners within EUCanImage. Radiomics quality score sets ideal standards for radiomics analysis that may be very difficult to fulfill. The requirements of RQS 1.0 such as imaging at multiple timepoints and phantom study on all scanners are difficult to satisfy for retrospective studies[10]. These considerations can be helpful for prospective studies and ultimately improve the overall robustness of the radiomics model to meet the standard required for clinical translation. The mean score of radiomics studies reported by several reviews is already low [7,9,21] using RQS 1.0. Due to recent advancements and to alleviate the shortcomings of RQS 1.0, RQS 2.0 sets even higher standards. RQS 2.0 is more comprehensive and takes into account technical, clinical and ethical challenges posed by the emergence of artificial intelligence. In particular, RQS 2.0 ensures that radiomics models can be trusted and accepted by healthcare professionals, and that they treat patients of different groups and backgrounds equally. RQS 2.0 is an effective tool to highlight the deficiencies of recent radiomics studies and to serve as an important consideration for future radiomics studies.

7 Future Directions for EuCanImage in light of RQS 2.0

Radiomics Quality Score 2.0 is an ideal benchmark that is difficult to achieve but it outlines the steps and practices that are necessary for clinical translation. This final version of the document will be disseminated to experts within and outside the consortium to incorporate their feedback. RQS 2.0 will be implemented as an online tool within the EuCanImage platform or on its separate website (https://www.radiomics.world/rqs2). This tool defines the best practices that the AI WGs will follow during their AI tool development lifecycle.

Radiomics Quality Score 2.0 is a qualitative assessment tool that advocates the use of quantitative measures for AI tool evaluation and validation. WP6 of EuCanImage is implementing software pipelines that will carry out the quantitative evaluation of the AI tools in light of the principles outlined by RQS 2.0. These software pipelines will be integrated into the OpenEBench provided by Barcelona Supercomputing Center. This will allow the visualisation of the important assessment metrics satisfying the RQS 2.0 on OpenEBench platform. RQS 2.0 will lead to the development of AI tools within the EuCanImage consortium that are closer to the goal of clinical translation.

8 References

1. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014; 5: 4006. 2014.

2. Gardin I, Grégoire V, Gibon D, Kirisli H, Pasquier D, Thariat J, et al. Radiomics: Principles and radiotherapy applications. Crit Rev Oncol Hematol. 2019;138: 44–50.

3. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48: 441–446.

4. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14: 749–762.

5. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology. 2013;269: 8–15.

6. Pinto dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol. 2021;31: 1–4.

7. Park JE, Kim D, Kim HS, Park SY, Kim JY, Cho SJ, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol. 2020;30: 523–536.

8. Lee S, Han K, Suh YJ. Quality assessment of radiomics research in cardiac CT: a systematic review. Eur Radiol. 2022. doi:10.1007/s00330-021-08429-0

9. Stanzione A, Gambardella M, Cuocolo R, Ponsiglione A, Romeo V, Imbriaco M. Prostate MRI radiomics: A systematic review and radiomic quality score assessment. Eur J Radiol. 2020;129: 109095.

10. Spadarella G, Calareso G, Garanzini E, Ugga L, Cuocolo A, Cuocolo R. MRI based radiomics in nasopharyngeal cancer: Systematic review and perspectives using radiomic quality score (RQS) assessment. Eur J Radiol. 2021;140: 109744.

11. Won SY, Park YW, Ahn SS, Moon JH, Kim EH, Kang S-G, et al. Quality assessment of meningioma radiomics studies: Bridging the gap between exploratory research and clinical applications. Eur J Radiol. 2021;138: 109673.

12. Park JE, Kim HS, Kim D, Park SY, Kim JY, Cho SJ, et al. A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features. BMC Cancer. 2020;20: 29.

13. Fornacon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. Lung Cancer. 2020;146: 197–208.

14. Ugga L, Perillo T, Cuocolo R, Stanzione A, Romeo V, Green R, et al. Meningioma MRI radiomics and machine learning: systematic review, quality score assessment, and meta-analysis. Neuroradiology. 2021;63: 1293–1304.

15. Ponsiglione A, Stanzione A, Cuocolo R, Ascione R, Gambardella M, De Giorgi M, et al. Cardiac CT and MRI radiomics: systematic review of the literature and radiomics quality score assessment. Eur Radiol. 2021. doi:10.1007/s00330-021-08375-x

16. Ursprung S, Beer L, Bruining A, Woitek R, Stewart GD, Gallagher FA, et al. Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. Eur Radiol. 2020;30: 3558–3566.

17. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. Radiother Oncol. 2018;127: 349–360.

18. Wang Q, Li C, Zhang J, Hu X, Fan Y, Ma K, et al. Radiomics Models for Predicting Microvascular Invasion in Hepatocellular Carcinoma: A Systematic Review and Radiomics Quality Score Assessment. Cancers . 2021;13. doi:10.3390/cancers13225864

19. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med. 2021;140: 105111.

20. Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. J Pers Med. 2021;11. doi:10.3390/jpm11090842

21. Zhong J, Hu Y, Si L, Jia G, Xing Y, Zhang H, et al. A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. Eur Radiol. 2021;31: 1526–1535.

22. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155: 529–536.

23. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. BMJ Open. 2019;9: e025611.

24. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol. 2016;61: R150–66.

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

26. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162: 55–63.

27. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295: 328–338.

28. Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R, Mottaghy FM, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. PLoS One. 2021;16: e0251147.

29. Ibrahim A, Refaee T, Primakov S, Barufaldi B, Acciavatti RJ, Granzier RWY, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. Cancers . 2021;13. doi:10.3390/cancers13081848

30. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. Eur Radiol. 2017;27: 4498–4509.

31. Prayer F, Hofmanninger J, Weber M, Kifjak D, Willenpart A, Pan J, et al. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. Methods. 2021;188: 98–104.

32. Lee J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, et al. Radiomics feature robustness as measured using an MRI phantom. Sci Rep. 2021;11: 3973.

33. Peerlings J, Woodruff HC, Winfield JM, Ibrahim A, Van Beers BE, Heerschap A, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. Sci Rep. 2019;9: 4800.

34. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based bias field correction of MR images of the brain. IEEE Trans Med Imaging. 1999;18: 885–896.

35. Masson I, Da-ano R, Lucia F, Doré M. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal …. Medical. 2021. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14948?casa_token=4OwnRHSh-CoAAAAA:RgduV-2HMf2Qkof-MOFgRQvjRonz12AtyaGd1rY2i_7XKjD_d_WQVJFKEOtxbyYrRXgH8eH5ExwP0GV3

36. Lucia F, Visvikis D, Vallières M, Desseroit M-C, Miranda O, Robin P, et al. External

validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019;46: 864–877.

37. Duron L, Balvay D, Vande Perre S, Bouchouicha A, Savatovsky J, Sadik J-C, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One. 2019;14: e0213459.

38. Lee H, Huang C, Yune S, Tajmir SH, Kim M, Do S. Machine Friendly Machine Learning: Interpretation of Computed Tomography Without Image Reconstruction. Sci Rep. 2019;9: 15540.

39. Gallardo-Estrella L, Lynch DA, Prokop M, Stinson D, Zach J, Judy PF, et al. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. Eur Radiol. 2016;26: 478–486.

40. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. Phys Med Biol. 2020;65: 24TR02.

41. Andrearczyk V, Depeursinge A, Müller H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. J Med Imaging (Bellingham). 2019;6: 024008.

42. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. Transl Oncol. 2014;7: 88–93.

43. Caramella C, Allorant A, Orlhac F, Bidault F, Asselain B, Ammari S, et al. Can we trust the calculation of texture indices of CT images? A phantom study. Med Phys. 2018;45: 1529–1536.

44. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288: 407–415.

45. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol. 2015;50: 757–765.

46. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49: 1012–1016.

47. Pfaehler E, van Sluis J, Merema BBJ, van Ooijen P, Berendsen RCM, van Velden FHP, et al. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. J Nucl Med. 2020;61: 469–476.

48. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. J Nucl Med. 2015;56: 1667–1673.

49. Pfaehler E, Beukinga RJ, de Jong JR, Slart RHJA, Slump CH, Dierckx RAJO, et al. Repeatability of 18 F-FDG PET radiomic features: A phantom study to explore sensitivity

to image reconstruction settings, noise, and delineation method. Med Phys. 2019;46: 665–678.

50. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. NeuroImage. 2017. pp. 149–170. doi:10.1016/j.neuroimage.2017.08.047

51. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med. 2018;59: 1321–1328.

52. Guha I, Nadeem SA, You C, Zhang X, Levy SM, Wang G, et al. Deep Learning Based High-Resolution Reconstruction of Trabecular Bone Microstructures from Low-Resolution CT Scans using GAN-CIRCLE. Proc SPIE Int Soc Opt Eng. 2020;11317. doi:10.1117/12.2549318

53. Dewey BE, Zhao C, Reinhold JC, Carass A, Fitzgerald KC, Sotirchos ES, et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. Magn Reson Imaging. 2019;64: 160–170.

54. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision. 2017. pp. 2223–2232.

55. Zhao F, Wu Z, Wang L, Lin W, Xia S, Shen D, et al. Harmonization of Infant Cortical Thickness Using Surface-to-Surface Cycle-Consistent Adversarial Networks. Med Image Comput Comput Assist Interv. 2019;11767: 475–483.

56. Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1703.05192

57. Yi Z, Zhang H, Tan P, Gong M. Dualgan: Unsupervised dual learning for image-to-image translation. Proceedings of the IEEE international conference on computer vision. 2017. pp. 2849–2857.

58. Goodfellow I. NIPS 2016 Tutorial: Generative Adversarial Networks. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1701.00160

59. Yan C, Lin J, Li H, Xu J, Zhang T, Chen H, et al. Cycle-consistent generative adversarial network: Effect on radiation dose reduction and image quality improvement in ultralow-dose CT for evaluation of pulmonary tuberculosis. Korean J Radiol. 2021;22: 983–993.

60. Graziani M, Andrearczyk V, Müller H. Regression Concept Vectors for Bidirectional Explanations in Histopathology. Understanding and Interpreting Machine Learning in Medical Image Computing Applications. Springer International Publishing; 2018. pp. 124–132.

61. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer; 2019.

62. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical

prediction models. J Clin Epidemiol. 2015;68: 279–289.

63. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol. 1982;115: 92–106.

64. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging. arXiv e-prints. 2021; arXiv:2109.09658.

65. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405: 442–451.

66. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30: 1105–1117.

67. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning. New York, NY, USA: Association for Computing Machinery; 2005. pp. 625–632.

68. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics BMC Med. 2019;17: 230.

69. Lad F. The Calibration Question. Br J Philos Sci. 1984;35: 213–221.

70. Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy Explainability Acceptance: A New Metric to Measure the Trustworthiness of Interpretable AI Medical Diagnostic Systems. Complex, Intelligent and Software Intensive Systems. Springer International Publishing; 2021. pp. 35–46.

71. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learn Syst. 2021;32: 4793–4813.

72. Temme M. Algorithms and transparency in view of the new general data protection regulation. Eur Data Prot Law Rev. 2017;3: 473–485.

73. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

74. Du R, Lee VH, Yuan H, Lam K-O, Pang HH, Chen Y, et al. Radiomics Model to Predict Early Progression of Nonmetastatic Nasopharyngeal Carcinoma after Intensity Modulation Radiation Therapy: A Multicenter Study. Radiol Artif Intell. 2019;1: e180075.

75. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med. 2020;26: 1229–1234.

76. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4: 4.

77. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, von Tengg-Kobligk H, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. Radiol Artif Intell. 2020;2: e190043.

78. Adebayo J, Gilmer J, Muelly M. Sanity checks for saliency maps. Adv Neural Inf Process Syst. 2018. Available: https://proceedings.neurips.cc/paper/8160-sanity-checks-for-saliency-maps

79. Orlhac F, Nioche C, Klyuzhin I, Rahmim A, Buvat I. Radiomics in PET Imaging:: A Practical Guide for Newcomers. PET Clin. 2021;16: 597–612.

80. Drummond DC. Replicability is not Reproducibility: Nor is it Good Science. Cogprints. 2009 [cited 6 Feb 2022]. Available: http://cogprints.org/7691/

81. Peng RD. Reproducible research in computational science. Science. 2011;334: 1226–1227.

82. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol. 2006;163: 783–789.

83. Shadrin SS, Ivanova AA. Analytical review of standard Sae J3016 «taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles» with latest updates. Avtomobil' Doroga Infrastruktura. 2019. Available: https://www.adi-madi.ru/madi/article/view/811?locale=en_US

84. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019. doi:10.1186/s41512-019-0064-7

85. Broome J. Qalys. J Public Econ. 1993;50: 149–167.

86. van Wijk Y, Ramaekers B, Vanneste BGL, Halilaj I, Oberije C, Chatterjee A, et al. Modeling-Based Decision Support System for Radical Prostatectomy Versus External Beam Radiotherapy for Prostate Cancer Incorporating an In Silico Clinical Trial and a Cost-Utility Study. Cancers . 2021;13. doi:10.3390/cancers13112687