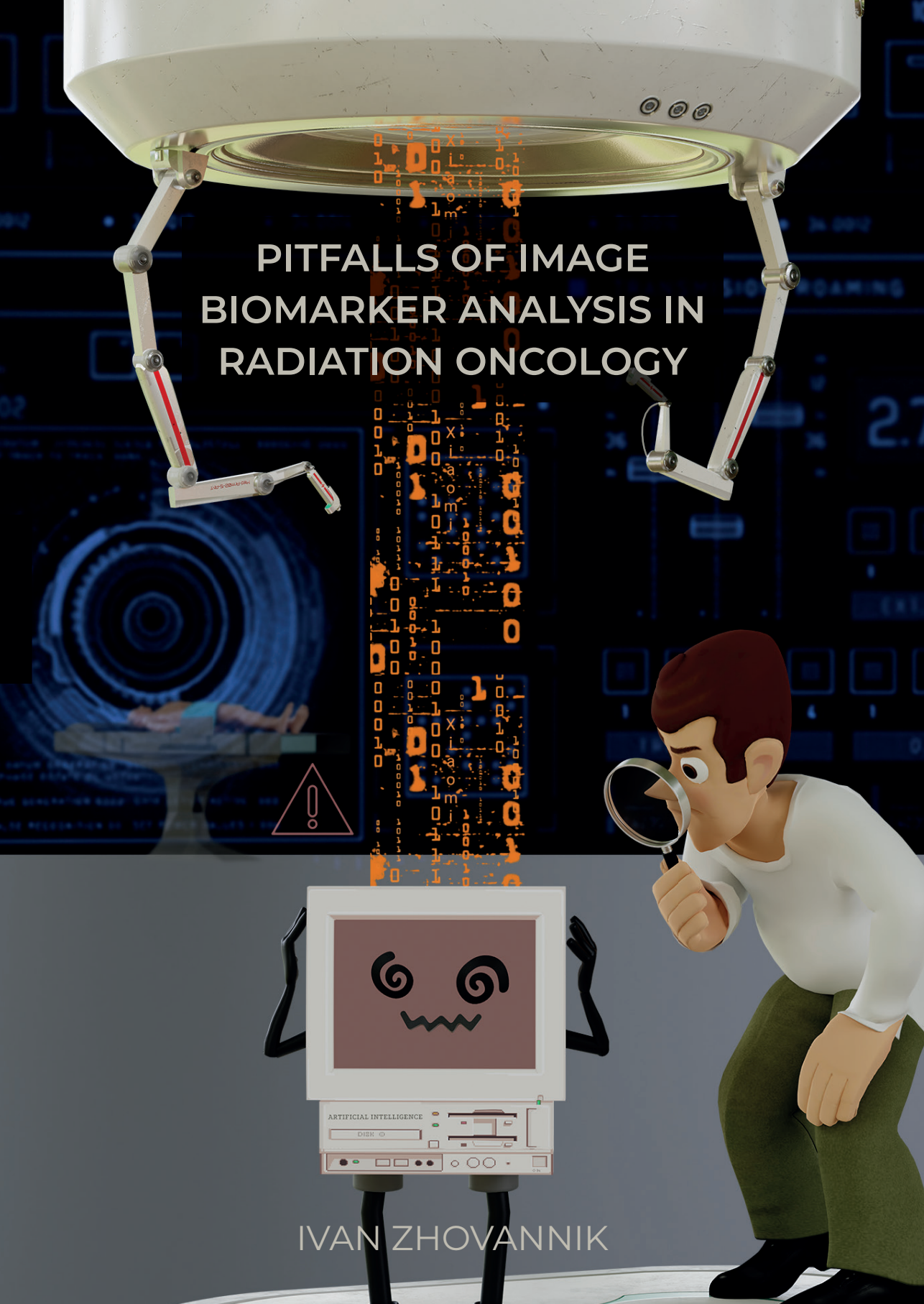
PITFALLS OF IMAGE BIOMARKER ANALYSIS IN RADIATION ONCOLOGY



IVAN ZHOVANNIK

Pitfalls of Image Biomarker Analysis in Radiation Oncology

Ivan Zhovannik

Cover design: Kate Lediakhova | www.linkedin.com/in/kate-lediakhova
Lay-out: Publiss | www.publiss.nl
Print: Ridderprint | www.ridderprint.nl

© Copyright 2022: Ivan Zhovannik, The Netherlands

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, by photocopying, recording, or otherwise, without the prior written permission of the author.

Pitfalls of Image Biomarker Analysis in Radiation Oncology

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 30 maart 2023
om 12.30 uur precies

door

Ivan Yevgenyevich Zhovannik
geboren op 15 juli 1992
te Arzamas-16, Rusland

Promotoren:

prof. dr. J. Bussink

prof. dr. A.L.A.J. Dekker (Maastricht University)

Copromotoren:

dr. R. Monshouwer

dr. R.R.R. Fijten (Maastro)

Manuscriptcommissie:

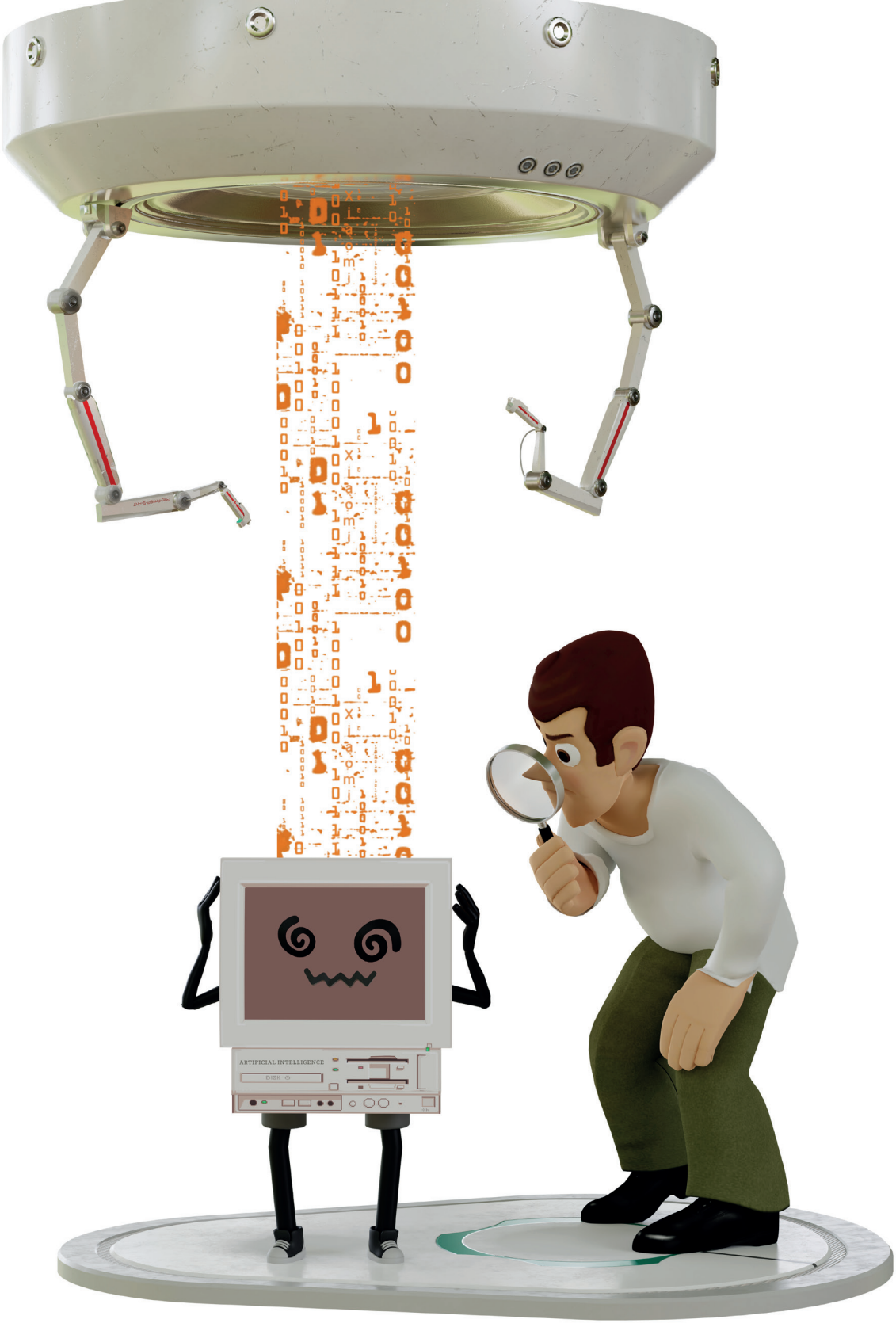
prof. dr. B. van Ginneken

prof. dr. R. Beets-Tan (Maastricht University)

prof. dr. M.S. Hoogeman (Erasmus Universiteit Rotterdam)

Table of contents

CHAPTER 1	General Introduction and Chapter Outline	7
CHAPTER 2	The state of Reproducibility and Repeatability in Radiomics	15
CHAPTER 3	Uncertainty Estimation for Robust Image Biomarkers – Segmentation Uncertainty Example	35
CHAPTER 4	Image Acquisition Uncertainty as a Source of Bias in Radiomics	59
CHAPTER 5	Gross Tumor Volume as a Confounding Factor in Radiomics	71
CHAPTER 6	Plausibility and Redundancy Analysis to Select FDG-PET Textural Features in Non-Small Cell Lung Cancer	91
CHAPTER 7	Radiomics Integration into a Picture Archiving and Communication System (PACS)	121
CHAPTER 8	Multicenter Federated Learning in Radiomics	131
CHAPTER 9	General Discussion and Future Perspectives	151
APPENDIX	Summary	162
	Dutch Summary	163
	Data Management Statement	164
	Portfolio	165
	CV	166
	Acknowledgements	167


ARTIFICIAL INTELLIGENCE

Chapter 1

General Introduction and Thesis Outline

RADIATION ONCOLOGY

Cancer is a disease in which genetically altered (mutated) cells grow and divide in an uncontrolled way, potentially spreading to other body parts. Each cancer case is unique due to the set of both genetic and epigenetic changes acquired by the cancer cells during their development and evolution. This disease heterogeneity requires an individualized treatment design for each patient. The most common treatment modalities include, but are not limited to, chemotherapy, surgery, immunotherapy, and radiotherapy. These therapies are often combined to get the most effective treatment.

Radiation Oncology or radiotherapy is a cancer treatment, which utilizes ionizing radiation to induce damage to tumor cells. The ionizing radiation induces various degrees of damage in cells on DNA level. The damage to the DNA can eventually lead to cell death. The effect of radiation is not specific: both cancerous and healthy tissues are affected – though they have different capabilities to repair the DNA damage. The therapy should be rigorously planned in order to deliver large enough dose to the tumor cells, so the cancer cells die or stop proliferating, and as low as possible dose to the healthy tissues to inflict as insignificant damage as possible to reduce treatment toxicity.

In order to plan, deliver, and monitor radiation therapy, different sources of data are utilized in clinical decision making. Among these sources, imaging is by far the largest. Clinical data-driven insights might bring benefit clinical processes and decision making in prognosis, toxicity modeling, and tissue phenotype characterization. To derive insights from the data efficiently, good data management and analysis practices must be implemented – both are covered in this thesis. Data analysis is usually performed with various statistical techniques, such as machine learning and statistical hypothesis testing.

MACHINE LEARNING AND STATISTICS FOR DATA ANALYSIS

Statistical methods are essential in deriving insights from data. Machine learning, a statistical discipline, provides statistical insights into data by fitting a mathematical model, which should theoretically approximate training and upcoming unseen data well. Machine learning methods can be roughly divided into supervised (ground truth is provided to train the model) and unsupervised (no labeled data is used) methods.

Another important statistical method concerns hypothesis testing. Statistical hypothesis testing helps to derive conclusions from data by assessing statistical

significance. In this thesis, hypothesis testing is used, for instance, to assess the quality of machine learning models for survival stratification [1].

Bayesian statistical techniques help to assess causal relationships in data, find confounding factors, and estimate uncertainties. In the following chapters, we further highlight the importance of assessing causal relationships in clinical data and give examples of hidden confounders, which may mislead research conclusions. Markov chain Monte Carlo is a Bayesian technique to sample from a probability distribution, that can be used to estimate the uncertainties in data acquisition, feature extraction, and model performance described in CHAPTER 3.

IMAGING IN RADIATION ONCOLOGY

One of the largest sources of patient data acquired during radiation therapy is imaging data. Imaging data is acquired routinely during cancer diagnosis, treatment planning, delivery, and follow-up.

There are various imaging modalities, the most common being Computed Tomography (CT) – used to measure tissue density contrast, Magnetic Resonance Imaging (MRI) - used to measure tissues' magnetic properties and metabolism, and Positron Emission Tomography (PET) – designed to assess tissue metabolism by measuring uptake of various radioactive tracers.

In order to estimate the data size per patient, an open DICOM data source can be considered, such as the LUNG1 dataset as an example [2]. The dataset shows 33 GB of well-processed image and delineation data for 422 patient subjects – resulting in approximately 80 MB of Computed Tomography series accompanied by delineations. In a realistic radiation oncology setting, tens of DICOM series are acquired during diagnostics, treatment planning, treatment delivery, monitoring, and follow-up. In addition to CT, PET, and MRI, typically, there are cone-beam CT images, RTstruct delineations, RTplan treatment planning files, RTdose dose distribution series generated. This increases the realistic data size of a subject at least two fold, to approximately 1 GB per patient.

With improving image-guidance, the size of data per subject will continue to grow. For instance, for an MR-Linac treatment (MR-guided radiotherapy), four or more MRI images per treatment session are routinely acquired. These data potentially contain valuable information for clinical decision making, but in order to mine this information (such as treatment response monitoring), the data must first be reduced to a clinically utilizable format. This formatting requires appropriate mathematical methods to transform the imaging data.

RADIOMICS AS QUANTITATIVE IMAGE BIOMARKERS

As imaging data is acquired routinely for each oncological patient, this data can be reused in clinical machine learning algorithms in order to assist clinical routines and decision making. As machine learning algorithms require a unified data format to interpret the data correctly, the imaging data must be pre-processed to match the input format (data standardization) and input variable distribution (data harmonization). The pre-processing may include but not limited to removing or regrouping metadata, filtering, resampling, binning, contouring, feature engineering etc. The generalized workflow is shown in Figure 1

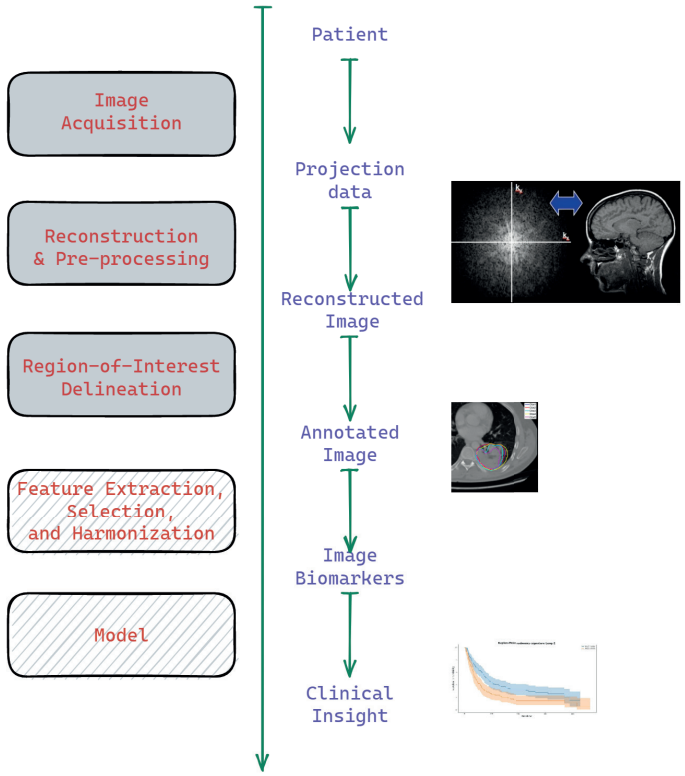


Figure 1. *The image biomarker workflow: Various parts of the workflow can be merged or split depending on implementation (see data-driven vs expert-driven approaches below)*

As it was pointed out, image data are up to 1 GB in size per subject and carry much latent information, not all of which is relevant for a specific clinical and machine learning task. Therefore, these data should be compressed into some representation – a process referred to as 'feature extraction'. The extraction process can be either performed in

a hand-crafted manner with an analytical algorithm (expert-driven) or by using an optimization algorithm (data-driven). Expert- driven or handcrafted feature extraction is commonly referred to as radiomics. Data-driven approaches are overwhelmingly based on deep learning optimization of convolutional neural networks (CNNs)

Although data-driven feature extraction based on deep learning can generate features that better describe the training data, those generated features' definition is harder to trace and their value might still carry some degree of uncertainty based on how the data was acquired. In this thesis, we focus on the analytical hand-crafted features, commonly referred to as radiomics, as it allows us to control for feature engineering bias: the handcrafted radiomic features are always extracted according to fixed mathematical definition.

Radiomics as a term for quantitative image biomarkers started gaining moment in 2012 with only three studies per year according to PubMed, five years later the number of studies crossed 254 per year. In 2021, PubMed reported 2270 radiomic studies per year.

FEDERATED NATURE OF RADIOMICS

Commonly, radiomics and image analysis generally is performed by firstly aggregating the data on a centralized repository and then analyzing it. Such approach is commonly referred to as centralized learning. However, imaging data is bulky and acquired in a decentralized manner, where every radiation oncology treatment site hosts and maintains at least one source of data. Imaging data storage is usually taking place in a so-called picture archiving and communication system (PACS), which performs image storage, retrieval, and sometimes analysis. In order to build a robust image biomarker, much training data is needed to derive the relationship between imaging data and the outcome that we are trying to predict.

Due to various reasons, including data privacy (e.g. General Data Protection Regulation, GDPR), legal aspects of data ownership, size of the data, and labor intensity to de-identify and transfer the data, it is not feasible to perform scalable image analysis studies in a centralized manner. Federated learning is an approach to develop machine learning models for federated data sources without the need for data sharing. Federated learning requires the data to be prepared locally so that a model can find and access the data. Findability and Accessibility, together with Interoperability and Reusability formed the FAIR concept for data management [3]. This thesis covers the topics of radiomics integration in a PACS (CHAPTER 7) and federated learning for radiomics (CHAPTER 8).

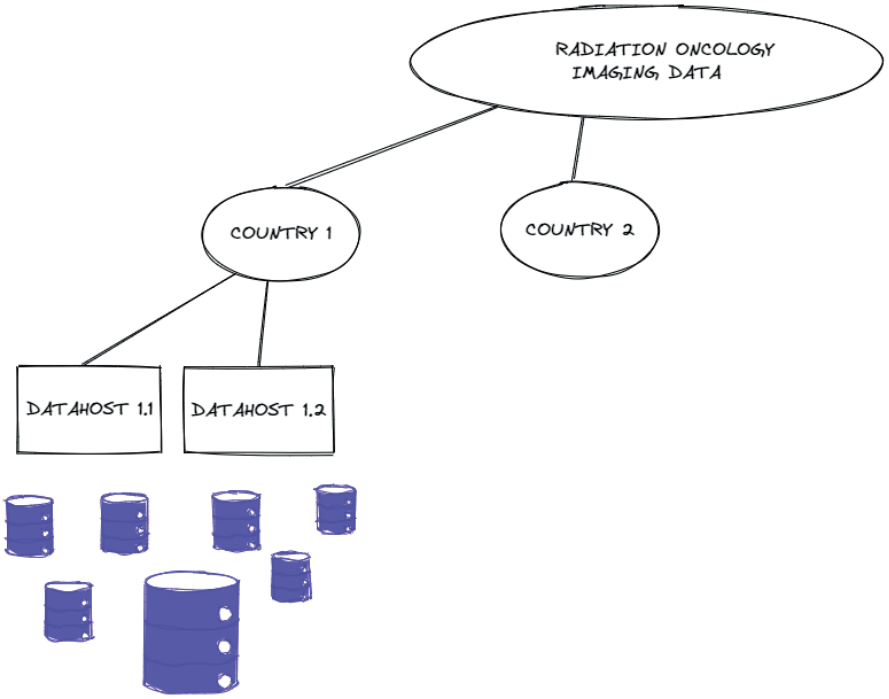


Figure 2. *The federated nature of imaging studies in radiation oncology: In order to execute a large scale study, multiple data sources must be utilized.*

THE OUTLINE OF THIS THESIS

Even though the utilization of radiomic features as biomarkers for clinical prognostic and predictive modeling sounds promising, there is still a major issue with radiomics to be addressed, namely, its reproducibility and repeatability. In *CHAPTER 2*, we assess the current state of radiomics reproducibility and repeatability. Most of the studies in this and preceding reviews report on poor reproducibility of the radiomic features in various imaging modalities, disease sites, and radiomic classes. The uncertainties in the protocol settings, namely, in image acquisition, processing, and segmentation induce uncertainty in radiomic values.

We proposed the generalized framework for uncertainty estimation in radiomics in CHAPTER 3 and used segmentation uncertainty estimation as an example to show how the clinically adequate variability in tumor delineation induces uncertainty in model performance. We also proposed some applications for the uncertainty reduction.

The first major pitfall of image biomarker analysis is the underestimation of uncertainty induced by the measurement of image biomarkers. In order to characterize and subsequently harmonize the radiomic uncertainty in Computed Tomography, we performed a phantom-based study in CHAPTER 4. Both CHAPTERS 3 and 4 investigate the influence of different sources of uncertainty in domain (related to image biomarker measurement equipment) parameters on radiomic features.

The second major pitfall of image biomarker analysis is the presence of confounding factors. CHAPTER 5 shows how to identify the influence of the major confounder for texture and first-order radiomics – region-of-interest (ROI) volume. In order to show the clinical application and generalize over various imaging modalities, we performed FTG-PET radiomic analysis in CHAPTER 6. This chapter presents a sample FTG-PET radiomic study with some of the sanity checks for confounder detection implemented.

In order to breach the space between the research and the clinic, technical advancements must be met, therefore, we describe technical applications for radiomics in CHAPTERS 7 and 8. Namely, CHAPTER 7 presents radiomics integration in a clinical Picture Archiving and Communication System (PACS) and CHAPTER 8 gives an example of a federated radiomics analysis.

In the last chapter, CHAPTER 9, we provide the general discussion and the future perspective for the image biomarker research.

[1] Z. Shi *et al.*, "Distributed radiomics as a signature validation study using the Personal Health Train infrastructure," *Scientific Data*, vol. 6, no. 1, 2019, doi: 10.1038/s41597-019-0241-0.

[2] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014, doi: 10.1038/ncomms5006.

[3] M. D. Wilkinson *et al.*, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016, doi: 10.1038/sdata.2016.18.

Chapter 2

The state of Reproducibility and Repeatability in Radiomics

Elisabeth Pfaehler*, Ivan Zhovannik*, Lise Wei, Ronald Boellaard, Andre Dekker, René Monshouwer, Issam El Naqa et al. "A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features." *Physics and imaging in radiation oncology* 20 (2021): 69-75.

ABSTRACT

Purpose: Although quantitative image biomarkers (radiomics) show promising value for cancer diagnosis, prognosis, and treatment assessment, these biomarkers still lack reproducibility. In this systematic review, we aimed to assess the progress in radiomics reproducibility and repeatability in the recent years.

Methods and Materials: Four hundred fifty-one abstracts were retrieved according to the original PubMed search pattern with the publication dates ranging from 2017/05/01 to 2020/12/01. Each abstract including the keywords was independently screened by four observers. Forty-two full-text articles were selected for further analysis. Patient population data, radiomic feature classes, feature extraction software, image preprocessing, and reproducibility results were extracted from each article. To support the community with a standardized reporting strategy, we propose a specific reporting checklist to evaluate the feasibility to reproduce each study.

Results: Many studies continue to under-report essential reproducibility information: all but one clinical and all but two phantom studies missed to report at least one important item reporting image acquisition. The studies included in this review indicate that all radiomic features are sensitive to image acquisition, reconstruction, tumor segmentation, and interpolation. However, the amount of sensitivity is feature dependent, for instance, textural features were, in general, less robust than statistical features.

Conclusions: Radiomics repeatability, reproducibility, and reporting quality can substantially be improved regarding feature extraction software and settings, image preprocessing and acquisition, cutoff values for stable feature selection. Our proposed radiomics reporting checklist can serve to simplify and improve the reporting and, eventually, guarantee the possibility to fully replicate and validate radiomic studies.

1. Introduction

"Radiomics", the automated extraction of imaging biomarkers from patients' scans has gained an increasing interest in the last decade. Several radiomics studies have reported promising results for cancer diagnosis, prognosis, or evaluation of treatment response [1–3]. Radiomics studies span common volumetric imaging modalities such as Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI). The numbers of studies investigating applications of radiomics have increased dramatically in the last years. In 2019 alone, 728 studies were indexed in PubMed relating to radiomics studies. However, there remains a translational gap between academic study and clinical utilization [4]. One challenge that makes a clinical implementation of radiomics difficult is the problem of replicating published results. These difficulties are due to the unavailability of input images and software used for computations, poor reporting of study design, and lack of metadata associated with radiomic studies [5]. External validation of radiomic models has been hampered by models trained on small institutional cohorts, prevalence of overfitting and high false positive discovery rates [6]. Additionally, there is a dissonance between the study design methodology and guidelines from TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) [7] that strongly recommends the validation of prediction models on independent (non-randomly partitioned) datasets and full transparency in reporting.

In a previous review aiming to find a consensus in literature on reproducible and repeatable features, the authors indicated a lack of clear consensus in radiomics methodology. Moreover, the review concluded that deficiencies in reporting of study design, methodology and results were hampering the transparency and reproducibility of radiomic studies [5]. However, the previous effort to identify a subset of features likely to be generally reproducible and repeatable, via a quantitative meta-analysis, was encumbered by two issues identified in the literature: (i) incomplete reporting of radiomic analysis procedure and (ii) heterogeneity in metrics used to report feature repeatability and/or reproducibility. Thus, it was not possible in previous work to perform a conclusive quantitative meta-analysis for every imaging modality (CT, PET, MRI).

Considering the increase in radiomic studies, this work is a revised review about the repeatability and reproducibility of radiomic features, with multiple aims: 1. to identify a set of repeatable/reproducible features for each imaging

modality; 2. to verify if consensus has recently emerged regarding a list of major factors impacting on reproducibility/repeatability; 3. to isolate a set of repeatable/reproducible features across different modalities in both human and phantom data; 4. to verify if one of the largest issues identified in previous work, being the poor quality of reporting, has been addressed in new studies. To address aim (4) and support high-quality reporting of radiomic studies, we propose a checklist, which includes all necessary steps to fully reproduce a radiomic study.

2. Methods and Materials

2.1. Eligibility criteria

Peer-reviewed full-text articles in the English language eligible for this review must have been published between 2017/05/01 to 2020/12/01. One electronic database (PubMed) was used to search for records (see supplementary material for query filter). Only studies investigating radiomic features extracted A) from one of the imaging modalities CT, PET, or MRI and B) from radiologic phantoms or from human persons suffering from at least one primary tumor were eligible for review. Included articles had to report on the repeatability/reproducibility of radiomic features at least oneof the following aspects: image acquisition/reconstruction parameters, effect of image pre-processing such as smoothing, or segmentation method. Studies must report a statistical metric assessing the degree of robustness (such as the Interclass- Correlation Coefficient (ICC)).

2.2. Study records

Selection process : After the literature search, titles and abstracts were checked for matching the described criteria by four independent observers. Each reviewer voted if an article was eligible for review. In case of disagreement amongst reviewers, a consensus was obtained by joint discussion. EP and IZ reviewed the phantom studies, while AT and LIW reviewed the patient studies.

Data extraction: We extracted information about the datasets used for radiomics (e.g. primary tumor type, or phantom details), details about the segmentation method used, details about radiomic feature extraction such as the interpolation method, details about the statistical analysis and the summary of results. Details of the reviewer form are given in Table 1a. DICOM attributes about the important imaging details are summarized in Table 1b for the imaging modalities.

Creation of the radiomic reporting checklist: In order to generate a radiomic reporting checklist, we attempted to literally reproduce each step of a study by following the description in each manuscript. The checklist complements the points mentioned by Vallieres et al. [8] and the IBSI reference manual [9,10] and necessary information to reproduce a patient study (i.e. patient inclusion, statistical analysis). This checklist was also inspired by the QUADAS-2, a tool for the quality assessment of diagnostic accuracy studies [9]. The development of "signaling questions" has been carried out by the authors of this paper, who have at least 5 years of expertise in the radiomic domain and are members of tasks forces for radiomic standardization such as the IBSI [10]. Each of the authors defined a list of signaling questions, which were fundamental to reproduce a study. After a round of discussion, the final list of signaling questions was obtained. If a step was not reported and could therefore not be replicated, this step was noted, and we attempted to continue to the next step with a "best guess". This procedure was followed until all steps of the radiomics workflow was completed or guessed.

The risk probability was estimated as the number of "no's divided by the number of "signaling" questions, expressed in percentages.

$$risk\ of\ bias = \frac{|questions\ answered\ with\ no|}{|questions|} \cdot 100\%$$

E.g. if two out of four questions were answered with a "no", the risk of bias is calculated as:

$$risk\ of\ bias = \frac{2}{4} \cdot 100\% = 50\%$$

Supplementary tables 2a-2b show the risk assessment sheets for human and phantom studies, respectively. The only differences between human and phantom studies can be found in section A. This radiomic checklist was proposed in order to tackle the limitations of using only "quality scores" to evaluate such a complex question as methodological quality.

2.3. Outcomes and prioritization

The primary outcome of this review was the degree of repeatability/ reproducibility of a radiomic feature. The secondary outcomes were the

impact of image acquisition and reconstruction settings, preprocessing steps, and tumor segmentation on the reliability/reproducibility of radiomic features. Additional outcomes were the metrics used for reporting on reliability/reproducibility. Finally, the radiomic reporting checklist was used to evaluate the quality of reporting of analyzed studies.

2.4. Risk of bias in individual studies

Two reviewers independently reviewed the studies. In a discussion round, both observers merged their results. This step was performed in order to avoid that the results were biased based on single reviewer's judgement. Forced consensus was used.

3. Results

3.1. Literature search

A total of 451 abstracts not retrieved from the previous review were found while searching PubMed with the aforementioned search filter. After reviewing the abstracts, 42 studies fulfilled the inclusion criteria and seven additional studies had been included as prior knowledge in the field. The PRISMA flowchart illustrating the selection process is shown as Figure 1.

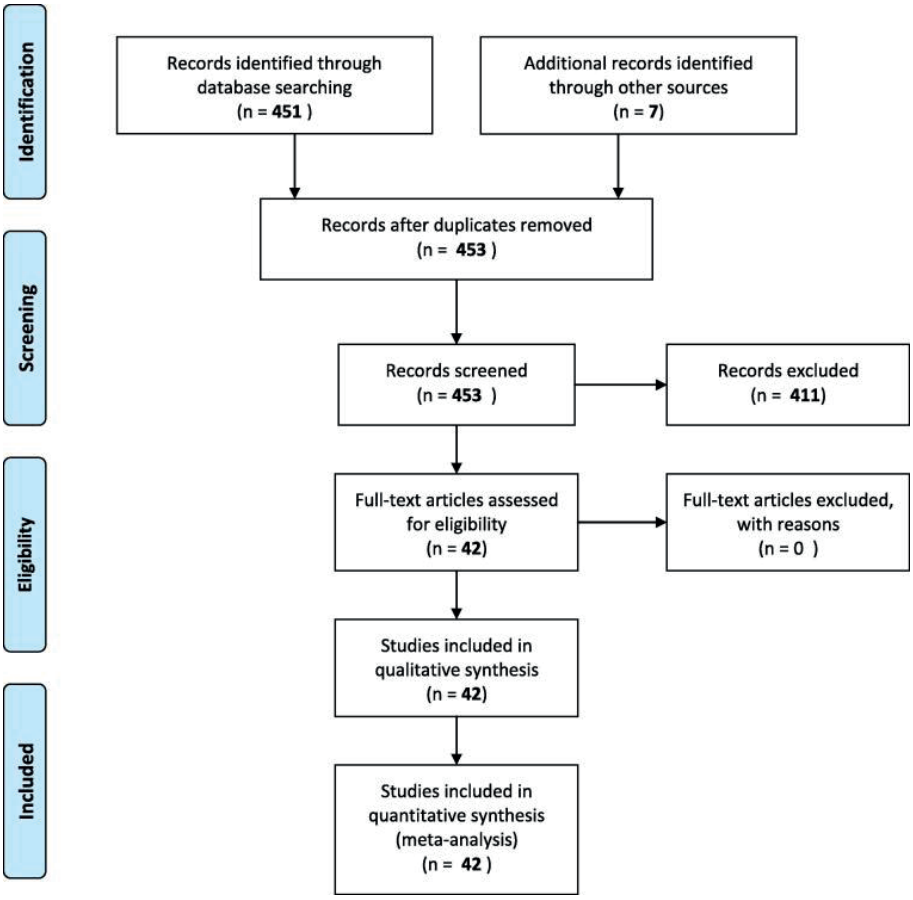


Figure 1. *Preferred Reporting Items for Systematc Reviews and Meta-Analyses (PRISMA) flow chart. The primary PubMed search returned 168 studies. 7 studies were added from primary knowledge. After abstract screening and full-text analysis, a total of 26 studies were included in the qualitative synthesis.*

Of these 42 studies, 29 studies were clinical (human subjects) studies. A summary on the study characteristics are displayed in Table 3a. Eight of these studies reported on PET, nine on MR, and twelve on CT images, with only one prospective study [12]. The number of patients included in the studies ranged from a minimum of 14 to a maximum of 465. Five consisted of multi-institutional studies. Most of the studies were focused on reproducibility, with one investigating both reproducibility and repeatability [13].

The other thirteen studies used imaging phantoms to assess repeatability/ reproducibility. The general characteristics are summarized in Supplementary

table 3b. Two studies reported on MR, six on PET and five on CT images. All studies were retrospective, with one study that provided a publicly available dataset [14]. Six phantom articles reported on feature reproducibility and seven on repeatability.

Five clinical studies [15–18] and two phantom studies [14] made their image dataset publicly available or used a publicly available dataset. A second phantom study used a digital phantom that was already publicly available [19].

17 patient and five phantom studies used publicly available software to calculate radiomic feature values. Pyradiomics was used in eight human studies and one phantom study, and was the most frequently used software [20]. Three patient studies used the open-source software CGITA [21], two patient studies used MaZda [22], one phantom study used LifeX [23], and one phantom and one patient study used IBEX [24]. Furthermore, three studies used open-source code written in Matlab. Six human and three phantom studies failed to mention the programming language used [25].

The most frequently used metric to assess reproducibility was the ICC which was used in 19 human studies and four phantom studies. Five human studies and one phantom study used the Concordance Correlation Coefficient (CCC), however a total of four human studies and two phantom studies used other metrics not described above. Some studies used more than one metric to assess feature stability. In general, the cut-off value to dichotomize stable versus unstable features were highly heterogeneous across studies. One study failed to mention the cut-off value for the ICC, while the threshold for “excellent feature” stability differed from 0.5 to 0.85 in the other studies. One study used the half-width of the ICC confidence interval as threshold for repeatable features. CCC values above 0.8, above 0.7 or equal or higher than 0.85 were considered as robust.

3.2. Factors impacting radiomic feature values

The variety of analyzed settings makes it difficult to draw a general conclusion. In summary, all patient studies confirm that differences in image acquisition, reconstruction, preprocessing, and discretization have an impact on radiomic feature values [12,16,25,27–33]. While the sensitivity to these factors is feature dependent, this sensitivity is found for all cancer types and imaging modalities. To mitigate this effect, Park et al. demonstrated that CT feature reproducibility can be improved by using a CNN-based super resolution algorithm [34]. For CT studies, Erdal et al. pointed out that a slice thickness of 2 mm leads to the most accurate shape features [35]. It remains unclear if radiomic features extracted from images

acquired under different conditions also lead to different conclusions. In one head and neck PET study, strong dependencies of radiomic features with respect to digital image pre-processing parameters was shown, but these differences were not important enough to affect the prognostic power of radiomic features. Same conclusion holds in the study about nasopharyngeal carcinoma [27].

Many studies also reported on the sensitivity of radiomic features to differences in segmentation [13,15,17,36–46]. Also here, the robustness is feature dependent. Some studies demonstrated that the majority of features was stable to differences in tumor segmentation for lung [42] and oesophageal cancer. However, in one lung PET study [41], the shape metric sphericity was found to change its prognostic value when using different delineations. Also in CT studies, contours delineated by different clinicians impacted the prognostic value of radiomic features. For head and neck cancer, in one PET and one MR study features were found to be very sensitive to differences in tumor delineation [35,36,40,47]. Overall, semi-automated and automated algorithms produced more stable results.

Individual studies showed that the interpolation method, image discretization, voxel size and motion blurring had an effect on radiomic feature values [35,48–51] as listed in detail in the Supplementary material.

Also results from PET phantom studies agreed that image reconstruction protocols (in particular matrix size) had strong impact on feature reproducibility, but with different level of sensitivity per feature categories. In [52], the authors pointed out that smaller volumes seem to result in lower repeatability of feature values. In contrast to these results, Ger et al. reported that most features resulted in a good or excellent reliability when the phantom was scanned on the same scanner with different acquisition protocols [53].

All CT phantom studies focused on differences in acquisition settings such as tube current and agreed that difference in acquisition protocols strongly impact feature reproducibility. However, it is difficult to draw a consensus on which features are stable: The authors in [54] demonstrated that the use of delta radiomic features, i.e. differences of feature values between two different scans of the same patient, increases the repeatability of features .

In one MR phantom study, Baessler et al. [55] investigated both reproducibility and repeatability of radiomic features using a physical phantom scanned using different sequences. The investigators showed that radiomic features extracted from FLAIR (Fluid Attenuated Inversion Recovery) images were more repeatable than features from T1- and T2-weighted images.

3.3. Stable feature categories

As stated above, the variety of tumor types and investigated parameters makes it difficult to identify features which are in general robust. The feature groups found to be stable by the majority of studies were statistical and morphological features, as well as GLCM and GLRLM features, while GLSZM and NGLDM features were found to be less robust. However, a few studies reported the contrary: Yang et al. found in simulated PET lung cancer data that NGLDM features were the most robust, while GLCM features were the least robust feature group [56]. One CT phantom study showed that statistical features are less robust than textural features [57]. Baessler et al. also showed in their MR phantom study that GLSZM features were more robust than GLCM features [55]. One study showed that features extracted from Fourier transformed images are the most robust, a feature group none of the other papers investigated. The authors in [54] demonstrated that the use of delta radiomic features, i.e. differences of feature values between two different scans of the same patient, increases the repeatability of features .

3.4. Radiomic reporting checklist

Supplementary tables 4a-5b summarizes the radiomic reporting checklist risk assessment for clinical and phantom studies. None of the studies scored a zero-risk bias probability. An example of the check-list can be found in the Supplementary material.

For clinical studies, the lowest median risk was achieved in the “study design” domain. Information about the creation of the binary mask (question B4-imaging domain) was the least reported item with only one study providing detailed information. Twelve patient studies missed to report on feature specific parameters (questions C5-C6-radiomic pipeline domain) such as feature aggregation and. In contrast, only three patient studies missed to provide any detail of the software which is a clear improvement when compared with the previous review.

For phantom studies, there were no risks of biases in the “study design” and “statistical analysis” domains. Compared to clinical studies a) all but one phantom study included a table reporting all statistical results (signaling question E3- data and metadata availability domain), and b) all but two studies failed to report on how the binary mask was created from the segmentation (question B4-imaging domain).

4. Discussion

The studies included in this review confirm almost all findings of the previous published review. Major issues remain the little number of studies that made their data publicly available, the heterogeneity of used metrics and cut-off values for the assessment of feature robustness, and the lack of detailed reporting. To ease the way to reproduce a study, we invite again the radiomics community to make their data and metadata publicly available. The variability in metrics and cut-offs used to categorize the features into good/poor reproducibility/repeatability, makes it difficult to compare the results of the studies. While there is no evidence that a specific metric should be used for analysis, a description about the metrics, as well as statistical hypothesis underlying the data analysis and specific cut-offs applied should be reported to guarantee the transparency and the reproducibility of the study. We also strongly recommend the users to append as supplementary material the raw results of the analysis to facilitate meta-analyses. We strongly advise to follow guidelines provided by the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement for a transparent reporting of model design, development and evaluation.

Even though it was hard to get a consensus on which features are repeatable/reproducible, all studies agreed that reconstruction settings, image noise, and segmentation method have a high impact on radiomic feature values for all imaging modalities. This implies that multi-center radiomic studies require harmonized images in terms of image reconstruction setting and signal to noise ratio. Using images from different centers without harmonizing the images itself can lead to wrong conclusions. Further harmonization can be achieved with correction methods applied before feature extraction such as e.g. resampling the images to cubic voxels or by standardizing images via post-processing such as by histogram equalization of MR images or post- reconstruction smoothing [59,60]. Moreover, radiomics harmonization can be achieved by image domain adaptation to reduce the influence of image acquisition settings as Chen et al showed in a simulation study [61]. However, it still has to be validated if these methods can be used for radiomic analysis as it might be that by standardizing the images important textural information gets lost.

Additionally to the standardization of images, there are methods aligning radiomic features. One of the most popular methods is the so-called ComBat, which has been applied to CT [62] and PET features [60], but still requires a large-

scale validation. However, even though, these algorithms to correct for multi-center effects are being developed, it is still important to keep this correction as small as possible. Therefore, when using multi- centric data, it is essential that the images are as comparable as possible in terms of image acquisition.

Moreover, patient cohorts across institutions/scanners should be comparable, i.e. the number of patients with a positive/negative outcome should be comparable across the datasets of each institution/scanner type.

Otherwise the findings using a radiomic model can be due to inter-scanner differences of radiomic features and not to differences caused by variability in tumor characteristics.

To ensure a valid and reproducible analysis of PET studies, it should be carefully checked if reported tracer dose and uptake time are correct and the conversion from image data in Bq/ml to SUV units is accurate. If the liver is displayed in the image, this can be checked by drawing a 3 cm^2 in the liver and verifying that the mean SUV inside the 'liver' sphere is in the range between 1.5 and 2.5. Higher/lower values are an indication for calibration or other errors and these images should be verified, corrected or excluded from the analysis. If the liver is not displayed in the image such as for brain images, a digital reference object can be used to verify the correct conversion to SUV as proposed by Pierce et al. [63]

To minimize the effect of different segmentations, a (semi-) automatic segmentation method might be preferred, as automatic approaches reduce inter-observer variability and yield a higher reproducibility than manual segmentations [64]. The most suitable segmentation method for radiomic analysis has to be identified what has to be done for each imaging modality and cancer type separately. Likely, several segmentation methods will be a good candidate as they yield similar accuracy and repeatability.

Regarding the quality of reporting, we invite the users to provide not only their software but also the metadata associated with it such as information about the programming language. In general, we recommend that each software used for feature calculation should be tested if it complies with the benchmarks provided by IBSI. In this way, feature values extracted by different software packages become comparable which is one important step in the standardization of radiomic feature values.

However, also many studies included in this review missed to report details of preprocessing steps. This missing information has the consequence that the study itself becomes non reproducible and the results are not comparable with other

studies. In conclusion, to make radiomic studies comparable across centers, pre-processing steps should be standardized for each imaging modality as suggested by Park et al. [34]. This includes the discretization method as well as the bin number/bin width of choice as it has an impact on radiomic feature values [65,66]. Since radiomic features can be sensitive to differences in voxel size, it is recommended to interpolate the images before feature extraction to an isotropic voxel size. This step and the used interpolation method should be reported if applied and the radiomic community should agree on which kind of resampling is the preferred one. Almost all studies did not report any information related to the generation of the binary mask from the original data. However, different software tools are available to go from a contour to the final binary mask, therefore it is important to state it.

Most studies reported on the robustness of first order and local textural features such as GLCM and GLRLM features, while global textural features (such as GLSZM features) were found to be less robust.

One limitation of the current review is that it was not possible to draw a general conclusion on which features are reproducible and can be used in the clinic what was one of the original aims of this review. The variety of analyzed settings and used metrics makes it impossible to perform a quantitative synthesis of the analyzed articles. Unfortunately, since the last review, the radiomics community did not came to a consensus on which metric is the most adequate to use in a radiomics setting.

Supplementary Materials

References

[1] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006. https://doi.org/10.1038/ncomms5006.

[2] Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6. https://doi.org/10.7554/eLife.23421.

[3] Tixier F, Hatt M, Valla C, Fleury V, Lamour C, Ezzouhri S, et al. Visual Versus Quantitative Assessment of Intratumor 18F-FDG PET Uptake Heterogeneity: Prognostic Value in Non-Small Cell Lung Cancer. J Nucl Med 2014;55:1235–41. https://doi.org/10.2967/jnumed.113.133389.

[4] Buvat I, Orlhac F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. J Nucl Med 2019;60:1543–4. https://doi.org/10.2967/jnumed.119.235325.

[5] Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol 2018;102:1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.

[6] Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. PLoS One 2015;10:e0124165. https://doi.org/10.1371/journal.pone.0124165.

[7] Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1. https://doi.org/10.1186/s12916-014-0241-z.

[8] Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med 2018;59:189–93. https://doi.org/10.2967/jnumed.117.200501.

[9] Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med 2011;155:529. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

[10] 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 2020:191145. https://doi.org/10.1148/radiol.2020191145.

[11] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med 2014;11:e1001744. https://doi.org/10.1371/journal.pmed.1001744.

[12] Zhuang M, García DV, Kramer GM, Frings V, Smit EF, Dierckx R, et al. Variability and repeatability of quantitative uptake metrics in 18F-FDG PET/CT of non–small cell lung cancer: Impact of segmentation method, uptake interval, and reconstruction protocol. J Nucl Med 2019;60:600–7. https://doi.org/10.2967/jnumed.118.216028.

[13] Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. Radiother Oncol 2019;135:107–14. https://doi.org/10.1016/j.radonc.2019.03.001.

[14] Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clin Transl Radiat Oncol 2019;19:33–8. https://doi.org/10.1016/j.ctro.2019.07.003.

[15] Xia W, Chen Y, Zhang R, Yan Z, Zhou X, Zhang B, et al. Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data—a preliminary study. Phys Med Biol 2018;63:035044. https://doi.org/10.1088/1361-6560/aaa609.

[16] Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. Sci Rep 2019;9:1–16. https://doi.org/10.1038/s41598-019- 45766-z.

[17] Haarburger C, Müller-Franzes G, Weninger L, Kuhl C, Truhn D, Merhof D. Radiomics feature reproducibility under inter-rater variability in segmentations of CT images. Sci Rep 2020;10:12688. https://doi.org/10.1038/s41598-020-69534-6.

[18] Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys 2020;21:179–90. https://doi.org/10.1002/acm2.12795.

[19] Midya A, Chakraborty J, Gönen M, Do RKG, Simpson AL. Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility. J Med Imaging 2018;5:1. https://doi.org/10.1117/1.jmi.5.1.011020.

[20] Griethuysen JJM Van, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype 2017;77:104–8. https://doi.org/10.1158/0008- 5472.CAN-17-0339.

[21] Fang Y-HD, Lin C-Y, Shih M-J, Wang H-M, Ho T-Y, Liao C-T, et al. Development and Evaluation of an Open-Source Software Package "CGITA" for Quantifying Tumor Heterogeneity with Molecular Images. Biomed Res Int 2014;2014:1–9. https://doi.org/10.1155/2014/248505.

[22] Strzelecki M, Szczypinski P, Materka A, Klepaczko A. A software tool for automatic classification and segmentation of 2D/3D medical images. Nucl Instruments Methods Phys Res Sect A Accel Spectrometers, Detect Assoc Equip 2013;702:137–40. https://doi.org/10.1016/j.nima.2012.09.006.

[23] Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res 2018;78:4786–9. https://doi.org/10.1158/0008-5472.CAN-18-0125.

[24] Zhang L, Fried D V, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys 2015;42:1341–53. https://doi.org/10.1118/1.4908210.

[25] Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J Appl Clin Med Phys 2017;18:32–48. https://doi.org/10.1002/acm2.12170.

[26] Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS. Image biomarker standardisation initiative 2016. https://doi.org/10.17195/candat.2016.08.1.

[27] Lv W, Yuan Q, Wang Q, Ma J, Jiang J, Yang W, et al. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. Eur Radiol 2018;28:3245–54. https://doi.org/10.1007/s00330-018-5343-0.

[28] Traverso A, Kazmierski M, Shi Z, Kalendralis P, Welch M, Nissen HD, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre- processing. Phys Medica 2019;61:44–51. https://doi.org/10.1016/j.ejmp.2019.04.009.

[29] Carles M, Bach T, Torres-Espallardo I, Baltas D, Nestle U, Martí-Bonmatí L. Significance of the impact of motion compensation on the variability of PET image features. Phys Med Biol 2018;63. https://doi.org/10.1088/1361-6560/aab180.

[30] Loi S, Mori M, Benedetti G, Partelli S, Broggi S, Cattaneo GM, et al. Robustness of CT radiomic features against image discretization and interpolation in characterizing pancreatic neuroendocrine neoplasms. Phys Medica 2020;76:125–33. https://doi.org/10.1016/j.ejmp.2020.06.025.

[31] Meyer M, Ronald J, Vernuccio F, Nelson RC, Ramirez-Giraldo JC, Solomon J, et al. Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings. Radiology 2019;293:583–91. https://doi.org/10.1148/radiol.2019190928.

[32] Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. J Appl Clin Med Phys 2020;21:179–90. https://doi.org/10.1002/acm2.12795.

[33] Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, et al. Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. Eur Radiol 2020;30:195–205. https://doi.org/10.1007/s00330-019-06381-8.

[34] Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol 2019;20:1124. https://doi.org/10.3348/kjr.2018.0070.

[35] Erdal BS, Demirer M, Little KJ, Amadi CC, Ibrahim GFM, O'Donnell TP, et al. Are quantitative features of lung nodules reproducible at different CT acquisition and reconstruction parameters? PLoS One 2020;15:e0240184. https://doi.org/10.1371/journal.pone.0240184.

[36] Kocak B, Yardimci AH, Bektas CT, Turkcanoglu MH, Erdim C, Yucetas U, et al. Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation. Eur J Radiol 2018;107:149–57. https://doi.org/10.1016/j.ejrad.2018.08.014.

[37] Qu J, Shen C, Qin J, Wang Z, Liu Z, Guo J, et al. The MR radiomic signature can predict preoperative lymph node metastasis in patients with esophageal cancer. Eur Radiol 2019;29:906–14. https://doi.org/10.1007/s00330-018-5583-z.

[38] Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. Med Phys 2019;46:3582–91. https://doi.org/10.1002/mp.13624.

[39] Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. Eur Radiol 2018;28:2255–63. https://doi.org/10.1007/s00330-017-5154-8.

[40] Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. Quantifying the robustness of [18 F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Medica 2018;49:105–11. https://doi.org/10.1016/j.ejmp.2018.05.013.

[41] Hatt M, Laurent B, Fayad H, Jaouen V, Visvikis D, Le Rest CC. Tumour functional sphericity from PET images: prognostic value in NSCLC and impact of delineation method. Eur J Nucl Med Mol Imaging 2018;45:630–41. https://doi.org/10.1007/s00259-017-3865-3.

[42] Guan Y, Li W, Jiang Z, Chen Y, Liu S, He J, et al. Whole-Lesion Apparent Diffusion Coefficient-Based Entropy-Related Parameters for Characterizing Cervical Cancers. Acad Radiol 2016;23:1559–67. https://doi.org/10.1016/j.acra.2016.08.010.

[43] Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, et al. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. Radiol Phys Technol 2018;11:27–35. https://doi.org/10.1007/s12194-017-0433-2.

[44] Takeda K, Takanami K, Shirata Y, Yamamoto T, Takahashi N, Ito K, et al. Clinical utility of texture analysis of 18F-FDG PET/CT in patients with Stage I lung cancer treated with stereotactic body radiotherapy. J Radiat Res 2017;58:862–9. https://doi.org/10.1093/jrr/rrx050.

[45] Bektas CT, Kocak B, Yardimci AH, Turkcanoglu MH, Yucetas U, Koca SB, et al. Clear Cell Renal Cell Carcinoma: Machine Learning-Based Quantitative Computed Tomography Texture Analysis for Prediction of Fuhrman Nuclear Grade. Eur Radiol 2019;29:1153–63. https://doi.org/10.1007/s00330-018-5698-2.

[46] Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, et al. Machine learning-based quantitative texture analysis of CT images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma. Eur Radiol 2018;28:1625–33. https://doi.org/10.1007/s00330-017-5118-z.

[47] Zhang X, Zhong L, Zhang B, Zhang L, Du H, Lu L, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: evaluation with two disease groups. Cancer Imaging 2019;19:89. https://doi.org/10.1186/s40644-019-0276-7.

[48] Park S, Lee SM, Do K-H, Lee J-G, Bae W, Park H, et al. Deep Learning Algorithm for Reducing CT Slice Thickness: Effect on Reproducibility of Radiomic Features in Lung Cancer. Korean J Radiol 2019;20:1431. https://doi.org/10.3348/kjr.2019.0212.

[49] Whybra P, Parkinson C, Foley K, Staffurth J, Spezi E. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. Sci Rep 2019;9:9649. https://doi.org/10.1038/s41598-019-46030-0.

[50] Lee S-H, Cho H, Lee HY, Park H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging 2019;19:54. https://doi.org/10.1186/s40644- 019-0239-z.

[51] Lafata K, Cai J, Wang C, Hong J, Kelsey CR, Yin F-F. Spatial-temporal variability of radiomic features and its effect on the classification of lung cancer histology. Phys Med Biol 2018;63:225003. https://doi.org/10.1088/1361-6560/aae56a.

[52] 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 2018. https://doi.org/10.1002/mp.13322.

[53] Ger RB, Meier JG, Pahlka RB, Gay S, Mumme R, Fuller CD, et al. Effects of alterations in positron emission tomography imaging parameters on radiomics features. PLoS One 2019;14:e0221877. https://doi.org/10.1371/journal.pone.0221877.

[54] Nardone V, Reginelli A, Guida C, Belfiore MP, Biondi M, Mormile M, et al. Delta-radiomics increases multicentre reproducibility: a phantom study. Med Oncol 2020;37:38. https://doi.org/10.1007/s12032-020- 01359-9.

[55] Baeßler B, Weiss K, Santos DP Dos. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Invest Radiol 2019;54:221–8. https://doi.org/10.1097/RLI.0000000000000530.

[56] Yang F, Simpson G, Young L, Ford J, Dogan N, Wang L. Impact of contouring variability on oncological PET radiomics features in the lung. Sci Rep 2020;10:369. https://doi.org/10.1038/s41598-019-57171-7.

[57] Varghese BA, Hwang D, Cen SY, Levy J, Liu D, Lau C, et al. Reliability of CT-based texture features: Phantom study. J Appl Clin Med Phys 2019;20:155–63. https://doi.org/10.1002/acm2.12666.

[58] Johnson PB, Young LA, Lamichhane N, Patel V, Chinea FM, Yang F. Quantitative imaging: Correlating image features with the segmentation accuracy of PET based tumor contours in the lung. Radiother Oncol 2017;123:257–62. https://doi.org/10.1016/j.radonc.2017.03.008.

[59] Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One 2017;12:e0178524. https://doi.org/10.1371/journal.pone.0178524.

[60] 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;59:1321–8. https://doi.org/10.2967/jnumed.117.199935.

[61] Chen J, Zhang C, Traverso A, Zhovannik I, Dekker A, Wee L, et al. Generative models improve radiomics reproducibility in low dose CTs: a simulation study. Phys Med Biol 2021;66:165002. https://doi.org/10.1088/1361-6560/ac16c0.

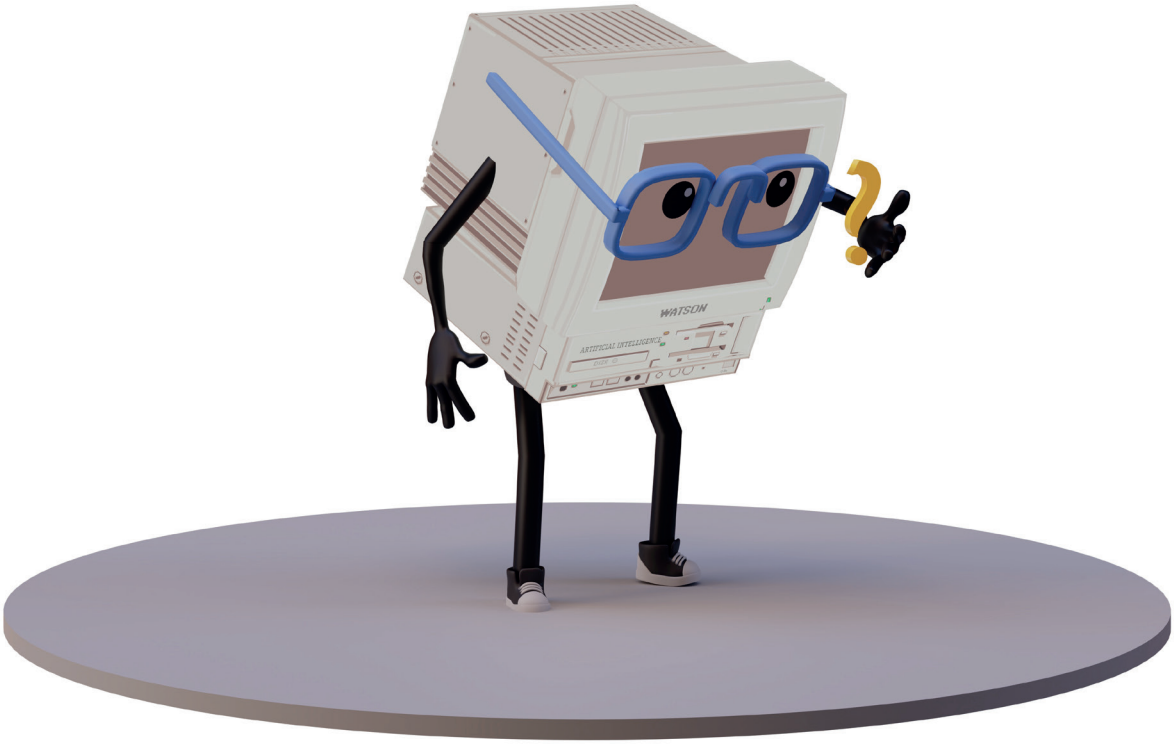
[62] Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 2019;291:53–9. https://doi.org/10.1148/radiol.2019182023.

[63] Pierce LA, Elston BF, Clunie DA, Nelson D, Kinahan PE. A Digital Reference Object to Analyze Calculation Accuracy of PET Standardized Uptake Value. Radiology 2015;277:538–45. https://doi.org/10.1148/radiol.2015141262.

[64] Kolinger GD, Vállez García D, Kramer GM, Frings V, Smit EF, de Langen AJ, et al. Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. EJNMMI Res 2019;9:14. https://doi.org/10.1186/s13550-019-0481-1.

[65] 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. https://doi.org/10.1371/journal.pone.0213459.

[66] Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep 2015;5:11075. https://doi.org/10.1038/srep11075.


WATSON
ARTIFICIAL INTELLIGENCE

Chapter 3

Uncertainty Estimation for Robust Image Biomarkers – Segmentation Uncertainty Example

Zhovannik, Ivan, Dennis Bontempi, Alessio Romita, Elisabeth Pfaehler, Sergey Primakov, Andre Dekker, Johan Bus- sink, Alberto Traverso, and René Monshouwer. “Segmentation uncertainty estimation as a sanity check for image bi- omarker studies.” *Cancers* 14, no. 5 (2022): 1288.

Abstract

Problem. Image biomarker analysis, also known as radiomics, is a tool for tissue character- ization and treatment prognosis that relies on routinely acquired clinical images and delineations. Due to the uncertainty in image acquisition, processing, and segmentation (delineation) protocols, radiomics often lack reproducibility. Radiomics harmonization techniques have been proposed as a solution to reduce these sources of uncertainty and/or their influence on the prognostic model performance. A relevant question is how to estimate the protocol-induced uncertainty of a specific im- age biomarker, what the effect is on the model performance, and how to optimize the model given the uncertainty. *Methods*. Two non-small cell lung cancer (NSCLC) cohorts, composed of 421 and 240 patients, respectively, were used for training and testing. Per patient, a Monte Carlo algorithm was used to generate three hundred synthetic contours with a surface dice tolerance measure of less than 1.18 mm with respect to the original GTV. These contours were subsequently used to derive 104 radiomic features, which were ranked on their relative sensitivity to contour perturbation, ex- pressed in the parameter η. The top four (low η) and the bottom four (high η) features were selected for two models based on the Cox proportional hazards model. To investigate the influence of seg- mentation uncertainty on the prognostic model, we trained and tested the setup in 5000 augmented realizations (using a Monte Carlo sampling method); the log-rank test was used to assess the strati- fication performance and stability of segmentation uncertainty. *Results*. Although both low and high η setup showed significant testing set log-rank p-values ($p = 0.01$) in the original GTV delineations (without segmentation uncertainty introduced), in the model with high uncertainty, to effect ratio, only around 30% of the augmented realizations resulted in model performance with $p < 0.05$ in the test set. In contrast, the low η setup performed with a log-rank $p < 0.05$ in 90% of the augmented realizations. Moreover, the high η setup classification was uncertain in its predictions for 50% of the subjects in the testing set (for 80% agreement rate), whereas the low η setup was uncertain only in 10% of the cases. *Discussion*. Estimating image biomarker model performance based only on the original GTV segmentation, without considering segmentation, uncertainty may be deceiving. The model might result in a significant stratification performance, but can be unstable for delineation variations, which are inherent to manual segmentation. Simulating segmentation uncertainty using the method described allows for more stable image biomarker estimation, selection, and model

de- velopment. The segmentation uncertainty estimation method described here is universal and can be extended to estimate other protocol uncertainties (such as image acquisition and pre-processing).

Keywords: uncertainty; image biomarkers; radiomics; radiomics harmonization; prognostic modeling

1. Introduction

1.1. Background

Image biomarker analysis, also known as radiomics, is a technique to extract quanti- tative information (radiomic features or image biomarkers) from routinely acquired med- ical images, such as computed tomography (CT), positron emission tomography (PET), or magnetic resonance imaging (MRI). Radiomics extraction organically fits clinical routines if annotated data is already present in the clinic. For example, in radiation oncology, where tumors and organs-at-risk are routinely delineated as a part of treatment planning, radiomics can be extracted and stored automatically in the clinical picture archiving and communication system (PACS) [1].

Currently, the image biomarker research focuses either on its direct implementation in prognostic or diagnostic clinical models [2,3] or on image biomarker harmonization and reproducibility [4,5]. Whereas the first research field focuses on the final step of the image biomarker workflow (Figure 1), the latter field works on harmonizing uncertainties in- duced in the image biomarker values in various stages of Figure 1's workflow. Each of the workflow steps in Figure 1 could be approached with a range of tools and methods: from purely expert-knowledge-driven hand-crafted to data-driven machine learning ap- proaches.

Recent radiomics studies show that radiomic features lack reproducibility against image acquisition, pre-processing, and segmentation protocols [5]. Therefore, a set of methods referred to as radiomics harmonization is proposed to calibrate radiomic values based on those protocol differences (e.g., scanner signal-to-noise ratio, voxel size, segmen- tation algorithm, etc.) [4,6,7]. Protocol uncertainty propagates through the Image Bi- omarker Analysis workflow (Figure 1) and induces uncertainty in image biomarker and prognostic model performance.

A fundamental research question of our manuscript is how to estimate the uncer- tainty induced by the protocol differences and decide whether a model can benefit from harmonization. We answer this question by defining statistical metrics to estimate image acquisition, pre-processing, or segmentation protocol-induced uncertainty in prognostic modeling. As a specific example, we show how the segmentation uncertainty inherent in clinical cohorts influences model performance.

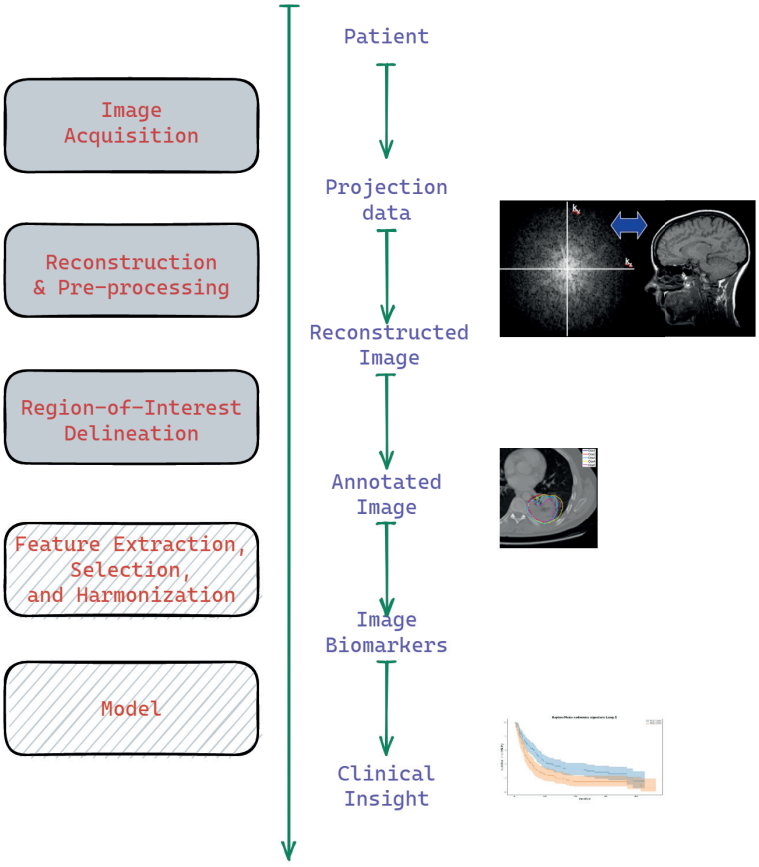


Figure 1. The Workflow of Image Biomarker Analysis: from image acquisition to modeling—each procedure can be performed using a different protocol (set of parameters), thus inducing uncer- tainty in the image biomarker model performance. Partially adapted from: https://med.stan- ford.edu/bmrgroup/Research/AcqRecon.html, https://www.nature.com/articles/srep03529.

1.2. Problem

Let us start with a concrete example: take into consideration a biomarker (for example, a radiomic feature) given a statistical binary classification problem that has two pop- ulations A and B (for instance, short versus long survival stratification). Each population has its population mean (expected value μ) and standard deviation σ, as shown in Figure 2. A biomarker value is sampled from the population distribution and the sample will have its sample mean $\bar{X}$ and sample standard deviation s.

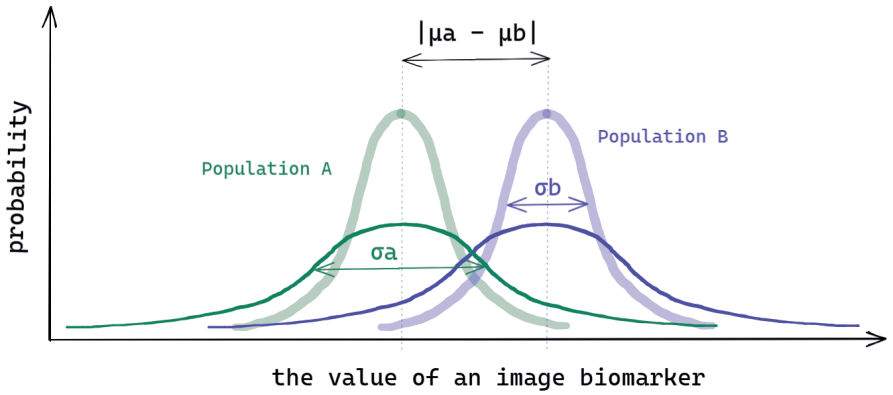


Figure 2. Image biomarker value distributions in populations A and B given relatively low (low opacity curves) and high (full opacity curves) intra-population variance.

Larger intra-population variance makes two distributions overlap more and, there- fore, two populations become less distinguishable. The cause of the variance can be in- trinsic (clinical) or due to the measurement (i.e., protocol uncertainty). With the effect, the distance between the groups, and the intra-population variance, an uncertainty measure, matters. One conventional way of quantifying the effect versus uncertainty is Cohen's *d* (1) effect size measure, defined as a standardized mean difference between groups A and B [8]:

$$d_{AB} = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}} \quad (1)$$

This effect size measure is proportional to the distance $|\bar{X}_A - \bar{X}_B|$ between the groups and is inversely related to the two intra-group variances s_A^2 and s_B^2. This measure, however, does not directly assess uncertainty, induced by the protocol parameters of the workflow shown in Figure 1 for a specific patient, but rather the cumulative effect of both intrinsic (clinical) and measurement (protocol) uncertainties. The latter may also be re-ferred to as intra-patient uncertainty. Measuring the biomarker's intra-patient uncertainty relative to the effect is crucial, but rarely done due to how the conventional radiomics modeling is performed.

In the conventional process of radiomics modeling, a coupled image-mask pair is introduced to a radiomics extraction software (e.g., pyradiomics [2]). In such an approach, the intra-patient sample size (the number of measurements) equals

one (one image-mask pair), which can lead to a poor estimation of a radiomic feature value. In the ideal setting, the intra-patient sample size can be increased by extracting radiomics from several image- mask pairs acquired using different image acquisition (scanner brand, signal-to-noise, re- construction kernel), segmentation (intra-observer variability), and pre-processing (bin- ning, filtering, resampling) protocols. It is important to mention that *the protocol variation should be adequate for the realistic clinical protocol uncertainty*. The increased sample size then will improve the quality of the estimates (mean biomarker value and its variance). There- fore, an estimate for intra-patient uncertainty in a radiomic feature can be then defined as the total variance (2):

$$s^2_{intrapat} = s^2_{acq} + s^2_{seg} + s^2_{proc} + 2COV_{acq,seg} + 2COV_{acq,proc} + 2COV_{seg,proc} \quad (2)$$

where

- s^2_a — the intra-patient variance due to protocol 'a' uncertainty;
- $COV_{a,b}$ — the covariance of protocol 'a' and 'b' uncertainties.

This estimate of intra-patient variance can be introduced into an effect size measure derived from (1), in order to measure the ratio between uncertainty and the effect in this specific radiomic feature (3):

$$\eta = \frac{E[s_{intrapat}]}{\Delta_{effect}} = \frac{E[s_{intrapat}]}{IQR_{0.25-0.75}(X_i)} \quad (3)$$

where the denominator (effect) can be measured as the interquartile range in the dataset, because whenever the two classes are even in size (e.g., the median stratification is used), the medians of both high- and low-risk groups can be estimated as the dataset's 0.25 and 0.75 quantiles. The numerator is the estimated value of intra-patient variance for this ra- diomic feature. If the value of η is close to 1, the intra-patient variance is comparable to the effect, which reduces the feature usefulness in the detection of the effect unless har- monization is applied to reduce the uncertainty.

This study was limited to segmentation-induced uncertainty estimation in radiomic values, so a simplified version of Equations (2) and (3) was reduced to Equation (4). The η values were estimated with a median interquartile range per

patient subject in the nu- merator (uncertainty estimation) and with the interquartile range of the median estimated per subject in the denominator (effect estimation)

$$\eta = \frac{I\widetilde{QR(X_i)}}{IQR\left(\widetilde{X_i}\right)} \quad (4$$

2. Materials and Methods

2.1. Datasets

The two non-small cell lung cancer (NSCLC) cohorts in this study were previously labeled by Aerts et al. as “Lung1” and “Lung2” and utilized in a replication study by Shi et al. [9,10]. The patients in both cohorts were previously treated with (chemo-) radiother- apy and the latest clinical parameters of the cohorts were provided in Shi et al. [10]. For our analysis, 420 Lung1 and 227 Lung2 patient images and respective GTV contours were selected from 422 and 267 records, respectively: the excluded cases contained either no primary tumor contours or missed survival data. The overall survival was used as the outcome. Similarly to the two abovementioned studies, the Lung1 set was used for train- ing, feature ranking and selection, whereas the Lung2 set was used for testing and evalu- ation. All the DICOM images and RTSTRUCT contour files were converted to SimpleITK- compatible images and binary masks with a spacing of 1 mm^3 using Plastimatch v 1.9.0 (https://plastimatch.org/, last access on October 1st, 2021).

2.2. Estimation of Segmentation Uncertainty and the Effect Size of the Radiomic Features

In order to estimate the segmentation uncertainty in radiomic values, we generated binary masks within a clinically relevant surface dice tolerance [11] using a Monte Carlo sampling technique described below. For every patient in the training and testing sets, 300 binary deformed GTV masks were generated using a max surface dice tolerance of 1.18 mm relative to the original contour. The value of 1.18 was taken as a clinically reasonable maximum value following a comprehensive lung tumor (auto-) segmentation study by Primakov et al., where the interobserver variability experiment found the surface dice tol- erance of 1.18 mm in the Maastro Interobserver Lung cancer dataset (https://xnat.bmia.nl, last access on October 1st, 2021) [12].

Each generated mask was used alongside the reference 3D image to extract radiomic features with 25 HU binning and a 1 mm^3 voxel size using PyRadiomics v3.0.1 [2]. First- order statistics, shape, gray level co-occurrence (GLCM), gray level run length (GLRLM), gray level dependence (GLDM), gray level size zone (GLSZM), and Neighboring Gray Tone Difference (NGTDM) matrix-based features were extracted—104 radiomic features in total.

The extracted radiomic values for the 300 augmented + one original binary mask for each of 420 subjects in the training set were used to calculate η according to Equation (4). The 104 features were then ranked in descending order according to their respective η values, as shown in Supplementary Table S1.

The Segmentation UnceRtainty Estimation (sure) project code for image and mask processing, radiomics extraction, and subsequent clinical model setup is available online: Segmetation UnceRtainty Estimation repository – https://github.com/Maastro-CDS-Imag- ing-Group/sure.

2.3. Prognostic Model Simulation

In order to evaluate segmentation uncertainty and its influence on radiomic values and clinical prognostic models, we selected the top four and bottom four features based on their η ranking (Tables 1 and S1). For each setup, two identical analysis pipelines were executed based on a Cox proportional hazards model with median stratification similar to the previous studies [2,10,13].

The *pipeline* consisted of the following steps: (1) A Cox proportional hazards model was trained in the Lung1 training set using the four features as the covariates and overall survival as the outcome. (2) Median hazard ratio and model weights were derived based upon the training set and applied in the Lung2 testing set. (3) Performance metrics in the testing set were collected (log-rank statistic and *p*-value). Prior to analysis, all four features were log-scaled and standardized, and the outliers were removed (absolute value > 3σ), similar to Ref. [10], in both augmented (300 + 1 masks per patient) datasets.

In the baseline model, the feature values extracted from the original GTVs were used as the covariates of the model. To measure the pipeline's sensitivity for segmentation un- certainty, one record was sampled from 301 radiomic values per patient (generated during the Monte Carlo sampling, Figure 3) with replacement 5000 times in both training and testing sets—resulting in a total of 5000 pipeline simulations.

Table 1. Two sets of radiomic features selected for analysis.

Setup A—Low η		Setup B—High η	
Feature	η	***Feature***	η
first-order Maximum	0.0	glszm GrayLevelVariance	0.2944
gldm GrayLevelNonUniformity	0.0108	glrlm RunEntropy	0.3066
glrlm GrayLevelNonUniformity	0.0115	glcm MCC	0.3216
ngtdm Coarseness	0.0129	glszm GrayLevelNonUniformityNormalized	0.4005

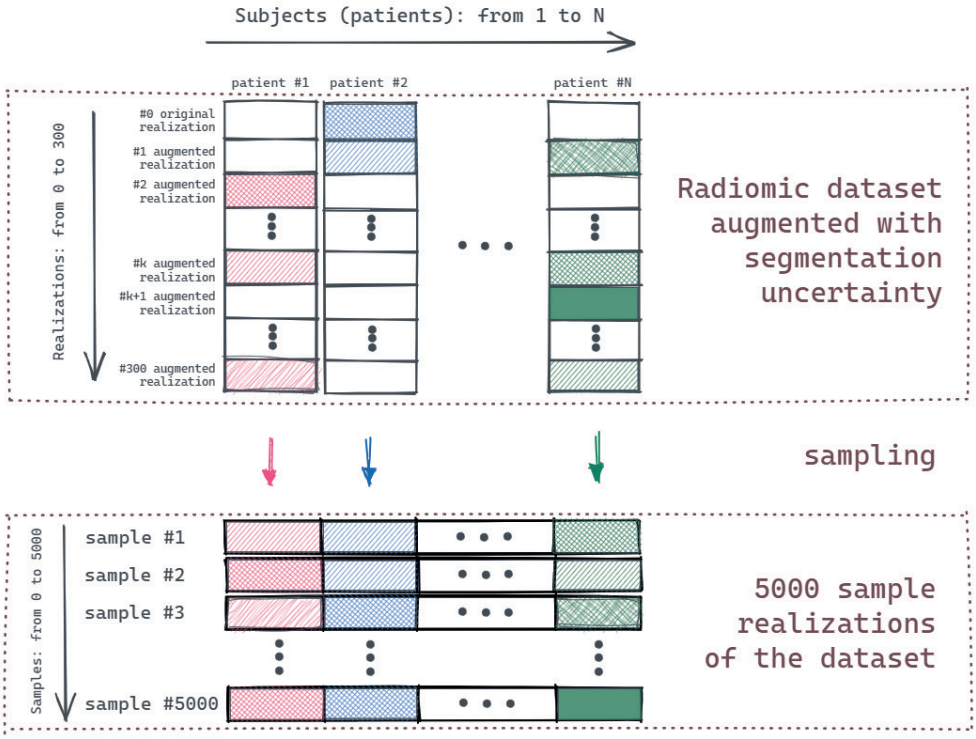


Figure 3. Monte Carlo sampling.

Out of these 5000 pipeline simulations, one was selected based on the least squares of its covariate coefficients and medians of the coefficients; the resulting averaged model was then scored in the abovementioned 5000 realizations of the Lung2 testing set. The 5000 scores were averaged per patient and the final "ensembled" model scores were de- rived. The general outline of the experiment is shown in Figure 4.

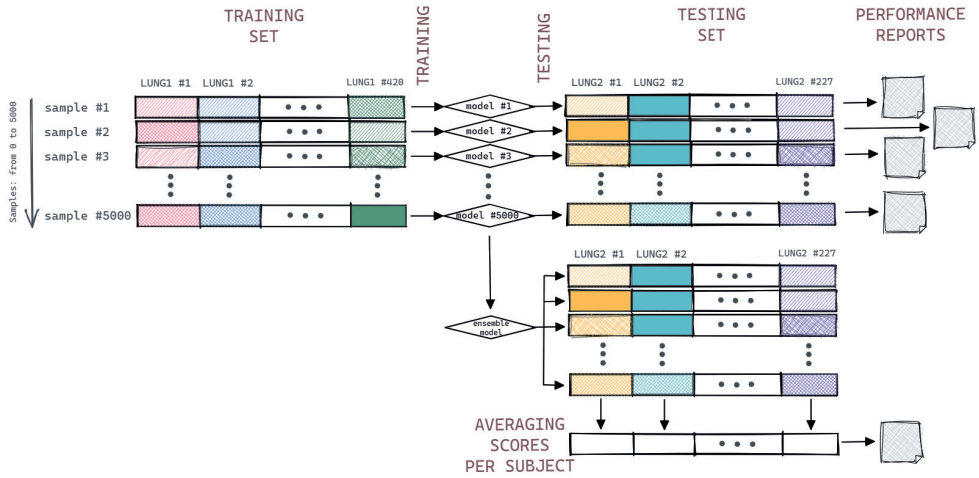


Figure 4. The general outline of the experiment.

Although *p*-values give a good estimation of statistical performance of both setups, it is clinically relevant to measure how confident the model is in its prediction when seg- mentation uncertainty is introduced. Using the ensembled Cox model for each patient of the testing set, we measured the agreement of the model by measuring misclassified sam- ples using 301 augmented mask realizations. This was done by scoring the ensembled model in 301 augmented realizations for every subject in the Lung2 testing set according to their stratification mislabeling δ (5):

$$\delta = \frac{\#MajC - \#MinC}{\#MajC + \#MinC} \tag{5}$$

where

- $\#MajC$ — number of samples of the majority class for a subject;
- $\#MinC$ — number of samples of the minority class for a subject.

3. Results

The original GTV segmentation-based models A and B were performed with a log- rank p = 0.011 and 0.012, respectively. After the introduction of segmentation uncertainty in both datasets by sampling 5000 times randomly from the augmented realizations, setup B showed that only 1428 out of 5000 scenarios achieved a log-rank $p < 0.05$ in the test set— a 28.5% chance of getting a significant result while sampling from the clinically adequate range of segmentation uncertainty.

After the ensembled model scoring in test data, a log-rank $p = 0.1$ was achieved in the test set. The setup B performance summary is shown in Figure 5 (right). Setup A, on the other hand, showed robust performance in the testing set's augmented realizations: its 4508 out of 5000 realizations (90%) showed a log-rank $p < 0.05$ and its ensembled model performed with a log-rank $p = 0.012$, which was similar to the original setup A perfor- mance ($p = 0.011$).

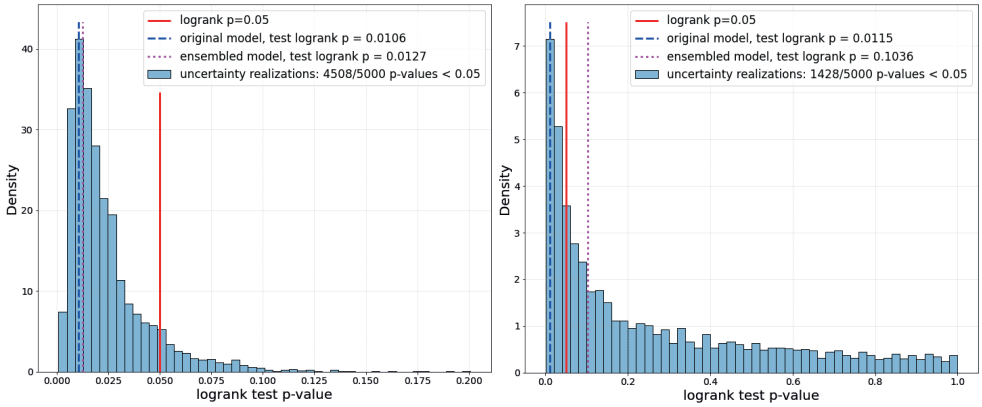


Figure 5. Log-rank *p*-value distribution in the test set for setups A (left, low η) and B (right, high η).

Figure 5 shows an integral picture of how segmentation uncertainty influences the statistical performance of both setups A and B. However, it is more illustrative to check a specific realization. Figure 6 shows two sample performances of the same core setup (en- sembled model, Figure 5) in two different realizations of the testing set: although exactly the same images of Lung2 subjects were used, using radiomic features that are sensitive for segmentation uncertainty (Setup B) can lead to distinctly worse performance.

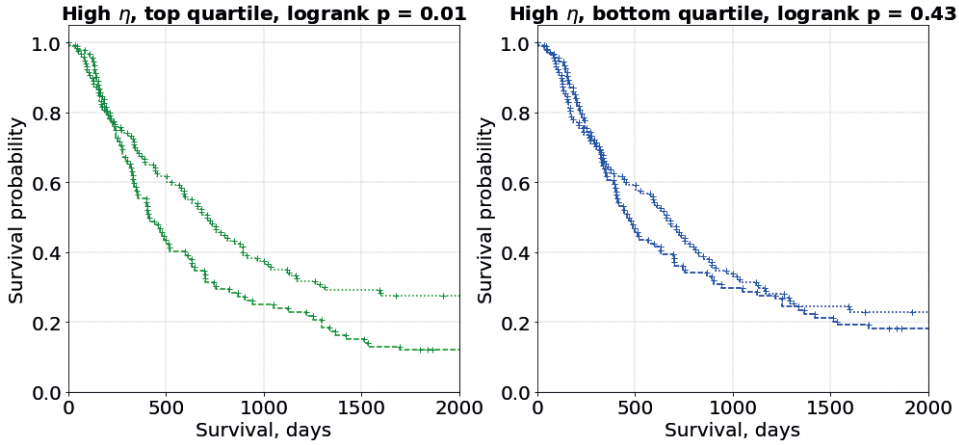


Figure 6. Segmentation uncertainty influence in the high η (setup B): two sample testing set realiza- tions result in different performances.

In order to measure the agreement of the setup's scoring different segmentation re- alizations of the same subject, we scored every subject's segmentation realization with the ensembled model in both setups. Figure 7 shows that segmentation uncertainty causes setup B to have an 80% agreement only for 50% of patients in the testing set, whereas there is an 80% agreement for 90% of subjects' setup A.

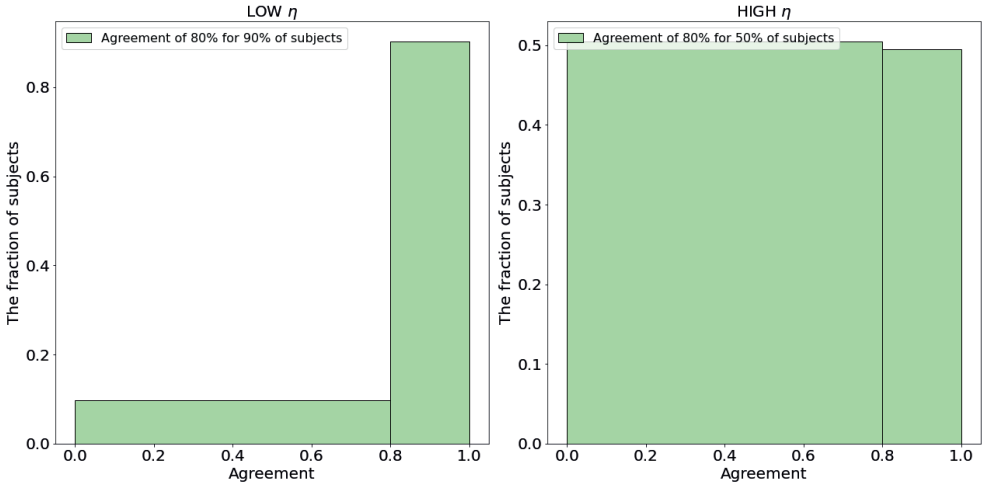


Figure 7. Segmentation uncertainty and patient stratification agreement δ (5) in setups A (left, low η) and B (right, high η).

4. Discussion

4.1. General Discussion

Image and contour data acquired during diagnostic or treatment planning in the clinic are inherently subject to stochastic (for instance, noise) as well as systematic devia- tions (different scanners, delineation protocols, analysis methods). When using the data for modeling, we therefore can consider the given dataset as one representation sampled from a broader image protocol uncertainty distribution. An alteration or variability in protocol settings will enlarge the uncertainty in radiomic values, ultimately impacting the image biomarker's prognostic performance. These protocol variations should be reduced as much as possible, driving the need for radiomics harmonization across the Image Bi- omarkerAanalysis Workflow (Figure 1). As the uncertainty propagates through the Image Biomarker Analysis Workflow (Figure 1), it influences all subsequent stages of the work- flow and, eventually, causes uncertainty in the image biomarker performance. This prop- agation may cause the terminological ambiguity between the terms "protocol uncer- tainty", "protocol-induced uncertainty", and "image biomarker uncertainty". We used the first term to describe the variations in the protocol (e.g., variations in delineation), whereas the latter terms were used to describe the variance in the image biomarker per- formance.

The segmentation uncertainty is an important source of uncertainty in radiomics as, until now, tumor delineation is a manual process in clinical practice and can vary substan- tially from person to person [12]. In this study, we investigated segmentation uncertainty influence in radiomics and proposed a generalizable method to estimate protocol uncer- tainty influence in image biomarker-based prognostic models. Even though the frame- work is described in terms of segmentation uncertainty, the method can be applied to other types of uncertainty. *We would like to strongly emphasize the importance of measuring the uncertainty alongside the performance metrics whenever possible in image biomarker studies.* In this study, we added three aspects to the uncertainty estimation for image bi- omarker studies. Firstly, we have added a framework (Equations (2)–(4)) to estimate un- certainty versus effect in radiomic values with regard to a specific parameter(s) varia- tion—in our case—segmentation uncertainty. Although the standard error might be a use- ful rough estimate for overall uncertainty, it does not measure the influence of a specific protocol parameter. Secondly, we have shown how to model uncertainty propagation through the image biomarker analysis workflow (Figure 1) using Monte Carlo sampling techniques for binary

mask sampling and dataset realization aggregation. Thirdly, we showed that with even a simple harmonization technique, the ensembled model with score averaging gives above median performance—the uncertainty estimates can be used to improve model performance.

In our study, we made the assumption that for each patient, the given original con- tour has uncertainty, and could be contoured slightly differently, for instance, by another radiation oncologist. Figure 5 shows the distribution of performance (expressed as log- rank *p*-value) of 5000 models in different realizations of the testing set—this distribution's parameters (e.g., the interquartile range or 95% interval) can be used as an estimate of segmentation-induced uncertainty in the image biomarker model. Figure 5 can be consid- ered as an integral picture of Figure 6, where the two sample models' performances would lead to absolutely different conclusions because of the segmentation uncertainty of max 1.18 mm surface dice tolerance. On the other hand, Figure 7 shows how one model can be uncertain while performing low- and high-risk survival stratification: high uncertainty leads to the model being uncertain in 50% of the subjects (given the mislabeling rate of 20%—Equation (5)).

4.2. Limitations and Future Perspective

In our manuscript, we wanted to avoid the topic of evaluating image and binary mask processing's influence on radiomic values. We tried to keep all the parameters for pre- and post-processing in the pipeline constant by using a standard pipeline so that they do not cause additional uncertainty in data. In other words, we tried to isolate the three sources of uncertainty, namely, acquisition, segmentation, and processing, in order to es- timate the individual contribution of each term. One example of image pre-processing influence can be the raw original mask delineation and interpolation. This protocol varia- tion causes different coarseness of the ROI surface.

Image acquisition-uncertainty is hard to directly access in the clinical setting since this requires acquiring a large set of image data samples for each subject, which is not realistic due to time and, in the case of computed tomography, radiation dose constraints. Phantom data lifts the abovementioned constraints and can be used to measure proxy estimates of image acquisition-uncertainty. However, a phantom set does not necessarily represent the complexity and textural patterns of biological tissues, therefore, phantom- estimated image acquisition-uncertainty should be used with precautions [6].

To estimate segmentation uncertainty, we performed 300 binary mask augmenta- tions using the so-called “volume-front” method (see the GitHub repository). However, we expect that other methods for inducing binary mask uncertainty can also be used. For instance, Zwanenburg et al. investigated various contour perturbation techniques and their influence on radiomic values [14]. The augmented radiomic data were distributed normally, therefore, we considered that 300 binary mask augmentations gave a good es- timation of segmentation uncertainty in radiomics. Using binary masks with 1 mm^3 voxel size limits the number of augmented binary masks, as we sample from a finite set of voxels in the uncertainty ring. To mitigate this, segmentation uncertainty might be based on raw RTSTRUCT contour coordinates and not on binary masks [15]. We do not, however, be- lieve that this will fundamentally change the analysis of this paper.

Although in our study we considered every segmentation realization as equally pos- sible (uniformly distributed), this might not be the case in a realistic scenario: a data- driven approach to assign a non-uniform probability to measure a segmentation realiza- tion could make an interesting study. In addition, the 5000 number of core setup iterations was selected based on the memory limitation of the computer—more computations could be done, however, we do not expect them to be beneficial, as similar results could be ac- quired for a much lower number of iterations (~1000).

The proposed generic framework of protocol uncertainty can be used to measure im- age acquisition, processing, and segmentation uncertainty combined using (2) and (3) to estimate η. This framework is not bound to radiomics only: any model setup that relies on image data with uncertainty can be investigated on its stability versus protocol uncer- tainty. In this case, instead of radiomic values, the model predicted outcome or embed- dings (latent representations) could be benchmarked against protocol uncertainty.

5. Conclusions

Image biomarker value estimated only in the original segmentation might be deceiv- ing, as this estimation does not account for segmentation uncertainty. Although a prog- nostic model might accidentally give a significant stratification performance, this can be a matter of chance given a set of segmentations provided by the clinic. Without segmenta- tion uncertainty introduced and estimated, the model will not be able to accommodate for it. Segmentation uncertainty estimation not only gives clearer insight into model perfor- mance but is essential to improve model robustness.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Table S1: radiomic features ranked according to their η values in Lung1 set. The top four and the bottom four features were used in the model setup (highlighted in green).

Author Contributions: Conceptualization, I.Z., D.B., A.R., E.P., J.B., A.T. and R.M.; data curation, I.Z., A.R. and S.P.; formal analysis, I.Z.; funding acquisition, A.D., J.B. and R.M.; investigation, I.Z., D.B. and R.M.; methodology, I.Z., D.B., E.P., S.P., A.T. and R.M.; project administration, I.Z., A.T. and R.M.; software, I.Z., A.R. and R.M.; supervision, A.D., J.B. and R.M.; validation, I.Z.; visualiza- tion, I.Z., D.B. and R.M.; writing—original draft, I.Z. and R.M.; writing—review and editing, I.Z., D.B., A.R., E.P., A.D., J.B., A.T. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the research program STRaTeGy with project number 14930, which is (partly) financed by the Netherlands Organization for Scientific Research (NWO).

Institutional Review Board Statement: Ethical review and approval were waived for this study due to open and/or previously reported data.

Informed Consent Statement: Patient consent was waived due to open and/or previously reported data.

Data Availability Statement: The Lung1 dataset is available at https://wiki.cancerimagingar- chive.net/display/Public/NSCLC-Radiomics (last access October 1st, 2021). The code is public and available at https://github.com/Maastro-CDS-Imaging-Group/sure.

Acknowledgements: The authors would like to express their gratitude to *Erik van der Bijl* from Rad- boudumc, *Ray Sheombarsing* and *Jan-Jakob Sonke* from The Netherlands Cancer Institute-Antoni van Leeuwenhoek hospital, and *Suraj Pai* and *Chinmay Rao* from the Maastro Clinic for their comments, support and ideas on the present study design improvement. We would like to thank the *Excalidraw* (https://excalidraw.com/) developers for empowering us with better visualization tools.

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

References

[1] Zhovannik, I.; Pai, S.; Santos, T.A.d.S.; van Driel, L.L.; Dekker, A.; Fijten, R.; Traverso, A.; Bussink, J.; Monshouwer, R. Radiomics integration into a picture archiving and communication system. *Phys. Imaging Radiat. Oncol.* **2021**, *20*, 30–33. https://doi.org/10.1016/j.phro.2021.09.007.

[2] Van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104– e107. https://doi.org/10.1158/0008-5472.can-17-0339.

[3] Hosny, A.; Aerts, H.J.; Mak, R.H. Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *Lancet Digit. Health* **2019**, *1*, e106–e107. https://doi.org/10.1016/s2589-7500(19)30062-7.

[4] Mali, S.A.; Ibrahim, A.; Woodruff, H.C.; Andrearczyk, V.; Müller, H.; Primakov, S.; Salahuddin, Z.; Chatterjee, A.; Lambin, P. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J. Pers. Med.* **2021**, *11*, 842. https://doi.org/10.3390/jpm11090842.

[5] Pfaehler, E.; Zhovannik, I.; Wei, L.; Boellaard, R.; Dekker, A.; Monshouwer, R.; El Naqa, I.; Bussink, J.; Gillies, R.; Wee, L.; et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. *Phys. Imaging Radiat. Oncol.* **2021**, *20*, 69–75. https://doi.org/10.1016/j.phro.2021.10.007.

[6] Zhovannik, I.; Bussink, J.; Traverso, A.; Shi, Z.; Kalendralis, P.; Wee, L.; Dekker, A.; Fijten, R.; Monshouwer, R. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin. Transl. Radiat. Oncol.* **2019**, *19*, 33–38. https://doi.org/10.1016/j.ctro.2019.07.003.

[7] Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J. Nucl. Med.* **2018**, *59*, 1321–1328. https://doi.org/10.2967/jnumed.117.199935.

[8] Cohen, J. *Statistical Power Analysis for the Behavioural Sciences/Jacob Cohen*, 2nd ed.; Routledge: New York, Abingdon, UK, 1988.

[9] Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe- Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **2014**, *5*, 4006. https://doi.org/10.1038/ncomms5006.

[10] Shi, Z.; Zhovannik, I.; Traverso, A.; Dankers, F.J.W.M.; Deist, T.M.; Kalendralis, P.; Monshouwer, R.; Bussink, J.; Fijten, R.; Aerts, H.J.W.L.; et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Sci. Data* **2019**, *6*, 1–8. https://doi.org/10.1038/s41597-019-0241-0.

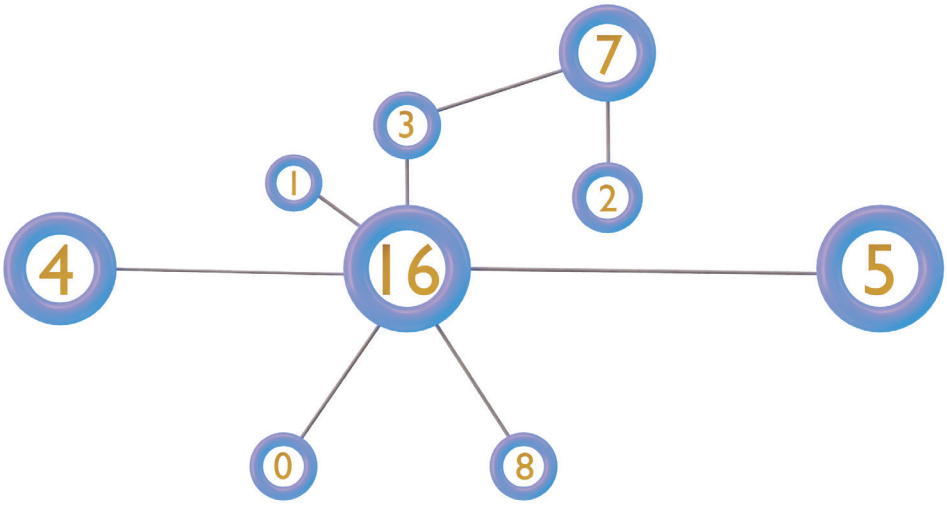
[11] Nikolov, S.; Blackwell, S.; Zverovitch, A.; Mendes, R.; Livne, M.; De Fauw, J.; Patel, Y.; Meyer, C.; Askham, H.; Romera-Paredes, B.; et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Devel- opment and Validation Study. *J. Med. Internet Res.* **2021**, *23*, e26151. https://doi.org/10.2196/26151.

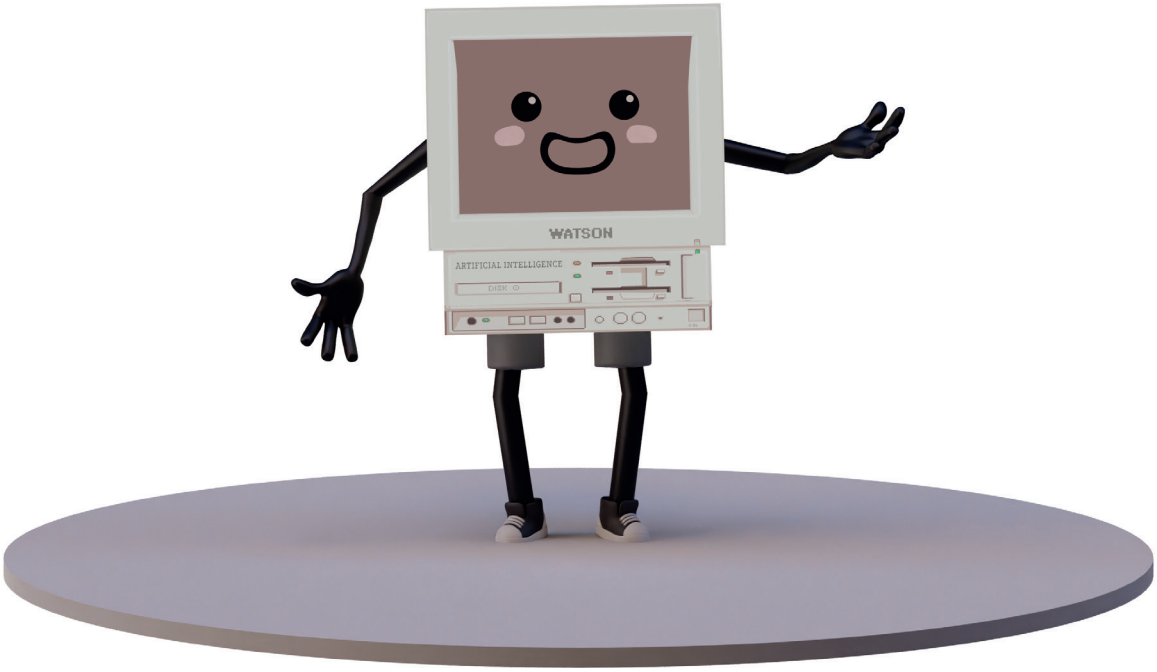
[12] Primakov, S.; Ibrahim, A.; van Timmeren, J.; Wu, G.; Keek, S.; Beuque, M.; Granzier, R.; Scrivener, M.; Sanduleanu, S.; Kayan, E.; et al. OC-0557 AI-based NSCLC detection and segmentation: Faster and more prognostic than manual segmentation. *Radi- other. Oncol.* **2021**, *161*, S441–S443. https://doi.org/10.1016/s0167-8140(21)06964-4.

[13] Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.; O'Sullivan, B.; Aerts, H.J.; et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother. Oncol.* **2019**, *130*, 2–9. https://doi.org/10.1016/j.radonc.2018.10.027.

[14] Zwanenburg, A.; Leger, S.; Agolli, L.; Pilz, K.; Troost, E.G.C.; Richter, C.; Löck, S. Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* **2019**, *9*, 1–10. https://doi.org/10.1038/s41598-018-36938-4.

[15] Sheombarsing, R.; Moriakov, N.; Sonke, J.-J.; Teuwen, J. Subpixel object segmentation using wavelets and multi resolution anal- ysis. *arXiv* **2021**, arXiv:2110.15233.


7
3
1
2
4
16
5
0
8


WATSON
ARTIFICIAL INTELLIGENCE

Chapter 4

Image Acquisition Uncertainty as a Source of Bias in Radiomics

Zhovannik, Ivan, Johan Bussink, Alberto Traverso, Zhenwei Shi, Petros Kalendralis, Leonard Wee, Andre Dekker, Rianne Fijten, and René Monshouwer. "Learning from scanners: Bias reduction and feature correction in radiomics." *Clinical and translational radiation oncology* 19 (2019): 33-38.

Abstract

Purpose: Radiomics are quantitative features extracted from medical images. Many radiomic features depend not only on tumor properties, but also on non-tumor related factors such as scanner signal-to-noise ratio (SNR), reconstruction kernel and other image acquisition settings. This causes undesirable value variations in the features and reduces the performance of prediction models. In this paper, we investigate whether we can use phantom measurements to characterize and correct for the scanner SNR dependence.

Methods: We used a phantom with 17 regions of interest (ROI) to investigate the influence of different SNR values. CT scans were acquired with 9 different exposure settings. We developed an additive correction model to reduce scanner SNR influence.

Results: Sixty-two of 92 radiomic features showed high variance due to the scanner SNR. All these 62 features showed significant (and 47 of them showed at least a factor 2) standard deviation reduction by using the additive correction model. We assessed the clinical relevance of radiomics instability by using a 221 NSCLC patient cohort measured with the same scanner.

Conclusions: Phantom measurements show that roughly two third of the radiomic features depend on the exposure setting of the scanner. The dependence can be modeled and corrected significantly reducing the variation in feature values. More complex models will likely increase the correctability. Scanner SNR correction will result in more reliable radiomics predictions in NSCLC.

1. Introduction

Imaging is an essential part of the radiation oncology workflow: images are used for cancer staging and treatment planning and verification. Medical images contain a large amount of data, which enables their use in clinical practice to personalize radiation therapy for each patient by deriving quantitative features from these images, referred to as radiomics [1]. Radiomics describe tumor phenotype using shape, statistical, and textural features extracted from images of different modalities: Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET). Subsequently, machine learning algorithms use these radiomic features to predict patient survival time [2,3], treatment toxicity [4], tumor habitat characterization [5].

Although the radiomics approach shows promising results, different feature definitions, image pre- processing methods, and imaging instruments make cross-institutional learning difficult [6-10]. The Image Biomarker Standardization Initiative (IBSI) standardized radiomics mathematical definitions and image pre-processing [11]. Still, imaging scanners are not designed for high quality radiomics, but for the best possible image quality for visual (human) interpretation. In daily practice, oncology institutions use their CT scanners with different imaging settings (reconstruction kernel, voxel spacing, X-ray tube exposure, etc) for each patient to optimize subsequent diagnosis and delineation. This lack of inter-scanner (scanner-to-scanner), intra-scanner (various settings within one scanner), and even test-retest (with exact the same settings) reproducibility makes the radiomics approach fragile [6-10]. The inter- and intra-scanner effects induce a non-tumor related variation in the measurements which can be described as bias in the radiomic features. Eventually, this bias may lead to misinterpretation of the radiomics data.

One of the main intra-scanner variations in the CT images is the X-ray tube exposure related to the scanner signal-to-noise ratio (SNR). In our study, we use phantom measurements to quantify how scanner SNR variation results in biasing the extracted features. We hypothesize that the SNR dependent bias can be characterized and quantified, providing the opportunity to correct for it.

2. Materials and Methods

2.1. Phantom

To investigate the influence of scanner SNR on radiomic features we used a commercial phantom (Gammex 467 CT phantom, Middletone, WI, USA). The phantom was used in the standard configuration with its 16 inserts of different tissue-like densities. We performed five sessions of scans with each 9 exposure settings (from 30 to 460 mAs) with a Brilliance Big Bore CT (Philips, Best, The Netherlands) using the Thorax protocol. The images were reconstructed with the B reconstruction kernel with pixel resolution 512×512. To extract radiomics, we delineated regions of interest (ROI) in all the 16 inserts and the phantom center (total of 17 ROIs) as equally-sized cylinders using the *Pinnacle 16.0.2* treatment planning system (Philips Healthcare, Fitchburg, WI, USA). To avoid edge effects, we delineated the ROI smaller than the inserts as shown in figure 1. For radiomics extraction, we used open-source *pyradiomics 2.1.2* software with 25 HU binning and no re-sampling [12].

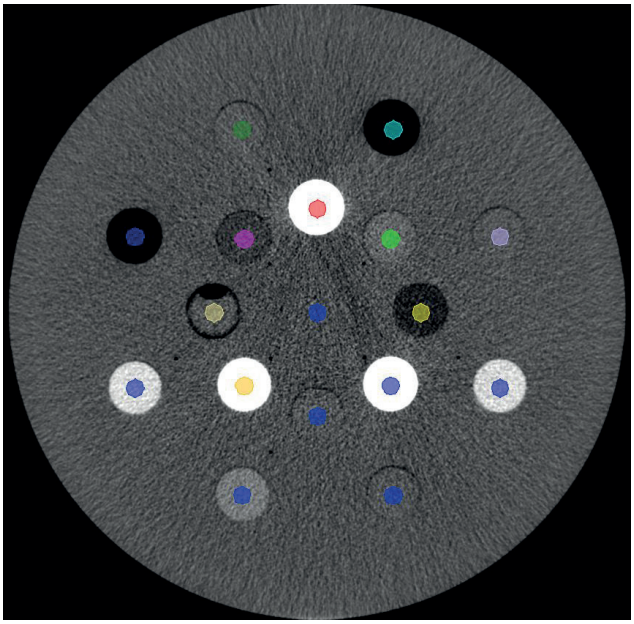


Figure 1. *Gammex phantom configuration with cylindrical delineations. The 17 plug descriptions are in the supplementary A.*

2.2. Patient cohort

To relate our phantom study to clinical applications, we used images of a 221 non-small cell lung cancer (NSCLC) cohort (supplementary table B1) previously treated with (chemo)-radiotherapy and scanned with the same scanner as the phantom set. The data consists of radiotherapy treatment planning DICOM CT images with various scanner settings and physician-delineated primary NSCLC tumors as RT structure sets. The median X-ray tube exposure was 300 mAs. Radiomic features were extracted from the gross tumor volume (GTV) of the primary tumor with the same *pyradiomics* extraction settings as in the phantom set.

2.3. Correction method

Using the five repeated measurements, we calculated mean and standard deviation for each exposure value and every ROI. We arbitrarily defined the target radiomic value (TRV) as the mean value of the radiomic feature measured with the 200 mAs exposure. The aim of the correction was to correct all exposure values to the value observed at 200 mAs as that was the median exposure value in the phantom set. Further data processing included: 1) TRV calculation (for 200 mAs) for each ROI in raw data (figure 2A), 2) Subtracting TRV from radiomic feature's data, isolating the SNR trend in the data (figure 2B), 3) fitting the correction function (figure 2B), 4) Correcting the raw data (figure 2C).

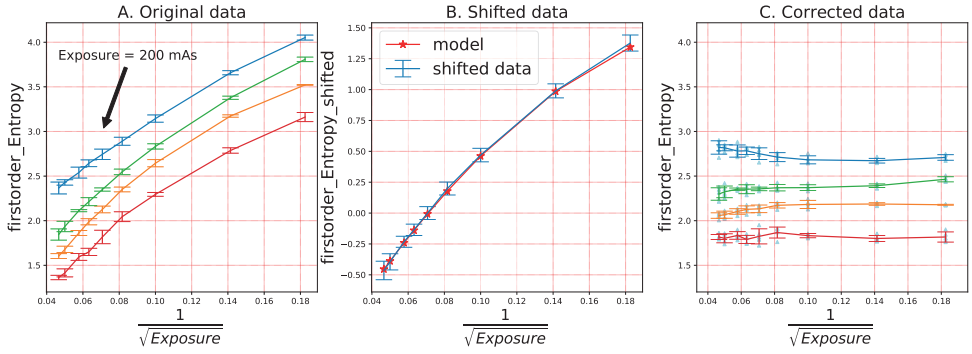


Figure 2. radiomics correction model in three steps: A) shift original data to 0 with TRV, B) fit the model using the shifted data, C) correct the original data using the model.

As scanner SNR in CT images is proportional to the square root of number of photons, and therefore, to $\sqrt{E}$; we analyzed the relationships between radiomics values and $\frac{1}{\sqrt{E}}$. To avoid overfitting, we trained a regression model with the

only two predictors (excluding intercept): $\frac{1}{\sqrt{E}}$ and $(\frac{1}{\sqrt{E}})^2$. We used no predictor scaling. Eventually, we defined the correction model as by formula (1), where w – model weights, b – intercept, E – exposure, Δ - correction factor. We developed the model using *scikit-learn* package for python, version 0.19.1 [13].

$$RF_{corrected} = RF_{measured} + \Delta(\mathrm{E}),$$
$$\Delta(\mathrm{E}) = w_1 \times \left(\frac{1}{\sqrt{E}}\right) + w_2 \times \left(\frac{1}{\sqrt{E}}\right)^2 + b. \quad (1)$$

2.4. Radiomic feature correctability

We defined correctability as the ability to reduce scanner SNR influence on a radiomic feature. To assess correctability of a feature, we defined the correctability score (CS) as in formula (2). To derive the score, we used TRV-shifted data (figure 2B). The correctability score is a ratio: the numerator describes variability due to exposure (variance in means), the denominator describes intrinsic repeatability variance; $\Delta\boldsymbol{RF}$ stands for TRV-shifted radiomic feature values. For each exposure value in the range [30-460 mAs], numerator calculates mean and denominator calculates standard deviation of $\Delta\boldsymbol{RF}$ values. Then, numerator calculates standard deviation of means and denominator calculates mean standard deviation across the 9 exposure values. A value of 1 denotes that the correction is of the order of the noise and therefore is not very relevant. The correctability becomes more relevant at increasing values of CS. Eventually, the CS parameter is a measure of how correctable a feature is based on the phantom scans.

$$CS = \frac{std\left(mean_{\,while\,\{exposure=E\ mAs\}}(\Delta RF)\right)}{mean\left(std_{\,while\,\{exposure=E\ mAs\}}(\Delta RF)\right)}. \quad (2)$$

2.5. Correction evaluation

The final aim of the correction is to reduce the variance of the RF values due to the variation of noise, for this purpose, we defined the evaluation score (ES) as ratio of standard deviations before and after the correction calculated for each ROI and every radiomic feature (RF), where values above 1 indicate a gain of the correction mechanism, by formula (3):

$$ES(ROI) = \frac{StD\left(RF_{before\ correction}\right)}{StD\left(RF_{after\ correction}\right)}. \quad (3)$$

2.6. Clinical relevance of the phantom

Phantom radiomics studies should be applicable in clinical data. To assess clinical relevance, we evaluated 1) distribution overlap in features to test if a radiomic feature distribution in phantom set present absolute values of the same magnitude as values in clinical studies; 2) investigate how scanner SNR distorts feature values of clinical data by simulating (adding) noise to the scans.

When comparing distributions between the phantom and patient cohorts, note that all 17 phantom ROIs had the same shape in the phantom set, while in the patient cohort shape delineations differ between subjects [14]. Therefore, we performed the distribution comparison only for 4 volume-normalized features: 'gldm DependenceNonUniformityNormalized', 'glrlm GrayLevelNonUniformityNormalized', 'glszm SizeZoneNonUniformityNormalized', and 'glrlm RunLengthNonUniformityNormalized'.

We cannot scan a patient with different exposure settings, therefore, we modeled scanner SNR in patient images by adding Poisson noise. The magnitudes of the Poisson noise were initially calibrated in phantom set to be adequate to real exposure settings (30-460 mAs) by applying Poisson noise of different magnitudes to the phantom images with the maximum exposure of 460 mAs (supplementary figure B3). As the next step, Poisson noise with the magnitude calibrated for - 160mAs SNR reduction was applied in patient images. We used those generated images to extract radiomics and evaluate the relative shift in features. The relative shift is defined in formula (4) and evaluates how large the difference between feature values in original ($RF_{original}$) and SNR-influenced ($RF_{\text{-}160mAs}$) images is if compared to the interquartile range in the feature distribution ($IQR_{0.75\text{-}0.25}(RF_{original})$):

$$relative\ shift(patient_i, RF) = \left|\frac{RF_{160mAs,i} - RF_{original,i}}{IQR_{0.75-0.25}(RF_{original})}\right| \times 100\% \qquad (4)$$

3. Results

3.1. Radiomic feature correctability

We calculated the correctability score (CS) for each radiomic feature – 92 scores in total. If the CS of a radiomic feature is close or less than one, the intrinsic reproducibility variance is equal to the scanner SNR-caused variation; that makes the feature uncorrectable. Therefore, we chose for the correctability threshold of

CS>2, meaning that the correctable scanner SNR variance is 2 times higher than the intrinsic reproducibility in a radiomic feature. Based on this threshold criterion, we selected 62 features for further analysis. The upper panel of figure 3 shows CS for each selected radiomic feature as the step blue line.

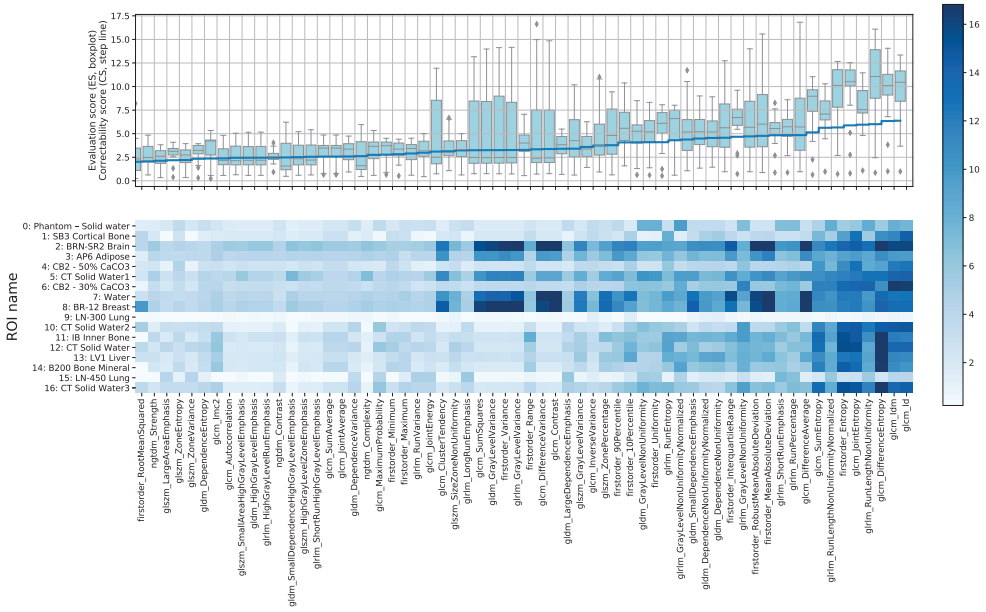


Figure 3. Correctability (2)and Evaluation (3) scores for the selected 62 radiomic features and 17 ROIs. The color bar represents the evaluation score (ES): the darker, the larger reduction in standard deviation was obtained.

3.2. Correction evaluation

To assess whether the exposure dependence could be corrected with our model we calculated the evaluation score (ES). All 62 selected with the CS>2 threshold criterion radiomic features showed significant (ES versus 1 Wilcoxon signed-rank test $p<0.01$) reduction in standard deviation (averaged across the ROIs) using our additive model. Forty-seven out of 62 radiomic features showed significant (ES versus 2 Wilcoxon test $p<0.05$) at least 2 times standard deviation reduction. In summary, the upper panel boxplot (figure 3) describes ES distribution across 62 radiomic features and 17 ROIs.

We evaluated how different materials react on the scanner noise by calculating 17 ROIs' ES for each radiomic feature and placed the scores in lower panel of figure

3. Interestingly, ROIs 9 and 15 (low density plugs, < -600 HU) have low correctability, on the other hand, ROIs 2 and 8 (28 and -45 HU mean density respectively) have good correctability. These results show that different materials react differently on scanner SNR in radiomic features: some materials are more dependent on scanner SNR than others are.

3.3. Clinical relevance of the phantom

In our study, we used phantom measurements to simulate and characterize the acquisition of radiomic features for clinical scans. Figure 4 shows how large the relative shift (4) in radiomic features is while applying Poisson noise of the magnitude equivalent of -160mAs scanner SNR reduction. For example, relative shift of 10% means that -160 mAs reduction in a patient scan causes feature value to change 10% relative to the feature distribution width in the patient cohort.

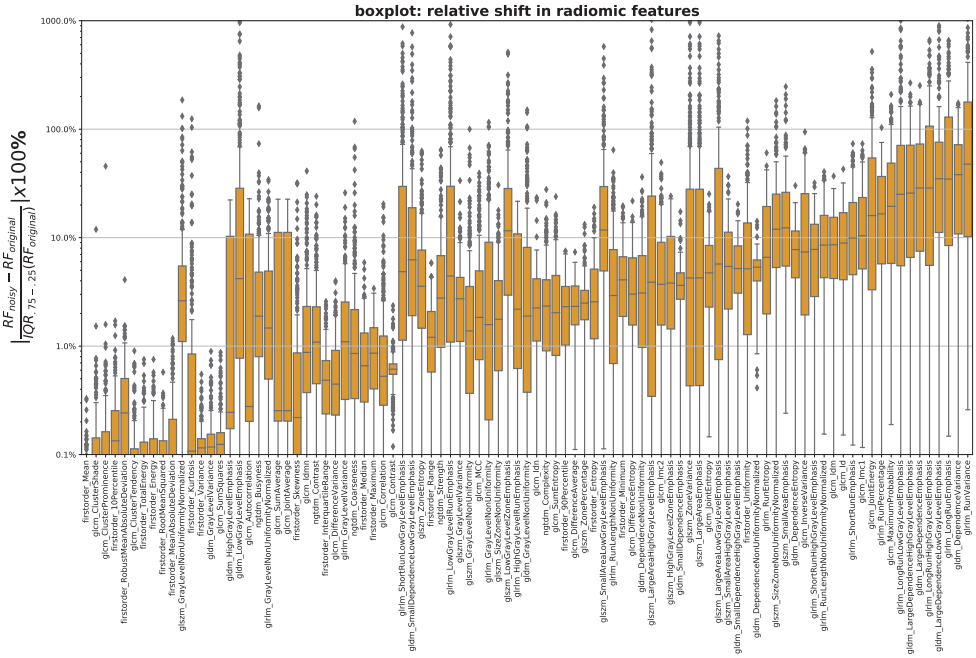


Figure 4. relative shift (4) in radiomic features (in ascending order) versus feature names while applying Poisson noise (equivalent to decreasing scanner SNR, mAs) in the images of the NSCLC cohort.

In addition, we evaluated overlap between the clinical and phantom sets in 4 normalized feature distributions (supplementary figure B1). We found that the distributions have clear overlap; therefore, phantom radiomics are at least partly relevant for clinical scans.

4. Discussion

We systematically investigated the dependence of radiomic feature values on scanner SNR using a commercial phantom and a patient cohort of lung cancer patients. The phantom measurements were obtained using a standard clinical protocol, where the SNR was varied by changing exposure settings from 30 to 460 mAs. We showed that many radiomic features form a trend with the scanner SNR, making the value of the feature not only dependent on the tumor, but also on a specific scanner setting. To remedy this effect, we developed a method to correct the radiomic features for scanner SNR.

4.1. Radiomics correctability

We used correctability score (CS) to separate radiomic features which are biased and correctable in terms of scanner SNR from those that are not. Although Spearman correlation is a reliable criterion for trend detection, it does not include the intrinsic repeatability of the measurement. For instance, the statistical radiomic feature 'Energy' (supplementary C) has a high Spearman correlation with scanner SNR, but the feature's correctable trend variance is smaller than its intrinsic repeatability making correction not effective. Therefore, we defined CS that assesses both intrinsic repeatability and correctable trend variance. Of the 92 features considered, 62 show a CS>2, indicating that they have a dependence on scanner SNR that dominates the repeatability. Note that stability for different exposure settings (CS<1) does not mean a radiomic feature is stable for other scanner settings (image reconstruction kernel, voxel spacing, etc).

4.2. Correction model

Given that there is a trend of the feature value with exposure, we hypothesize that it is possible to correct for the variation. We chose an additive quadratic regression model and used X-ray tube exposure as the predictor. Adding more variables (e.g. uncorrected feature values and/or its intersection term with exposure) might

benefit the correction for some features where additive terms cannot explain trends for different ROIs perfectly. For instance, for the feature 'glrlm GreyLevelVariance' (see supplementary C), the correction seems to depend on the density of the plug, suggesting that a model incorporating the exposure and the mean HU as predictors could improve the correction significantly. We did not pursue developing more complicated correction models in this paper since our main goal was give a proof of principle regarding correctability, and since other issues such as overfitting must be considered when making the model more complex. Supplementary C shows the scanner SNR correction in all the 92 features and intraclass correlation coefficients before and after correction for the 62 selected features.

4.3. Clinical relevance of the phantom and correction model

In using phantom measurements to study scanner dependence of clinical scans, it is paramount that the phantom (material) is representative for the patient case [8]. We compared the distribution of radiomic features in a clinical cohort with the distribution in a phantom. Ideally, the distribution of the features in both phantom and patient cohorts should be identical for all features. Firstly, as has been described before, a part of the 'texture' features are dependent on the shape or the size of the ROI [14]. Comparing the distribution of these is not relevant since we use artificial (cylindrical) regions, therefore only features insensitive to volume or shape could be used. Some typical examples of these features are given in the supplementary data (supplementary B). Overlap is present in for almost all features. This suggests that the properties of the phantom captured by the radiomic features are at least partly relevant for the patient cohort. Future work is needed to develop plugs that are identical to patient material, although a perfect match with the patient cohort for all features is unrealistic [8].

As a second method to test the applicability in the clinical situation, we simulated for each patient scan what the effect would have been if the scan was made with lower exposure. For this we applied Poisson noise to images, where the quantitative relation between the noise amplitude and the exposure was derived from the phantom scans. We found that scanner SNR results in change of the radiomics values for the clinical scans (figure 4). For a large part of the patients/features, a moderate change in the exposure resulted in more than 10% change of the radiomic feature compared to the width of the distribution of the whole cohort. When using the radiomic features as an input for a personalized outcome prediction, this will clearly affect the value of the prediction for individual patients.

Fave et al also investigated the effect of noise in patient CT's on radiomic features by adding noise to the scans. Their findings is in line with ours, namely that the effect is significant, leading to the conclusion that scanning with a range of patient dose should be avoided [15]. Our finding is however in contrast with the conclusion of Mackin et al. [8]. Their measurements were done using the Credence Cartridge Radiomics phantom, and reached the conclusion that SNR of the scan was not likely to be of significant influence since for the rubber insert (which was taken to be most representative for tumor tissue) the effect of the changing tube current was small. Their argument is that the addition of the noise to the scan negligible due to the tumor inhomogeneity. However, the added noise simulations by Fave et al. and us show that for the patient scans involved (in both cases NSCLC patients) the noise indeed affects feature values significantly.

5. Conclusion

We found that 62 out of 92 radiomic features strongly depend on scanner SNR. Due to this dependence, non-tumor related variation is added to the features' values, seriously limiting the use of radiomics in clinical applications. We showed that a simple additive model effectively corrects the undesired variation for 47 out of 62 features. By comparing a NSCLC cohort with the phantom set, we showed that variation in scanner SNR is a reality in a typical clinical cohort, and thus is an actual problem in using radiomics for prediction modeling and personalized medicine.

Supplementary materials

References

[1] Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: Images Are More than Pictures, They Are Data.* Radiology, 2016. **278**(2): p. 563-77.

[2] Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. **5**: p. 4006.

[3] Hosny, A., et al., *Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study.* PLoS Med, 2018. **15**(11): p. e1002711.

[4] Abdollahi, H., et al., *Rectal wall MRI radiomics in prostate cancer patients: prediction of and correlation with early rectal toxicity.* Int J Radiat Biol, 2018. **94**(9): p. 829-837.

[5] Sala, E., et al., *Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging.* Clin Radiol, 2017. **72**(1): p. 3-10.

[6] Mackin, D., et al., *Measuring Computed Tomography Scanner Variability of Radiomics Features.* Invest Radiol, 2015. **50**(11): p. 757-65.

[7] Mackin, D., et al., *Harmonizing the pixel size in retrospective computed tomography radiomics studies.* PLoS One, 2017. **12**(9): p. e0178524.

[8] Mackin, D., et al., *Effect of tube current on computed tomography radiomic features.* Sci Rep, 2018. **8**(1): p. 2354.

[9] Shafiq-Ul-Hassan, M., et al., *Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra.* J Med Imaging (Bellingham), 2018. **5**(1): p. 011013.

[10] Shafiq-Ul-Hassan, M., et al., *Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels.* Med Phys, 2017. **44**(3): p. 1050-1062.

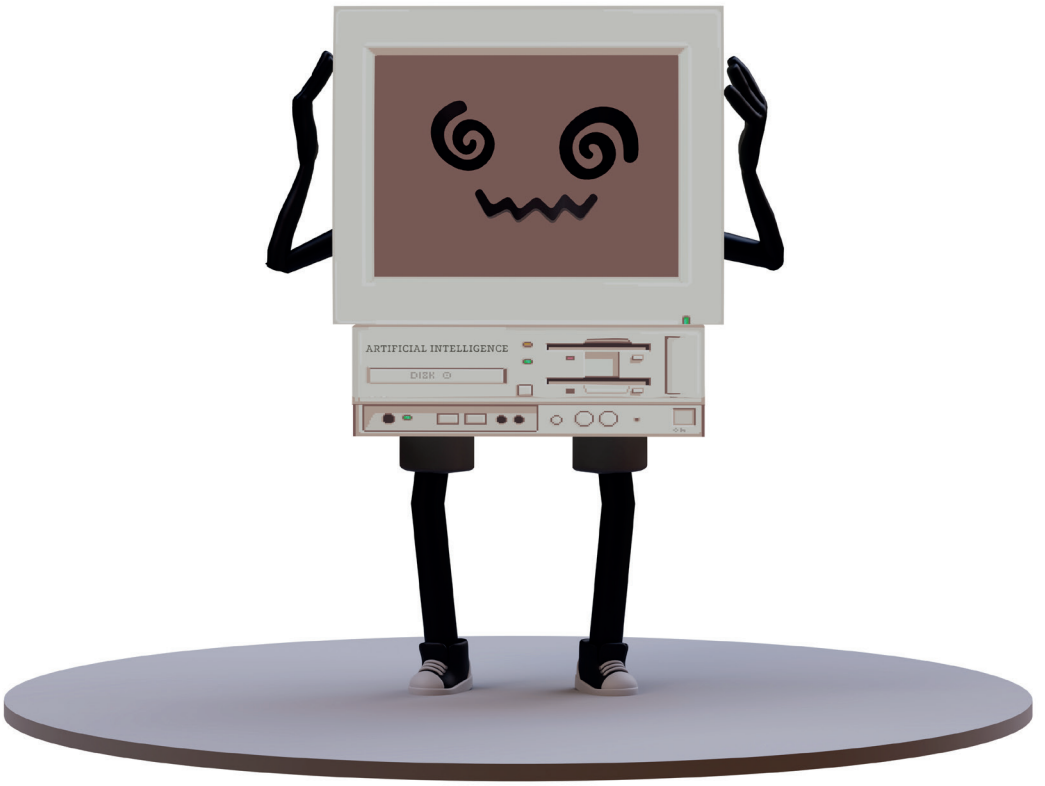
[11] Alex Zwanenburg, S.L., Martin Vallières, Steffen Löck, *Image biomarker standardisation initiative.* arXiv preprint arXiv:1612.07003, 2016.

[12] van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.

[13] Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

[14] Welch, M.L., et al., *Vulnerabilities of radiomic signature development: The need for safeguards.* Radiother Oncol, 2018.

[15] Fave, X., et al., *Preliminary investigation into sources of uncertainty in quantitative imaging features.* Computerized Medical Imaging and Graphics, 2015. **44**: p. 54-61.


ARTIFICIAL INTELLIGENCE
DISK

Chapter 5

Gross Tumor Volume as a confounding factor in Radiomics

Alberto Traverso*, Michal Kazmierski*, Ivan Zhovannik, Mattea Welch, Leonard Wee, David Jaffray, Andre Dekker, and Andrew Hope. “Machine learning helps identifying volume-confounding effects in radiomics.” Physica Medica 71 (2020): 24-30.

Abstract

Purpose: Highlighting the risk of biases in radiomics-based models will help improve their quality and increase usage as decision support systems in the clinic. In this study we use machine learning-based methods to identify the presence of volume-confounding effects in radiomics features.

Methods: 841 radiomics features were extracted from two retrospective publicly available datasets of lung and head neck cancers using open source software. Unsupervised hierarchical clustering and principal component analysis (PCA) identified relations between radiomics and clinical outcomes (overall survival). Bootstrapping techniques with logistic regression verified features' prognostic power and robustness.

Results: Over 80% of the features had large pairwise correlations. Nearly 30% of the features presented strong correlations with tumor volume. Using volume-independent features for clustering and PCA did not allow risk stratification of patients. Clinical predictors outperformed radiomics features in bootstrapping and logistic regression.

Conclusions: The adoption of safeguards in radiomics is imperative to improve the quality of radiomics studies. We proposed machine learning (ML) – based methods for robust radiomics signatures development.

Keywords: Radiomics, Machine, Learning, Predictions, Lung, Head and neck

1. Introduction

Radiomics, the automated extraction of quantitative descriptors from medical images, has demonstrated promising prognostic and predictive results for overall survival [1], distant metastases [4] and cancer biology [6]. After an initial phase of enthusiasm related to the introduction of this technology in the medical domain, investigation of the weakness and drawbacks of the new methodology always follows. These discussions are constructive and represent part of the scientific process to mature a technology, especially if it is meant to be clinically applicable.

In the radiomics scenario, recent publications warned about the presence of biases and potential risks that could be associated with radiomics-based models. In Chalkidou et al. [3], the authors pointed out that the usage of an elevated number of features combined with arbitrary feature selection cut-offs, might produce the undesired problem of multicollinearity, which leads to model over-fitting, often related to false discovery rates. The problem is that all radiomics computational packages compute hundreds to thousands of radiomics features, which often do not differ in their definitions, but are the same formulas computed by perturbing the original image with digital filters. This hyperspace of correlated features is usually much larger than the outcomes of interest, leading to models that are prone to overfitting and exposed to false positive associations [8]. Moreover, some radiomics features embed in their definition hidden confounding factors, which drive their prognostic/predictive power, but it is not immediately understood by inspecting the mathematical definitions of the features. A recent paper showed the presence of a strong volume-confounding effect in some radiomics signatures based on texture or statistical features [14]. The authors showed the randomization of grey level values still produced radiomics features able to have strong predictive power. This paper was the first one to introduce the concept of “safeguards” in radiomics studies. Understanding and evaluating the correlations between radiomic features and clinical prognostic variables is fundamental to evaluate the added value of imaging features compared to the previously mentioned factors. In a recent study [7], the authors investigated the complementary nature of heterogeneity quantified by imaging features and tumour volume in FDG-PET from multi-site cancers. They showed that volume and imaging features were both independent prognostic factors for Non-small Cell Lung Cancers (NSCLC) for volumes above 10 cm(3), with complementary information increasing substantially for larger tumour volumes.

However, when smaller volumes were considered as in oesophageal cancers, the complementary value was degraded because of the presence of smaller volumes.

Again, another study in FDG-PET [2], but for cervical cancers investigated the effect of e small tumour volumes on studies of intertumoral heterogeneity of tracer uptake. The authors used a computer simulation to isolate the effects of tumour volume on the image local entropy. They concluded that inclusion of tumour volumes below 45 cm(3) can profoundly bias comparisons of intra-tumoral uptake heterogeneity metrics. From the cited studies, to fully exploit the complementary prognostic/predictive power of imaging features it is imperative to benchmark them with respect for example to tumour volume. In fact, additional prognostic factors should be added to an existing model, since the introduction of redundant information could be dangerously prone to overfitting. By taking the previous studies as support, in this paper we intent to provide the radiomics community with a machine-learning based framework to evaluate complimentary role of imaging features when benchmarking with other prognostic factors, such as for example tumour volume.

We investigated how machine learning techniques can be used to discover the presence of volume- confounded features, effectively applying radiomics safeguards.

Machine learning methods are often used in the form of supervised methods, where classifiers are trained to learn associations between radiomics features and outcomes (labels). Large efforts have been dedicated to tuning classifiers, but there is no guarantee that biases will be uncovered. On the contrary, unsupervised methods do not look at labels and only utilize the original radiomics features. These methods are very popular in genomics studies, but not often used in radiomics studies. In this work we show how a combination of unsupervised and supervised methods can be used to introduce safeguards to radiomics studies.

2. Methods

2.1. Datasets

We used two retrospective public data sets for the analysis:

1)

Lung1: 421 NSCLC (Non-Small Cell Lung Cancer) patients treated with concurrent chemo-radiotherapy. Computed Tomography (CT) scans of the patients and manually delineated contours of the primary Gross Tumor Volume (GTV) in form of DICOM and RTSTRUCT files were available. The dataset is available for download at the XNAT repository (https://xnat.bmia.nl) and on the TCIA archive (http://doi.org/10.7937/K9/TCIA.20%0A15.L4FRET6Z). The dataset is the same used in Aerts et al. [1], Shi et al., [10].

2)

HN1: 132 CT scans of oropharynx and larynx squamous cell carcinoma patients treated with concurrent chemo-radiotherapy and manually delineated contours of the primary gross tumor volume (GTV) in form of DICOM and RTSTRUCT files were available. The dataset is available for download at the XNAT repository (https://xnat.bmia.nl) and the TCIA (https://doi.org/10.7937/tcia.2019.8kap372n). The dataset is the same used in Aerts et al. [1].

Additionally, clinical variables including: TNM, AJCC staging information, age, sex, as well as overall survival (OS) with a 3-year follow up were available.

2.2. Radiomics features extraction

We used the open source software PyRadiomics v2.2.0 [15] to extract imaging features from each GTV. Pyrex (https://github.com/zhenweishi/Py-rex), an extension of PyRadiomics was used to handle DICOM RTSTRUCT files as input, by generating a binary segmentation mask from the contour data [9]. For LUNG1 and HN1 datasets we used extraction parameters suggested in Aerts et al. [1]. A detailed description of the computational settings is provided in the Supplementary material. To further evaluate the impact in the results of the aggregation method of texture features, we compared the default "3Daverage" with the other most commonly used "3Dmerging" method.

Following features classes were extracted: statistical first order (FO), shape metrics (SM), texture features (TA) including Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighbouring Gray Tone Difference Matrix (NGTDM), Gray Level Dependence Matrix (GLDM), wavelet features (WF) for all of the above features excluding shape, computed using all combinations of applying either a High or Low pass filter in each of the three dimensions. In total, we extracted 841 features from each image volume (18 FO, 13 SM, 23 GLCM, 16 GLRLM, 16 GLSZM, 14 GLDM, 5 NGTDM, and 736 WF). The shape feature volume of each GTV was approximated by multiplying the number of voxels in the region of interest (ROI) by the volume of one voxel.

2.3. Elimination of redundant features

Pairwise feature inter-dependencies were evaluated using the Spearman rank correlation coefficient (ρ) The ρ metric does not assume any a priori functional

dependence for the data (contrary, for example, to the Pearson coefficient) and therefore it is able to catch complex functional dependencies between features. The redundant features (with $|\rho| \geq t$, where t is a chosen threshold value) were eliminated by randomly dropping one of the two features. Thresholds from 0 to 1 with a 0.05 increment step were used.

2.4. Cluster analysis

We used hierarchical clustering to discover groups of patients with similar radiomics signatures. The optimal number of clusters (k) was determined using the consensus clustering method [16]. Briefly, clustering is repeated multiple times for different values of k using random sub-samples of the data. The value of k resulting in the most stable clusters (i.e. least change in cluster assignment for each observation across samples) is selected. We compared the distributions of clinical variables and GTV volume between clusters. In addition, we computed the Kaplan-Meier estimator of overall survival in each cluster. The log-rank test is a standard procedure to assess the statistical significance of difference between survival function estimates (with p-values corrected for multiple comparisons, using the FDR (False discovery rate) correction method (Benjamini-Hochberg procedure) [16].

2.5. Principal component analysis (PCA)

Principal component analysis (PCA) is an unsupervised method aiming to discover the sources of variance in the data. PCA identifies the directions of largest variability in the original dataset (called principal components). PCA is useful to determine if there is a confounding factor intrinsically present in the computed features, which is driving the variance in the data. The principal components are linear combinations of features and are ordered by the amount of total variance they explain. Thus, the first principal component represents the predominant pattern in the data and its strength is captured by explained variance. PCA can be used to identify latent variables sources of variability not observed directly but nevertheless captured by the features. For example, if a suspected confounding variable is highly 1correlated with the first principal component, it is likely to be the true source of variation. A more comprehensive overview of the technique can be found in Traverso et al. [12].

2.6. Feature selection and modelling

To investigate the predictive value of volume-independent ($\rho \leq 0.1$) imaging features, we used them in combination with clinical variables (age, T, N, M stages, AJCC stage and tumour volume) in a binary logistic regression. We applied the model to two-year overall survival prediction. To determine the relative importance of features, as well as the stability to perturbations in the input, we applied a bootstrap-based method as detailed in [17]. Briefly, the model is refit on multiple bootstrap re-samples of the data and the order in which a feature is important for a model is obtained using Recursive Feature Elimination (RFE). The importance of each feature, as well as correlations between features, can be identified easily by visualizing the resampling results. Furthermore, the overall importance of each feature can be identified by aggregating the results across bootstrap resamples. Bootstrap-based variable selection analysis increases the reliability of reported models. All the statistical analysis was performed in Python v3.7.5 using the statistical package scikit-learn v0.21.3. Statistical significance was set at $p < 0.05$.

The workflow is briefly summarized in Table 1 (further commented in the discussion section) and it is composed by the following sequential steps that were adopted in this study: 1) evaluation of pairwise correlations between tumour volume using Spearman rank analysis and drop highly correlated features; 2) use the reduced list to perform hierarchical clustering and evaluating distribution of clinical variables (or confounding factors) in the clusters. Use PCA to select components that explain the largest percentage of variance, but still evaluating correlations between components and confounding factors; 3) to address sample biases use bootstrap techniques with RFE (Recursive Feature Elimination) and force in the model the presence of clinical prognostic factors. Build the final signature by selecting the most selected features.

Table1. Suggested workflow for radiomics signature developments that incorporates safeguards.

Step #	Steps description	Method used	Address biases
1 – Redundancy and confounding factors analysis	Evaluate pairwise correlations between features by and tumour volume using ρ coefficients. Select a cut off for ρ and keep only non-redundant features	Spearman correlation coefficients (ρ)	Redundancy Confounding factors
2 – Clustering and PCA analysis	a) Use the new list of features as input for unsupervised hierarchical clustering b) Evaluate distributions of clinical variables in the clusters c) Select features that show significant differences between the clusters and explain most of the variance in the data (PCA)	Hierarchical clustering PCA	Dimensionality Reduction
3 – Outcome modeling	a) Use reduced list of features to develop the model. Choose the preferred classifier (e.g. SVM or logistic regression) combined with RFE b) Bootstrap the original dataset at least 1000 times and derive the most selected features c) Build the final model with most selected features	Bootstrapping RFE (Recursive feature elimination)	Sample bias

3. Results

3.1. Feature correlations

Pairwise Spearman correlation between features revealed a high level of inter-dependence in the NSCLC dataset, with over 80% correlating with at least one other feature at $|\rho| > 0.9$ (Fig. 1a). Furthermore, nearly 30% showed correlation with tumour volume greater than 0.75 (Fig. 1c). A similar correlation was observed in the HN1 dataset (Fig. 1b and 1d). [R1] SupplementaryTable S1 lists all the radiomic features that presented a Spearman correlation $|\rho| \geq 0.8$ with GTV.

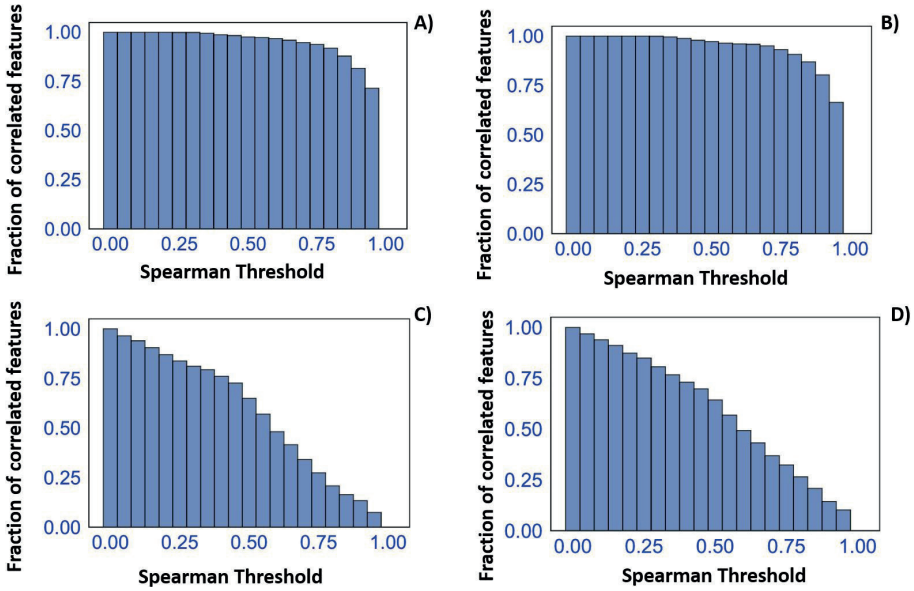


Figure 1. Proportion of correlated features as a function of Spearman rank correlation threshold. (a) shows the proportion of features with pairwise correlation greater than the threshold value in the lung dataset. Percentage of features correlated with volume at a given threshold is shown in (c). The results were similar in HN1 (b, d).

3.2. Clustering and PCA analysis

Using all the feature set, the patients could be stratified into two groups with significantly different survival times (log-rank $P < 10–6$, Fig. 2a). The difference in tumour volume distribution between clusters was highly significant (permutation $P < .001$, Fig. 2a). Removing features moderately correlated with volume (with Spearman $|\rho| > 0.6$) still allowed for cluster separation by OS ($P < 10–6$, Fig. 2b); however, the volume difference between clusters remained highly significant ($P < .001$, Fig. 2b). Using only volume-independent features ($|\rho| < 0.1$) the groups could not be separated by survival ($P = .8$, Fig. 2c) or tumour volume ($P = .9$, Fig. 2c). Groups could be separated by survival [R2] using as only input feature the computed GTV (log-rank $P < 10–6$) with no statistically significant differences between the full signature and GTV signature as shown in Fig. 2d.

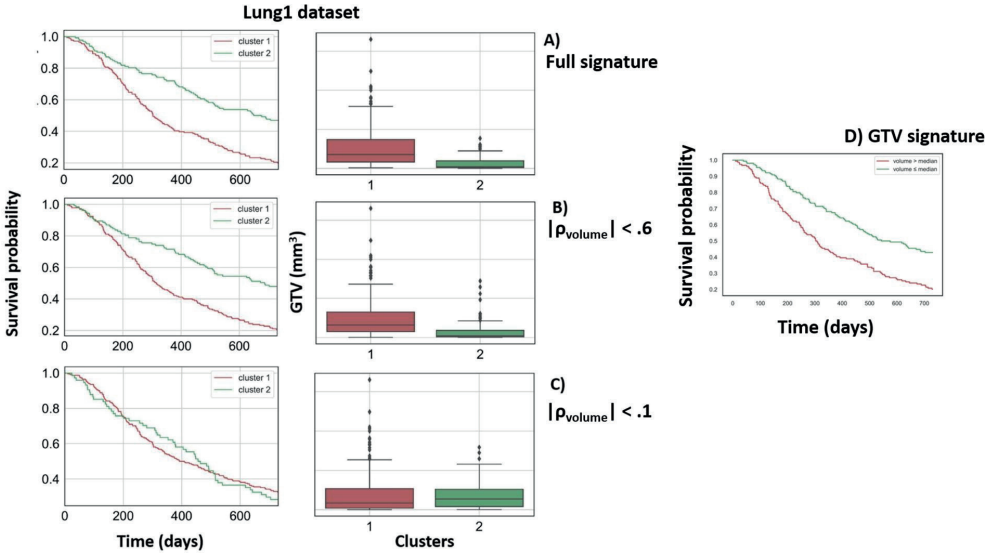


Figure 2. Kaplan-Meier OS estimates and volume distributions for each cluster identified in the NSCLC dataset. (a) all original and wavelet features, (b) features moderately correlated with volume (defined as Spearman |ρ| < 0.6 with volume), (c) features not correlated with volume (|ρ| < 0.1), d) GTV signature.

As per this experiment it is possible to appreciate a degradation of performances when slowly removing features that are highly correlated with tumour volume, finally reaching a point ((|ρ| < 0.1) where only volume independent features are left, but no stratification is possible.

The first principal component (PC) extracted from all feature signature correlated with volume (Spearman ρ = 0.78, Fig. 3a). The first 2 PCs explained over 50% of the total variance, reflecting the large number of volume-correlated features. The latent volume effect was still present when moderately (|ρ| < 0.6) correlated features were used (correlation with volume: ρ = − 0.37 for PC 1 and ρ = 0.79 140 for PC 2, Fig. 3b), explaining the significance between-cluster differences in tumour volume. Finally, there was no volume-dominant effect in features independent (|ρ| < 0.1) from volume (ρ = 0.01 for PC 1 and ρ = 0.05 for PC 2, 143 Fig. 3c). Due to a smaller number of cases in the HN1 dataset, only one cluster could be reliably identified. In PCA, we found a dominant volume effect in full signatures (ρ = −0.91, Fig. 3a right) similarly to the NSCLC dataset. Crucially, the effect was present even in moderately correlated features (correlation with volume: ρ = −0.52 for PC 1 and ρ = −0.58 149 for PC 2, 3b right). Again, the effect was not present in non-correlated features 150 (ρ = −0.05 for PC 1 and ρ = 0.03 for PC 2, Fig. 3c right).

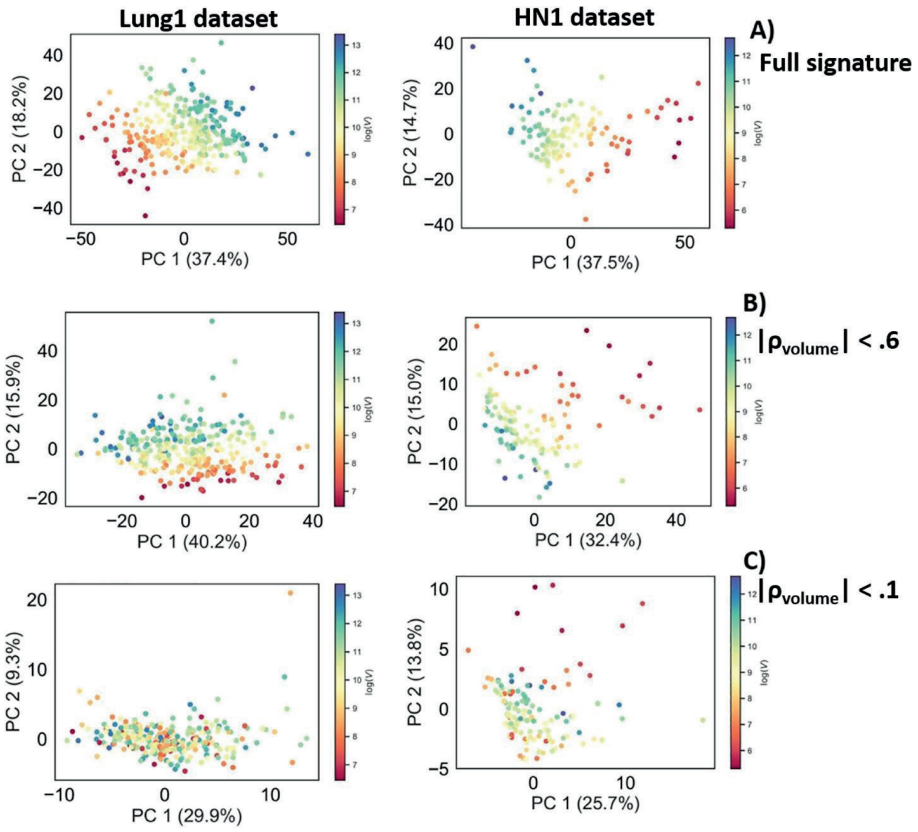


Figure 3. Principal component analysis of (a) full feature signatures, (b) features moderately correlated with volume (|ρ| < 0.6), (c) volume-independent features (|ρ| < 0.1) in lung and head & neck datasets. The data is shown projected on the first 2 principal components and the proportion of variance explained by each component is indicated. Colors correspond to tumor volume. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Feature selection and modelling

Fig. 4 shows the order of selection in each bootstrap dataset (1000 replications in total) alongside the frequency of each feature entering the model first. In Lung1, volume enters the model first in most resampling iterations (84%), followed by T stage (which carries partially overlapping, but not identical information) and M stage (both 10%). In HN1, it is worth noting that the number of volume-independent features is larger. The most frequently selected feature is the overall stage (44%), followed by N stage and tumour volume (17% and 15% respectively). This reflects the higher importance of nodal involvement in head & neck squamous carcinomas for OS [18]. Interestingly,

one imaging feature (GLDM-Gray Level Variance, correlation with volume 0.1) entered the model as first almost as frequently as volume (14%), indicating potentially complementary information. It is worth noting that our aim was not to create the optimal model, but rather to investigate the robustness of feature predictive performance. Results and conclusions remained unchanged when considering texture features computed with the "3Dmerging" aggregation approach.

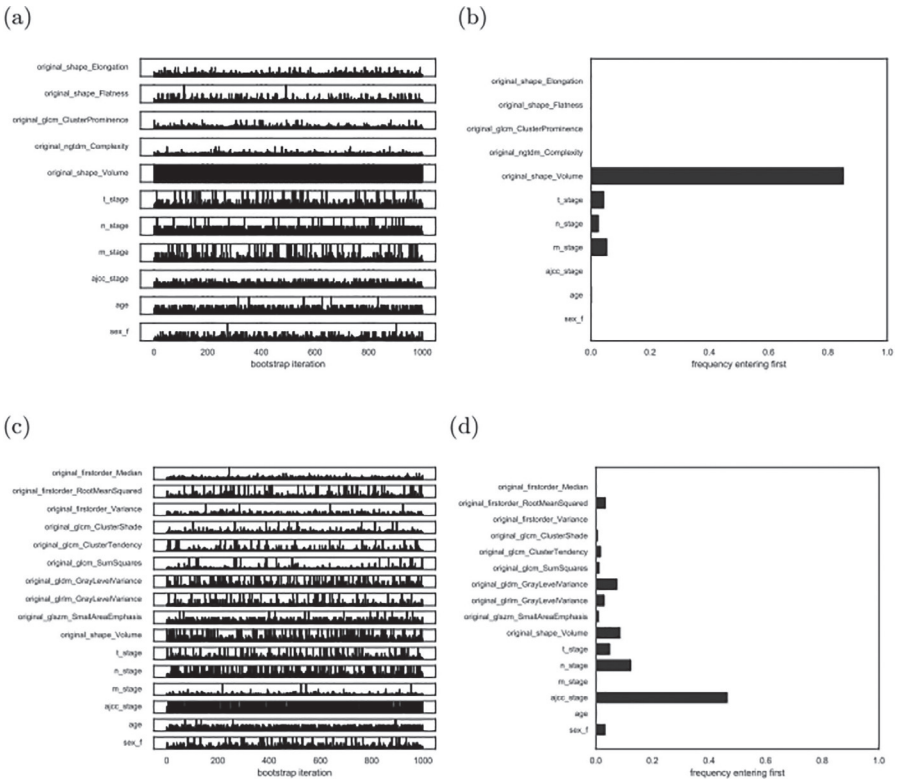


Figure 4. Bootstrap-based evaluation of predictive power and stability of imaging and clinical features. (a) and (c) show the order of each feature entering the model across 1000 resampling iterations in Lung1 and HN1, respectively. The height of the bar is inversely proportional to the order of selection (therefore, filled bar indicates higher importance). (b) and (d) show the frequency of each feature entering the model first in both datasets.

4. Discussion

The evaluation of radiomics features multicollinearity and their benchmark with respect to accepted clinical prognostic factors is a needed safeguard. Our results show that radiomics features present strong inter-correlations, where texture

features (TA) are usually more correlated between each other than first order (FO) features. Applying a wavelet filter augmented this problem, increasing therefore the dimensionality of problem to be solved and leading to a situation prone to overfitting.

Besides feature-feature correlations, a large percentage of radiomics features showed marked dependencies with tumour volume: 50% of total features had ρ volume $>|0.6|$, independently from the anatomical site considered (Fig. 1c and d). Again, TA features showed higher correlations with tumour volume than FO features. Three of highly correlated features were confirmed to be affected by strong volume correlations also in [14]. In HN1, texture features had slightly lower correlations with tumour volume than in Lung1.

Applying filtering decomposition of the original image did not eliminate the volume-effect. Our results confirmed that was no statistically significant difference of volume correlations between original and filtered features. Therefore, the usage of image filtering should carefully be adopted and justified to avoid an increase in the dimensionality of the features space to be reduced, without bringing any new information.

Tumour volume is a well-established and benchmarked prognostic factor for lung and head and neck cancers [5,11]. Therefore, correctly identifying if a feature or a combination of feature (e.g. signature) prognostic power is driven by a volume-confounded effect is fundamental to avoid spurious conclusions. With this evidence in mind we want again to clarify that the goal of our manuscript was not to discourage the community from building radiomics models and discard this effort since other predictors exist, but it was to provide this community with machine learning methods that could help achieving a good trade-off between explicability, transparency, parsimony, accuracy, and overfitting. Driver to reaching this trade-off, but still achieving good and robust performances is to identify the important explanatory variables.

Unfortunately, the majority of radiomics features come as complicated mathematical formulas, where identifying a direct and immediate dependence to tumour volume is far from trivial. In addition, two single features might not present strong dependencies to tumour volume, but their combination could.

In this study we showed how machine learning can be used to address the above-mentioned issue. Compared to the traditional radiomics workflow, we collocated machine learning at the top of the process, as a powerful instrument for exploratory analysis and acts as a safeguard against unanticipated cross-correlation with known prognostic features. The unsupervised methods of

clustering and PCA present the following advantages: a) searching for patterns in the data without assuming any a-priori distribution or condition (i.e. without looking at the 'labels'); b) providing an intuitive way to retain pertinent information in the analysis and verify the main driver of it. When we cluster patients using all the radiomics features, the separation in terms of overall survival was statistically significant. However, the main reason of splitting can be attributed to strong volume differences between the clusters (Fig. 2). When we drop features correlated with tumour volume using a cut-off of 0.6, it was still possible to separate the two clusters in terms of OS, but with worse statistic. However, the clusters still had a predominant volume difference (Fig. 2). Finally, when considering only volume-independent features ($\rho \leq 0.1$) there was not significant splitting and no statistically significant difference between tumour volumes in the clusters. The results confirm that most of the radiomics features, when combined, led to spurious associations with tumour volume. Furthermore, volume-independent features alone did not allow stratification of patients into bad and good prognosis groups.

To further prove that the volume-latent effect is present independently from the unsupervised algorithm chosen, we repeated the PCA analysis but using the tSNE (t-distributed Stochastic Neighbouring Entities) [13]. It is another well known visualization method for high dimensional data, but compared to PCA, it uses a probabilistic approach. In our study, these two techniques were used as complimentary to further verify the found results. In fact, the same volume-latent effect was confirmed (figures available in the Supplementary material) also with tSNE.

Finally, we showed how bootstrap methods can be combined with supervised machine learning to evaluate feature significance. Furthermore, since bootstrap methods consider different subsamples of the original datasets, the risk of spurious associations, due to sample effects, is reduced. It is then possible to rank features according to their importance for the model by Recursive Feature Elimination. If a feature is important and has high prognostic value, it will often be selected, despite the chosen sample. A recent submitted publication to this journal related to radiomics-based model in head and neck cancers [14], showed that combining radiomics and clinical predictors did not lead to an elevate increase of performances. Similar results are found in our analysis also for the lung dataset: when considering only volume-independent features, tumour volume and t-stage outperformed each of the imaging features (Fig. 4a and b). In HN1 one radiomics feature was selected as often as other traditional clinical factors, but still the most frequent feature was nodal, showing that information outside the GTV (e.g. nodal

involvement) plays a strong role in head and neck cancers. It is important to notice that it was out of this paper's scope to build the best model for predicting OS. Rather the aim was to provide the radiomics community with a method to benchmark radiomics predictors with accepted clinical factors and evaluate their stability with respect to a particular splitting of the datasets.

Finally, we provide safeguarding recommendations for signature developments in radiomics studies that build upon Welch et al. [14]: a) unsupervised learning methods (e.g. clustering and PCA) are preferable for exploratory analysis and dimensionality reduction with respect to traditional univariate and multivariate analysis; b) bootstrapping of radiomics predictors with accepted clinical factors provides a method to benchmark radiomics features and check the stability with respect to different sample sizes.

Table 1 summarizes a list of radiomics safeguards with suggestions of machine learning-based methodology for their applications.

While the results presented in this study remain valid only for the investigated clinical outcome (2-year OS), for the imaging modality (CT) and for the anatomical sites of lung and head and neck, the workflow presented in Table 1 can be extended as standard methodology for radiomic studies. We encourage the radiomic community to consider using unsupervised methods and the benchmarking of radiomic features with bootstrap techniques in their studies. Additional proven evidence of results found in this paper (e.g. degradation/contamination of prognostic power as a function of GTV/ feature dependencies) will help improving the quality of radiomic studies as well as re-thinking the definitions/role of some radiomic features.

Finally, it is worth mentioning some limitations of this study: a) due to limited availability we focused only on OS; b) the bootstrap modelling was limited only to logistic regression, but it could have extended also to other classifiers; c) the stated conclusions only apply to the studied anatomical sites (lung and head and neck) and for non CE (Contrast Enhanced) CT. The same conclusions might not be valid when different modalities are considered (e.g. PET/CECT/MR) or applied to other anatomical sites, posing the urgent need to validate and share our methods with the radiomics community**.** Future works include addressing points a) and b) as well as considering volume-correction methods for improving signature developments in radiomics. We are planning to release the code as open source to incentive the community to adopt the presented methodology as benchmark for their studies.

5. Conclusion

In available datasets, volume confounds common radiomics analysis approaches. Volume, or other parameters which confound analysis, should be recognized during any radiomics workflow and dedicated safeguards should be built into analysis pipelines to identify and mitigate these risks.

Our study showed that by only using volume-independent features it was not possible to cluster patients in different survival groups.

Supplementary materials

References

[1] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5. https://doi.org/10.1038/ncomms5006.

[2] Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med 2014;55:37–42. https://doi.org/10.2967/jnumed.112.116715.

[3] Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One 2015;10:e0124165 https://doi.org/10.1371/journal.pone.0124165.

[4] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345–50. https://doi.org/10.1016/j.radonc.2015.02.015.

[5] Dejaco D, Steinbichler T, Schartinger VH, Fischer N, Anegg M, Dudas J, et al. Prognostic value of tumor volume in patients with head and neck squamous cell carcinoma treated with primary surgery. Head Neck 2018;40:728–39. https://doi.org/10.1002/hed.25040.

[6] 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. https://doi.org/10.1016/j.critrevonc.2019.03.015.

[7] Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med 2015;56:38–44. https://doi.org/10.2967/jnumed.114.144055.

[8] Matzner-Lober E, Suehs CM, Dohan A, Molinari N. Thoughts on entering correlated imaging variables into a multivariable model: application to radiomics and texture analysis. Diagn Intervent Imaging 2018;99:269–70. https://doi.org/10.1016/j.diii.2018.04.011.

[9] Shi Z, Traverso A, Soest J, Dekker A, Wee L. Technical note: ontology-guided radiomics analysis workflow (O-RAW). Med. Phys. 2019. https://doi.org/10.1002/mp.13844. mp.13844.

[10] Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the personal health train infrastructure. Sci Data 2019;6:218. https://doi.org/10.1038/s41597-019-0241-0.

[11] Takenaka T, Yamazaki K, Miura N, Mori R, Takeo S. The prognostic impact of tumor volume in patients with clinical stage IA non-small cell lung cancer. J Thorac Oncol 2016;11:1074–80. https://doi.org/10.1016/j.jtho.2016.02.005.

[12] Traverso A, Dankers FJWM, Osong B, Wee L, van Kuijk SMJ. Diving deeper into models. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science Cham: Springer International Publishing; 2019. p. 121–33. https://doi.org/10.1007/978-3-319-99713-1_9.

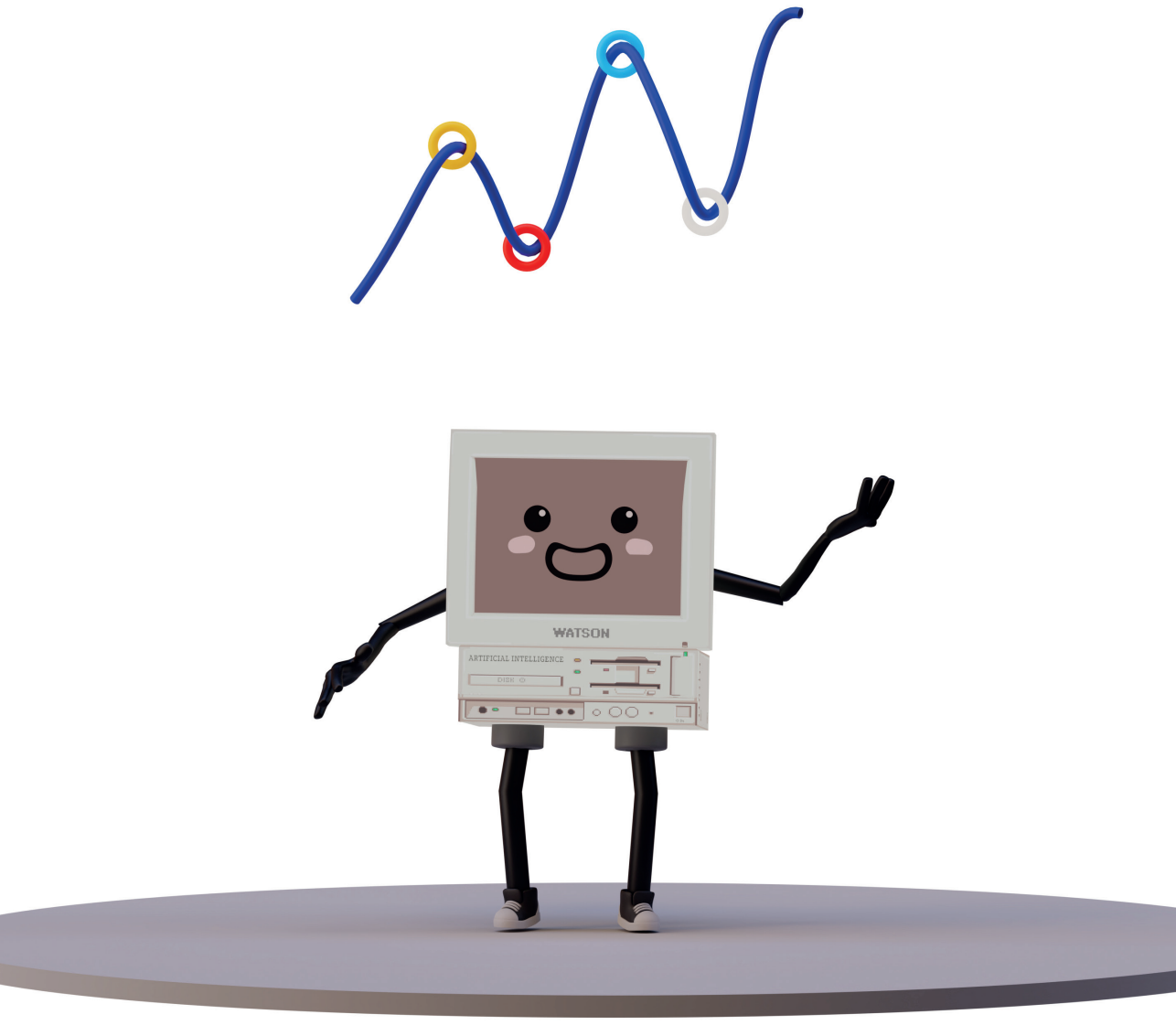
[13] Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. Distill 1 2016. https://doi.org/10.23915/distill.00002.

[14] Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2018. https://doi.org/10.1016/j.radonc.2018.10.027.

[15] van Griethuysen JJM, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[16] Vega-Pons S, Ruiz-Shulcloper J, et al. A survey of clustering ensemble algorithms. Int J Pattern Recogn Art Intel 2011;25(3):337–72. https://doi.org/10.1142/ S0218001411008683.

[17] El Naqa I, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys 2006;15(64):1275– 86.https://doi.org/10.1016/j.ijrobp.2005.11.022.


WATSON
ARTIFICIAL INTELLIGENCE

Chapter 6

Plausibility and Redundancy Analysis to Select FDG-PET Textural Features in Non-Small Cell Lung Cancer

Pfaehler, Elisabeth, Liesbet Mesotten, Ivan Zhovannik, Simone Pieplenbosch, Michiel Thomeer, Karolien Vanhove, Peter Adriaensens, and Ronald Boellaard. "Plausibility and redundancy analysis to select FDG-PET textural features in non-small cell lung cancer." *Medical physics* 48, no. 3 (2021): 1226-1238.

Abstract

Background: Radiomics refers to the extraction of a large number of image biomarker describing the tumor phenotype displayed in a medical image. Extracted from Positron Emission Tomography (PET) images, radiomics showed diagnostic and prognostic value for several cancer types. However, a large number of radiomic features are non-reproducible or highly correlated with conventional PET metrics. Moreover, radiomic features used in the clinic should yield relevant information about tumor texture. In this study, we propose a framework to identify technical and clinical meaningful features and exemplify our results using a PET Non- Small-Cell Lung Cancer (NSCLC) dataset.

Materials and Methods: The proposed selection procedure consists of several steps. A priori, we only include features that were found to be reproducible in a multicenter setting. Next, we apply a voxel randomization step to identify features that reflect actual textural information, i.e. that yield in 90% of the patient scans a value significantly different from random texture. Finally, the remaining features were correlated with standard PET metrics to further remove redundancy with common PET metrics. The selection procedure was performed for different volume ranges, i.e. excluding lesions with smaller volumes in order to assess the effect of tumor size on the results. To exemplify our procedure, the selected features were used to predict 1-year survival in a dataset of 150 NSCLC patients. A predictive model was built using volume as predictive factor for smaller, and one of the selected features as predictive factor for bigger lesions. The prediction accuracy of the both models were compared with the prediction accuracy of volume.

Results: The number of selected features depended on the lesion size included in the analysis. When including the whole dataset, from 19 features reflecting actual texture only two were found to be not strongly correlated with conventional PET metrics. When excluding lesions smaller than 11.49 mL and 33.10 mL (25 and 50 percentile of the dataset), 4 out of 27 features and 13 out of 29 features remained after eliminating features highly correlated with standard PET metrics. When excluding lesions smaller than 103.9 mL (75 percentile) 33 out of 53 features remained. For larger lesions, some of these features outperformed volume in terms of classification accuracy (increase of 4-10%). The combination of using volume as predictor for smaller and one of the selected features for larger lesions also improved the accuracy when compared with volume only (increase from 72% to 76%).

Conclusion: When performing radiomics analysis for smaller lesions, it should be first carefully investigated if a textural feature reflects actual heterogeneity information. Next, verification of absence of correlation with all conventional PET metrics is essential in order to assess the additional value of radiomic features. Radiomic analysis with lesions larger than 11.4 mL might give additional information to conventional metrics while at the same time reflecting actual tumor texture. Using a combination of volume and one of the selected features for prediction yields promise to increase accuracy and reliability of a radiomic model.

Keywords: Radiomics, feature selection, clinical value

1. Introduction

Quantitative interpretation of Positron Emission Tomography (PET) may be used for diagnosis, prognosis, and treatment response assessment for cancer patients[1], [2]. To date, the maximum intensity value in the tumor (SUVMAX), the metabolic active tumor volume (MATV) or the total lesion glycolysis (TLG) are often used for this purpose [3,4]. Recently, other metrics describing the textural properties of the tumor (i.e. tumor heterogeneity) have gained increasing interest [5,6]. These so called radiomic features might yield additional value to conventional metrics and might help to determine prognosis and treatment efficacy more accurately and reliably than conventional metrics. Several studies commented on the promising value of radiomic features for different tumor types [6-8].

While a large number of studies reported on the additional value of radiomics, more and more studies address its challenges and pitfalls [9,10]. The sensitivity of radiomic features to all steps in the radiomics workflow (i.e. image acquisition and reconstruction, tumor delineation, image discretization, and image processing) has been discussed in several reports [8-12]. In a clinical workflow, only features that result in comparable values when extracted from a patient scanned several times under the same conditions (repeatable features) and that are yielding only small differences when extracted from scans of the same patient acquired under different conditions (e.g. different scanners) (reproducible features) should be used in a radiomics workflow [13]. Both characteristics are essential in order to guarantee a reliable treatment assessment and an accurate diagnosis independent of e.g. on which PET scanner the patient was examined. Moreover, many radiomic features are highly correlated with conventional metrics such as MATV what makes their additional value questionable [14]. Additionally, a radiomic feature used in the clinic should be relevant and explainable, i.e. should reflect the tumor heterogeneity observed in the medical image accurately [15].

It is of utmost importance that all these requirements are fulfilled as it might happen that a feature is found by chance to be predictive for the required task without yielding any meaningful information about tumor heterogeneity. Regarding the large number of radiomic features and the relatively small number of patients included in the majority of studies, this can happen as illustrated in [16] where the authors used randomly generated feature values for the prediction of overall survival and achieved a maximum cross-validation accuracy under the curve (AUC) of 0.79. Therefore, each radiomic feature should be carefully checked if it really reflects the tumor heterogeneity observed in the image.

In this study, we propose a procedure to identify and select only those features that may yield technical and clinical meaningful information before using them in a prognostic model in an example dataset. We only include radiomic features which were found to be repeatable and reproducible in a previous multi-center study. Out of these robust features, we further select features using a method by Welch et al. who randomly shuffled tumor intensity values of CT images in order to destroy the underlying tumor texture. By comparing the features extracted from the original image with the features extracted from randomly redistributed voxel value distributions, we can identify features describing tumor heterogeneity beyond randomness. Hereby, we consider a feature as describing actual texture information when the feature value extracted from the original image is significantly different from the feature value of the randomly redistributed voxel intensities. If this is not the case, the feature does not seem to describe actual texture information. Moreover, we eliminate those features which are strongly correlated to conventional PET metrics. Previous studies demonstrated that the correlation to conventional PET metrics depends on the lesion size [17]. To cover this aspect, we apply the proposed feature selection procedure on different tumor volume ranges. The proposed feature selection procedure does not only avoid false positive findings and overfitting, but can also be used for feature space reduction allowing to study radiomics performance in more realistically large sample size (typically for PET up to a few hundred at most). We perform a detailed investigation which radiomic features are meeting all described requirements. As an example of the proposed selection procedure, the selected features are used for classifying 1-year survival using a logistic regression model. Hereby, we do not aim to build an optimal prognostic model by including e.g. additional clinical parameters. The aim of this study is to illustrate if we can identify features that describe actual tumor texture and to investigate if these features would then also have additional clinical value in the example dataset.

2. Materials and Methods

2.1. Dataset

The study was registered at clinical trials.gov (NCT02024113). All patients gave informed consent for study participation and use of their data for scientific research. The dataset consists of 150 patients with Stage I – IV NSCLC. Details about the patient cohort can be found in table 1. Additional clinical parameters such as the

smoking status are displayed in Supplemental Table 1. All patients fasted at least six hours before image acquisition. Time between PET scan and tracer injection was around 60 minutes. All images were corrected for attenuation, scatter, random coincidences, and normalization. Images were acquired on a Gemini TF Big Bore (Philips Healthcare, Cleveland, OH, USA) and reconstructed using the BLOB-OS-TOF reconstruction method provided by the vendor. The reconstructed images yielded a cubic voxel size of 4 mm and are compliant with the benchmarks of the European Association of Nuclear Medicine Research Ltd. (EARL).

Table 1: Patient characteristics

Cancer stage	IA	28 patients
	IB	13 patients
	IIA	10 patients
	IIB	8 patients
	IIIA	29 patients
	IIIB	15 patients
	IV	47 patients
Age	mean	67 years
	std	9.3 years
Sex	men	93
	women	57

2.2. Tumor segmentation

Image analysis was performed using an in-house developed tool designed for the analysis of PET images [18] used in previous works [19,20]. For the segmentation of the tumor volume of interest (VOI), a semi-automatic segmentation was performed including all voxels yielding a SUV above 2.5. High-uptake regions such as the heart or inflammations in the lung were eliminated by the observer. For lesions close to the heart or other high-uptake regions, a bounding box was manually drawn around the tumor to avoid inclusion of physiologically high uptake normal tissues. Every segmentation was manually checked and corrected if necessary. If some voxels of a close high-uptake region were still included in the VOI, these voxels were excluded manually from the VOI. Only the primary tumor was included in the current analysis, which was defined as the lung lesion with the largest volume.

2.3. Radiomic feature calculation

Radiomic features were calculated using the open-source software RaCaT which complies with the benchmarks provided by the Image Biomarker Standardization Initiative (IBSI) [21,22]. Prior to feature calculation, images and corresponding VOIs were resampled to a cubic voxel size of 2 mm using tri-linear interpolation as recommended by Hatt et *al.* [10]. A resampling to a cubic voxel size leads to a larger number of reproducible radiomic features as some features depend on the number of voxels included in the VOI [21]. After interpolation, all voxels yielding a value above 0.5 in the resampled segmentation mask were included in the VOI. Before textural feature calculation, all images were first converted to standardized uptake values (SUV) and then discretized using a fixed bin size of 0.25 SUV as recommended by various studies [11,23-25]. Moreover, these settings were chosen as it has been demonstrated that they lead to the largest number of reproducible features [26]. Exact feature definitions as well as details about the feature calculations are described elsewhere [21].

2.4. Data Analysis

All data analysis was performed in Python 3.4 using the packages numpy, scipy, and scikit-learn.

2.5. Feature selection procedure

Only features that were identified to be robust and reproducible in previous work were included in the analysis [26]. In this work, a phantom containing 3D printed phantom inserts reflecting realistic tumor heterogeneity was scanned on various PET systems. The inserts were segmented on each scan separately and radiomic features were extracted. The features that only yielded small differences between the different PET systems and delineations were identified to be reproducible and are included in the present study. However, we recommend to make use of semi-automated segmentation approaches, as much as possible, to avoid observer variability in tumor delineations which in turn will affect the reproducibility of radiomic analysis. Morphological and statistical features were not considered in the feature selection as they remain constant when performing step 1 (randomly shuffling the intensity values within the VOI). This lead in total to 92 radiomic features. A list containing the names of all features included in this study is provided in the supplemental material (supplemental Table S1).

The proposed feature selection procedure consists of three steps:

Step 1.) *Randomized voxel assignment:* Use a voxel randomization method to identify features that are reflecting the tumor heterogeneity observed in the PET image beyond randomness

Step 2.) *Correlation with conventional metrics:* Identify the features yielding additional value to conventional PET metrics

Step 3.) *Mutual correlation:* From the remaining features, select only the features yielding complementary information to each other

The described selection steps will be detailed in the next paragraphs.

Step 1: Randomized voxel assignment

In order to use a radiomic feature in the clinic, it is important that the feature reflects the actual heterogeneity displayed in the PET image. This means that a lesion with the same shape but with a different texture should also have a different feature value. To check if a feature is fulfilling this assumption, the intensity values in a VOI were randomly reshuffled similar to a method proposed by Welch et al [14]. In this way, the original texture of a lesion is destroyed. However, instead of shuffling the intensity values of the whole 3D dataset as proposed by Welch et al., we only shuffled the intensity values inside the VOI and preserved in this way the first order statistics. Radiomic features from these randomly generated voxel redistributions were calculated. The procedure was repeated 50 times per image. An example illustrating the original and two randomly generated distributions are displayed in Figure 1. Features extracted from the randomly generated voxel distributions will be called random features hereafter.

Figure 1. Original tumor (left), two examples (middle and right) of the same tumor after randomly shuffling the intensity values in the VOI

All features calculated from the original voxel distribution within the tumor VOI were compared with features values from the random distributions. Features yielding different values for the original voxel distributions, were regarded as reflecting the actual texture displayed in the image. For this purpose, the mean and standard deviation of the 50 random features were calculated for each tumor separately. A feature calculated from the original voxel distribution was considered to describe random texture when its feature value was within the 95% confidence interval:

$$[mean_{rand} - 1.96 * std_{rand};\ mean_{rand} + 1.96 * std_{rand}]$$

With $mean_{rand}$ being the mean value and std_{rand} being the standard deviation calculated from all 50 random features of one tumor. Every image and tumor was evaluated separately.

Features considered for further analysis needed to yield a value outside the proposed confidence interval for 90% of the patients (note that 90% is used in this paper for illustration purposes, but other thresholds can be applied as deemed appropriate. For example a lower threshold would result in a more liberal selection- thus more features passing this selection step).

Step 2: Correlation with conventional metrics

A radiomic feature should yield additional value to conventional PET metrics. Therefore, all features were checked for their correlation with MATV, SUV_{PEAK}, and SUV_{MEAN} using the Spearman's Rank Correlation Coefficient. The Spearman's Correlation Coefficient is a non-parametric metric describing the relationship between two variables. By comparing the statistical dependence of the rank of the variables, it also captures non-linear relationships. A correlation above 0.9 was regarded as very strong [27]. Features with a very strong correlation to one conventional metric were regarded as redundant and were therefore discarded from the analysis.

Step 3: Mutual correlation between features

Finally, it is also important that the remaining features are reflecting different tumor characteristics and are therefore not highly correlated among each other. Therefore, all remaining features were checked for a very strong correlation among each other. If two or more features yielded a very strong correlation (correlation coefficient > 0.9), the feature that yielded the lowest correlation with the conventional metrics was kept and tested for its clinical value.

2.6. Summary of the feature selection procedure

In summary, from the repeatable and reproducible features selected a priori the features that were considered for further analysis were the features that:

1. described the texture observed in the image beyond randomness: yielded different feature values for the original tumor voxel distribution than for randomly redistributed voxel intensities within the tumor VOI
2. have additional value to conventional PET metrics: yielded a correlation coefficient below 0.9 with tumor volume, SUV_{PEAK}, and SUV_{MEAN}
3. yielded complementary information among each other: if features were strongly correlated among each other (0.9 or more), the feature resulting in the lowest correlation with conventional PET metrics was kept

As it has been reported previously that the tumor volume has an impact on the the correlation of a feature with conventional metrics [17], all steps were performed for five volume ranges separately and results were compared:

- tumors yielding more than 50 voxels in the original image (>3.2 mL)
- tumors larger than the 25% percentile of the dataset (>11.48 mL)
- tumors larger than the median of the dataset (>33.04 mL)
- tumors larger than 45 mL as indicated by Brooks et *al*. [28]
- tumors larger than the 75% percentile (>103.94 mL).

The results of the feature selection procedure will be detailed as function of tumor volume range in the remainder of the paper.

2.7. Comparison with other feature selection methods

The proposed feature selection procedure was compared with two automatic feature selection methods: maximum relevance and minimum redundancy (MRMR) and RELIEF. Both methods aim to select only non- redundant (i.e. non-correlated) as well as features relevant for the outcome. The MRMR algorithm is selecting the features showing the highest dependence with the outcome variable and at the same time the lowest dependence with other selected features [29]. While the Relief algorithm assigns weights to each feature [30]. The given weight

increases when a feature is highly similar between patients in the same group (i.e. alive after one year), while the weight decreases when a feature is highly similar to the feature of patients belonging to the inverse group (i.e. dead after one year). Features with the highest weights are selected.

Both feature selection methods were performed for each volume range separately. The ten most important features identified by the two approaches were compared with the features selected by our procedure before eliminating features with a high mutual correlation. The MRMR feature selection was performed in the programming language R using the package mRMRe [31], while the RELIEF algorithm was implemented in Python 3.6.4.

2.8. Features found in previous studies

Features found to have clinical value in previous studies were analyzed if they fit the described criteria. This included zone percentage (GLSZM) and entropy (GLCM) [6,32], as well as high intensity large area emphasis (GLSZM) [32]. Moreover, coarseness, contrast, and busyness (NGTDM) were analyzed [33,34].

2.9. Clinical value of selected features

In order to illustrate the clinical value of features matching the described criteria, three different models were considered:

Model 1: Each feature was used separately for the prediction of 1-year survival using stratified cross-validation and a logistic regression classifier (a detailed description of the classification process is given below). The mean accuracy under the curve (AUC) of all cross-validation folds was compared with the mean AUC when using the feature volume.

Model 2: In order to assess their additional value, features found to yield a clinical value in Model 1 were, in a second step, used in combination with MATV in the cross-validation and the logistic regression model. Also here, the mean AUC of the combined classification was compared with the mean AUC of volume alone.

Model 3: For features yielding a clinical value in Model 1, receiver operating characteristic curves (ROC) were drawn for the whole dataset. For lesions with volumes below the different thresholds, volume was used as prognostic factor, while for lesions above the thresholds, the selected feature was used. This was done for each selected feature separately. The AUC of the combined prognosis (volume + feature) was compared with the AUC of volume only.

For all models, the clinical value of each feature was determined using the whole dataset and for each volume range separately. Even if a feature was only found to match all criteria for larger lesions, classification was performed for all volume ranges in order to compare the predictive value of the feature for datasets with smaller and larger volume ranges. The distribution of positive/negative outcomes for each volume range is displayed in supplemental Figure 1.

As the number of patients decreases with the exclusion of smaller lesions, classification was also performed with a subsampled number of patients (=number of patients with volumes >133.04 mL) for the different volume ranges. This procedure was performed to identify if the number of training samples had an impact on classification accuracy.

2.10. Description of classification process

As the range of radiomic features differs widely from feature to feature, all features were z-transformed in order to normalize feature ranges before the start of the classification process. To guarantee that in the training data the number of positives and negatives outcomes (alive/dead after one year) are equally distributed, the under- represented class was up sampled using the Synthetic Minority Oversampling Technique (SMOTE) [35]. The testing was performed on the original data distributions without any over- or undersampling. For comparison, the minority class was also randomly oversampled. In order to assess the clinical value of the classifier on independent test datasets, stratified cross validation with 5 folds and 10 repetitions was performed. Hereby, 80% of the patients were used as training and 20% were used as testing dataset. This procedure was repeated until each data part served once as testing set.

3. Results

3.1. Features remaining after feature selection process

Step 1: Randomized voxel assignment

The number of features describing actual texture (i.e. features outside the random range) depended on the volume range included in the analysis: While 19 features were found to be outside the random range when including the whole dataset, the number of features increased to 27 when excluding lesions smaller than 3.2 mL. Eliminating lesions smaller than 11.48 mL, 33.10 mL, 45 mL, and 103.9 mL, led to 29,

45, 45, and 53 features describing actual textural information. All features outside the random range and their correlation coefficients with conventional metrics are listed in supplemental Table S2-S7 for the different volume ranges.

Step 2: Correlation with conventional metrics

The tumor volume had also an effect on the correlation with conventional metrics and therefore on the additional value of the features: The larger the volume range, the more features were found to be highly correlated with standard PET metrics. From the 19 features remaining after Step 1, only two features remained when eliminating features highly correlated with conventional metrics when analyzing the whole dataset. When excluding lesions smaller than 3.2 mL, four of the 27 features remained were not highly correlated with MATV, SUV_{MEAN}, and SUV_{PEAK}. The number of remaining features increased to 13 out of 29, 25 out of 45, 26 out of 45, and 33 out of 53 features.

The behavior of two features as a function of tumor volume and SUV_{MEAN} are displayed in Figure 2 a and b. As illustrated, for smaller lesions the original features are inside the random range and therefore not describing actual textural information. With increasing volume the features are found to be outside the proposed random range. The correlation with MATV is very strong for smaller lesions but decreases for larger lesions. While yielding a low correlation with MATV, some features result in a very strong correlation with SUV_{MEAN} as illustrated in Figure 2 b.

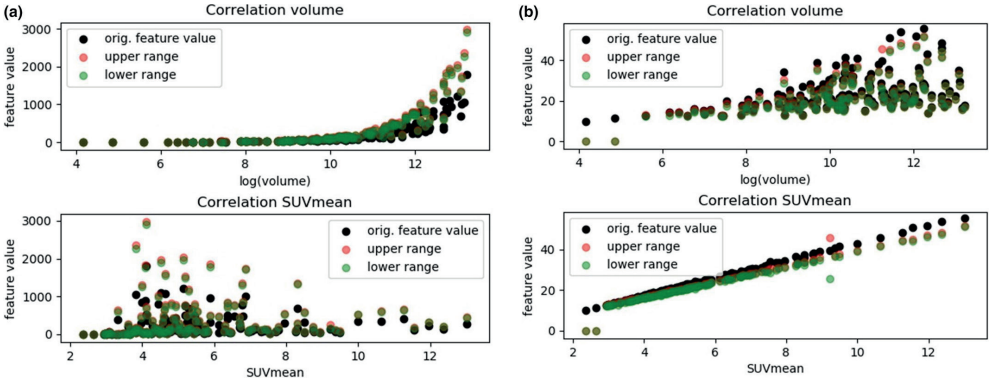


Figure 2. Illustration of the correlation between features and MATV or SUV_{MEAN}: Left: the feature Grey level non uniformity GLSZM2D yielding a correlation coefficient with MATV of 0.97 (upper row) and a correlation with SUV_{MEAN} of 0.34 (lower row). Right: The feature joint average GLCM 3D avg yielding a correlation of 0.45 with MATV (upper row) and a correlation of 0.99 with SUV_{MEAN}.

Step 3: Mutual correlation between features

When identifying those features yielding complementary information, i.e. when eliminating features highly correlated between each other, one feature remained when including the whole dataset as well as lesions larger than 3.2 mL. 2, 9, 11, and 15 features remained when excluding volumes less than 11.48, 33.1, 45, and 103.9 mL, respectively.

The described increase in number of selected features for each step of the feature selection procedure is illustrated in Figure 3. All features remaining after Step 3 as well as their correlation coefficients with the conventional metrics are displayed in Table 2.

Table 2: Selected features for different volume ranges

Feature name	Selected by which volume range
Grey_level_non_uniformity (GLCM, 2Davg)	all volume ranges
Grey_level_non_uniformity (GLCM, 2Dmrg)	all volume ranges >=3.2 mL
long_runs_emphasis (GLRLM, 2Davg)	all volume ranges >=11.4 mL
Run_percentage (GLRLM 2DWmrg)	all volume ranges >=33.4 mL
Run_length_variance (GLRLM 2Dvmrg)	all volume ranges >=33.4 mL
Grey_level_non_uniformity (GLRLM, 3Davg)	all volume ranges >=33.4 mL
Grey_level_non_uniformity (GLDZM 2Davg)	all volume ranges >=33.4 mL
Zone_percentage (GLDZM 2Davg)	all volume ranges >=33.4 mL
Grey_level_non_uniformity (GLDZM 3D)	all volume ranges >=33.4 mL
Zone_distance_non_uniformity (GLDZM 2Davg)	all volume ranges >=45 mL
Zone_distance_non_uniformity (GLDZM 2Dmrg)	all volume ranges >=45 mL
coarseness (NGLDM 2Dmrg)	all volume ranges >=103.9 mL
small_distance_emphasis (GLDZM 2Davg)	all volume ranges >=103.9 mL
Dependence_count_non_uniformity (NGTDM 2Dmrg)	all volume ranges >=103.9 mL
Dependence_count_entropy (NGTDM 2Davg)	all volume ranges >=103.9 mL

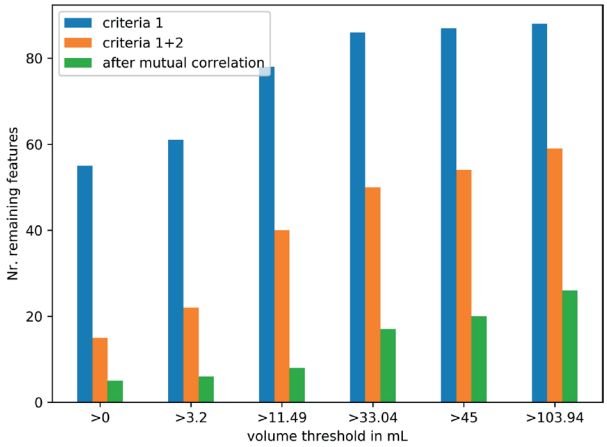


Figure 3. Number of features outside the random range (criteria 1) and not correlated with conventional metrics (criteria 1+2), and number of features after eliminating features highly correlated between each other (after mutual correlation) for the different volume ranges (left: all volumes are included, from left to right: more and more smaller lesions are excluded from the analysis until on the very right only tumors with a MATV > 103944 are left)

3.2. Features selected by automatic feature selection methods

The features that were identified by the feature selection method RELIEF are listed in Table 3. When all tumor volumes were included in the feature selection, all features selected by RELIEF were either highly correlated with conventional PET metrics or were not describing actual texture.

Table 3: Features selected by the RELIEF algorithm. Features that were also found by our procedure are displayed in bold, while features not reflecting actual texture are marked with *(NoTex)*, features highly correlated with volume or SUV_{MEAN} are marked with *(CORR)*.

Volume > 0 mL	Volume > 3.2 mL	Volume > 11.48 mL	Volume > 33.04 mL	Volume > 45 mL	Volume > 103.94 mL
(CORR) Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* Large zone high grey level emphasis (GLSZM 3D)
(NoTex) coarseness (NGTDM 2Dmrg)	*(CORR)* Run length variance (GLRLM 2Dvmrg)	*(NoTex)* Zone distance non uniformity normalized (GLDZM2Davg)	*(CORR)* Zone size entropy (GLSZM 2Davg)	*(CORR)* Zone size entropy2Davg	**Zone distance non uniformity normalized (GLDZM 2Davg)**
(NoTex) coarseness (NGTDM 3D)	*(NoTex)* long runs emphasis (GLRLM 2Davg)	difference variance (GLCM 2Dvmrg)	**Zone distance non uniformity normalized GLDZM2Davg**	*(CORR)* Dependence count entropy (NGTDM 2Davg)	*(CORR)* Zone size entropy (GLSZM 2Davg)
(NoTex) difference variance (GLCM 2Dvmrg)	*(NoTex)* Zone distance non uniformity normalized (GLDZM2Davg)	*(NoTex)* difference variance (GLCM 2Dmrg)	*(CORR)* Dependence count entropy (NGTDM 2Davg)	**Zone distance non uniformity normalized (GLDZM 2Davg)**	**Run length variance (GLRLM 2Dvmrg)**
(NoTex) difference variance (GLCM 2Dmrg)	*(NoTex)* joint maximum (GLCM 2Davg)	*(NoTex)* joint maximum (GLCM 2Davg)	**Run length variance (GLRLM 2Dvmrg)**	*(NoTex)* joint maximum (GLCM 2Davg)	**Grey level non uniformity (GLSZM 3D)**
(NoTex) joint maximum (GLCM 2Dmrg)	**Grey level non uniformity (GLCM 2Davg)**	SUV_{PEAK}	*(NoTex)* joint maximum (GLCM)2Davg	*(NoTex)* joint maximum (GLCM 2Dmrg)	**Grey level non uniformity (GLDZM 3D)**
(NoTex) joint maximum (GLCM 2Davg)	*(NoTex)* joint maximum (GLCM 2Dmrg)	*(NoTex)* small distance emphasis (GLDZM 2Davg)	*(NoTex)* joint maximum (GLCM 2Dmrg)	**Run length variance2Dvmrg**	**Grey level non uniformity (GLCM 3Dmrg)**

Volume > 0 mL	Volume > 3.2 mL	Volume > 11.48 mL	Volume > 33.04 mL	Volume > 45 mL	Volume > 103.94 mL
(CORR) contrast (NGTDM 2Dmrg)	**Grey level non uniformity (GLCM 3Dmrg)**	*(NoTex)* joint maximum (GLCM 2Dmrg)	*(NoTex)* Dependence count non uniformity (NGTDM 2Davg)	*(NoTex)* small distance emphasis (GLDZM 2Davg)	**Grey level non uniformity(GLCM 2Dmrg)**
(CORR) contrast (NGTDM 2Davg)	**Grey level non uniformity (GLCM 2Dmrg)**	*(NoTex)* Dependence count non uniformity 2Davg	**difference variance (GLCM 2Dvmrg)**	**coarseness (NGLDM 2Dmrg)**	**long runs emphasis (GLRLM 2Davg)**
(CORR) Low dependence high grey level emphasis (NGTDM 2Dmrg)	small distance emphasis (GLDZM 2Davg) *(CORR)*	Zone size entropy (GLSZM 2Davg) (CORR)	**long runs emphasis (GLRLM 2Davg)**	**long runs emphasis (GLRLM 2Davg)**	**Run length non uniformity (2DWmrg)**

When excluding smaller lesions, some features were selected by both RELIEF and our proposed selection procedure. When only including very large lesions (i.e. with a volume > 103.94) eight of ten features selected by RELIEF were overlapping with the features found by our procedure. The other two features did not reflect actual texture for the majority (more than 95%) of the patients.

Features selected by the MRMR algorithm are listed in Table 4, respectively. Similarly to the RELIEF feature selection, when including the whole dataset only features eliminated by our procedure were selected (leading to no overlap in selected features). By excluding smaller lesions, more and more features selected by MRMR were also selected by our procedure. A maximum of six overlapping features was found when only including very large lesions. Some features were selected by RELIEF as most representative even though they were identified to not reflect actual texture.

Table 4: Features selected by the MRMR algorithm for the different volume ranges. Features that were also found by our procedure are displayed in bold, while features not reflecting actual texture are marked with *(NoTex)*, features highly correlated with volume or SUV_{MEAN} are marked with *(CORR)*. Features found in previous studies

Volume > 0 mL	Volume > 3.2 mL	Volume > 11.48 mL	Volume > 33.04 mL	Volume > 45 mL	Volume > 103.94 mL
(NoTex) Grey level non uniformity (GLSZM 2Davg)	**Grey level non uniformity (GLSZM2Davg)**	*(CORR)* Short run low grey level emphasis (GLRLM 3Davg)	*(CORR)* Short run low grey level emphasis (GLRLM 3Davg)	**long runs emphasis (GLRLM 2Davg)**	**Zone distance non uniformity (GLDZM 2Davg)**
(NoTex) long runs emphasis (GLRLM 2Davg)	*(NoTex)* long runs emphasis (GLRM 2Davg)	*(NoTex)* Zone distance non uniformity (GLDZM 2Davg)	**Grey level non uniformity (GLSZM 3D)**	**Zone distance non uniformity (GLDZM 2Davg)**	**Grey level non uniformity (2DWmrg)**
(NoTex) Zone distance non uniformity normalized (GLDZM 2Davg)	*(NoTex)* Zone distance non uniformity normalized (GLDZM 2Davg)	*(NoTex)* long runs emphasis (GLRLM 2Davg)	**long runs emphasis (GLRLM 2Davg)**	*(NoTex)* contrast (NGTDM 2Dmrg)	**Grey level non uniformity (GLSZM 3D)**
(CORR) Grey level non uniformity (3Dmrg)	*(CORR)* Grey level non uniformity (3Dmr)g	**Grey level non uniformity (GLCM 2Dvmrg)**	*(NoTex)* Run length non uniformity (GLRLM 2Davg)	*(NoTex)* joint maximum (GLCM 2Dmrg)	**Run percentage (GLRLM 2Davg)**
(NoTex) contrast (NGTDM 2Dmrg)	*(NoTex)* contrast (NGTDM 2Dmrg)	*(NoTex)* small distance emphasis (GLDZM 2Davg)	*(NoTex)* contrast (NGTDM 2Dmrg)	**Grey level non uniformity (GLSZM 2Davg)**	**Grey level non uniformity (GLDZM 2Davg)**
(NoTex) Morans I	*(CORR)* Short run low grey level emphasis (GLRLM 3Davg)	*(NoTex)* contrast (NGTDM 2Dmrg)	**Grey level non uniformity (GLSZM 2Davg)**	*(NoTex)* Zone size entropy (GLSZM 2Dmrg)	*(CORR)* Run length non uniformity (GLRLM 2Davg)
(CORR) Short run low grey level emphasis (GLRLM 3Davg)	*(NoTex)* Run length non uniformity (GLRLM 2Dvmrg)	*(NoTex)* Small distance emphasis (GLDZM 2Davg)	*(NoTex)* joint maximum (GLCM 2Dmrg)	*(NoTex)* Large zone high grey level emphasis (GLSZM 3D)	*(NoTex)* Large zone high grey level emphasis (GLSZM 3D)
(NoTex) Run length non uniformity (GLRLM 2Dvmrg)	*(NoTex)* Large zone high grey level emphasis (GLSZM 3D)	*(NoTex)* Large zone high grey level emphasis (GLSZM 3D)	*(NoTex)* Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* coarseness (NGTDM 3D)	*(CORR)* Idifference variance (GLCM 2Dvmrg)

Volume > 0 mL	Volume > 3.2 mL	Volume > 11.48 mL	Volume > 33.04 mL	Volume > 45 mL	Volume > 103.94 mL
(NoTex) Large zone high grey level emphasis (GLSZM 3D)	*(CORR)* difference variance (GLCM 2Dvmrg)	*(CORR)* difference variance (GLCM 2Dvmrg)	*(CORR)* difference variance (GLCM 2Dvmrg)	*(CORR)* coarseness (NGTDM 3D)	*(NoTex)* joint maximum (GLCM 2Davg)
(CORR) difference variance (GLCM 2Dvmrg)	*(NoTex)* joint maximum (GLCM 2Davg)	*(NoTex)* joint maximum (GLCM 2Davg)	**Zone distance non uniformity normalized (GLDZM 2Davg)**	**Zone distance non uniformity normalized (GLDZM 2Davg)**	**Zone distance non uniformity (GLDZM 2Dmrg)**

When analyzing features identified to have clinical value in previous studies, the feature busyness (NGTDM) was found to be non-reproducible and was therefore a priori excluded from the analysis. None of the remaining features was passing Step 1 of the feature selection procedure when analyzing the whole dataset. When excluding patients with lesions smaller than 33.4 mL, coarseness (NGTDM) and zone percentage (GLSZM) were fulfilling the requirements of Step 1, i.e. yielded actual textural information. Coarseness correlated with MATV showing a correlation coefficient of 0.96 even for bigger lesions. Zone percentage yielded a lower correlation to the conventional metrics for larger volumes (MATV > 33.4 mL) and was the only feature that passed all steps. Therefore, this feature was checked for its possible clinical value in the present dataset.

3.3. Clinical value of selected features

Model 1: The clinical value of the selected features depended on the volume range included in the classification process. Some features were e.g. selected to describe actual texture for lesions with volumes above 33.4 mL but did not yield any clinical value for this volume range (mean AUC around 0.55). However, when including also smaller lesions, i.e. when performing the classification process for the whole dataset, their clinical value was comparable to the value of MATV. Nevertheless, in these cases, these features were also highly correlated to MATV (Figure 4).

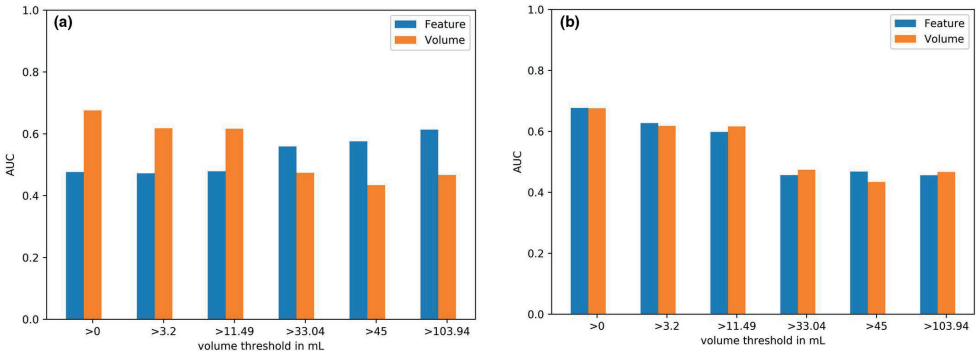


Figure 4. Model 1: Mean cross-validation AUC of : (a) the feature run percentage which shows an increasing AUC with larger volumes included in the analysis; (b) the feature run length non uniformity results in a reasonable accuracy when including the whole dataset, while the accuracy is decreasing with decreasing volume range and decreasing correlation to volume.

Features that were found to be highly correlated to SUV_{MEAN} or SUV_{PEAK} did not yield clinical value (mean AUC around 0.5) when used in Model 1. SUV_{MEAN} and SUV_{PEAK} yielded a comparable (low) accuracy when used for classification and yielded therefore no clinical value in the example dataset.

For the rest of the features, the prediction accuracy depended on the feature: The features contrast (NGTDM 2Dmrg) and Zone size entropy (GLSZM 2Davg) were fulfilling all described characteristics but were not yielding a clinical value for 1-year survival in Model 1, i.e. resulted in a mean cross-validation AUC of around 0.5. The low accuracy was observed for all volume ranges.

The features run length variance (GLRLM 2Dvmrg), run percentage (GLRLM 2Dmrg), difference entropy (GLCM 2Dmrg), Grey level non uniformity (GLRLM 2Davg), long runs emphasis (GLRLM 2Davg), and zone percentage (GLDZM 2Davg) yielded an accuracy around 0.5 when also smaller lesions were included in the classification process. The accuracy increased when only including larger volumes

(>33.4 mL) for which they were also found to match all described criteria (Figure 4). For these larger volumes, they outperformed MATV in terms of accuracy (mean AUC 0.6 vs. 0.48).

Model 2: For larger lesions, the features leading to an accuracy improvement in Model 1 led also to an improvement when used together with MATV in the logistic regression model (Figure 5).

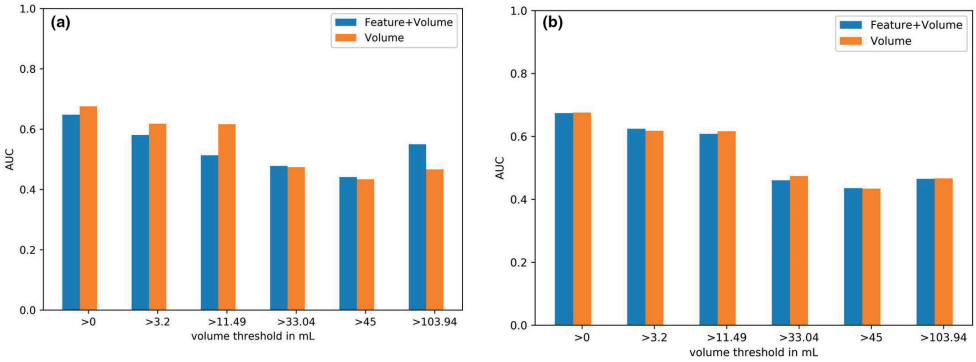


Figure 5. Model 2: Mean cross-validation AUC of : (a) the feature run percentage, (b) the feature run length non uniformity used together with MATV in the logistic regression model

Model 3: For the features outperforming volume for larger lesions, also the combined ROC curves yielded higher AUCs than volume. The combined prediction led to an increase in AUC from 72.5% for volume only to up to 76% for the combined prognosis (Figure 6). For the feature Grey level non uniformity (GLRLM 2Davg) an increase in AUC was already observed when using the feature for lesions with volumes above 11.4 mL for which the feature also reflected actual heterogeneity. All AUCs of Model 1, 2, and 3 are listed in supplemental Table S8.

The oversampling technique had only minimal impact on the accuracy of the classifier as illustrated in the supplemental material. The accuracy when using all patients or a subsample of patients was for all volume ranges comparable, indicating that the difference in accuracy for the different data subsets (by excluding more and more patients) is not due to the different size of training sets.

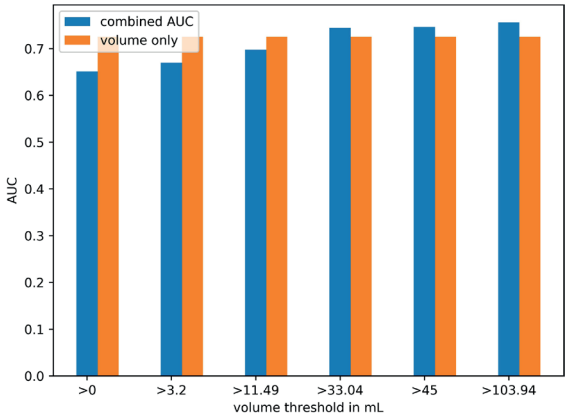


Figure 6. Model 3: Combined AUCs: for lesions above threshold, the feature run percentage is used for prediction, while for features below the threshold, volume is used for prediction.

4. Discussion

In this study, we proposed a feature selection procedure to identify which reproducible radiomic features are at the same time describing the tumor texture beyond randomness and yielding additional value to conventional PET metrics. The impact of lesion size on fulfilling these requirements was assessed.

Our results suggest that the number of plausible radiomic features depends on the lesion size. For smaller lesions, the majority of textural features did not describe actual tumor texture (did not yield feature values significant different from randomness). This result indicates that textural features extracted from small lesions should be handled with care. As small lesions consist only of a few voxels, randomly shuffled and original image are similar and result therefore in similar radiomic features. Hence, it is questionable how accurate a textural feature extracted from a small lesion can reflect underlying lesion heterogeneity. For each textural feature family, a matrix is composed describing e.g. how often a discretized intensity value is appearing consecutively in the VOI or how many connected voxels yield the same intensity value [36]. These textural features might result in a certain value due to the small expansion of the lesion but not due to tumor heterogeneity. The low number of selected radiomic features might be related to the low spatial resolution of PET images which has a high impact on smaller lesions. The low resolution is a general drawback of PET images. Using modern scanners yielding a better spatial resolution might lead to more radiomic features reflecting actual texture. Therefore, if a radiomic model including small

lesions is used, it should be carefully checked if a textural feature is describing the underlying tumor characteristics beyond randomness. However, conventional PET metrics such as MATV and SUVmax could be used for the analysis of smaller lesions, while for larger lesions textural features might be used, as was shown with our combined model 3. Additionally, features extracted from smaller tumors showed also a higher correlation with conventional metrics and might not yield additional value what is in line with the results of Hatt et *al*. and Brooks et *al*. [28,37].

As larger lesions yield more complex anatomical properties, it is natural that for larger lesions more features were selected. However, also for larger lesions, around one third of the features failed to describe texture different from randomness and contain therefore no relevant information about tumor heterogeneity. Especially for small datasets, it can happen that one of these features results by chance in a high correlation with the outcome and results therefore in a good AUC. Thus, a careful check if there is a relationship between a predictive feature and the heterogeneity observed in the medical image is necessary. Moreover, the use of cross-validation and an external testing set is essential in order to assess the value of a feature in independent test scenarios and lowers the risk of identifying a feature that does not describe relevant information [13,15,38].

Due to image noise, partial volume effect, and other intrinsic factors, the heterogeneity displayed in a medical image does not reflect the real underlying tumor heterogeneity accurately as was shown previously for MR and PET images [39], [40]. However, several studies demonstrated that a tumor displayed heterogeneously in a PET image is an indicator for a lower survival chance and higher treatment resistance [10,41]. The features identified by the proposed selection procedure are not describing the real tumor heterogeneity, but the heterogeneity observed in the PET image and thus are reflecting the PET tumor phenotype.

The clinical value of the radiomic features selected by our procedure was in our dataset relatively low. Some features resulted in a reasonable mean cross-validation accuracy when analyzing the whole dataset (i.e. including also smaller volumes), but were in this case also highly correlated to conventional PET metrics. In addition, in this case, no feature yielded additional value to MATV. This fact indicates that the accuracy of these features when applied to the whole dataset is due to the high correlation with volume. With a decrease of correlation with volume (by excluding smaller lesions), the accuracy also decreases what is in line with previous studies reporting a low accuracy when eliminating features that were highly correlated to MATV [14,42].

However, when excluding smaller lesions, some features yielded complementary value to MATV, improving the accuracy of around 4 - 10%. Moreover, using the combination of tumor volume as prognostic factor for smaller lesions and one selected features as prognostic factor for larger lesions led to an increase in AUC up to 76% when compared with using only volume as prognostic factor for the whole dataset (AUC 72.7%). This indicates, that the selected features had additional value only for the larger lesions for which they were also selected. Therefore, it is worthwhile to explore the use of different prognostic imaging biomarkers (e.g. MATV versus a radiomic feature) for different lesion size ranges in future studies.

Brooks et *al*. indicated that the correlation with MATV decreases with lesions yielding a MATV above 45 mL and radiomic analysis including smaller lesions might be questionable [28]. However, our findings demonstrate that the most adequate threshold is feature dependent what is in line with the findings of Hatt et *al*. [43]. Additionally, Hatt et *al*. indicated that a more appropriate volume threshold might be 10 mL as this threshold already led to a lower correlation with MATV for some radiomic features. Our results support this finding. Some radiomic features yielded relevant information and were at the same time not highly correlated to conventional metrics when excluding lesions with a volume below 11.4 mL.

One source of uncertainty in the quantitative analysis of PET images results from uncertainties in tumor delineation. As manual segmentations suffer from a high inter- and intra-observer variability, a semi- or ideally fully automatic segmentation method should preferably be used. Yet, segmentation of tumors in PET images still requires supervision and sometimes manual correction [44], resulting in intra- and inter-observer variability. It is, however, highly recommended to keep the level of user-interaction as low as possible, as we recently showed in [45]. By using a semi-automated segmentation approach and workflow, which required delineation adjustment in a few cases (when the tumor was located to another high uptake region), we tried to limit the amount of user-interaction and we found that this reduces the observer variability as much as possible.

While previous studies concentrated on the correlation of radiomic features with volume, we also investigated the correlation with SUV_{MEAN} and SUVPEAK and found that a large number of radiomic features yields a high correlation to these two parameters as well. Hence, even though some features might yield additional value to MATV, they yield no additional value to other conventional PET metrics. Therefore, the additional value of a feature to all conventional PET metrics (MATV, SUV_{MEAN}, SUV_{PEAK}) should be investigated and features resulting in a very strong correlation with conventional metrics should be discarded from the analysis.

As automatic feature selection methods are data-driven, the two automatic feature selection methods used for comparison in our study could not identify features describing non-relevant or random texture. As our feature selection method does not only check the value of a feature but also its plausibility, it adds an important aspect to automatic feature selection methods. Therefore, it could be used in addition to automatic methods as it can identify features which are found 'by chance' to yield clinical value and can therefore help to identify false- positive findings.

The selected features as well as the volume thresholds used in this study should be validated in a larger, independent patient cohort and are limited to the NSCLC dataset included in this study. Moreover, the used thresholds (i.e. including only features reflecting actual texture for more than 90% of the patients) were chosen as example. In other datasets, these thresholds could be adapted as deemed appropriate or optimal and a more liberal threshold, thus resulting in more selected features, may be feasible when larger datasets are available.

Moreover, the retrospective nature of our study is a clear limitation and our findings should be validated in a larger, prospective cohort. However, the aim of our study was to demonstrate our proposed radiomic feature selection procedure on a clinical dataset. We showed that, especially for smaller lesions, a large number of radiomic features might not reflect actual texture information. Our study demonstrates the need for a more thoughtful performed radiomic analysis and feature selection procedure. The combination of using conventional PET metrics for smaller and textural features for larger lesions might be a solution that should be explored in other datasets. Only with good statistical methods (i.e. using cross-validation), the use of external testing datasets, as well as a careful check which textural features are repeatable, reproducible, and contain relevant textural information, a radiomic study becomes transferable to other datasets and opens the way for a clinical implementation of radiomics.

5. Conclusion

In this study, we proposed a feature selection procedure which identifies reproducible textural that are (a) describing relevant texture and (b) are not highly correlated with conventional PET metrics. Our results show that the larger the lesions the more features are selected. Our results illustrate that when performing textural analysis for small lesions (<11.4 mL), it should be carefully investigated if a textural feature reflects relevant tumor characteristics and yields additional

value to conventional metrics. Using tumor volume as prognostic value for smaller lesions and the identified textural features as prognostic value for larger lesions, yields promising value for a more accurate and reliable radiomic analysis.

Supplementary materials

Financial support: This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This study was financed by the Dutch Cancer Society, POINTING project, grant 10034.

Acknowledgments: We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

Disclosure of Conflicts of Interest: The authors have no relevant conflicts of interest to disclose.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

[1] J. F. Vansteenkiste *et al.*, "Prognostic Importance of the Standardized Uptake Value on 18 F-Fluoro-2- Deoxy-Glucose–Positron Emission Tomography Scan in Non–Small-Cell Lung Cancer: An Analysis of 125 Cases," *Journal of Clinical Oncology*, vol. 17, no. 10, pp. 3201–3206, Oct. 1999, doi: 10.1200/JCO.1999.17.10.3201.

[2] S. Hammerschmidt and H. Wirtz, "Lung Cancer," *Deutsches Aerzteblatt Online*, Dec. 2009, doi: 10.3238/arztebl.2009.0809.

[3] S. Larson, "Tumor Treatment Response Based on Visual and Quantitative Changes in Global Tumor Glycolysis Using PET-FDG Imaging The Visual Response Score and the Change in Total Lesion Glycolysis," *Clinical Positron Imaging*, vol. 2, no. 3, pp. 159–171, May 1999, doi: 10.1016/S1095-0397(99)00016-3.

[4] K. Vanhove *et al.*, "Prognostic value of total lesion glycolysis and metabolic active tumor volume in non- small cell lung cancer," *Cancer Treatment and Research Communications*, vol. 15, pp. 7–12, 2018, doi: 10.1016/j.ctarc.2017.11.005.

[5] T. A. Data, R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics : Images Are More than," vol. 278, no. 2, 2016.

[6] F. Tixier *et al.*, "Visual Versus Quantitative Assessment of Intratumor 18F-FDG PET Uptake Heterogeneity: Prognostic Value in Non-Small Cell Lung Cancer," *Journal of Nuclear Medicine*, vol. 55, no. 8, pp. 1235–1241, 2014, doi: 10.2967/jnumed.113.133389.

[7] D.-H. Kim *et al.*, "Prognostic Significance of Intratumoral Metabolic Heterogeneity on 18F-FDG PET/CT in Pathological N0 Non–Small Cell Lung Cancer," *Clinical Nuclear Medicine*, vol. 40, no. 9, pp. 708–714, Sep. 2015, doi: 10.1097/RLU.0000000000000867.

[8] M.-C. Desseroit *et al.*, "Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non–Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort," *Journal of Nuclear Medicine*, vol. 58, no. 3, pp. 406–411, 2017, doi: 10.2967/jnumed.116.180919.

[9] V. Kumar *et al.*, "Radiomics : the process and the challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012, doi: 10.1016/j.mri.2012.06.010.

[10] M. Hatt *et al.*, "Characterization of PET / CT images using texture analysis : the past , the present ... any future ?," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 44, pp. 151–165, 2017, doi: 10.1007/s00259-016-3427-0.

[11] E. Pfaehler *et al.*, "Repeatability of 18 F-FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method," *Medical Physics*, Dec. 2018, doi: 10.1002/mp.13322.

[12] I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, and A. Bitarafan-Rajabi, "The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies," *European Radiology*, vol. 27, no. 11, pp. 4498–4509, Nov. 2017, doi: 10.1007/s00330-017- 4859-z.

[13] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," *International Journal of Radiation Oncology*Biology*Physics*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018, doi: 10.1016/j.ijrobp.2018.05.053.

[14] M. L. Welch *et al.*, "Vulnerabilities of radiomic signature development: The need for safeguards,"*Radiotherapy and Oncology*, vol. 130, pp. 2–9, Jan. 2019, doi: 10.1016/j.radonc.2018.10.027.

[15] I. Buvat and F. Orlhac, "The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results," *Journal of Nuclear Medicine*, vol. 60, no. 11, pp. 1543–1544, Nov. 2019, doi: 10.2967/jnumed.119.235325.

[16] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review," *PLOS ONE*, vol. 10, no. 5, p. e0124165, May 2015, doi: 10.1371/journal.pone.0124165.

[17] M. Hatt *et al.*, "18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort," *Journal of Nuclear Medicine*, vol. 56, no. 1, pp. 38–44, Jan. 2015, doi: 10.2967/jnumed.114.144055.

[18] R. Boellaard, "Quantitative oncology molecular analysis suite: ACCURATE.," 2018.

[19] G. D. Kolinger *et al.*, "Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients," *EJNMMI Research*, vol. 9, no. 1, p. 14, Dec. 2019, doi: 10.1186/s13550-019-0481-1.

[20] G. M. Kramer *et al.*, "Repeatability of Quantitative Whole-Body 18F-FDG PET/CT Uptake Measures as Function of Uptake Interval and Lesion Selection in Non-Small Cell Lung Cancer Patients," *Journal of Nuclear Medicine*, vol. 57, no. 9, pp. 1343–1349, 2016, doi: 10.2967/jnumed.115.170225.

[21] A. Zwanenburg *et al.*, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology*, p. 191145, Mar. 2020, doi: 10.1148/radiol.2020191145.

[22] E. Pfaehler, A. Zwanenburg, J. R. de Jong, and R. Boellaard, "RaCaT: An open source and easy to use radiomics calculator tool," *PLOS ONE*, vol. 14, no. 2, p. e0212223, Feb. 2019, doi: 10.1371/journal.pone.0212223.

[23] F. H. P. van Velden *et al.*, "Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG- PET/CT Studies: Impact of Reconstruction and Delineation," *Molecular Imaging and Biology*, vol. 18, no. 5, pp. 788–795, Oct. 2016, doi: 10.1007/s11307-016-0940-2.

[24] R. T. H. Leijenaar *et al.*, "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis," *Scientific Reports*, vol. 5, no. 1, p. 11075, Sep. 2015, doi: 10.1038/srep11075.

[25] F. Orlhac, M. Soussan, K. Chouahnia, E. Martinod, and I. Buvat, "18F-FDG PET-Derived Textural Indices Reflect Tissue-Specific Uptake Pattern in Non-Small Cell Lung Cancer," *PLOS ONE*, vol. 10, no. 12, p. e0145063, Dec. 2015, doi: 10.1371/journal.pone.0145063.

[26] E. Pfaehler *et al.*, "Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts," *Journal of Nuclear Medicine*, vol. 61, no. 3, pp. 469–476, Mar. 2020, doi: 10.2967/jnumed.119.229724.

[27] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research.," *Malawi medical journal : the journal of Medical Association of Malawi*, vol. 24, no. 3, pp. 69–71, Sep. 2012.

[28] F. J. Brooks and P. W. Grigsby, "The Effect of Small Tumor Volumes on Studies of Intratumoral Heterogeneity of Tracer Uptake," *Journal of Nuclear Medicine*, vol. 55, no. 1, pp. 37–42, Jan. 2014, doi: 10.2967/jnumed.112.116715.

[29] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[30] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning Proceedings 1992*, Elsevier, 1992, pp. 249–256. doi: 10.1016/B978-1-55860-247-2.50037-1.

[31] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "mRMRe: an R package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365– 2368, Sep. 2013, doi: 10.1093/bioinformatics/btt383.

[32] M.-C. Desseroit *et al.*, "Development of a nomogram combining clinical staging with 18F-FDG PET/CT image features in non-small-cell lung cancer stage I–III," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 43, no. 8, pp. 1477–1485, Jul. 2016, doi: 10.1007/s00259-016-3325-5.

[33] G. J. R. Cook *et al.*, "Are Pretreatment 18F-FDG PET Tumor Textural Features in Non-Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy?," *Journal of Nuclear Medicine*, vol. 54, no. 1, pp. 19–26, Jan. 2013, doi: 10.2967/jnumed.112.107375.

[34] H. K. Ahn, H. Lee, S. G. Kim, and S. H. Hyun, "Pre-treatment 18F-FDG PET-based radiomics predict survival in resected non-small cell lung cancer," *Clinical Radiology*, vol. 74, no. 6, pp. 467–473, Jun. 2019, doi: 10.1016/j.crad.2019.02.008.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over- sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[36] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, and for the I. B. S. Initiative, "Image biomarker standardisation initiative," 2016, doi: 10.17195/candat.2016.08.1.

[37] M. Hatt, F. Tixier, C. Cheze Le Rest, O. Pradier, and D. Visvikis, "Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 40, no. 11, pp. 1662–1671, Oct. 2013, doi: 10.1007/s00259-013-2486-8.

[38] A. Zwanenburg and S. Löck, "Why validation of prognostic models matters?," *Radiotherapy and Oncology*, vol. 127, no. 3, pp. 370–373, Jun. 2018, doi: 10.1016/j.radonc.2018.03.004.

[39] F. Yang, L. A. Young, and P. B. Johnson, "Quantitative radiomics: Validating image textural features for oncological PET in lung cancer," *Radiotherapy and Oncology*, vol. 129, no. 2, pp. 209–217, 2018, doi: 10.1016/j.radonc.2018.09.009.

[40] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification.," *Magnetic resonance imaging*, vol. 22, no. 1, pp. 81– 91, Jan. 2004, doi: 10.1016/j.mri.2003.09.001.

[41] C. Bailly *et al.*, "Exploring Tumor Heterogeneity Using PET Imaging: The Big Picture," *Cancers*, vol. 11, no. 9, p. 1282, Aug. 2019, doi: 10.3390/cancers11091282.

[42] A. Traverso *et al.*, "Machine learning helps identifying volume-confounding effects in radiomics,"*Physica Medica*, vol. 71, pp. 24–30, Mar. 2020, doi: 10.1016/j.ejmp.2020.02.010.

[43] M. Hatt *et al.*, "18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort," *Journal of Nuclear Medicine*, vol. 56, no. 1, pp. 38–44, 2015, doi: 10.2967/jnumed.114.144055.

[44] M. Hatt *et al.*, "The first MICCAI challenge on PET tumor segmentation," *Medical Image Analysis*, vol. 44, pp. 177–195, 2018, doi: 10.1016/j.media.2017.12.007.

[45] E. Pfaehler *et al.*, "PET segmentation of bulky tumors: Strategies and workflows to improve inter- observer variability," *PLOS ONE*, vol. 15, no. 3, p. e0230901, Mar. 2020, doi: 10.1371/journal.pone.0230901.

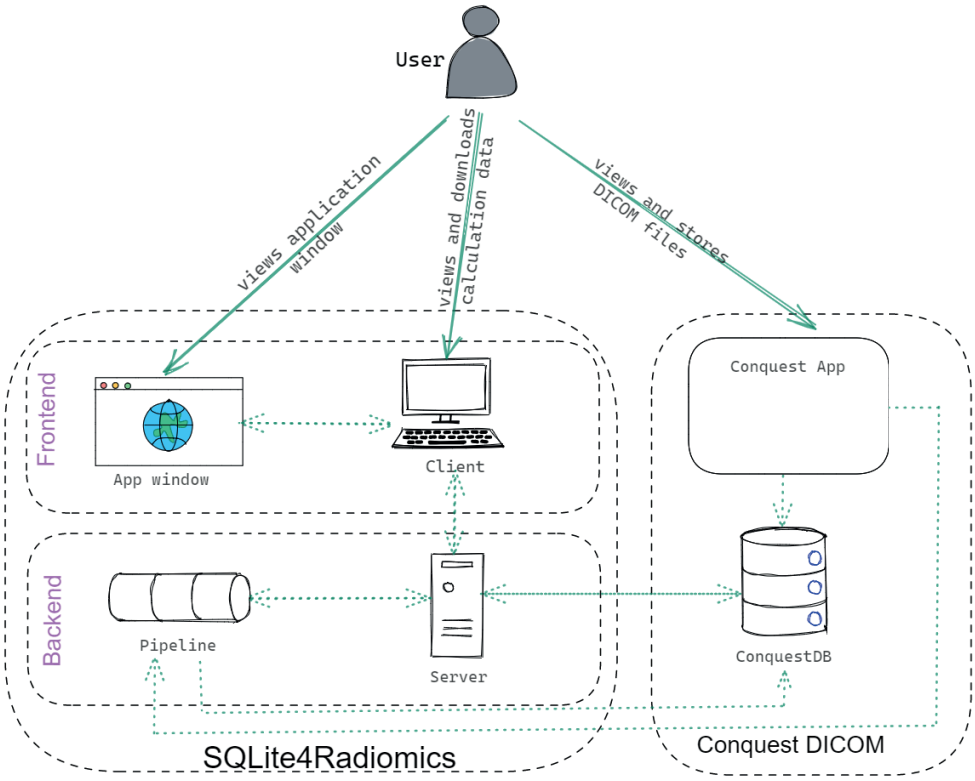
MISTER DATA
7
3
1
2
4
16
5
0
8
WATSON
ARTIFICIAL INTELLIGENCE

Chapter 7

Radiomics Integration into a Picture Archiving and Communication System (PACS)

Zhovannik, Ivan, Suraj Pai*, Talia A. da Silva Santos*, Lars LG van Driel*, Andre Dekker, Rianne Fijten, Alberto Traverso, Johan Bussink, and René Monshouwer. “Radiomics integration into a picture archiving and communication system.” *Physics and Imaging in Radiation Oncology* 20 (2021): 30-33.

Abstract



Radiomics is referred to as quantitative imaging of biomarkers used for clinical outcome prognosis or tumor characterization. In order to bridge radiomics and its clinical application, we aimed to build an integrated solution of radiomics extraction with an open-source Picture Archiving and Communication System (PACS). The integrated SQLite4Radiomics software was tested in three different imaging modalities and its performance was benchmarked in lung cancer open datasets RIDER and MMD with median extraction time of 10.7 (percentiles 25-75: 8.9-18.7) seconds per ROI in three different configurations.

1. Introduction

Radiological images have been acquired routinely for decades in the process of radiation treatment. Commonly used imaging modalities are (cone beam) computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). The aim of these imaging modalities is to give either anatomical (CT and MRI) or functional (MRI and PET) information.

Generally, the interpretation of medical images is performed by visual inspection. With the advent of advanced computer vision, it was hypothesized that algorithms could add quantitative and objective measurements to visual interpretation, as for example, in Computer-aided diagnosis (CAD) systems [1].

Recently, *radiomics*, where properties of the image, such as local textures, can be used as *imaging biomarkers* was proposed. Radiomics has been extensively applied to quantify biological properties of the tumor for better prognostication or treatment response predictions [2]. However, many studies have shown that radiomic values can be influenced by specific acquisition settings [3], leading to a poor generalizability of radiomic signatures on multi-center data.

Radiomic feature computations need to be combined with the corresponding image acquisition settings (Digital Imaging and Communications in Medicine, DICOM, metadata) so that the feature values can be harmonized against these settings using appropriately calibrated algorithms [3,4]. *Conquest* is an open-source Picture Archiving and Communication System (PACS) that stores DICOM metadata as database tables, opening the possibility of combining imaging data and metadata hosted on a PACS with radiomics. There are several radiomics extraction solutions available [5-8], however, from the best of our knowledge, current existing radiomics extraction software solutions do not support direct PACS integration and database-centered radiomic feature storage.

So far, the calculation of radiomic features required some knowledge of programming. As all the imaging data is already stored in a PACS, we viewed radiomics-PACS integration as a logical step to facilitate the introduction of radiomics in a clinical environment. We believe that due to the DICOM connectivity of Conquest, its radiomics-powered version can be used to form a pipeline within the hospital wide PACS system, enabling routine analysis of the data.

The *aim* of this study was to develop a solution for wider radiomics adoption in radiation oncology by integrating a publicly available library, *pyradiomics*, with the open-source *Conquest* DICOM PACS software. In this technical note, we described

the architecture, workflow, and benchmarking of an integration software for pyradiomics and Conquest DICOM. The code and the user manual are available open-source.

2. Materials and Methods

2.1. SQLite4Radiomics

Conquest (or *Conquest DICOM*) is open-source PACS software (https://ingenium.home.xs4all.nl/dicom.html) that stores imaging data as DICOM files and the DICOM metadata in a SQL database. It runs a DICOM server and has full DICOM functionality, which makes it suitable for integration in a radiation oncology department. *Pyradiomics* is an open-source Python package (https://github.com/Radiomics/pyradiomics) for radiomics extraction [5] from medical images as defined in the Image Biomarker Standardization Initiative (IBSI) [9]. Our proposed solution *SQLite4Radiomics* further broadens Conquest's functionality by integrating pyradiomics feature extraction into the PACS. The supplementary materials describe SQLite4Radiomics application customization, pipeline, graphical user interface (GUI) frontend and backend.

2.2. Case study: datasets

Open data from the Maastro LUNG1 cohort [10,11] was used to develop and test the software. These data included CT scans with manual delineations stored in the RTSTRUCT format. The LUNG1 dataset with the detailed cohort description is publicly available at the XNAT repository (https://xnat.bmia.nl/). Individual users collected their internal CT, MR, and PET data with RTSTRUCT delineations to test SQLite4Radiomics application in those modalities.

To benchmark SQLite4Radiomics performance, two open lung cancer cohorts were selected: Interobserver MMD PET-CT dataset and RIDER CT dataset, both available at the XNAT repository (https://xnat.bmia.nl/). The former consists of 22 unique PET-CT lung cancer images with multiple gross tumor volume delineations, and it was originally described and used in *Aerts et al* [10]. The latter, consists of 32 pairs of test-retest scans of lung cancer patients. The ROI delineations in both the datasets were provided as DICOM RTSTRUCT files suitable for SQLite4Radiomics. These data were previously used to benchmark the performance of the O-RAW software, which we compared SQLite4Radiomics performance to [6].

2.3. Case study: benchmarking and performance evaluation

The *MMD PET-CT* and *RIDER CT* datasets were used to evaluate the performance of SQLite4Radiomics. The benchmarking pyradiomics parameter file corresponds to the default parameter file of SQLite4Radiomics stored on GitHub. With this file, 107 features including shape, first order statistics, and texture (GLCM, GLRLM, GLDM, GLSZM, NGTDM) categories were extracted from each of two ROIs per DICOMSeries. For instance, an MMD patient instance contains two series (PET and CT) – for both of those two, we extracted radiomics from two ROIs – the total of four extractions per patient.

Three system configurations were used to evaluate the performance of SQLite4Radiomics including SQLite query, plastimatch conversion, pyradiomics feature extraction, feature storage in the Conquest database. The *first* system configuration was represented with a HP EliteDesk 800 G2 TWR workstation (Windows 7 Enterprise, 16 GB RAM, Processor Intel(R) Core(TM) i7-6700). The *second* configuration was represented with HP EliteBook 840 G4 laptop (Windows 10 Enterprise, 8 GB RAM, Processor Intel(R) Core(TM) i5-7200U). The *third* configuration was represented by a Lenovo ThinkPad L480 laptop (Windows 10 Pro, 16 GB RAM, Processor Intel(R) Core(TM) i5-8250U). All the three system configurations were benchmarked using Novabench System Benchmarking Software (version 4.0.9 – January 2021).

A total of 107 features were extracted from two ROIs per DICOMSeries (either CT or PET) with the default parameter and configuration files listed on SQLite4Radiomics repository. The extraction time in seconds per ROI was chosen as a performance benchmarking metric in our study. There is, however, the possibility to customize the SQLite4Radiomics performance evaluation by, for instance, extraction time per voxel of ROI – this might be beneficial when there is a high ROI volume variation in a dataset.

2.4. User Testing

In addition to the developers, two users independently tested the standalone pipeline application, while five users tested the GUI version. Two of the users are clinical physicists, one is a radiobiologist, and two are researchers. The user tests were performed by observation (a user performs a set of tasks, while a developer observes the process without interfering and takes notes) and in a remote fashion (a user freely tested and used SQLite4Radiomics tool in their own time). In both cases, feedback was given by the users at the end of each test.

3. Results

In system configuration I, it took 4057, 4074, and 4204 seconds to process the RIDER dataset with a median extraction time per ROI of 10.1 seconds and it took 1592, 1537, and 1547 seconds to run the MMD dataset with a median extraction time per ROI of 6.3 seconds. In system configuration II, it took 7492, 7527, and 8745 seconds to run the RIDER dataset with a median extraction time per ROI of 20.5 seconds and it took 3472, 3843, and 3727 seconds to run the MMD dataset with a median extraction time per ROI of 17.4 seconds. In system configuration III, it took 4687, 4817, and 4776 seconds to run the RIDER dataset with a median extraction time per ROI of 9.7 seconds and it took 1632, 1596, and 1696 seconds to run the MMD dataset with a median extraction time per ROI of 6.2 seconds. All three configurations resulted in the total of 18 successful runs of SQLite4Radiomics. The results are represented in Figure 1.

Each of the three configurations were benchmarked with the Novabench System Benchmarking Software. The configuration I received a total score of 1264 (CPU 920, RAM 257, Disk 87, GPU unavailable); the configuration II received 766 points (CPU 395, RAM 169, Disk 55, GPU 147), and the configuration III received 1330 (CPU 821, RAM 254, Disk 52, GPU 203). The GPU score was irrelevant for this specific study because the radiomics calculation and SQLite4Radiomics operation were CPU- based.

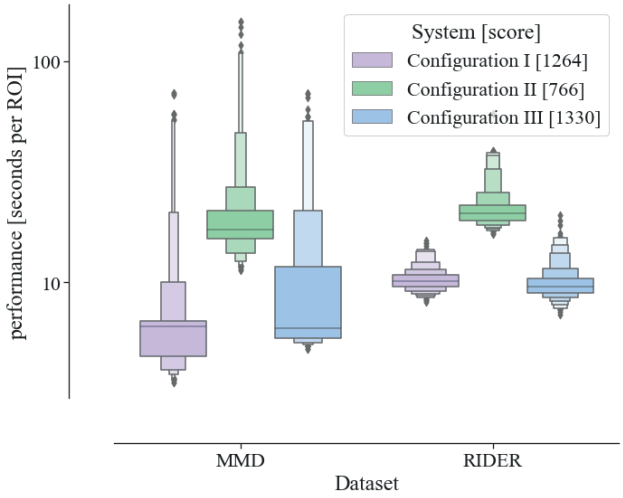


Figure 1 – performance letter-value plot (generalization of the boxplot [15])
The graph shows the performance in seconds per region-of-interest in both MMD PET-CT and RIDER CT datasets for the three system configurations we used to benchmark SQLite4Radiomics' performance in. The letter-value plot generalizes the classical boxplot by visualizing more percentiles (in %): [(3.125, 96.875), (6.25, 93.75), (12.5, 87.5), (25.0, 75.0)]. The box area is proportional to the percentage of covered data. Each system configuration contains its Novabench score for reference.

4. Discussion

Radiomics currently lacks integration in clinical pipelines. At the same time, we routinely generate, archive, and store images routinely in hospital PACS systems. In order to bridge radiomics and its clinical application, we presented SQLite4Radiomics – an integration software of pyradiomics and Conquest DICOM – two popular open-source tools of both worlds of radiomic analysis imaging and clinical imaging. SQLite4Radiomics can receive data through the standard DICOM frameworks present as a part of treatment planning workflow in radiation oncology.

We have built-in plastimatch (https://plastimatch.org/) conversion that removes the burden from the user regarding the conversion of the contour data of DICOM RTSTRUCT into binary mask files. The volumes (ROIs) selection is customizable, which allows for better match of local ROI labeling in a particular clinic. The calculated radiomic data is stored in a alongside the DICOM metadata and images. Therefore, the data can be easily coupled to statistical environments such as Python and R, which have database connection facilities. Due to the integration of the DICOM metadata and the radiomics output, the data can be combined to examine dependencies between the two [3,4].

Although SQLite4Radiomics allows for simpler radiomics extraction and storage, it also gives the users opportunity to investigate the relationship between a radiomic feature and image acquisition settings. We would like to encourage the radiomics community to use this opportunity to further improve the reporting quality on radiomics reproducibility [12]. In addition, our approach will lead to simpler integration of radiomics and conventional clinical variables, such as performance and tumor stage, into cancer prognostic models [13].

We found that the computer performance benchmarking score (e.g. with Novabench) may be a relevant predictor in SQLite4Radiomics time efficiency approximation as the score is based on the processor and RAM performance. The time efficiency of SQLite4Radiomics is comparable to that of pyradiomics and it outperforms O-RAW. The main difference in the execution times between O-RAW and SQLite4Radiomics may be due to O-RAW's RDF-conversion of radiomic data, which is not present in SQLite4Radiomics [6]. Yet the extraction can be performed without the user input if the RTSTRUCT trigger is enabled, which makes the access time to radiomic data comparable with executing an SQLite query only.

The radiomics extraction and storage process was automated using SQLIte4Radiomic tool and can be customized according to research or clinical

requirements. The software was developed with the possibility to be extended upon by the research community and is open-source. With regards to the future functionality, we see the following possible developments. Currently, SQLite4Radiomics is hosted on a local machine, therefore, an extension may include scaling up with authentication and proper online hosting. SQLite4Radiomics can possibly be extended by converting the SQLite radiomics tables to SPARQL RDF triples to match Radiomics Ontology [6]. This integration with RO and machine learning toolkits will allow to perform rapid learning to automatically re-adjust prognostic models whenever new data comes in, as was proposed by Deist et al [14].

Interestingly, SQLite4Radiomics is based on pyradiomics, which was not originally intended to mimic IBSI's image processing exactly. This causes a mismatch while executing feature extraction in the Lung Cancer CT phantom. This issue was addressed on pyradiomics github page (https://github.com/AIM-Harvard/pyradiomics/issues/498). For the purpose of matching the pre- processing of IBSI and pyradiomics, we listed IBSI-compliant pre-processing methods in the *IBSI- pyradiomics_discrepancies* branch.

SQLite4Radiomics reduces the entrance threshold for clinical researchers and makes radiomics extraction from organic: whenever new images arrive in a PACS, radiomic features can be extracted and stored as SQLite tables. Radiomics extraction is based on a popular open-source pyradiomics library and DICOM-RT conversion is performed by plastimatch – reliable open-source ITK software. SQLite4Radiomics makes radiomics data easy to store, query and combine with clinical data at any time with no need for additional wrappers.

Acknowledgements

The authors thank Paul Rijken, Guy Warmerdam, and Theo Lam for helping with SQLite4Radiomics quality assurance and for giving their feedback.

Conflict of interest statement

The authors have no conflicts of interest to declare.

References:

[1] K. Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput. Med. Imaging Graph.*, vol. 31, no. 4–5, 2007, doi: 10.1016/j.compmedimag.2007.02.002.

[2] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, vol. 278, no. 2, 2016, doi: 10.1148/radiol.2015151169.

[3] I. Zhovannik *et al.* Learning from scanners: Bias reduction and feature correction in radiomics. *Clin. Transl. Radiat. Oncol.*, vol. 19, 2019, doi: 10.1016/j.ctro.2019.07.003.

[4] F. Orlhac *et al.* A postreconstruction harmonization method for multicenter radiomic studies in PET. *J. Nucl. Med.*, vol. 59, no. 8, 2018, doi: 10.2967/jnumed.117.199935.

[5] J. J. M. Van Griethuysen *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.*, vol. 77, no. 21, 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[6] Z. Shi, A. Traverso, J. van Soest, A. Dekker, and L. Wee. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). *Med. Phys.*, vol. 46, no. 12, 2019, doi: 10.1002/mp.13844.

[7] E. Pfaehler, A. Zwanenburg, J. R. de Jong, and R. Boellaard. RACAT: An open source and easy to use radiomics calculator tool. *PLoS One*, vol. 14, no. 2, 2019, doi: 10.1371/journal.pone.0212223.

[8] P. M. Szczypiński, M. Strzelecki, A. Materka, and A. Klepaczko. MaZda-A software package for image texture analysis. *Comput. Methods Programs Biomed.*, vol. 94, no. 1, 2009, doi: 10.1016/j.cmpb.2008.08.005.

[9] A. Zwanenburg *et al.* The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, vol. 295, no. 2, 2020, doi: 10.1148/radiol.2020191145.

[10] H. J. W. L. Aerts *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.*, vol. 5, 2014, doi: 10.1038/ncomms5006.

[11] P. Kalendralis *et al.* FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. *Med. Phys.*, vol. 47, no. 12, 2020, doi: 10.1002/mp.14805.

[12] A. Traverso, L. Wee, A. Dekker, and R. Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 102, no. 4, 2018, doi: 10.1016/j.ijrobp.2018.05.053.

[13] K. G. Foley *et al.* External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. *Radiother. Oncol.*, vol. 133, 2019, doi: 10.1016/j.radonc.2018.10.033.

[14] T. M. Deist *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin. Transl. Radiat. Oncol.*, vol. 4, 2017, doi: 10.1016/j.ctro.2016.12.004.

[15] H. Hofmann, H. Wickham, and K. Kafadar. Letter-Value Plots: Boxplots for Large Data. *J. Comput. Graph. Stat.*, vol. 26, no. 3, 2017, doi: 10.1080/10618600.2017.1305277.

MISTER DATA
WATSON
ARTIFICIAL INTELLIGENCE

Chapter 8

Multicenter Federated Learning in Radiomics

Zhenwei Shi*, Ivan Zhovannik*, Alberto Traverso, Frank JWM Dankers, Timo M. Deist, Petros Kalendralis, René Monshouwer et al. “Distributed radiomics as a signature validation study using the Personal Health Train infrastructure.” *Scientific data* 6, no. 1 (2019): 1-8.

Abstract

Prediction modelling with radiomics is a rapidly developing research topic that requires access to vast amounts of imaging data. Methods that work on decentralized data are urgently needed, because of concerns about patient privacy. Previously published computed tomography medical image sets with gross tumour volume (GTV) outlines for non-small cell lung cancer have been updated with extended follow-up. [A0-3] In a previous study, these were referred to as Lung1 (n = 421) and Lung2 (n = 221). [A0-4] The Lung1 dataset is made publicly accessible via The Cancer Imaging Archive (TCIA; https://www.cancerimagingarchive.net). We performed a decentralized multi-centre study to develop a radiomic signature (hereafter "ZS2019") in one institution and validated the performance in an independent institution, without the need for data exchange and compared this to an analysis where all data was centralized. [A0-1] [A1-1] The performance of ZS2019 for 2-year overall survival validated in distributed radiomics was not statistically different from the centralized validation (AUC 0.61 vs 0.61; p = 0.52). [A0-2] [A1-1] Although slightly different in terms of data and methods, no statistically significant difference in performance was observed between the new signature and previous work (c-index 0.58 vs 0.65; p = 0.37). Our objective was not the development of a new signature with the best performance, but to suggest an approach for distributed radiomics. Therefore, we used a similar method as an earlier study. We foresee that the Lung1 dataset can be further re-used for testing radiomic models and investigating feature reproducibility.

1. Introduction

Images from radiological examinations are presently one of the largest underutilized resources in healthcare "big data" [1]. *Radiomics* refers to computerized extraction of quantitative image metrics, known as "features". In 2014, Aerts et al. [2] showed that radiological features from Computed Tomography (CT) scans might encode additional information about phenotypic differences between tumours that lie beyond the grasp of the unaided human eye. The hypothesis is that multifactorial prediction models incorporating selected radiomic features may better inform individually personalized treatment strategies [3-6]. Radiomic data have now been investigated in CT [7-9], magnetic resonance imaging (MRI) [10,11] and positron emission tomography (PET) [12,13].

The availability of commercial and open source software for radiomic feature extraction has made this line of inquiry accessible to a large number of investigators [14-17]. However, multi- institutional development and validation of radiomic-assisted prediction models is slowed down due to privacy concerns about sharing of individual patients' medical images. [A0-4] Significant efforts are under way to make image sets used in radiomic investigations openly accessible via centralized repositories such as The Cancer Imaging Archive (TCIA; https://www.cancerimagingarchive.net) [18], however, many data owners remain cautious about sharing individual patient images publicly online.

A privacy-preserving distributed learning infrastructure based on World Wide Web Consortium "Semantic Web" data sharing standards [19], known as Personal Health Train (PHT; https://vimeo.com/143245835) [20] has been successfully used to develop and validate models on non-image clinical data [21-23]. To extend the PHT approach to radiomics, we first need to publish our radiomic features in a manner that is Findable, Accessible, Interoperable and Re- useable (FAIR) [24]. We have developed a pragmatic and extensible Radiomics Ontology (RO) that is publicly accessible via NCBO BioPortal (https://bioportal.bioontology.org/ontologies/RO). With the RO, we can describe over 430 class objects and 60 predicates between objects to publish radiomic features (with some relationships and dependencies) according to Semantic Web standards. The class objects include unique feature identifiers that are aligned with the Image Biomarker Standardization Initiative (IBSI) [25].

In this article, we show that the PHT infrastructure supports exchange of cross-institutional radiomic-based clinical data without material transfer of individual-level patient clinical data or images. Our primary objective was to show that external validation of a radiomic signature can be done with entirely decentralized data.

The specific use case was to learn a radiomic signature "ZS2019" for non-small cell lung cancer (NSCLC) overall survival at one institution and validate it at a remote

institution in a distributed fashion. We included two of the NSCLC subject cohorts used by Aerts et al. [2], however, with independently reviewed annotations (tumour delineations) and extended follow-up times for overall survival. We did not select new radiomic features, and instead used the four features corresponding to those described previously in the original publication, but using a different software implementation (see materials and methods). [A0-5] The first of these datasets (hereafter referred to as "Lung1") [26] was generated at Maastricht University, which was used exclusively for model training, thus obtaining coefficients for a four-feature signature in ZS2019. The second of these datasets (hereafter "Lung2") was generated at Radboud University remains in a private hospital collection that could not be shared publicly for privacy reasons; Lung2 was used exclusively for model validation.

[A1-2] **Table 1.** The clinical case-comparison for the training cohort (Lung1) and the validation cohort (Lung2). The abbreviations are: (GTV) is Gross Tumour Volume delineated on the radiotherapy treatment planning computed tomography image, (Clinical T) is the tumour staging, (Clinical N) is the node staging and (Clinical M) is the metastasis staging, respectively, according to the TNM tumour classification system.

		Lung1 (n=421)	Lung2 (n=221)
Median age (range) at diagnosis in years		68.5 (34-92)	66.0 (36-87)
Median GTV size (range) in cm³		39 (0-660)	88 (1-860)
Clinical T stage			
	Less than 3	249 (59%)	119 (54%)
	3 or greater	171 (41%)	85 (38%)
	Unknown	1 (0%)	17 (8%)
Clinical N stage			
	0	170 (40%)	49 (22%)
	1	22 (5%)	16 (7%)
	2 or greater	229 (55%)	137 (62%)
	Unknown	0 (0%)	19 (9%)
Clinical M stage			
	0	416 (99%)	200 (90%)
	1 or greater	5 (1%)	21 (10%)
Histology			
	Adenocarcinoma	51 (12%)	64 (29%)
	Large-cell	143 (34%)	22 (10%)
	Squamous cell carcinoma	152 (36%)	82 (37%)
	Other, or not otherwise specified	63 (15%)	47 (21%)
	Unknown	12 (3%)	6 (3%)
Outcomes			
	Median follow-up in days	546	595
	Median survival time in days	478	500
	2-year overall survival rate	40%	41%

2. Results

Cohort summary information was exchanged through private discussion between the collaborating investigators, prior to performing this study. This was to confirm that general characteristics were comparable between the updated cohorts. This is shown in **Table 1**.

None of the information contained in Table 1 was used in the model. There was a slightly higher proportion of patients with metastatic disease in Lung2 (10% vs 1%) compared to Lung1. The most common histology types in Lung1 were large-cell and squamous-cell carcinomas, whereas adenocarcinoma and squamous-cell carcinoma were most common in Lung2. The median follow-up time, the median survival time and the overall 2-year survival rate were similar in both cohorts.

We evaluated ZS2019 for 2-year overall survival using multivariable logistic regression. The area under the receiver operating characteristic curve (AUC) discrimination metric was 0.61 (95% confidence interval: 0.54 to 0.69) in the Lung2 validation cohort.

Distributed learning code for Cox regression in MATLAB (MATLAB 2016a, Mathworks, Natick MA, USA) was deployed via the PHT infrastructure connecting MAASTRO Clinic and Radboudumc. We retrieved anonymous event timepoints and thus compiled Kaplan-Meier curves for overall survival in each of the training and validation cohorts (in **Figure 1**). Within each cohort, the subjects were stratified into two risk groups, based on the median of the risk score distribution in Lung1. [A1-1] Stratification of survival curves by ZS2019 in the validation cohort was quantified via a Harrell Concordance Index (HCI) of 0.58, and a 95% confidence interval from 0.51 to 0.65. The discrimination was statistically significantly different from random ($p < 0.0001$) based on a bootstrapped Wilcoxon estimation. We performed the same bootstrapped Wilcoxon estimation between the mean HCI of model ZS2019 (0.58) and the HCI previously published by Aerts et al (0.65) [2], and found no evidence of significant divergence ($p = 0.37$).

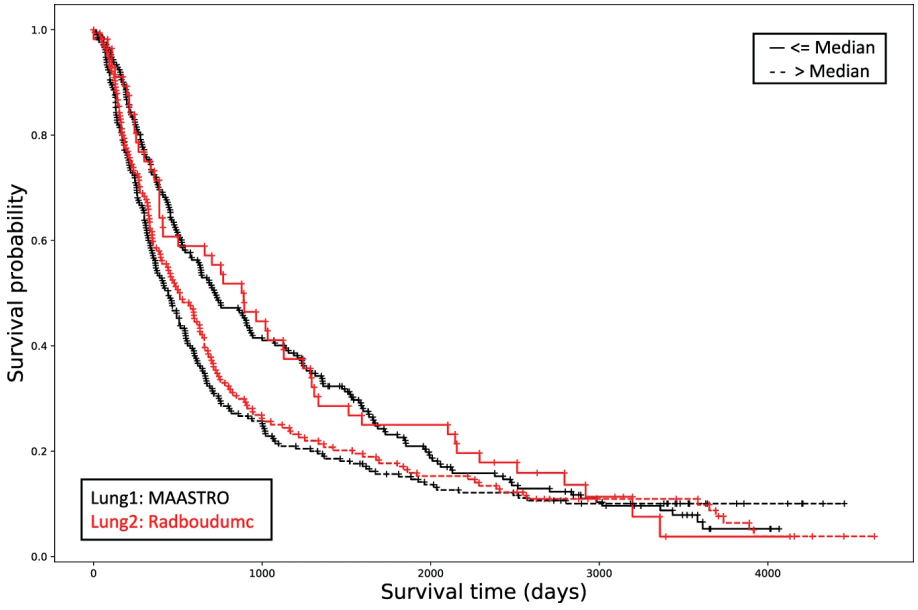


Figure 1. *The performance of radiomic signature ZS2019 according to Kaplan-Meier survival analysis. The signature was developed in Lung1 (MAASTRO; black line) and then distributedly validated in Lung2 (Radboudumc; red line). The upper and lower survival curves were split according to the median of the Cox regression linear predictor from the Lung1 data, and applied to both Lung1 and Lung2 data. The Harrell concordance index in the test cohort was 0.58, the log-rank test yielded a p-value of 0.09 and the Wilcoxon test gave p-value < 0.0001.*

We confirmed that the same ZS2019 result was obtained when trained centrally on Lung1 and validated in Lung2. The analysis is given in a Python v3.6 JuPyter notebook that is made publicly available (https://gitlab.com/UM-CDS/distributedradiomics). [A1-3] The central data approach yielded a HCI of 0.58 with a 95% confidence interval estimated by bootstrap sampling to be 0.53 to 0.64.

3. Discussion

In this paper, a model (ZS2019) derived from radiomic features and overall survival locally within one institution was able to be exchanged interoperably with an external institution, without mandating any transfer of either images, feature values or clinical outcomes at the individual subject level. This is an essential and unique contribution to radiomic investigations, because we hereby demonstrate the concept for carrying out multi-centre radiomic studies with fully decentralized data. The results obtained with decentralized data were the same as if all the data

had been brought into the same location. However, the unique advantage of our approach is that no one party needs to risk breaking patient confidentiality by exposing the original data to another party. Each institutional data owner retains complete control over their privacy-sensitive patient data, and decides what they wish to share for a collaborative project.

We foresee that public access to the updated Lung1 dataset, accessible together with open source radiomics software code, encourages re-use of the data for validating models, investigating radiomic feature generalizability and deep-learning for image analysis.

To learn effectively across institutions, it is essential that the investigation should be led by clinical experts. Our approach does not bypass the need for human experts to communicate extensively before commencing a study, in order to establish consensus on: (i) what is the clinical question to be addressed, (ii) relevant inclusion and exclusion criteria, (iii) which datasets are appropriate for answering the question and (iv) how to define the radiomic features and outcome concepts.

With respect to handling errors and discrepancies for a distributed radiomics study, it is essential that each data owner takes responsibility for curation and quality assurance of the data, such that it conforms to the agreed consensus. Where errors are detected, it is only the owners of the data that are able to review, contextualize and correct their own data.

[A0-7] In this study, both sites used the same feature extraction software, PyRadiomics. We retained the step of attaching metadata to the features using the Radiomics Ontology so that, in future, sites might be able to use different software but can still understand each other because features having the same metadata labels from this ontology will be unambiguously defined as being semantically identical. Besides applying an ontology, this also requires the different Radiomics feature extraction software to use the (exact) same feature calculation method.

The approach of making data FAIR using semantic ontologies has the benefit of allowing each data owner to keep their own native language and annotation conventions in the original data. No syntactic harmonization of the data below the level of the FAIR station needs to be enforced, and no data code-books need to be exchanged. The only prerequisite here is that partnering institutions must follow their consensus agreement to label the comparable outcomes and equivalent radiomic features with the same unique identifier from the same domain ontology.

To develop ZS2019, we attempted to follow, as closely as possible, the approach adopted in the original publication. The HCI and AUC results we reported above

were built using radiomic features that might not be optimal for the updated datasets, because we chose to use the four features with names corresponding to those described previously in the supplementary material of the prior study [27]. Development of an optimal radiomic signature for NSCLC overall survival would require a detailed re-examination of features and feature selection in the updated datasets, which is not the primary objective of the present study.

The PHT approach utilises existing data to answer key questions in personalised healthcare, preventive medicine and value-based healthcare. PHT is one of a number of innovative approaches (DataSHIELD [28] and WebDISCO [29]) where the research question is coded as machine-learning algorithms sent to wherever data may reside, instead of centralising all of the data at one location. This is achieved by (i) creating FAIR data stations, (ii) creating "trains" containing the research question as a machine-learning algorithm and (iii) establishing "tracks" to regulate the trains and securely transmit them to data stations. The PHT is thus a "privacy-by-design" architecture, since it enables controlled access to heterogeneous data sources for clinical research. This respects data protection and personal privacy regulations, and requires active engagement of data owners in the process.

We used Semantic Web standards to make radiomic features and outcome data available as FAIR stations in keeping with our trains metaphor. This included locally storing radiomic features and outcome states in Resource Description Format (RDF), and allowing semantic interoperability using a combination of the Radiomics Ontology and Radiation Oncology Ontology. The benefit of Semantic Web is to make distributed learning possible even if the underlying implementation of data extraction and storage differs between sites. The RDF standard makes it unnecessary to first know the internal structural organization of a remote database in order to successfully execute a local data retrieval query. Furthermore, as the diversity and complexity of the data within the FAIR stations increases in the future, an RDF triple store approach is sufficiently flexible to describe arbitrarily complex concepts without the need to redesign the database.

[A0-8] Use of the Varian Learning Portal (VLP; Varian Medical Systems, Palo Alto, USA) was of benefit for distributed radiomics, because the software had already implemented the essential technical overheads (logging, messaging and internet security) required for such distributed studies. This included underlying legal agreements between the parties and Varian, that makes distributed radiomics more scalable since one does need to revisit these common aspects above for each project. [A0-9] The VLP system had no effect on the mathematical results of our

study because it was purely a way for us to securely transmit learning algorithms and trained models. Alternatives to VLP such as DataSHIELD (http://www.datashield.ac.uk) [28], WebDisco (https://omictools.com/webdisco-tool) [29] and ppDLI (https://distributedlearning.ai/blog) may also be used for distributed radiomics. The differences between the present study and the original study may be traced to : (i) the original Matlab code is commercial confidential and not available to the authors, so we used PyRadiomics developed by Aerts et al. [2] as a practical alternative and (ii) we tried our best to replicate the original method using the documented steps in the original manuscript, but we also improved the survival follow-up such that many right-censored events were now confirmed deaths.

4. Conclusion

This study demonstrates the proof of concept for multi-centre distributed radiomics investigation without exchanging individual-level data or medical images using the PHT infrastructure. The results showed that the proposed decentralized approach achieved the identical results as the fully centralized approach. Moreover, we performed a radiomics study where data was stored in the FAIR station at the institute rather than publishing as open-source. Finally, the work of this study may be used as the basis for other types of radiomics studies such as binary classification or regression, not only limiting to survival analysis.

5. Methods

Patients

Subjects in this replication study were from the same cohorts of non-small cell lung cancer (NSCLC) patients previously treated with (chemo-)radiotherapy at MAASTRO Clinic (MAASTRO) and Radboud University Medical Centre (Radboudumc). These were previously labelled by Aerts et al. [2] as cohorts “Lung1” and “Lung2”, respectively, and the same nomenclature is followed in this study. The Lung1 cohort (n = 421) was used only for fitting of model coefficients, and Lung2 (n = 221) was exclusively used for external validation.

Tumour delineations

Radiotherapy treatment planning DICOM CT images and physician-delineated primary NSCLC tumours as RT structure sets were used. From 422 available, 34

cases were found to have a reference frame translation between the image and delineation due to incorrect coding of the treatment couch height offset in the planning system; these have been rectified for the TCIA collection. Only 1 patient was post-operative radiotherapy, so this case was excluded from any further analysis, leading to 421 eligible cases in Lung1 for model training.

In the Lung2 cohort, there were initially 267 subjects available. A check against delineation criteria found 221 eligible primary tumours for radiomic analysis. The other 46 patients had either gross tumour volumes including lymph nodes, or were cases with neoadjuvant treatment or had no primary tumour in the list of structures.

Outcomes

Updated follow-up intervals in early 2018 with recent dates of death were obtained with ethics board permission from the Dutch citizens registry. As expected, the number of registered deaths in Lung1 and Lung2 had increased significantly since the original publication. The time intervals from date of first radiotherapy fraction to date of either registered death or last known survival were updated in both Lung1 and Lung2.

Data processing

The study steps are shown schematically in Figure 2 for MAASTRO and Radboudumc. The core of the radiomic feature extraction process utilizes free and open-source PyRadiomics [15] (v1.3) libraries. Software wrapper extensions collectively known as O-RAW (https://gitlab.com/UM-CDS/o-raw) were used to convert DICOM objects into numerical arrays as inputs for PyRadiomics; these were based on the SimpleITK (v1.0.1) [30] toolkit.

Figure 2. *A schematic diagram explaining the primary methodology for survival analysis used in this study. Details have been provided in the text. Briefly, radiomics features were extracted locally by each institution and then labelled with the radiomics ontology. We then trained a Cox regression model on Lung1 (MAASTRO) and then validated on Lung2 (Radboudumc) by distributing the learning algorithm through the Varian Learning Portal (VLP). Only the event coordinates required to plot a Kaplan-Meier survival curve was returned to MAASTRO, without any identifiable patient-level data.*

The original MATLAB scripts used by Aerts et al. were not accessible to the current authors. The open source PyRadiomics was developed independently of this MATLAB code, and was based on the original study from Aerts et al. The PyRadiomics community has documented and standardized the feature calculation formulae (https://pyradiomics.readthedocs.io)

The image pre-processing methodology was the same as in the original publication [2]; an extraction intensity bin width was set at 25 Hounsfield Units with no image resampling and no image intensity normalization. The coif1 wavelet package from the pywavelets library (v0.5.2, https://github.com/PyWavelets/pywt) was used to generate wavelet features with a starting bin edge of 0. All of these settings are the default in PyRadiomics.

For the development of ZS2019 we did not select new radiomic features, and instead used the four features with names corresponding to those described previously in the supplementary material [27] that accompanied the original publication:

i. energy from the intensity histogram feature class, which estimates the overall density of the region of interest,

ii. compactness from the morphological feature class, which describes the volume of the object relative to that of a perfect sphere,

iii. grey level run-length matrix (GLRLM) non-uniformity from the textural feature class, which is a measure of intensity heterogeneity averaged over 13 different directions in a 3D matrix of values, and

iv. wavelet-filtered (HLH) GLRLM non-uniformity, which was the same as (iii) after applying a wavelet decomposition filter over the original image.

In our work, the feature “compactness” had been deprecated in PyRadiomics, so we derived the mathematical equivalent of compactness by taking the cube of the shape feature “sphericity” (see formulae in Table A of Supplementary Materials).

Semantic web ontologies

Semantic Web technologies and ontologies play a key role in distributed learning by enabling semantic interoperability between data from multi-centres. In this study, radiomic features and clinical data were defined by a Radiomics Ontology v1.3 (https://bioportal.bioontology.org/ontologies/RO) and a Radiation Oncology Ontology [31] , respectively.

We elected to use the published open access Radiomics Ontology, that identifies radiomic features via a globally persistent unique identifier and allows us to attach important dependencies, such as digital image pre-processing steps, directly to each given feature. Though radiomic features definitions have been defined by previous investigators, our contention is that human-readable labels alone may not always be easily extensible to define dependencies such as software versions, image pre-processing steps and mathematical implementation of the feature. For example, to avoid conflation between features labelled "entropy", the IBSI distinguishes between Intensity Histogram Entropy (unique ID = TLU2) and the textural feature Joint Entropy (unique ID = TU9B). The Radiomic Ontology allows extensible and adaptable declaration of radiomic feature provenance by publishing it as a data graph object. Therefore, independent researchers (in the aforementioned example) who have computed Joint Entropy may use the SPARQL federated query language (https://www.w3.org/TR/rdf-sparql-query) on feature graphs to also probe for similarities in imaging setting, pre-processing methods, and suchlike. We hypothesise that the data graph based approach is more scalable than pairwise cross-referencing of multiple dictionaries of feature definitions.

Distributed approach

The VLP distributed learning architecture has been described in deep detail elsewhere [21-23]. In brief, VLP consists of (i) a global web-based clinical learning environment that spans across any number of participating institutes for a given learning project, and (ii) a local connector application that runs exclusively inside the IT firewall of each institute. The former coordinates access permission, asynchronous messaging, web security and site privacy protocols across the learning network, while the latter hosts a local FAIR data repository.

Radiomic feature values were hosted in the respective VLP local connector application (v2.0.1) as RDF.

Authenticated and verified (e.g. encrypted digital signature) machine learning packages are distributed via the global part of VLP, then picked up and executed on the RDF data via the local connector part. Only the statistical summary result of the computation, not any identifiable patient data, is thereafter passed back to the instigator via the global VLP part. Any process that had executed within local firewalls remain permanently quarantined from the global part.

Model training

The Lung1 radiomic feature values were log-transformed and then scaled to z-scores. A multivariable Cox proportional hazards model for overall survival (with removal of right censored subjects not yet deceased) was then fitted using all of the available subjects in the training cohort. The median risk score in the training cohort was recorded and thus used to stratify the training population into two risk groups. The fitted Cox model coefficients, the median risk score and the z-score transformations from the training cohort were packaged as self-contained validation application, which was then transmitted via VLP to Radboudumc.

At Radboudumc, the application queried the local RDF repository for the radiomic features, then applied the same log-transform of raw feature values and the same z-score scaling as had been executed on Lung1. For each available validation subject in Lung2, the risk score was computed and stratified according to the median risk score of Lung1. A flat table of individual timepoints and death/censor events was sent back via VLP to MAASTRO.

Cox model evaluation

Anonymous timepoints for Kaplan-Meier survival curves [32] were retrieved over the PHT infrastructure. Risk scores were stratified into two strata according to the median value in the Lung1 population. A Harrell concordance index (HCI) [33] implemented using the python lifelines package (v0.14.4) was used to quantify discrimination performance using the retrieved timepoints. The log-rank method [34] was used to calculate a chi-squared test statistic and p-value for the significance of the discrimination. To assess if the survival model had any value beyond random discrimination (null hypothesis: c-index = 0.5), we used a two-sided Wilcoxon test with a bootstrap approach on 100 repeated sub-samples of 100 patients per repetition from Lung2.

2-year overall survival

A multivariable logistic regression model for 2-year overall survival was developed on Lung1 then validated on Lung2 using the aforementioned four features. The area under the curve of the receiver operating characteristic was used to assess the discrimination. The bootstrap method (1000 times) was used to estimate a 95% confidence interval around the mean AUC.

6. Data Availability

The Lung1 images, primary tumour delineations (from Method: tumour delineations) and clinical outcomes with updated follow-up (from Method: outcomes) has been approved for open access publication, and is curated as the collection called "NSCLC-Radiomics" via The Cancer Imaging Archive (TCIA) (https://wiki.cancerimagingarchive.net/display/Public/NSCLC- Radiomics) [26]. The clinical data for Lung1 that support the findings of this study are also available in TCIA with the data identifier (http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI). Further information regarding the Lung1 data may be obtained from the authors responsible, A Dekker (email: andre.dekker@maastro.nl; address: Doctor Tanslaan 12, 6229 ET; Maastricht, The Netherlands; phone: +31 88 445 5600) and L Wee (email: leonard.wee@maastro.nl; address: Doctor Tanslaan 12, 6229 ET; Maastricht, The Netherlands; phone: +31 88 445 5600)

The Lung2 datasest that support the findings of this study are available by request from the authors R Monshouwer (email: rene.monshouwer@radboudumc.nl; address: Radboud university medical center, Department of Radiation Oncology, Geert Grooteplein 32, 6525 GA, Nijmegen, The Netherlands; phone: +31 24 361 4515) and J Bussink (email:jan.bussink@radboudumc.nl; address: Radboud university medical center, Department of Radiation Oncology, Geert Grooteplein 32, 6525 GA, Nijmegen, The Netherlands; phone: +31 24 361 4515). This part of data are not publicly available due to the data containing information that could compromise research participant privacy.

7. Code availability

The code used in this study is made publicly available on the Maastricht University Clinical Data Science (UM-CDS) GitLab repository (https://gitlab.com/UM-CDS/distributedradiomics). The code repository has the following organization:

a. D2RQ folder: contains the raw feature value to RDF mapping (D2RQ) script and the SPARQL query used to retrieve the local data into the local VLP connector application.

b. VLP folder: contains the MATLAB codes submitted by the user into VLP, which then transmits it to the participating site for model validation and analysis.

c. Analysis Centralized Learning folder: contains the Jupyter notebook from Radboudumc for model development and evaluation on the aggregated datasets.

The open-access Radiomics Ontology (RO) is published via the National Center for Biomedical Ontology (NCBO) ontology registry. It is available to download in a range of formats from the following URL: https://bioportal.bioontology.org/ontologies/RO. As a domain ontology, the RO defines histogram-based, morphology-based and texture-based radiomic features, including (since v.1.6, 08 November 2018) all feature entities presented in the International Biomarker Standardization Initiative. The ontology also defines software properties, digital imaging filter operations and feature extraction settings, together with relational predicates to link these to each feature entity.

Supplementary materials

References

[1] McKnight, J., Babineau, B. & Gahm, J. North American Health Care Provider Information Market Size & Forecast. *ESG-Enterprise Strategy Group* (2011).

[2] Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5**, 4006 (2014).

[3] Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563-577 (2015).

[4] Kumar, V. *et al.* Radiomics: the process and the challenges. *Magnetic resonance imaging* **30**, 1234-1248 (2012).

[5] Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **14**, 749 (2017).

[6] Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* **48**, 441-446 (2012).

[7] Coroller, T. P. *et al.* CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology* **114**, 345-350 (2015).

[8] Huang, Y.-q. *et al.* Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *Journal of Clinical Oncology* **34**, 2157-2164 (2016).

[9] Parmar, C. *et al.* Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports* **5**, 11044 (2015).

[10] Nie, K. *et al.* Rectal cancer: assessment of neoadjuvant chemo-radiation outcome based on radiomics of multi-parametric MRI. *Clinical cancer research* **22.21**, 5256- 5264 (2016).

[11] Zhang, B. *et al.* Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clinical Cancer Research* **23.15**, 4259-4269 (2017).

[12] Foley, K. G. *et al.* Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. *European radiology* **28**, 428-436 (2018).

[13] Leijenaar, R. T. *et al.* Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica* **52**, 1391-1397 (2013).

[14] Apte, A. P. *et al.* Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. *Medical physics* **45.8** 3713- 3720 (2018).

[15] van Griethuysen, J. J. *et al.* Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research* **77**, e104-e107 (2017).

[16] Zhang, L. *et al.* IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. *Medical physics* **42**, 1341-1353 (2015).

[17] Nioche, C. *et al.* LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer research* **78**, 4786-4789 (2018).

[18] Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* **26**, 1045-1057 (2013).

[19] Berners-Lee, T., Hendler, J. & Lassila, O. The semantic web. *Scientific american* **284**, 28-37 (2001).

[20] van Soest, J. *et al.* Using the Personal Health Train for Automated and Privacy- Preserving Analytics on Vertically Partitioned Data. *Studies in health technology and informatics* **247**, 581-585 (2018).

[21] Jochems, A. *et al.* Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. *International Journal of Radiation Oncology· Biology· Physics* **99**, 344-352 (2017).

[22] Jochems, A. *et al.* Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–A real life proof of concept. *Radiotherapy and Oncology* **121**, 459-467 (2016).

[23] Deist, T. M. *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and translational radiation oncology* **4**, 24-31 (2017).

[24] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **3** (2016).

[25] Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative-feature definitions. *Preprint at arXiv: https://arxiv.org/abs/1612.07003* (2016).

[26] Aerts, H. J. W. L., Rios Velazquez, Emmanuel, Leijenaar, Ralph T. H., Parmar, Chintan, Grossmann, Patrick, Carvalho, Sara, ... Lambin, Philippe. *(2015). Data From NSCLC-Radiomics. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI*.

[27] Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. (Supplementary). *Nature communications* **5**, 4006 (2014).

[28] Wolfson, M. *et al.* DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology* **39**, 1372-1382 (2010).

[29] Lu, C.-L. *et al.* WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association* **22**, 1212-1219 (2015).

[30] Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleITK. *Frontiers in neuroinformatics* **7**, 45 (2013).

[31] Traverso, A., van Soest, J., Wee, L. & Dekker, A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. *Medical physics* **45.10**, e854-e862 (2018).

[32] Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**, 457-481 (1958).

[33] Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15**, 361-387 (1996).

[34] Peto, R. & Peto, J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)* **135**, 185-198 (1972).

Acknowledgements

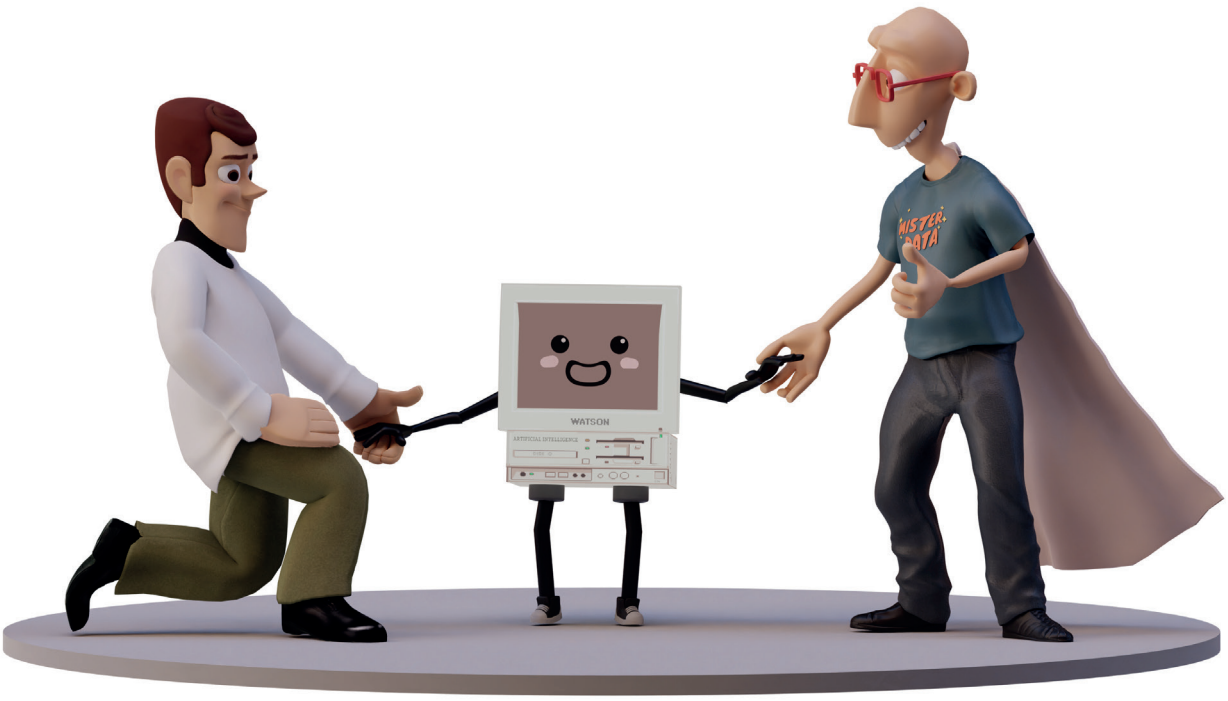
This work has been supported by a Dutch STW-Perspectief grant: Radiomics STRaTegy (file number 14930). The authors acknowledge Carlotta Masciocchi and Johan van Soest for productive conversations about distributed learning.

Author contributions

ZS was implemented the distributed MATLAB code via VLP, performed analysis on the results using python and made a major contribution to the writing of the manuscript as joint first author.
IZ was responsible for data preparation and analysis done at Radboudumc, and made a major contribution as joint first author to the writing of the manuscript.
FD assisted with clinical outcomes and feature mapping to RDF, and the SPARQL query for feature retrieval.
TD assisted with code implementation via VLP.
AT assisted with mapping of features to the Radiomics Ontology. PK assisted with data preparation at MAASTRO Clinic.
RM and JB updated the clinical follow-up at Radboudumc and were the primary supervisors of IZ.
RF assisted with manuscript proofreading and was co-supervisor of IZ.
HJLWA, AD and LW acted in the capacity of joint senior authors who motivated the study, set the general methodology and had overall scientific responsibility for this investigation.
All co-authors contributed to proof-reading of the manuscript.

Competing interests

MAASTRO Clinic receives institutional research support from Varian Medical Systems.
AD receives speaking and consultancy honoraria from Varian Medical Systems.
AD holds a patent on radiomics (US Patent 9721340 B2).


WATSON
ARTIFICIAL INTELLIGENCE

Chapter 9

General Discussion and Future Perspectives

General Discussion

Radiomics as an image biomarker method has gained traction in radiation oncology and medical image analysis research since 2012 [1,2]. In these ten years, various radiomic studies were performed for different disease sites and multiple image modalities [3]. Most of the studies consider a proof-of-concept radiomic model for specific disease sites, imaging modalities, and specific patient populations. Yet, there is no widespread conversion of radiomics-based machine learning tools into concrete implementation in clinical routines. In this thesis, among others, we explored some of the aspects of why so few radiomic studies end up in a working and deployable clinical solution.

In order to understand the possible cause of the low conversion of radiomic tools into clinical practice, we must first consider how imaging data is generated. In general, the generation of imaging data can be described as encoding clinical and biological information (C) together with domain information (D). Here, we assume all the information related to the data acquisition process by domain. There can be other sources of information encoded, we refer to them as residual parameters (R). The imaging data generation process is illustrated in Figure 1. The clinical confounders *C* may include semantic factors like tumor size, tumor stage, disease progression over time; the domain confounders *D* include the parameters of how the image was acquired, reconstructed, processed, and delineated.

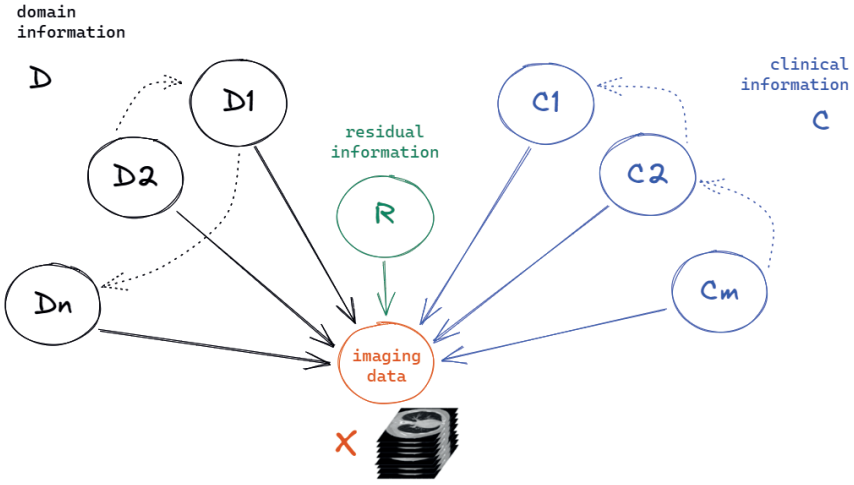


Figure 1. *the imaging data generation process can be described encoding (solid lines) clinical, domain, and other (residual) information into a multidimensional image grid. Note that there can be causal relationships between different components of either domain and clinical spaces. The illustration was inspired by Ilse et al* [4].

Problems and Solutions for Radiomics

This thesis started with an assessment on the state of radiomics repeatability and radiomics reproducibility. The first term, radiomics repeatability addresses the question: "How similar will a feature value be over time?" Or, in other words, "Will the feature value keep its value when the measurement is repeated with the exact measurement apparatus when there are no deterministic changes (shifts) in domain parameters D?" As time passes, we cannot guarantee that clinical parameters will stay the same and, therefore, time can be considered as a confounder, when it comes to repeatability studies.

The repeatability of a radiomic feature in patient data might depend on internal tissue movement, which can be attributed to clinical confounders C, and intrinsic radiomics uncertainty due to stochastic variations (for instance acquisition noise), both in clinical and domain parameters. The stochastic variations in D include, but are not limited to stochastic variations in image acquisition, reconstruction, and (re-) segmentation. Stochastic variations in image acquisition may, for instance, be caused by the physics of particle movement. The stochastic variations in segmentation may be caused by human factor (in the case of a human annotator) and technology (e.g. brush size). The radiomic data extracted from solid physical phantoms, on the other hand, only depends on domain uncertainties, therefore, phantom radiomics have a higher degree of repeatability [5]

Radiomics reproducibility concerns the question: "How similar will a feature value be over different domain shifts?" Here by a domain shift, we assume a set of deterministic (non-stochastic) changes in domain parameters D the data X was acquired with (see Figure 1). The domain parameters include, but are not limited to image acquisition, reconstruction, and delineation protocols. One relevant example of domain-induced acquisition uncertainty in radiomics is the use of various scanners and scanning protocols, which is unavoidable in large multi-institutional datasets. Ideally, a *robust image biomarker* should be relatively *domain-invariant* (independent with regards to domain uncertainty) and *bring new information* with regard to well-known clinical confounders.

This thesis raises the clinical confounder problem in CHAPTERS 4 and 5. For instance, it is well- known, that tumor size is related to cancer progression, and in turn, to patient survival. Figure 2 illustrates how the tumor volume (V) hidden confounder influences both the imaging data (V ➔ X) and the outcome (V ➔ Y). If a radiomic feature is confounded by an already known clinical parameter, it might

cause an illusion that this texture feature is a new prognostic biomarker, whereas in reality, no new information was discovered. Techniques like Markov Chain Monte Carlo (MCMC) simulation [CHAPTER 3] and correlation analysis [CHAPTERS 5, 6] can be used to find and mitigate the clinical confounders in radiomics.

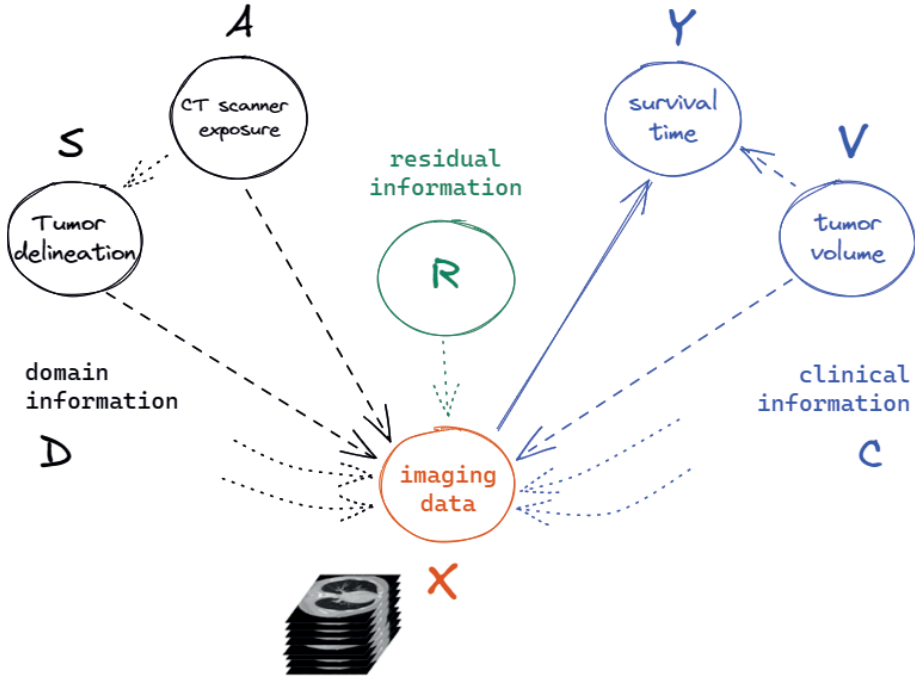


Figure 2. *the schema of the thesis contributions: The solid blue line illustrates the main clinical objective: survival prognosis using imaging data (X→Y); the dashed lines show the contributions in domain-induced uncertainty estimation (S→X & A→X) and clinical confounder identification (V→X & V→Y).*

Domain-induced uncertainty estimation is the corner stone of this work. In CHAPTER 3, we showed how uncertainty in clinical delineation of gross tumor volume (GTV) in lung cancer patients propagates into uncertainty of radiomic values (S→X) and, in turn, reduces prognostic model (X→Y) performance. CHAPTER 4 describes how acquisition uncertainty can be estimated and corrected in order to make radiomic data invariant to acquisition uncertainty (A→X). In summary, stochastic and deterministic domain uncertainty in D induces variation in the imaging data X of a patient. In terms of survival prognosis, a model should distinguish patients with poor and good prognosis. The model should be invariant to domain uncertainty and sensitive to the effect, which is survival prognosis

stratification in our case. Therefore, in CHAPTER 3, we proposed uncertainty to effect ratio η as an estimation for how insensitive a model or a biomarker is to domain-induced uncertainty relative to the effect.

The invariance to domain uncertainty or domain invariance may be achieved with domain adaptation techniques [6], which can be also referred to as data harmonization. In this thesis, we proposed two harmonization techniques: model ensemble in CHAPTER 3 and trend correction (de- trending) in CHAPTER 4. It is important to mention, that understanding of the domain shifts and uncertainty is essential to developing better data harmonization strategies and, therefore, comes before data harmonization.

In order to develop domain-invariant image biomarkers and machine learning models, first and foremost, it is paramount to ask the following questions:

1. Identify and Prioritize – Which domain confounders induce uncertainty for a given imaging modality and disease site?

2. Think about possible gains – What will be the gain in signal-to-noise ratio after the domain- uncertainty is reduced?

If the domain confounders are identified and prioritized by the possible gain, data harmonization might be performed and future data collection might be also standardized reducing the domain- induced uncertainty for the identified confounders.

Here we would like to highlight the importance of sharing knowledge, code, and data whenever possible. Lack of publicly available knowledge with regards to domain uncertainty for your modality and disease site. Results in a full analysis has to be performed to “map” the space in order to develop robust biomarkers. With knowledge available already for the modality and/or disease site of your interest, the questions above might be answered and the domain confounders identified. If, moreover, code and data is available and well-maintained, a new project can skip the software development phase and jump directly into the analysis. This would save the precious time and allow for further excelling the understanding of domain-invariance for the given modality and the disease site.

The Future Perspectives

There are two extremes in image biomarker research: modeling without considering uncertainty and uncertainty-only research. Based on CHAPTERs 3 and 4 we argue

that omitting the uncertainty will lead to poor image biomarker performance. In order to properly estimate all the uncertainty factors for a given modality and disease site, multiple well-thought experiments should be conducted and much data is required. Since there might usually be multiple domain uncertainty factors to account for; therefore, it is important to prioritize and investigate the most relevant (large gain in prognostic power) over less relevant factors. There is a need for improvement in the direction of domain- invariance, but it should be incremental, we believe. Therefore, it is paramount to report data, share knowledge, code and data, whenever possible – this way our common goal of developing reliable image biomarkers can get closer.

First possible future research direction might be to estimate uncertainty to effect ratio for a given modality and disease site. Take, for instance, the project and data described in CHAPTER 3. We can use the knowledge regarding the segmentation uncertainty and combine it with the acquisition uncertainty, estimated in an experiment similar to the one in CHAPTER 4. These can be separate studies performed to measure the two domain-induced uncertainty factors independently. The two uncertainty estimates can be then combined according to formula (2) in CHAPTER 3 if we assume zero covariance between the two factors.

In a prognostic study, it is desirable, but not obligatory, to perform an uncertainty estimation experiment. One may also reuse the knowledge of uncertainty for specific radiomic features in a given modality and disease site, if previously reported, as a proxy estimate. Using this proxy estimate, we can measure η as uncertainty proxy to effect ratio, thus assessing the signal-to-noise ratio for the image biomarker. In order to have those proxy estimates, there is a need for data studies for interobserver variability in segmentation for various disease sites and better phantom data sets to estimate image acquisition and reconstruction uncertainty.

A second future direction concerns the interoperability and reusability of the reported uncertainty factors. In order to precisely estimate all the domain-induced uncertainty and confounding factors, a meticulous study must be performed to accommodate and quantify all the factors. Also, proxy estimates for different factors derived from similar modality and/or disease site studies. Those proxies can be combined in order to estimate total uncertainty to effect ratio for the study, which will accommodate various factors of domain-induced variations.

Thirdly, it is important to keep in mind the end goal. As Yan LeCun mentioned in one of his LinkedIn posts, in 1980s-2000s, machine learning research was heavily focused on generative models and uncertainty estimation and not on actual

predictions [7]. The uncertainty estimation, generative models, and Markov Chain Monte Carlo methods give very valuable insight into data. However, just estimating the uncertainty to effect or signal-to-noise ratios in the data will not directly lead to our final objective – reliable image biomarkers, therefore, the focus on the final objective must be kept and studies should aim to breach the gap between imaging and clinical decision making.

We should also keep in mind the fourth future direction for both prognostic modeling and data harmonization – data structuring, FAIRness, open code and knowledge importance for accelerating imaging in radiation oncology and beyond. Due to the federated nature of radiomics described in CHAPTER 1, data sharing is problematic. When data sharing is possible, it still requires data sharing agreements (DTA), terminology harmonization and preprocessing in order to make centralized data analysis possible. Both federated and centralized radiomic studies would benefit from term harmonization by using standards such as DICOM and semantic DICOM (SeDi) for imaging data, Radiation Oncology Ontology(ROO) for clinical data [citation: ROO].

To summarize the vision for the future, image biomarker research would benefit from terminology and data harmonization, still the main focus should be kept on the main objective – clinical prognostic and predictive modeling and decision making.

Last but not least, lacking data quality causes poor radiomics implementation in the clinic and vice- versa as shown in Figure 3. The data is originally not acquired and annotated to be used for radiomic studies. Most of the radiomics studies are so far proof-of-concept and do not lead to the clinical decision making support. There are two factors, which could potentially contribute to breaking this vicious circle: the ever increasing data size in radiation oncology and the number of overall radiomic studies. If we focus on depth (clinical implementation) of a radiomic study versus breadth (more modalities, more datasets), we can turn the circle into an upwards positive feedback loop. More rigorous implementation can trigger the need for better data, and the better data, in turn, will lead to even wider implementation.

In conclusion, there are still steps to be made before we will see a wide implementation of image biomarkers in clinical decision making. Previous years of research and development in medical imaging showed that a straightforward implementation of image-based machine learning tools as “X to Y” or “data to clinical insights” mapping is often naïve and falls into the pitfalls described in this thesis.

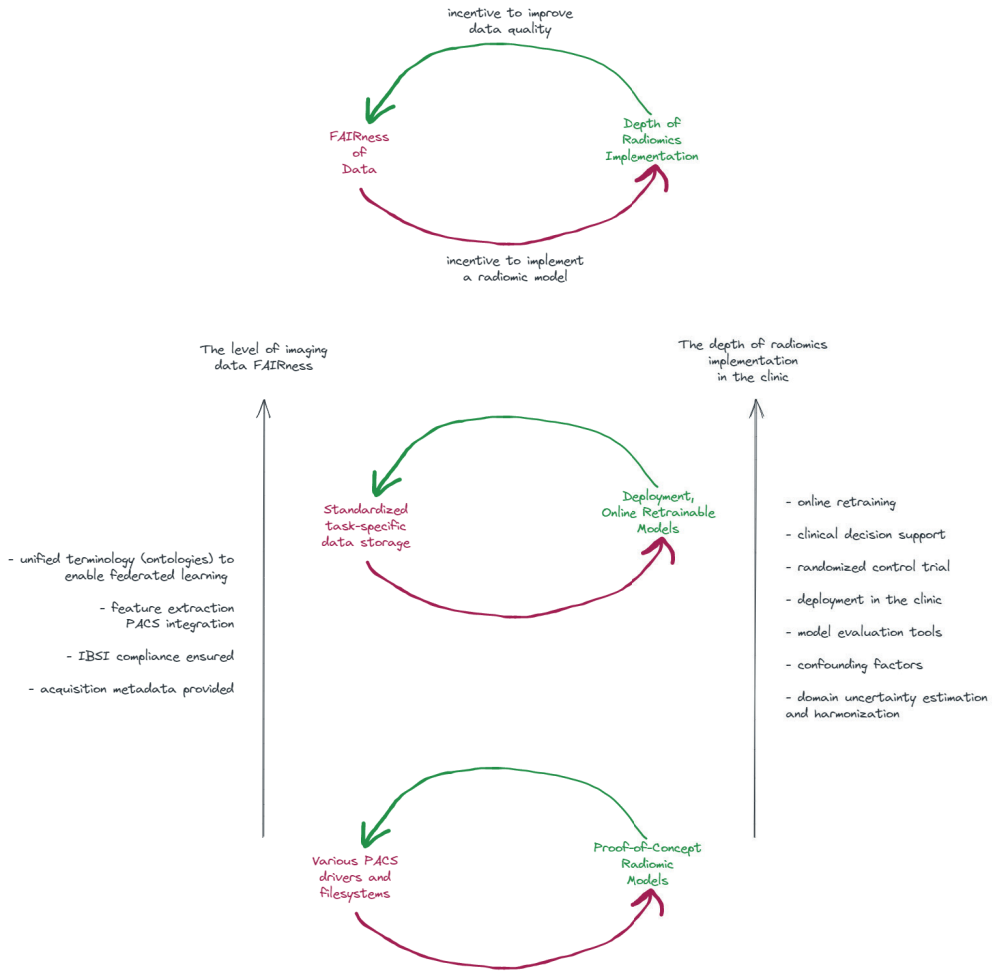


Figure 3. *we believe that incremental improvement in both data and radiomics implementation can break the "poor data – poor implementation" deadlock*

References

[1] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014, doi: 10.1038/ncomms5006.

[2] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, 2016, doi: 10.1148/radiol.2015151169.

[3] A. Traverso, L. Wee, A. Dekker, and R. Gillies, "Repeatability and Reproducibility of Radiomic Features: A Systematic Review," *International Journal of Radiation Oncology Biology Physics*, vol. 102, no. 4, 2018, doi: 10.1016/j.ijrobp.2018.05.053.

[4] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoder," 2019.

[5] I. Zhovannik *et al.*, "Learning from scanners: Bias reduction and feature correction in radiomics," *Clinical and Translational Radiation Oncology*, vol. 19, 2019, doi: 10.1016/j.ctro.2019.07.003.

[6] H. Zhao, R. T. des Combes, K. Zhang, and G. J. Gordon, "On learning invariant representations for domain adaptation," in *36th International Conference on Machine Learning, ICML 2019*, 2019, vol. 2019-June.

[7] Yan LeCun, "Yan LeCun LinkedIn post on Uncertainty," May 14, 2022. https://www.linkedin.com/posts/yann-lecun_yann-lecun-activity-6931321709572603904-_Mj3v/ (accessed Jun. 22, 2022).

Appendix

Summary

Dutch Summary

Data Management Statement

Portfolio

CV

Acknowledgements

Summary

In radiation oncology, imaging data is readily available for prediction modeling of treatment outcome. Even though the imaging data contain a large amount of latent information, the adoption of image-based models in, for instance, treatment response prognosis, is low. This thesis addresses why there is a bottleneck in clinical adoption of image biomarkers in radiation oncology and proposes a set of solutions to these complications.

In CHAPTER 2, the problem was formulated in the form of review on the state of radiomics reproducibility and repeatability. The uncertainty in image acquisition, reconstruction, processing, and delineation protocols propagates into the radiomic values, which leads to poor reproducibility of the radiomic values calculated from images acquired with different protocol settings. Moreover, some stable textural radiomic features are confounded with other clinical predictive factors, such as tumor volume.

CHAPTER 3 provided a theoretical basis for uncertainty estimation in radiomics. Segmentation- induced uncertainty estimation by Markov Chain Monte Carlo delineation sampling was considered. CHAPTER 4 described the image acquisition uncertainty using phantom and human imaging data acquired with different acquisition settings, scanner signal-to-noise ratios. This chapter also describes how the radiomic values could be harmonized against the uncertainty.

In CHAPTER 5, we showed a method to identify and remove the influence of the tumor volume as a clinical confounder for textural features. CHAPTER 6 gives insights on how to handle confounders and biases while developing a clinical prediction model.

CHAPTER 7 and CHAPTER 8 concerned technical challenges and solutions for future radiomics valorization in the clinical setting. CHAPTER 7 described the integration of radiomics into a PACS and CHAPTER 8 gave an example of a federated radiomics study between two cancer centers.

Dutch Summary

Bij patiënten die zijn behandeld met radiotherapie zijn beeldvormingsgegevens direct beschikbaar voor het ontwikkelen van voorspellende modellen. Hoewel de beeldvormingsgegevens een grote hoeveelheid latente informatie bevatten, is de acceptatie van op beelden gebaseerde predictiemodellen in bijvoorbeeld de prognose van de behandelingsrespons laag. Dit proefschrift gaat in op de redenen waarom klinische adoptie van deze beeld- of imaging-biomarkers (ook wel radiomics) in radiotherapie zo laag is, en stelt een reeks oplossingen voor.

In HOOFDSTUK 2 wordt het probleem geformuleerd in de vorm van een studie naar de reproduceerbaarheid en herhaalbaarheid van radiomics. De onzekerheid in beeldacquisitie-, reconstructie-, verwerkings- en inteken protocollen heeft invloed op de gemeten imaging- biomarkers, wat leidt tot een slechte reproduceerbaarheid van deze waarden als ze zijn berekend op basis van afbeeldingen die zijn verkregen met verschillende protocolinstellingen. Bovendien zijn sommige kenmerken van bijvoorbeeld de textuur sterk gecorreleerd aan klinische voorspellende factoren, zoals het tumorvolume.

HOOFDSTUK 3 beschrijft een theoretische basis voor het schatten van onzekerheid van imaging biomarkers. De invloed van segmentatie-geïnduceerde fouten wordt gesimuleerd met behulp van een Markov Chain Monte Carlo algoritme om model onzekerheid te schatten. HOOFDSTUK 4 beschrijft een studie naar de onzekerheid van radiomics met behulp van fantoom- en menselijke beeldgegevens verkregen met verschillende acquisitie-instellingen en scanner signaal- ruisverhoudingen. In dit hoofdstuk wordt ook beschreven hoe de radiomics waarden kunnen worden gecorrigeerd voor deze protocol variaties.

HOOFDSTUK 5 behandelt een methode om de invloed van het tumorvolume als een klinische confounder voor textuurkenmerken te identificeren en te verwijderen. HOOFDSTUK 6 geeft inzicht in hoe om te gaan met confounders en statistische bias bij het ontwikkelen van een klinisch voorspellingsmodel.

HOOFDSTUK 7 en HOOFDSTUK 8 hebben betrekking op de technische uitdagingen en oplossingen voor toekomstige toepassing van radiomics in de klinische setting. HOOFDSTUK 7 beschrijft de integratie van radiomics in een PACS en HOOFDSTUK 8 geeft een voorbeeld van een federated radiomics-onderzoek tussen twee ziekenhuizen waarbij de data zelf niet wordt uitgewisseld.

Data Management Statement

The thesis relied on open and closed sources of research data. The open data included Compute Tomography data sets: LUNG1, MMD – interobserver multi-delineation, RIDER, and GAMMEX phantom series *stwstrategyps4* – the first three sources were previously described in Aerts et al [1]. The phantom data can be accessed at https://xnat.bmia.nl/. The first closed source of data was the LUNG2 non-small cell lung cancer dataset, collected at Radboud University Medical Center between February 2004 and October 2011. The LUNG2 dataset was deidentified and analyzed within Radboudumc using an internal desktop station. The other closed source of data, an FTG-PET non- small cell lung cancer dataset, was collected by Hasselt University as a part of the clinical trial NCT02024113 registered at clinicaltrials.gov.

[1] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, 2014, doi: 10.1038/ncomms5006.

PhD portfolio of Ivan Zhovannik

Department: **Radiation Oncology**
PhD period: **01/01/2018 – 30/11/2021**
PhD Supervisors: **Prof. J. Bussink, Prof. A.L.A.J. Dekker**
PhD Co-supervisor(s): **Dr R. Monshouwer, Dr R.R.R. Fijten**

Training activities	Hours
Courses	
- Radboudumc - Introduction day (2018)	6
- RIHS - Introduction course for PhD candidates (2018)	15
- RU - Nederlands A1 (2018)	140
- Radboudumc - How to find grants via Research Professional (2018)	2
- Radboudumc - How to write a medical scientific paper (2018)	4
- Radboudumc – Infografics (2018)	2
- MIPT - Basic Combinatorics (2018)	28
- RU - Scientific Writing for PhD candidates (2018-2019)	84
- RU - The Art of Presenting Science (2018)	42
- RU - Nederlands A2 (2018-2019)	140
- RU - Grant Writing and Presenting for Funding Committees (2018)	18
- Radboudumc - Scientific integrity (2019)	28
- UMCU - Deep Learning in radiation therapy (2019)	8
- DeepLearning.ai - Deep Learning Specialization (2018-2019)	80
- RU - Effective Writing Strategies (2019)	75
Seminars	
- UMCU - radiotherapy research day – two oral presentations (2019)	8
- ESTRO physics workshop – section discussion (2019)	8
- ErasmusMC - radiotherapy research day – oral presentation (2020)	8
Conferences	
- Radiomics 2018 – oral presentation (2018)	20
- RIHS PhD retreat – oral presentation (2018)	20
- ESTRO38 – poster presentation (2019)	36
- ICCR2019 – two poster presentations (2019)	80
- Medical Imaging Symposium for PhD students (MISP) – oral presentation / best presenter award (2019)	8
- ASTRO2019 – oral presentation (2019)	80
- CMIMI2019 – poster presentation (2019)	20
- Radiomics 2019 – oral presentation (2019)	20
- Health-RI 2020 – student's oral presentation and software demonstration (2020)	8
- NACP2021 – oral presentation (2021)	20
- ESTRO2021 – oral presentation (2021)	80
- SIIM22 – oral presentation (2022)	80
Other	
- GoDataDriven Hackathon - Anomaly Detection on 4D Heart Scans (2018)	9
- GoDataDriven Hackathon - Scoliosis detection - 2/5 place (2019)	9
- SIIM - Kaggle Pneumothrax Competion – bronze medal / top 8% (2019)	120
- Award - ESTRO travel grant (2019)	
- Award - ZonMW translational research grant (2019)	
Teaching activities	
Supervision of internships / bachelor's or master's theses	
- Radboudumc - Supervision of 4 biomedical science students (2018)	20
- Radboudumc - Supervision of 2 software engineering BASc students (2019-2020)	80
- Maastro - Supervision of 4 data science MSc students (2019-2021)	456
Total	**1862**

Curriculum Vitae

Ivan Zhovannik was born on July 15th, 1992, in the research town of Arzamas-16 (now Sarov) in the family of two physics students of the Moscow Engineering Physics Institute – Yekaterina Kondrakhina and Yevgeny Zhovannik. He grew up as a curious child who constantly asked questions and liked reading books. Although Ivan did not study particularly well in middle school, he managed to secure a high school gold medal (magna cum laude analog) and win high school Olympiads on various subjects on the regional Moscow level. In 2010, Ivan was enrolled in the Biomedical Engineering diploma program of the Bauman Moscow State Technical University. Interestingly, the choice of the university was firstly influenced by the fact Ivan's friend from a calisthenics club was enrolled in the same university and not the university status itself.

In September 2015, Ivan spent one semester at the Czech Technical University in Prague on an exchange program. This trip was pivotal in his decision to continue as a PhD student, and he has accepted an opportunity to enroll in Radboud University in January 2018. The joint Maastro-Radboudumc PhD program was focused on data mining and machine learning applications in radiation oncology and had multiple collaborators around the world. During his PhD under the supervision of Prof. Jan Bussink, Prof. Andre Dekker, Dr. René Monshouwer, and Dr. Rianne Fijten, Ivan committed to research in the field of image biomarker analysis, performed several collaboration trips, and gave various conference talks. Since November 2021, he joined the Netherlands Cancer Institute as a researcher in Jan-Jakob Sonke's Adaptive Radiotherapy group.

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their unwavering support and guidance throughout my PhD journey, and I am deeply grateful for everything they have done for me. **René**, my daily supervisor, gave me something precious to any human being – his time. He always provided me with technical support, knowledge, and mental support during my work and life's hardships. René was always out there to help me! His expertise and encouragement have been invaluable to me.

As a freshman in my third month, I was given an opportunity to travel and work with our project partners at Moffitt Cancer Center in Florida. This would not have been possible without **André** as my supervisor. He always gave me opportunities and supported many of my undertakings. He helped me and Albi to hire and supervise wonderful machine learning researchers and engineers (then MSc students and my colleagues) Ibro, Suraj, Chinmay, Alessio, Julian, and others. I always felt I could rely on his support and talk to him honestly. His guidance and mentorship have been invaluable, and I am deeply grateful for everything he has done for me.

I would like to extend my heartfelt gratitude to **Jan**, a medical doctor, professor, and my university supervisor. His expertise in clinical oncology and science has been invaluable support throughout my PhD journey. Jan has always provided valuable guidance in making my research clinically relevant; fun fact: before pushing any new research idea, René asked me to first convince Jan, as a medical specialist. It was a pleasure to brainstorm and discuss my ideas with him. I am deeply grateful for his mentorship and support throughout my research.

I would like to express my gratitude to **Rianne**, my daily supervisor at Maastro. Despite not working on the same research topic, her guidance on how to organize my workflow, manage my time, and prevent burnout was invaluable. I very much appreciate that Rianne never intervened without need in my work and research, giving me much freedom to create new research findings. Her guidance and support have been crucial in my successful completion of my PhD.

My deep appreciation and gratitude to my colleagues from **Maastro clinic** and **Maastricht University**, including Clinical Data Science, Physics, Physics Innovation, Precision Medicine, Physicians, Physicists, RTTs, ICT, HR, and others. It was an incredible journey working together, and I have gained so much from our collaboration. The knowledge, support and encouragement I received from each and every one of them has been invaluable. It was a privilege to work with such a talented, dedicated and supportive group of individuals.

I am grateful to my colleagues at **Radboudumc**, including the Radboudumc graduation school, for providing me with the resources and support to complete my PhD. Special thanks to the Radboudumc Radiation Oncology department, including the ROI group, Physicists, my fellow PhD students Frank (who graduated in 2019) and Miguel, Physicians, RTTs, Technicians, and Administration, who have been instrumental in my research and education. Their expertise and dedication have been invaluable to me, and I am honored to have had the opportunity to work with you.

It would have been hard to conclude this work staying in the Netherlands, have not had I got the opportunity to join the group of **Jan-Jakob Sonke** at the **Netherlands Cancer Institute**. I want to express my gratitude to my collaborators and colleagues from all three research labs in RT, Physicists, RTT, Physicians, and Administration. I want to thank them for sharing their ideas with me, for helping me navigating my research problems, and for our cozy coffee and lunch breaks.

Alberto, Alessio, Albi, Chinmay, Carlotta, Dennis, Elli, Frank, Ibro, Jaileene, Julian, Lars, Leonard, Leonard, Mattea, Matt, Michal, Petros, Shamanie, Suraj, Talia, Timo, Yoga, Zhen, Zhenwei, Zhixiang. I would like to express my sincere gratitude to all of the above-mentioned individuals for their invaluable contributions and support during my PhD journey. Your guidance and collaboration were essential to the successful completion of my research. Thank you.

I would like to extend my heartfelt appreciation to all of my university friends and their families, particularly those who live close by - **Anya**, **Sergey**, **Kate**, **Natasha**, **Liza**, and **Valera**. Your support has been invaluable, you all helped my family to feel at home in the Netherlands, and I am grateful to have had the privilege of being part of the BMT (biomedical engineering) brotherhood. Thank you.

My gratitude extends to all the people I have met in my life, my friends, my students, the people I studied or worked with, even the strangers who helped me in a new town. Without our interactions, I would not be the person I am today, you all have become my teachers. I would like to express my appreciation for the role you have played in shaping me as a person and in my research. It is hard to mention everyone individually, but I thank you all for your support, guidance, and encouragement throughout the years. Your presence in my life has been invaluable and I am grateful for the lessons and memories we have shared.

My family, particularly my mom Katya and my dad Zhenya, thank you for raising me and investing your time and resources in my development. Your support has been a constant throughout my life and I am grateful for the sacrifices you have made for me. Being both students in early 1990s Moscow, you have not aborted me but instead decided to overcome all the challenges, united, and gave birth to me.

When my parents were young they got help from my grandma Zoya Kondrakhina - who is the great-grandma of Sophia now :), alongside my granddads Vladimir Kondrakhin (1944-2016), Viktor Zhovannik (1942-2015), and my great-grandma Ekaterina Golubeva (1915-2007). Their love, guidance, and support were essential in shaping who I am today. I am blessed to have such a supportive family and I am grateful for all that they have done for me.

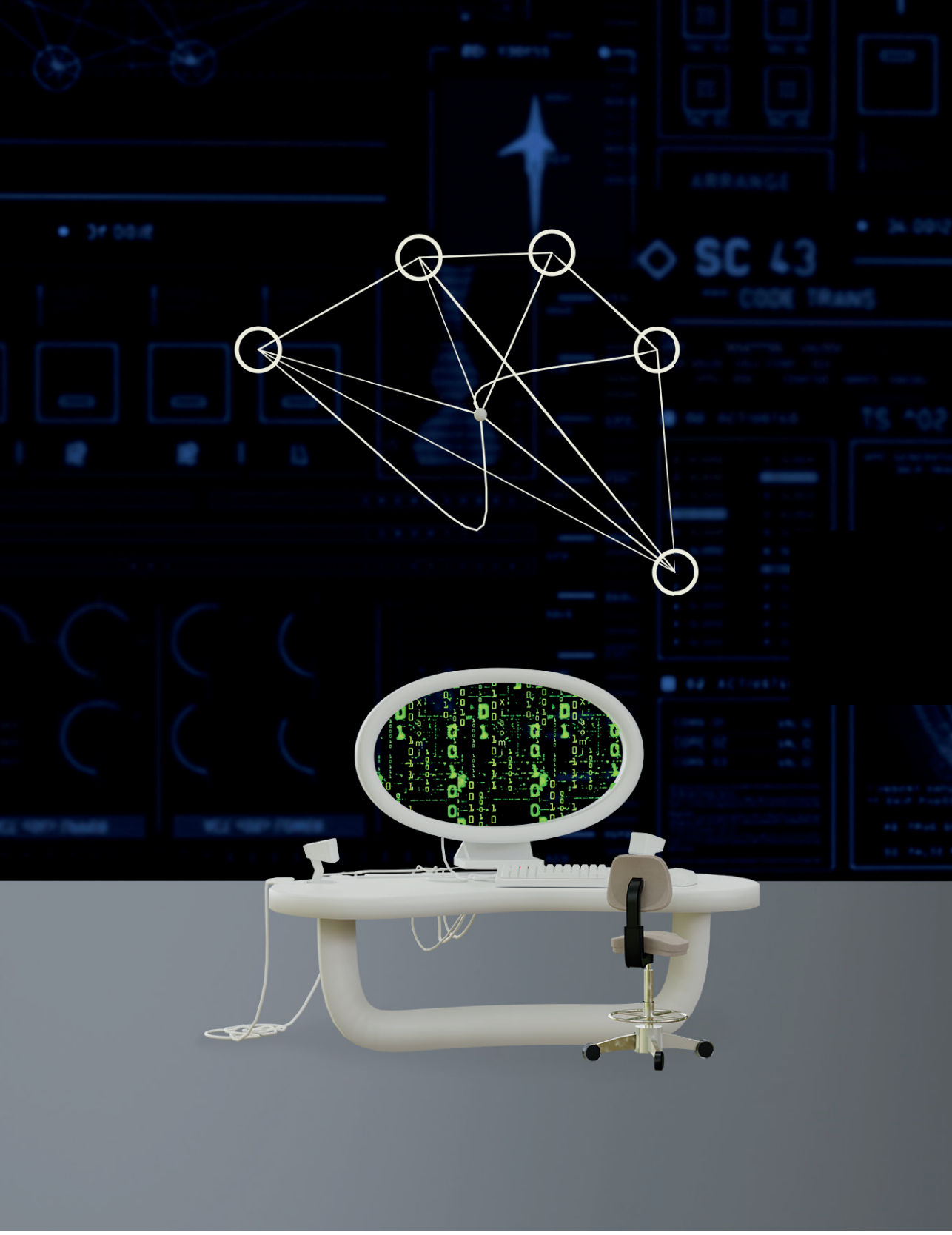
Моя семья, особенно моя мама Катя и мой папа Женя, я благодарен вам за то, что вырастили меня и вложили свое время и ресурсы в мое развитие. Ваша поддержка была постоянной на протяжении всей моей жизни, и я благодарен за жертвы, которые вы перенесли за меня. Будучи оба студентами в Москве начала 1990-х, вы не прервали беременность, а решили преодолеть все трудности, объединились и родили меня. Моим родителям помогали моя бабушка Зоя Кондрахина, которая сейчас стала прабабушкой Софии :), а также мои дедушки Владимир Кондрахин (1944-2016), Виктор Жованник (1942-2015) и моя прабабушка Екатерина Голубева (1915-2007). Их любовь, руководство и поддержка сыграли важную роль в формировании того, кем я являюсь сегодня. Я счастлив, что у меня есть такая поддерживающая семья, и я благодарен за все, что они сделали для меня.

My brothers Semyon and Stepan, my aunts and uncle and their families - thank you for being there for me. I am grateful for the role you have played in my growth and development as a person. Your presence in my life has been invaluable and I am grateful for the lessons and memories we have shared.

Мои братья Семен и Степан, мои тети и дядя и их семьи - спасибо вам за то, что вы со мной. Я благодарен за ту роль, которую вы сыграли в моем росте и развитии как личности. Ваше присутствие в моей жизни было бесценным, и я благодарен за уроки и общие воспоминания.

I would like to express my deepest gratitude to my parents-in-law, Irina and Dmitrii, and their families for giving me the greatest gift of all - my wife Natasha. Her love and support have been invaluable to me throughout my PhD journey. I am truly blessed to have her in my life. I want to thank the family of my step-father-in-law Henk, who sadly passed away in January of 2022, for considering us as the part of their family.

I am deeply grateful to my wife **Nataliya** for her unwavering support and encouragement throughout my journey to obtain my PhD. She has played a vital role in shaping me and helping me to grow as a person. Her open-mindedness and constant presence in my life have been invaluable. I also thank her for bringing our children into the world and for her efforts in raising them to become kind and compassionate individuals.



Radboudumc
university medical center

Radboud University