Quantitative imaging in radiation oncology

QUANTITATIVE IMAGING IN RADIATION ONCOLOGY

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. R.M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public
on Tuesday, 13th of April 2021, at 16.00 hours

by

Alberto Traverso

Promotor
Prof. dr. ir. A.L.A.J. Dekker

Copromotor
Dr. L.Y.L. Wee

Assessment Committee
Prof. dr. W. Backes (chair),
Dr. E. van Limbergen MAASTRO Clinic/Oncology Center
Maastricht UMC+, Maastricht (NL),
Prof. dr. C. Mayo, University of Michigan, (USA),
Dr. J. Teuwen, Radboud University Medical Center, Nijmegen

The work of this thesis has been funded through the NWO Perspective Strategy grant (number $P14 - 19.1.2$)

Cover Valeria Orsi, 2021

To my grandfather Livio, a great man willing to be little

Contents

1 Introduction

1.1 Medical imaging in oncology

Radiotherapy (RT) plays a pivotal role in the treatment of many cancers, considering that approximately 50 percent of all cancer patients can benefit from RT in the management of their disease [3]. RT is delivered using radiation produced by a linear accelerator (linac), optimizing the radiation to the tumour with the intent of killing the malignant cells, while preserving surrounding healthy tissues (referred to as OARs-organs at risk) and limiting the radiation-induced toxicity [10]. RT relies heavily on medical imaging to determine the extent of the disease, the spatial relation between target regions and neighbouring healthy tissues, the monitoring of RT delivery and subsequent follow-up to measure treatment effectiveness [2]. The RT workflow can be summarized into four distinct phases: diagnosis and staging, treatment planning, treatment delivery and post-treatment follow-up (Figure 1.1). Medical imaging intervenes in all of the shown phases. During diagnosis, patients' scans are acquired to identify the presence of suspicious cancerous lesions. These lesions are then classified using standard guidelines, such as the TNM classification of malignant

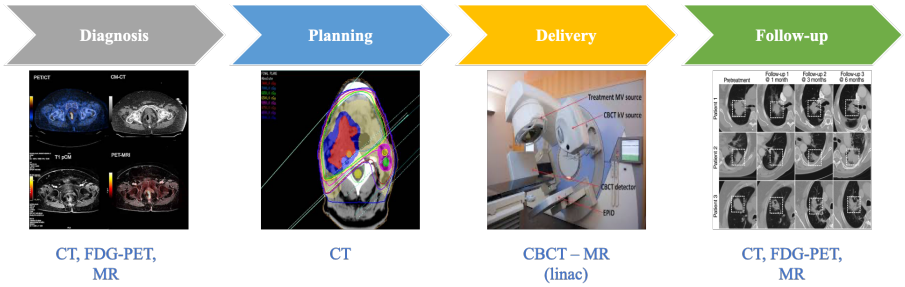


Figure 1.1: Overview of the major role of medical imaging in RT, from diagnosis to follow-up. Different imaging modalities are used in the different phases. Mostly commonly used are: CT (Computed Tomography), FDG-PET (Positron Emission Tomography), CB (Cone beam) CT, and MR (Magnetic Resonance).

tumours. The TNM is a globally recognized standard for classifying the extent of spread of cancer, based on visually identified properties [5]. Almost all solid cancers have their own TNM classification and according to the anatomical location of the lesions, different imaging modalities are employed for diagnosis staging, and molecular characterizations of the tumour. The most common modalities are CT (Computed Tomography), FDG-PET (Positron Emission Tomography) and MRI (Magnetic Resonance Imaging). Patients' scans acquired at this time-point are referred to as diagnostic scans. Diagnosis and staging, together with information about the patient (e.g. demographics, clinical data) is used to determine the treatment strategy. In RT, the treatment strategy involves the choice of the total amount of dose that target volumes (cancerous lesions) will receive, which will be delivered in smaller daily portions (fractions) over a period of several weeks, to take advantage of the higher repair capacity of normal tissues compared to tumour cells [10]. Additional clinical constraints determine the maximum amount of dose that OARs should receive, to reduce the risk of complication because of radiation. To plan the treatment, a scan of the patient is made around one week before the delivery of the first fraction. These images are referred to as treatment planning

scans. The image is used to manually delineate the target volumes and organ at risks. Treatment plans are always made on CT scans, because accurate determination of the dose due to ionising radiation is dependent on knowing the electron density spatial distribution. A treatment planning system is the software that is used to generate a radiation dosimetry plan using images, contours, and required clinical constraints. Treatment planning is an inverse optimization problem, with the goal of delivering the highest and most uniform possible dose to the target volumes, while reducing as much as possible the dose to surrounding healthy tissues. CBCT (Cone Beam CT) scanners integrated with linacs, are widely used in RT to capture the anatomy prior to treatment delivery. This facilitates tumour alignment to the original treatment planning position. CBCTs are also used to evaluate changes in patient anatomy such as tumour shrinkage or major changes in patient's anatomy, which might require a re-evaluation of the original treatment planning [16]. A recent development in RT has been the introduction of the MR-linac, in which MR is integrated with the linac and used for pre-treatment imaging. To monitor the control of the disease as well as to reduce the risk of disease relapse or the spread of tumor in other locations, follow up images are taken from a few months to up to years after treatment. Follow up times differ among cancer types and according to patients' need. Similar modalities as for the diagnosis and staging time-step are employed. The workflow described above emphasises the prominent role of medical imaging in the RT continuum. Furthermore, it becomes clear how advances in medical imaging will improve RT, offering better chances for cure, decreased side effects and extension of indications. Advances in medical imaging can be twofold: improvement of the hardware used to image the patients, and introduction of digital technologies to improve image analysis. Some of the most famous examples of the first category have been the introduction of CT scanners in RT, the integration of combined PET and CT imaging [18] and more recently the integration of combined PET/MR imaging [23]. Multi detector-row CT offered unparalleled speed of acquisition, spatial resolution, and anatomic coverage and they have provided the basis for IGRT (Image guided RT)

[17][21]. A recent application has been the integration of MRI devices with linacs in the treatment room, allowing a real time monitoring of the dose combined with large soft tissue contrast during irradiation (MR-linac) [25]. The second category of advances includes software applications related to medical imaging analysis. These applications can be aimed: A) at reducing the burden related to time consuming activities of clinicians and radiation technicians, such as OARs delineations or providing ultra-fast acquisition and image reconstruction; B) at augmenting our capability to interpret medical images besides visual inspection. The former is referred to as automation, while the latter has been named quantitative imaging.

1.2 The road to precision medicine

Both the presented advances will play a fundamental role in raising the bar of patient care in RT. There is an urgent demand in improving patient care towards precision medicine (PM), which is defined as the capability to tailor therapy with the best response and highest safety margin to ensure better patient care [8]. By enabling each patient to receive earlier diagnoses, risk assessments, and optimal treatments, PM is expected to improve health care while also lowering costs. Improvements toward PM in RT can can have an impact for ART (Adaptive RT). ART refers to the monitoring of treatment delivery across the fractions and the possible corrections that can be applied to the original plan [22]. In RT, ART strongly relies on patients' scans and it is often referred to as IGART (Image Guided ART), to stress the key role of medical imaging. The urgent demand of introducing PM in RT, requires us to augment the status of ART. Treatment strategy and possible adaptation should consider unique characteristics of the tumour of a patient. The tumour environment, as well as the anatomy of a patient, are dynamic and can change from diagnosis to and within fractions [4]. Objectively quantifying these changes during fractions will lead to better tailored treatments, as required by PM. To reach the

above-mentioned goal, there is the need of re-thinking medical imaging in RT. The new paradigm is to go beyond the traditional practice of treating medical images as pictures intended solely for visual interpretation. The advent of rapid high-throughput computing make it possible to extract quantitative descriptors, referred to as features, from tomographic images daily used in the RT workflow. This transition of patients' scans from digital medical images to mineable hyper dimensional data is known as 'radiomics' [13].

1.3 Re-thinking medical imaging: radiomics

The idea underlying radiomics is that medical images potentially embed information that reflects underlying biological properties of tumours, which are patient- and tumor-unique fingerprints. Being able to measure tumour biology via medical imaging is expected to help pave the road to precision medicine since it will improve tumour staging, treatment planning and monitoring. Although radiomics is a natural extension of CAD (Computer Aided Diagnosis) systems, there is a fundamental difference. CAD systems are meant to deliver a single answer (e.g. presence of a cancerous lesion), radiomics is a process aiming at extracting a vast amount (from hundreds to thousands) of quantitative features from digital images and subsequently mine the data for both hypothesis generation and testing. Radiomics by design is thought to develop decision support tools. Therefore, it requires to combine radiomic data with other patients' characteristics and information, as well as with genetic data extracted from the tumour ('radio-genomics'), generating an hyperspace of data which is much larger than the starting number of computed features [19]. Succinctly, radiomics has invited us to re-think medical imaging, translating the paradigm from visual inspection of medical images to quantitative precise measurements from medical images, shifting the attention from qualitative to quantitative image analysis. Although radiomics can be applied to several diseases and multiple conditions, it has reached its largest maturity in radiation oncology [11]. The strong

driver of this relies on the fact that RT had a high volume of very standardized (treatment planning) CT scans which were pre-delineated to identify the ROI. As we presented earlier, all the patients with cancer undergo imaging at some point, but more often, they are imaged multiple times during their care. Radiomics comes as a "cost-free" approach, since it is based on daily produced data from clinical care. With visual inspection only, images are analysed focusing on providing a (semi)qualitative evaluation of radiological findings and then discarded. Wasting the possibility of recovering precious information that can support better cancer detection, diagnosis, assessment of prognosis, prediction of treatment response and following disease status is a situation that collides with the aims of the previously presented concept of PM. Furthermore, radiomics opens the door for the development of non-invasive cancer-related biomarkers compared for example to genomics with the additional advantage that medical images offer to extract these biomarkers from the entire tumour (or tumours) rather than just from a sample. One of the most important concepts behind radiomics is that identifying biological properties of the tumour from medical imaging is a task that cannot be completed only via human visual inspection. As such, the word radiomics (or more in general quantitative imaging) cannot be separated from AI (Artificial Intelligence). AI is driven by machine learning (ML), a method of data analysis that automates analytical model building [1]. AI via ML is therefore the tool to fully exploit patients' daily scans as a valuable source of patient-centred quantitative information beyond human capabilities. Advances in ML applied to medical images is expected to boost our ability to introduce radiomics in the clinic as decision support tools. The typical radiomic workflow is depicted in Figure 1.2. Radiomics starts with the acquisition of images from patients. From these images a ROI (Region of Interest) containing either the whole tumour or sub regions within the tumour or in the surrounding tissues are identified. Most common regions include GTVs (Gross Tumour Volume), CTVs (Clinical Target Volume), but also, they extend to OARs for the evaluation of risks of radiation-induced side effects. Quantitative descriptors, called features, are then extracted from these regions.

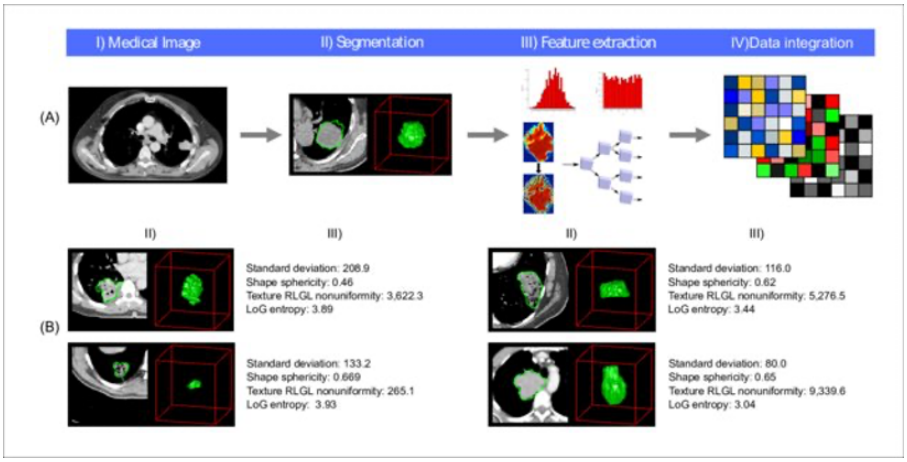


Figure 1.2: Radiomics approach depicted in the use case of lung cancer. (A) Workflow of extracting radiomic features: (I) A lung tumour is scanned in multiple slices. (II) Next, the tumour is delineated in every slice and validated by an experienced physician. This allows creation of a 3D representation of the tumour outlining phenotypic differences of tumours. (III) Radiomic features are extracted from this 3D mask, and (IV) integrated with genomic and clinical data. (B) Representative examples of lung cancer tumours. Visual and nonvisual differences in tumour shape and texture between patients can be objectively defined by radiomics features, such as entropy of voxel intensity values ('How heterogeneous is the tumour?') or sphericity of the tumour ('How round is the tumour?'). This image is taken from DOI: 10.7554/eLife.23421, under the license agreement CC BY 4.0.

These descriptors can be hand crafted features defined by mathematical formulas, or automatically determined by ML algorithms. The first type includes three main categories: shape descriptors looking at morphological properties (e.g. elongation, size) of the ROIs, first order statistical features describing the distribution of individual voxels without concerns related to their spatial distributions (e.g. the mean of grey level values within a selected ROI), and second order statistical descriptors, often referred to as 'textures', describing statistical interrelationships between voxels with similar or dissimilar contrast values. Textures are meant to provide a measure of intra-tumoral heterogeneity. The second type is mainly represented by DL (Deep Learning). Deep Learning (DL) has revolutionized medical image analysis and has essentially replaced all the older techniques. The idea behind DL is intuitive: building large (*deep*) artificial neural networks that can automatically mine the images and extract relevant features, referred to as *deep radiomics*, without the need of pre-defining them [14]. Extracted features are then combined using state of the art ML techniques and correlated with the outcome of interest to build diagnostic, prognostic, and predictive models. It is important for the reader to understand that radiomics involves several steps logically concatenated after each other, where feature values and ML only appear in the last steps of this process. Uncertainties in the early steps of this chain or inconsistencies in the input data (images and ROIs) will strongly impact the quality of the output. In fact, it has been recently shown how ML algorithms are very sensitive to the quality of input data, with lower quality meaning worse performances in the models [9]. Furthermore, it is of utmost importance to highlight that a robust and consistent methodology for feature computations is needed. A *small* fine tuning of the computational parameters can lead to conclusions that might not be consistent with previous results achieved under different settings. Robust and reliable radiomics, which can then produce decision support systems to be used in the clinic, relies on an absolute and meticulous effort in optimizing each of the computational steps in the chain shown in Figure 1.2.

1.4 Issues in radiomics

Looking at the radiomics literature, it emerges that this issue has not been tackled sufficiently enough for a rapid translation of all published radiomics models in the clinic. In the past ten years, radiomic research in tomographic imaging has dramatically increased. Radiomics has been deeply applied for A) enabling diagnosis allowing differentiation of cancerous from non-cancerous tissues as well as a quantification of tumour heterogeneity, B) tumour prognostication by showing relations between quantitative imaging features and gene expressions or clinical outcomes of interest (mainly overall survival or disease free survival), C) identification of imaging phenotypes that could guide the selection of therapy for individual tumours, and D) the development of imaging features to assess tumour response to treatment beyond the assessment merely based on shrinkages in tumour volume and the RECIST (Response Evaluation Criteria in Solid Tumours) [12][20]. This tremendous effort of scientific discoveries has been accompanied in the latest years by the need to investigate the reasons behind the previously mentioned unmet clinical need. However, radiomic literature still is imbalanced towards the idea of publishing radiomic models or positive discoveries, while being more reluctant to publish negative results or deeply investigating open methodological questions behind these discoveries. In all scientific discoveries there is always an excitement phase when a new technology is released, followed by a period of benchmarking, until it becomes more mature. Questions that arise during this period are part of science, since they offer the occasion to reconsider the weaknesses of a new technology and the unique perspective to build a better one by listening to all the stakeholders involved. In the radiomic community, we often forget that the primary stakeholders of our models are clinicians and (indirectly) patients. The presented clinical unmet need has brought a sense of frustration in the clinic with respect to AI and more specifically to radiomics [15]. But focusing only on accepting this frustration will never solve the problem. This is the time to raise the bar of radiomic studies to finally move from prototyping and academic exercise to a breakthrough application

with a meaningful clinical impact. The first step towards reaching the above goal is to deeply identify and investigate the methodological issues that have caused radiomics translation in the clinic to fail. This thesis poses itself in this context: highlighting and investigating these causes and proposing methods to mitigate these issues. The thesis focuses on three main issues identified in radiomic studies: A) lack of robustness of radiomic-derived biomarkers; B) methodological issues associated with a non-cautious use of ML in radiomics, and C) the lack of standardization and harmonization in radiomic studies.

1.5 Strategies to address the issues and structure of the thesis

The first issue is related to the fact that many radiomic models are extensively developed using only single-institutional data, but their performances degrade when validated on un-seen external datasets consisting of images often acquired with different scanner manufacturers or acquisition protocols [26]. Conversely, the need of generalizability of the models as well as the recommendations from the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) are calling for radiomic-models validated across several multi-centre datasets (TRIPOD-IV type models) [7]. Furthermore, if a radiomic feature (or a combination of them) is meant to measure a specific biological property of the tumour (i.e. biomarker) then this property should be stable and independent from the protocols used to acquire images. The presented problem is strongly related to the concepts of reproducibility and repeatability, which will be deeply investigated in the first part of this thesis. The second issue is related to the fact that radiomic prognostic and predictive models do not stand in a vacuum. Radiomic models need to be benchmarked against accepted clinical prognostic and predictive factors. An example of these is the TNM staging system. Radiomic models should show that they build upon previously published models, bringing new insights and

additional knowledge. Also, because of the complexity of radiomic feature definitions, hidden relations between radiomic features and clinical prognostic / predictive factors may exist and need to be discovered. In fact, introducing many correlated covariates in a ML classifier leads to the undesirable problem of over-fitting [6]. This problem as well as proposed solutions to use ML to benchmark radiomic models will be presented in the second part of this thesis. The third issue relates to the fact that radiomics lacks standardization and harmonization. Since the advent of radiomics, each institution involved in this research topic has developed its own radiomic computational package (i.e. software) or radiomic computational pipeline. Because of standalone naming conventions, it is impossible to compare and even perform radiomic experiments within multiple institutions. Furthermore, privacy-related issues and barriers to data sharing are a huge obstacle that require a re-think of the standard approach of centralized learning. Following the inspiring work carried out by the IBSI (Image Biomarker Standardization Initiative), the third part of this thesis focusing on a building a framework based on ontologies and semantic web technologies to allow the transition from radiomics to FAIR (Findable Accessible Interoperable Reusable) radiomics, as key enabler for multi-centre radiomic studies [24]. The overall structure of the thesis is as follows: Chapter 2 offers a broad overview of the concepts briefly mentioned in this introduction, with a dedicated focus on the current challenges and unmet clinical needs in radiomics. These challenges are then tackled in the work of this thesis. Chapter 3 provides a systematic review of the concept of radiomic reproducibility and repeatability identifying unsolved issues, such as for example the lack of robust methodology for radiomics in MRI. Therefore, Chapters 4-6 focus on investigating radiomic reproducibility and repeatability in MRI; while Chapter 7 proposes a method for radiomic harmonization in CT, which can be expanded to any other imaging modality. This first part of the thesis tackles the issue related to the need of improving the radiomic workflow steps related to image quality, image acquisition settings and lack of robustness of radiomic features. Chapters 8 and 9 are

related to the second aim of this thesis and they deeply investigate the role of ML in radiomic providing safeguards for responsible AI in radiomics. The results of the second part of this thesis contribute to solve the issue of optimizing the modelling part in the radiomic workflow presented in this introduction, as well as to reduce the risk of false discoveries. Chapters 10-13 investigate the third issue mentioned and offer a basis for FAIR quantitative imaging. The results of this last part of the thesis contributed to solve the presented issue of transparency in radiomic studies as well as proposing a solution for multi-centre radiomic studies. Finally, Chapter 14 provides a vision for the upcoming years of quantitative imaging research by engaging multiple stakeholders involved in the clinic. This last chapter offers working statements to re-think radiomics, not as a vacuum, but posing it in the context of big clinical data and arguing for the need of close collaboration with clinicians.

Bibliography

[1] null Amisha, Paras Malik, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, 8(7):2328–2331, July 2019.

[2] Mostafa Analoui, Joseph D Bronzino, and Donald R Peterson. *Medical imaging: principles and practices*. CRC Press, 2012.

[3] Michael B Barton, Michael Frommer, and Jesmin Shafiq. Role of radiotherapy in cancer control in low-income and middle-income countries. *The lancet oncology*, 7(7):584–595, 2006.

[4] Michael Baumann, Mechthild Krause, Jens Overgaard, Jürgen Debus, Søren M. Bentzen, Juliane Daartz, Christian Richter, Daniel Zips, and Thomas Bortfeld. Radiation oncology in the era of precision medicine. *Nature Reviews Cancer*, 16(4):234–249, April 2016.

[5] James Brierley, Mary Gospodarowicz, and Brian O'Sullivan. The principles of cancer staging. *Ecancermedicalscience*, 10:ed61, 2016.

[6] Anastasia Chalkidou, Michael J. O'Doherty, and Paul K. Marsden. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PLOS ONE*, 10(5):e0124165, May 2015.

[7] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*, 13(1):1, 2015.

[8] Michele De Palma and Douglas Hanahan. The biology of personalized cancer medicine: Facing individual complexities underlying hallmark capabilities. *Molecular Oncology*, 6(2):111–127, April 2012.

[9] Timo M. Deist, Frank J. W. M. Dankers, Gilmer Valdes, Robin Wijsman, I-Chow Hsu, Cary Oberije, Tim Lustberg, Johan Soest, Frank Hoebers, Arthur Jochems, Issam El Naqa, Leonard Wee, Olivier Morin, David R. Raleigh, Wouter Bots, Johannes H. Kaanders, José Belderbos, Margriet Kwint, Timothy Solberg, René Monshouwer, Johan Bussink, Andre Dekker, and Philippe Lambin. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics*, 45(7):3449–3459, July 2018.

[10] Elin Evans and John Staffurth. Principles of cancer treatment by radiotherapy. *Surgery (Oxford)*, 36(3):111–116, March 2018.

[11] I. Gardin, V. Grégoire, D. Gibon, H. Kirisli, D. Pasquier, J. Thariat, and P. Vera. Radiomics: Principles and radiotherapy applications. *Critical Reviews in Oncology/Hematology*, 138:44–50, June 2019.

[12] I. Gardin, V. Grégoire, D. Gibon, H. Kirisli, D. Pasquier, J. Thariat, and P. Vera. Radiomics: Principles and radiotherapy applications. *Critical Reviews in Oncology/Hematology*, 138:44–50, June 2019.

[13] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[14] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers. Guest Editorial Deep Learning in Medical Imaging: Overview

and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, May 2016.

[15] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36, January 2019.

[16] David Jaffray, Patrick Kupelian, Toufik Djemil, and Roger M Macklis. Review of image-guided radiation therapy. *Expert Review of Anticancer Therapy*, 7(1):89–103, January 2007.

[17] David A Jaffray. Image-guided radiotherapy: from current concept to future perspectives. *Nature Reviews Clinical Oncology*, 9(12):688, 2012.

[18] Vibhu Kapoor, Barry M McCook, and Frank S Torok. An introduction to pet-ct imaging. *Radiographics*, 24(2):523–543, 2004.

[19] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, November 2012.

[20] J. O. Park. Measuring Response in Solid Tumors: Comparison of RECIST and WHO Response Criteria. *Japanese Journal of Clinical Oncology*, 33(10):533–537, October 2003.

[21] Geoffrey D Rubin. Data explosion: the challenge of multidetector-row CT. *European Journal of Radiology*, 36(2):74–80, November 2000.

[22] Jan-Jakob Sonke and José Belderbos. Adaptive Radiotherapy for Lung Cancer. *Seminars in Radiation Oncology*, 20(2):94–106, April 2010.

[23] Drew A Torigian, Habib Zaidi, Thomas C Kwee, Babak Saboury, Jayaram K Udupa, Zang-Hee Cho, and Abass Alavi. Pet/mr imaging: technical aspects and potential clinical applications. *Radiology*, 267(1):26–44, 2013.

[24] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[25] Dennis Winkel, Gijsbert H. Bol, Petra S. Kroon, Bram van Asselen, Sara S. Hackett, Anita M. Werensteijn-Honingh, Martijn P.W. Intven, Wietse S.C. Eppinga, Rob H.N. Tijssen, Linda G.W. Kerkmeijer, Hans C.J. de Boer, Stella Mook, Gert J. Meijer, Jochem Hes, Mirjam Willemsen-Bosman, Eline N. de Groot-van Breugel, Ina M. Jürgenliemk-Schulz, and Bas W. Raaymakers. Adaptive radiotherapy: The Elekta Unity MR-linac concept. *Clinical and Translational Radiation Oncology*, 18:54–59, September 2019.

[26] Alex Zwanenburg. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(13):2638–2655, December 2019.

2

Quantitative radiomics in Radiation Oncology

Adapted from: **Quantitative radiomics in Radiation Oncology**. ML Welch, A Traverso, C Chung, and DA Jaffray. Book chapter in "The modern technology of radiation oncology", Volume 4. Edited by Jacob van Dyk. Medical Physics Publishing. ISBN 978-1-951134-03-7. (2020). Contribution: shared first authorship.

Abstract

Radiomics utilizes routine clinical imaging data for non-invasive quantification of tumour phenotypes. The ultimate goal is to use these features for prediction or prognostication of patient events or disease types. Radiomics pipelines share parallels with traditional quantitative imaging processes. This provides opportunities to align the processes and integrate radiomic features and QI metrics to advance the investigation of related mechanisms between radiomic features and QI metrics, thereby improving interpretability of radiomic features. Engagement of clinicians in radiomic studies is essential for progression of the field. It has shown potential; however, clinician involvement and comparison of performance to clinical standards is required to evaluate clinical relevance. Many radiomic features were designed for quantification of images and data not related to medical images. Therefore, caution is warranted when drawing conclusions about correlations between radiomic, clinical, and genomic features. Statistical knowledge, or collaboration with biostatisticians, is required during feature agglomeration, selection and model building. Extensive reporting of methodology is needed to safeguard against spurious results, as well as to increase understanding of the impacts of data and user biases. Data sharing, methodological refinement, and standardization is needed for radiomics to meet its full potential. These techniques are generalized methods from past pattern recognition research. It is, therefore, foreseeable that these methods could be applied to other data not previously considered for automated feature information generation.

2.1 INTRODUCTION

The importance of computer-aided diagnosis (CAD) and decision support systems for improved medical care was recognized almost 50 years ago with the influx of medical imaging data. This need is being highlighted by the rapid technological advancements in modern medical imaging, which enable precise, detailed, and quantitative images that are easier to collect before, during, and following treatment. The magnitude of imaging data collected per patient—combined with other confounding factors (disease staging, blood tests, genomic data, patient preferences, quality of life, etc.)—leads to a complex task that pushes the human cognitive capacity to its limit [1], but also has the potential to enable an elevated level of personalized care. As stated in [148], "radiomics involves the automatic conversion of medical imaging data to quantifiable features that can be mined in order to provide additional information that may assist with personalized medicine approaches" [157][45][26][147] (Figure 2.1). Radiomics is a field that is experiencing increased interest from both clinicians and researchers. Given the advancements of artificial intelligence (AI) methods for pattern recognition, computer vision, and model building, the utilization of radiomics for diagnostics, prognostics, and treatment decision processes is more promising than ever. By utilizing features that not only quantify information regarding the whole tumour, but also the various textures contained within the tumour, radiomics may serve as a useful method for quantification of disease heterogeneity [45][72]. This would provide an additional non-invasive approach for characterizing disease heterogeneity in addition to traditional quantitative imaging approaches and invasive targeted biopsies of sub-regions in tumours. Many of these concepts were introduced back in the 1970s by Hall et al. [50] and [55]. In these works, they described the use of extracted features, pattern recognition, and computer vision for radiographic image classification. These methods went on to demonstrate successfully that quantified imaging features could be correlated with tumour grade [157], histopathology [30], and treatment response

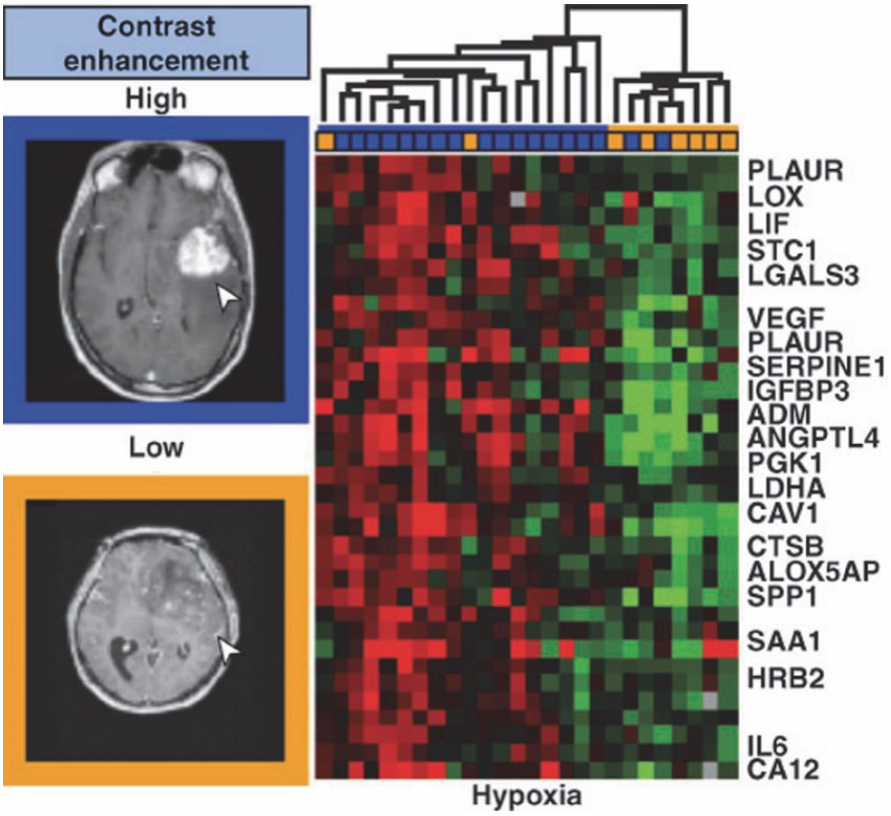


Figure 2.1: Radiomics introduces imaging signals or "image-based biomarkers" into the bioinformatics framework that was developed for genetic profiling and correlations with outcome. Imaging brings distinct spatial information—such as size, invasiveness and texture— to this framework and can be measured multiple times during the course of therapy. Diehn et al. (2008) examined MR images for gene-expression surrogates. Expert radiologists defined 10 MR image signal phenotypes, and hierarchical clustering of the gene expression profiles of 32 samples was performed and tested for statistical significance. Shown here are results specific to contrast accumulation in the brain and hypoxia. The box colour above the expression map corresponds to image traits for the different gene expressions in the tumours.

[108], but successful widespread integration into clinics has still not been achieved. With increasing computational power and automation, in conjunction with quantitative imaging (QI)—which is the extraction of quantifiable image-derived metrics from medical images that are associated with anatomical, physiological, and biological processes [66][127] —there have been substantial advances in this arena with image feature quantification re-branded as "radiomics" in 2012 [73]. Since the 1940s and 1950s, cancer staging using TNM classification was introduced by Pierre Denoix and the Union for International Cancer Control (UICC) [11] to help prognosticate and guide management of cancer patients. Rigorous and exhaustive clinical validation has been required prior to any changes or adjustments in these internationally accepted staging standards [97]. Although radiomic models have shown promise in providing added prognostic capability, the large majority of current radiomic models have not yet provided sufficient evidence of generalizability and accuracy to warrant clinical adoption or implementation as an international standard [148] [163]. A key limiting factor currently in radiomics is the lack of standardized quantitative approaches that hinders external validation and broader clinical applicability. Only through recognizing of the impact of heterogeneous input imaging data and developing approaches to address heterogeneities across images will the full potential of radiomics be realized. Also, the urgent need of a broader radiomic community, strongly bonded with the medical and QI efforts worldwide, is crucial to drive radiomics toward a product that can robustly support clinicians in patient care. This chapter will introduce the concepts of radiomics. The goal of this chapter is to familiarize readers to the ground-breaking research occurring in this field across different imaging modalities, while highlighting important concepts, challenges, and implications of QI, including the commonalities and differences of the QI and radiomic pipelines in the model building workflow. Additionally, the last part of this chapter discusses practical considerations that will enable reproducible results, such as software, data quantity and quality, contouring, and model reporting. As radiomics is a relatively new field for which

the potential clinical impact is just beginning to be understood, it is important for researchers to be diligent in their methodology and reporting to safeguard against spurious results. The authors' aim for the information included in this chapter is to prevent radiomics from being classified as *yet-another-omics* [147].

2.1.1 State of the art radiomics

Radiomics in computed tomography imaging

Interest in the application of radiomics was first spurred in analysis of computed tomography (CT) images [26]. The main idea behind this application was to provide quantitative information related to a particular region of interest (ROI) and the standard Hounsfield Unit (HU) values contained within the ROI. Radiomics in CT has largely been focused on head and neck (HN) and lung cancers. One of the most comprehensive studies investigating the role of radiomics in CT was published in 2014 [2]. The authors showed that textural and first-order statistical features were correlated with overall survival of HN and lung cancer patients. Following this study, different clinical end points besides overall survival were exploited. In particular, a study from 2015 [99] investigated relationships between radiomic features extracted from CT scans of HN and lung cancer patients, showing strongest associations for prognosis, histology, and staging. Other studies focused on predicting the probability of distant metastases in lung cancers [23], HN cancers[2], or investigated the association between radiomics and local regional failure [159]. Most recently, additional studies have investigated the role of radiomics in CT for oesophageal cancer [75][114] and liver metastases from other primary diseases [70][3]. There has also been recent interest in using cone beam CT (CBCT) images acquired at the end of each radiation treatment fraction to evaluate the pathological response to the treatment [5] or survival [143]. This is an area of research referred to as *delta-radiomics*, and it is discussed in more detail later in section

2.2.1. However, the CBCT images suffer from poor quantitative performance compared to CT, and this appears to limit the utility of radiomics in this particular modality [38]. Recent research into synthesized CT from CBCT using generative adversarial networks (GANs) [47] may provide an opportunity to address these challenges, as they have demonstrated improvements in voxel values, spatial uniformity, and artifact suppression [67]. However, tuning would be required for voxel-level analysis to ensure proper mimicking of details needed for radiomic analysis. Overall, it is important to note that daily acquired CBCT images are optimized for visual inspections by clinicians and not for automated pipelines (i.e., texture analysis). Parallel attention in CT radiomics has also been given to the investigation of reproducibility and repeatability of individual radiomic features. A recent review [138] analysed 22 studies explicitly investigating the reproducibility and repeatability of radiomic features in CT. Although there was no detailed consensus, first-order and shape CT features were generally more repeatable than textural features, with slice thickness re-sampling and different reconstruction algorithms strongly degrading feature reproducibility; the magnitude of this degradation was greater for textural features than for first-order features. This further emphasizes the need for QI as a pre-condition for robust radiomic-based predictions. In particular, it becomes clear that detailed knowledge of the acquisition parameters embedded in DICOM headers and surrogates for the physics underlying a specific modality is fundamental to guaranteeing the correct evaluation of reproducibility and repeatability results in radiomics. Again, this creates a need to form a broader community, with professionals from a variety of disciplines who have detailed knowledge of the computational details of the radiomic workflow, as well as QI knowledge. CT images are derived from the attenuation of x-ray radiation, and HU provide a measure of the linear attenuation coefficient of solid tissues relative to water. For this reason, CT is not likely to fully identify the presence of all texture variation associated with disease. For example, the contrast-to-noise might not be sufficient to capture differences in texture between different tissue

types, or to classify the state of aggressiveness of a tumour based on its heterogeneity. Due to these concerns, radiomic researchers have begun to look at different forms of CT imaging. Contrast Enhanced (CE) CT [3][160] imaging has recently been explored and is known to provide additional clinical benefit relative to non-contrasted CT by intravenous injection of an iodinated contrast agent just prior to imaging. Dual-energy CT (DECT) also improves tissue characterization and has been preliminarily studied for prognostics in lung cancer , characterization[19] of cervical lymphadenopathy [119], and distinguishing small-cell from non-small-cell lung cancer [154]. However, results between radiomic features computed in CT images might not be comparable with results achieved in CECT or DECT because of a different contrast-to-noise ratio, as well as differential agent accumulation at various time points post-injection.

Radiomics in magnetic resonance imaging

Compared to CT imaging, magnetic resonance imaging (MRI) provides greater soft tissue contrast [27], resulting in more consistent segmentations [145], as well as a means to measure physiological parameters and biochemical function of tissue. Besides the most common MRI techniques, such as T1 or T2 weighted images, diffusion-weighted imaging (DWI) in MRI enables measurement of water diffusivity via generation of apparent diffusion coefficient (ADC) maps, which is an established biomarker of tumour cell density and cell density changes post-therapy [76]. In addition, simple quantitative measurements of apparent diffusion coefficients (e.g., mean, standard deviation) extracted from diffusion-weighted MRI have shown potential prognostic value in cervical cancer, where MRI is the leading modality for staging and treatment evaluation [46]. It is, therefore, of interest to evaluate the role of radiomics as a complementary method to traditional statistical analysis approaches. In the literature, the main applications of radiomics in MRI refer to prostate and pelvic malignancies. For example, Stoyanova et al. [130]

highlighted the promising role of radiomic features combined with multiparametric MRI for measuring the aggressiveness of prostate tumours for risk stratification. This review shows how the majority of the studies are focused on the usage of radiomics for prostate cancer differentiations and correlations between images and histology. A similar approach was adopted for rectal cancer, where radiomic features were found to be correlated with tumour staging [132], while Nie et al. investigated associations between treatment outcomes and MRI-based radiomic features [92]. However, despite the effort that has been devoted to the investigation of the prognostic and predictive power of radiomics in MRI, the challenges of achieving quantitative performance of MRI raises several challenges that need to be addressed prior to pursuing clinical adoption. These include feature reproducibility challenges caused by inconsistent scanning protocols, image reconstruction processes, image post-processing, as well as absence of defined units associated with the signal in T1/T2 weighted sequences, which might open arguments about the validity of some texture feature analysis. This is further compounded for delta-radiomics as patients are often longitudinally scanned on different scanners, possibly at differing magnet field strengths and with differing imaging protocols. Unfortunately, there are few studies exploring the impact of these challenges on radiomic analysis in the literature. One study by Fiset et. al. investigated the reproducibility of radiomic features extracted from MRI images with respect to three different scenarios: (1) inter-observer variability in delineations, (2) test-retest stability, and (3) diagnostic vs. radiation oncology simulation MR scanners. Results revealed that different radiomic features are sensitive to various degrees based on the perturbations considered (i.e., no consensus was reached on the most stable features). In addition, the study revealed strong inter-dependencies between radiomic features, which should be considered when developing a radiomic signature [41]. Furthermore, it is common practice in MRI to pre-filter MR images prior to feature extraction. However, as pointed out in Traverso et al. (2019) [135], for ADC maps of rectal cancer patients, these image pre-processing steps

can alter feature values and can potentially degrade the underlying radiomic signature, alerting the need for further evaluation of the impact of image pre-processing on radiomic analysis. An interesting concept is to compare or even combine radiomic approaches with more traditional quantitative feature extraction that can occur using functional MRI (e.g., the volume transfer constant, Ktrans, which reflects the efflux rate of gadolinium contrast from blood plasma into the tissue and is perceived to reflect vascular permeability or average diffusion coefficient (ADC), which is associated with tumour cell density). As the quantitative parameters are usually determined using physical and biological models informed by MRI images, evaluating relationships between these parameters and radiomic features may reveal complementary information, as shown in [38]. As these traditional QI parameters share the same dependence on consistent image acquisition, reconstruction, post-pro cessing, and analysis, addressing these basic challenges will substantially advance the clinical implementation of imaging-based prognostic features.

Positron emission tomography and radiomics

Radiomics in positron emission tomography (PET) imaging has looked to supplement standard clinical measures that do not fully describe tumour heterogeneity. These measures include (1) conversion of PET-based voxel activity measurements to standard uptake values (SUV) and (2) metabolically active tumour volumes (MATV) that provide a single measure for the tumour and could be improved with measures of aggressiveness and metastatic potential. 18F-Fluorodeoxyglucose (18FFDG) PET, a measure of glucose cellular metabolism, has been the primary focus of PET radiomics. In a study by El Naqa et al [32], 18F-FDG PET images of 9 HN and 14 cervical cancer patients were successfully used to prove the potential utility of PET-based functional images for clinical prognostics. This study has motivated the development of standardized imaging methods that can assure consistent

inputs to assist in the challenges of PET radiomic analysis. PET images are challenging for radiomic analyses for many of the same reasons as other imaging modalities. These challenges arise because processes and imaging systems/parameters vary widely between institutions, and large voxel sizes, low contrast, and activity limiting signal-to-noise can result in weak features. Additionally, scanner type, injected activity, acquisition time after injection, acquisition time per bed position, attenuation correction with CT parameters, matrix sizes, and slice thickness are all potential sources of feature variation that can result from non-quantitative imaging practices found in PET imaging, as well as other modalities [57] [22]. In a study from 2017, Lovat et al. explored the impact of scan time after injection and found that first-order and texture features varied significantly between the two [80]. Reconstruction and post-reconstruction interpolation in clinical PET images also modulate features. This was demonstrated by Shiri et al., where 56% and 59% of image signals and texture features, respectively, were found to vary depending on the algorithms used [125]. Quantification of shape features in PET imaging is a challenging task as well, since there is no consensus on whether necrotic sections of the tumour should be included in contours and the subsequent quantification of the tumour phenotype. Additionally, although PET is one of the more quantitative imaging modalities, PET images are prone to nonlinearities and artifacts. Scattering events within the patient can displace the annihilation photon's line of response. This is particularly problematic in regions proximal to areas of high tracer accumulation (e.g., the bladder) and can result in substantial nonlinearities in the surrounding structures. The challenges of 18F-FDG have been recognized, and standardization protocols have been proposed. Suggestions have been made to improve quantization through proper consideration of blood glucose levels, image acquisition, reconstruction and uptake quantification, scanner quality control, and PET timing [7] [121]. Additionally, a recent study [96] looked at the potential of a post-reconstruction harmonization method to reduce multicentre differences. The method, ComBAT, was originally designed for genomic data and estimates the centre effect based on observed features. In Orlhac et al.'s work, they

successfully removed the multi-centre effect on textural and SUV features when differences in scanner brand, scan type (PET vs. PET/CT), and reconstruction protocols were present. This is important work that should be applied to all imaging modalities, but it is particularly important for PET imaging since data sets are often very small, requiring collaboration between institutions. Furthermore, biases generated by volume during PET radiomic studies have also been discovered. These discoveries have led to suggestions regarding the minimum size of a region of interest that should be considered during analysis. These size suggestions to reduce bias and increase complementary information have ranged from 10 cubic centimetres [56] to 45 cubic centimetres and require more analysis.

Radiomic phantoms

For ethical, safety, and logistic reasons, it is very challenging to test the influence of different acquisition protocols directly in patients. Imaging phantoms represent a useful tool to evaluate the reproducibility and repeatability of image acquisition and image signals, as well as the resulting radiomic features under different conditions. Investigators have used phantoms to study the influence of scanner manufacturers, slice thicknesses, and image reconstruction algorithms. In Traverso et al [138], the authors identified six phantom studies investigating the reproducibility of radiomic features. CT was the most common image modality (5 of 6), and PET was investigated in only one study. The consistency and quality of imaging data needs to be considered for proper evaluation and comparison of the performance of radiomic models. When images are obtained with different scanners, time points, and institutions, the imaging data is often inconsistently acquired, and the quality can vary widely. This is a similar challenge faced by traditional QI pipelines. For traditional QI, one approach taken to simulate and assess the quantitative impact of variances within the imaging data on the results of image analytic models has been the use of digital reference objects (DRO). Standard phantoms can help evaluate the

performance of imaging systems for particular imaging protocols, as shown for MR imaging when determining the stability of scans over time, and between subjects, sites, and vendors for the purposes of QI applications. The combined use of physical phantoms and DROs are being implemented to evaluate components that impact the quantitative capabilities across the QI pipeline. Similar practices in the generation of prospective data would assist in safeguarding radiomic studies against imaging variations that impact the repeatability and robustness of extracted features. To account for imaging variations in retrospective data, Zhovannik et al. used phantoms to evaluate and quantify the dependence of radiomic features with respect to CT technique (i.e., tube voltage and current) [162]. These dependencies were quantified to generate "calibration" curves that accounted for systematic uncertainties of radiomic features when computed with different settings and across different institutions. Phantoms could be similarly used to ensure consistent image acquisition and calibration. These techniques would provide a level of quality assurance that reflects the sources of errors across the overall imaging workflow. A full end-to-end radiomic phantom study would also provide excellent evaluation of an institution's ability to acquire consistent imaging that is adequate for meaningful radiomic analysis. However, one question about this type of work is how anatomy and simulated lesions (usually with inserts) produce textural features similar to values seen in patients. To the best of our knowledge, there is only one phantom that was specifically designed with different materials selected to simulate the textural distributions seen in human studies. The phantom was constructed of 10 layers (cartridges) representing different materials. It was scanned in different institutions, and the data was made publicly available [83]. It may also be possible to utilize more traditional quality assurance phantoms designed and used for scanner calibration and equipped with inserts dedicated to simulating lesions with uptakes similar to human studies [104]. Recent techniques, such as 3-D printing, can also create phantoms with inserts much more similar to lesions seen in human studies and will, therefore, enhance the usage and comparison of results from phantom studies to radiomic analyses in patient cohorts

[14].

2.1.2 Radiomic workflow and pipeline

Radiomic analysis and model development is a process involving a sequence of modular events (Figure 2.2). These events combine together to generate a pipeline that can be performed manually, semi-automatically, or fully automatically. Defining this pipeline is important to understanding the various aspects of radiomic analysis that can help or hinder the final results. It is traditionally thought of as a simple four-step process that involves (1) obtaining images, (2) contouring regions of interest, (3) extracting quantitative features, and (4) building a predictive or prognostic model. While it is true that these are the basic steps required for proper radiomic analysis, there are many details that need to be considered. Additionally, the radiomic pipeline shares many similarities to the traditional QI pipeline [59]. This provides opportunities to align processes and integrate radiomic features and QI metrics to advance the investigation of related mechanisms between radiomic features and QI metrics, thereby improving interpretability of radiomic features. This section of the chapter will introduce you to some of the details that require consideration during image acquisition, data preparation, feature extraction, and model building.

Image acquisition

In pursuit of personalized cancer care, quantitative imaging measurements have the potential to stratify patients based on prognosis, predict treatment response, and even spatially identify regions of higher risk to guide adaptive local therapies, such as radiation treatment. Image acquisition is the critical first step in generating the key data used for radiomic analysis or traditional quantitative image analysis. Although image post-processing methods are often applied to improve the consistency of the imaging

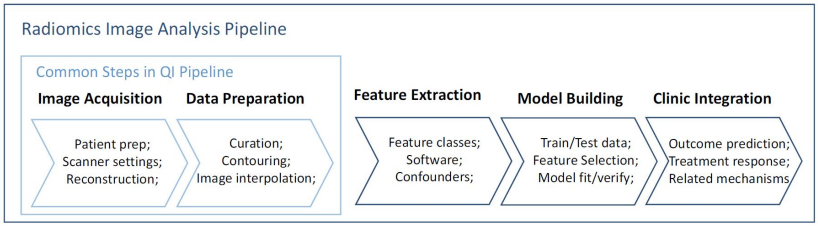


Figure 2.2: The radiomics image analysis pipeline is composed of a series of modular events designed to develop predictive models useful for clinical decision-making. Image acquisition, data preparation, feature extraction, model building, and clinical integration are the broad components of this pipeline. Within each of the pipeline modules there are many details that need to be considered to ensure positive progression of the field. Interestingly, this pipeline contains many of the same upstream components that make up a traditional quantitative imaging pipeline that focuses on measuring biophysical parameters on a ratio or interval scale. This provides opportunities for researchers to utilize the same input data in both a radiomics and traditional QI pipeline, thereby improving our ability to understand the underlying related mechanisms and, therefore, the driving mechanisms of selected radiomic signatures.

data used in the pipeline, the impact of variable image acquisition cannot entirely be corrected. Recognizing the importance of consistent image acquisition for quantitative image analysis, there have been a number of international efforts to establish guidelines to improve the repeatability and reproducibility of quantitative imaging measures including the Quantitative Imaging Biomarker Alliance (QIBA) of the Radiological Society of North America and the European Imaging Biomarkers Alliance (EIBALL). Quantitative imaging measures that have been addressed by these groups include: CT volumetry, MR biomarkers (diffusion weighted MR-DWI), dynamic contrast enhanced (DCE) MR, dynamic susceptibility-weighted MR), and PET. These groups aim to provide descriptions of image acquisition procedures with measured and characterized uncertainties based on test-retest data in the quantitative measures to guide the clinical implementation and appropriate interpretation of quantitative imaging biomarkers. However, a major common challenge between these quantitative imaging biomarkers and radiomics is the paucity of test-retest data to facilitate the implementation of quantitative approaches, as highlighted for DWI and DCE-MRI by Shukla-Dave et al. [128]. An additional challenge of generating quantitative radiomic models is the current practice of using conventionally acquired anatomical images, which generally have even less stringent oversight, even in the setting of clinical trials. The image acquisition protocol—including the selection of contrast agent, timing of contrast injection to image acquisition, and the image acquisition parameters—can impact the measured volume of the tumour.

Data preparation

The Imaging Biomarker Standardization Initiative [164] has provided the radiomic community with a set of guidelines for the standardization and harmonization of radiomic studies. In particular, the authors dedicated much attention to the computational steps that come prior to feature extraction, such as (a) contouring and segmentation of the

ROI, (b) image interpolation, and (c) image signal discretization. We will briefly describe the mentioned processes in the sections below.

Contouring and segmentation

The process of quantifying a patient image for prognostic and diagnostic analysis requires focusing on a specific region thought to be associated with the outcome of interest. Contouring and segmentation are used in this scenario to define the region of interest (ROI) that will be quantified using radiomic features. For application in radiation oncology, features are usually extracted from the gross tumour volume (GTV). The GTV is a convenient ROI to work with for radiomic studies because it is routinely contoured in clinical practice for radiation therapy (RT). Additionally, several machine-learning-based (ML) techniques have been proposed to automatically segment ROIs. This is beneficial because it permits previously un-contoured patient images to be used in studies, speeds up the contouring process, and reduces contour subjectivity. In one study by Lustberg et. al. [82], a deep learning algorithm that contoured the lung showed promising results when compared to existing solutions, and had marginal time required to "correct" them manually. The manual correction is there to ensure that an automatically generated ROI is still considered "clinically acceptable" and results in a semi-automated method. Semi-automated contouring methods were shown to be capable of reducing inter-observer variability and, therefore, producing a larger subset of reproducible features [100]. However, there has been a trend in the current literature to evaluate the accuracy of automated or semi-automated contours by looking at metrics such as DICE coefficient or Jaccard index [51], which may not be enough to state confidently that these contours are clinically relevant. To strengthen the conclusion that semi-automatically or automatically generated contours are clinically relevant, researchers could import machine-generated contours to treatment planning software and compare the dose-volume distributions obtained from original and machine-generated contours. Despite most radiomic studies

focusing on feature extraction from the GTV, recent papers have looked at using other ROIs. These ROIs can be alterations of a GTV, such as those used by Dou et al. , where the authors investigated radiomic features extracted from the peritumoral region, defined as an extension of 3 mm of the GTV [28]. They identified in this work that peritumoral radiomic features associated significantly with distant metastases in lung cancer patients. These results seem to suggest that additional quantitative information can be extracted from outside of the traditional GTV region, which may suggest that the traditional definitions of GTV may not encompass the full extent of tumour, or there may be additional radiomic changes in the tumour microenvironment. Other recent applications have also included the evaluation of radiomic features extracted from other ROIs, such as clinical target volumes (CTV) and organs at risk (OAR), showing promising results for predicting treatment response in breast [9] and glioblastoma [107]. However, limiting feature extraction OARs or target volumes may not be the optimal solution to predicting treatment side effects, such as toxicities. In fact, toxicity outcomes might be related to a sum of a set of features extracted from OARs, target volumes, and the morphological combinations of these two, such as the ROI defined as subtraction between the GTV/CTV and the anatomical primary site of the tumour (e.g., in lung, the volume of tissue of lung minus GTV/CTV). Most recent publications that use deep learning-based algorithms overcome the abovementioned problem by feeding the algorithm with the full set of slices of a patient's study, without any input ROI. For use with most radiomic platforms, the contour needs to be converted to a binary mask. Since in radiation oncology most of the contours are available as DICOM RTSRUCT files, an algorithm that converts the list points defining the contour to a binarized volume in the same coordinate system as the image is needed. The most common algorithm is ray casting. The ray casting algorithm belongs to the family of point in polygon problems of computational geometry, asking whether a given point in the plane lies inside, outside, or on the boundary of a polygon. The ray casting algorithm tests how many times a ray (e.g., the line connecting two points, starting from the point and going in any fixed direction) inter-

sects the edges of the polygon. If the point is on the outside of the polygon, the ray will intersect its edge an even number of times. If the point is on the inside of the polygon, then it will intersect the edge an odd number of times (Roth 1982). There is no specific recommendation on the particular choice of the algorithm; however, different algorithms can lead to different segmentations that can potentially affect the reproducibility of radiomic features [138]. For this reason, we suggest that researchers provide detailed explanations of the particular algorithm used for contour conversion. Best practice would be to use web-described, open-source platforms such as SimpleITK and the SlicerRT extension of 3-D Slicer [105], or to provide detailed descriptions of the algorithms upon publication. Although focusing our signal analysis in radiomics to an ROI may be beneficial for processing and directing our research, it is also a limiting factor. ROIs, such as GTVs and OARs, are generated for clinical usage and embed prior knowledge important for treatment planning purposes. This knowledge can involve past experience, adherence to institutional contouring guidelines, and uncertainties related to the treatment strategy (e.g. treatment margins to ensure coverage during treatment). It is, therefore, important to consider the inherent risk biases introduced into our analysis and conclusions as a result of ROI inclusion. Investigations on how to homogenize and reduce possible biases in contouring are on-going [129].

Image interpolation

The reconstructed resolution of an image is defined by the size of an individual pixel in each of the three dimensions, and this can vary with the imaging protocol used for image acquisition. To reduce image signal associated biases, pre-processing of images can be performed prior to feature quantification and extraction. In the specific case of texture features—which quantify local or distribution patterns of the image values, images should have harmonized resolution to avoid sampling a different underlying signal and a different number of image values

from one patient to the next [112]. That also makes extracted features invariant to different voxel dimensions, and it is used to guarantee reproducibility of features with respect to different voxel dimensions. The procedure of harmonizing the dimension of pixels (or voxels) to the same value is referred to as creating isotropic voxel representation. It should be noted that care should also be taken to harmonize the intrinsic frequency content of the underlying signal transfer, not just voxel dimensions, to assure consistent information flowing into the radiomic pipeline. Interpolation of voxels to a uniform size can be performed through up- or down-sampling. The choice of down-sampling (reducing to smaller voxel values) or up-sampling (increasing to larger voxel values) of the original images will alter the radiomic signatures. It has also been shown that re-sampling has an impact on feature reproducibility (Traverso et al. 2018). Different interpolation algorithms are used to map the original grid of voxel values to the interpolated grid. In such a grid, the voxels are spatially represented by their centre. Several algorithms are commonly used for interpolation, such as nearest neighbour, trilinear, tri-cubic convolution, and higher polynomials of the spline interpolation. In short, nearest neighbour interpolation assigns the signal value of the most nearby voxel in the original grid to each voxel in the interpolation grid. Alternatively, trilinear interpolation uses the signal values of the eight most nearby voxels in the original grid to calculate a new interpolated signal value using linear interpolation. Trilinear interpolation algorithm is a more conservative approach compared to the nearest neighbour algorithm since it does not lead to out-of-range signal values, which may occur due to overshoot with tri-cubic and higher-order interpolations. It is, of course, important to realize that these algorithms alter the underlying frequency content of the image signal. It is worth mentioning that the ROI mask will also require interpolation to match the dimensions of the newly interpolated image. To avoid possible partial volume effects (PVE), the nearest neighbour interpolation algorithm is suggested for ROI mask interpolation [164]. Additionally, the optimal algorithm for image interpolation might be modality or research-question-dependent, and care should be taken to evaluate the impact

of different re-sampling algorithms on the particular research question. However, caution is warranted, as the selection of the algorithms of interest will be biased by a particular user and put the research at risk of data leakage if proper holdout data sets are not used.

Discretization of image signal values

Image signal values are typically re-sampled to extract texture features. This can be a simple process, such as grouping or binning the values. In fact, texture matrices from which texture descriptors are computed are obtained by grouping together neighbouring voxels or regions of voxels. Furthermore, discretization is often employed to perform noise-suppressing prior to texture descriptor extraction, making the calculation more manageable, as explained by Yip and Aerts [156]. The procedure of image signal value discretization is usually referred as "quantization" or "binning." Two main quantization approaches exist: fixed bin width and fixed bin count. There is currently no evidence to prefer one method to the other, but some suggestions have been made [78]. When utilizing non-quantitative images with arbitrary units (e.g., MRI T1 or T2 weighted images for which the signal does not have a defined unit) there is a lack of linearity present in the signals. In these scenarios the Image Biomarker Standardisation Initiative (IBSI) recommends using a fixed bin count approach that breaks any possible relationships between the feature values and physiological meaning, but potentially introduces an intrinsic normalizing effect. However, recent publications focusing on radiomics in MRI have suggested the performance of a priori explicit normalization, e.g., normalizing the entire image against the values of a specific anatomical region to increase the stability of radiomic features. In the fixed bin size approach, a new bin is assigned for every signal value interval of fixed width, starting from a minimum value. The IBSI suggests assigning the minimum value of the quantization as the minimum value of a re-segmentation range, or by using the minimum value in the ROI.

In general, we believe the users should test and verify all the configurations presented above. This empirical approach will then allow extending the work by IBSI. However, as we have already mentioned and the quality of reporting of radiomic studies plays a fundamental role in the field's success and clinical impact. In fact, without an adequate standardized reporting system, it becomes impossible to compare and contrast two different radiomic pipelines and their underlying determining imaging signal sources.

Verifying limiting dependence on pipeline parameters

As it emerged from the previous subsections, all the available radiomic computational packages allow a full customization of the procedure for radiomic feature extraction, but available customizability without a consensus for a universally accepted, standardized strategy for radiomic features extraction impedes the ability to compare across studies or more broadly validate radiomic features. As the freedom of pipeline customization is left to the user, radiomic features could exhibit statistically significant differences that can impact their potential prognostic or predictive values, but some features might embed a signal that is independent from the particular configuration of the pipeline. We believe that the robustness of radiomic features should be tested with respect to all the parameters available in the computational pipeline. Of course, the combinations explode as the number of tuneable parameters in the pipeline increases. To start, we suggest the user evaluate a grid of parameters that were suggested from previous studies or within the IBSI document. For example, for MRI, consider a set of normalizations presented in Traverso et al. [136].

2.1.3 Quantitative radiomic features

Current features used in radiomic studies quantify the shape and signal value distribution at various scales (i.e., texture) in or of an ROI.

What have become known as "radiomic features" are defined quantitative features from across the research community. For feature equations and details, the authors direct readers to documents by the IBSI [164] and the developers of PyRadiomics [142], an open-source feature extraction platform.

First-order statistical features

First-order statistical features describe the overall voxel signal value distribution of an ROI. They can be as simple as the average signal value of the ROI, which could mean different things for different imaging modalities. For example, in a CT image it would describe the Hounsfield units (i.e., average linear attenuation coefficient relative to water), and in 18F-FDG it would describe the average glucose-associated activity uptake. First-order features can also describe heterogeneity of a voxel signal value histogram. For example, skewness is designed to quantify asymmetry of values, while entropy measures uncertainty/randomness in values, and kurtosis describes "peakedness," or the tendency of the distribution to be toward the tail or mean of the distribution. Other features that are found in this class are standard deviation, minimum, maximum, range, and mean absolute deviation.

Shape features

Features in the shape class are designed to quantify the 2-D or 3-D shape of the ROI and are independent of the image signal values found within the ROI. Many of these features have historical relevance in clinical prognostics. For example, volume may be considered a surrogate feature for TNM staging, while surface-to-volume ratio and sphericity (i.e., similarity of the ROI to a sphere) could be descriptors of tumour spiculation. These types of features are highly referenced outside of radiomics but would benefit from a more automated

and objective quantification. However, it should also be noted that one of the goals of radiomics is to improve upon current clinical standards, and highly redundant features should be avoided to prevent over-fitting in models. Care must be taken in understanding the impact of image resolution on shape features.

Texture features

The final feature class that has become synonymous with radiomic features is the texture class. These features quantify voxel signal value information, while also capturing relationships between two or more voxels. Texture features can be calculated from a variety of different matrices (e.g., co-occurrence, run length, or size zone) designed to describe the distances and relationships between voxels [134][63]. Relationships between pairs of voxels can be quantified using entropy, homogeneity, and contrast features (not to be confused with features of the same name in the first-order class), while relationships between voxels in neighbouring planes can be quantified with features such as busyness and complexity. Texture features have been the focus of much of the radiomic field since they are not easily quantified and differ from standard clinical features like tumour volume or average glucose metabolism-associated activity in 18F-FDG PET imaging. However, caution is warranted as these features were originally intended for evaluation of aerial photographs that were uniform in size [53]. This poses a challenge for radiomics since the features are quantifying ROIs of varying sizes and cause a confounding effect with respect to the volume feature. This has been noted by many research groups, and suggestions have been made to correct these features or remove them from analysis [140][148].

Filtered images

Another common practice in radiomic feature quantification is to filter the original image and re-extract first order and texture features from the newly generated image volume. A filter modifies the information in the original image and can bring forward features of interest, such as Laplacian of Gaussian (LOG) filtering, which is designed for edge detection and identifies areas of rapid change in images. By applying a LOG filter and re-extracting first-order and texture features, homogeneous regions within the heterogeneous ROI, may be more easily identified. There are many options for filtering methods, but they often result in less reproducible features [138]. Additionally, when binning a filtered image, the width or number of bins used should be altered to represent the range of values present in the image to avoid bins with single voxel counts. Volume as a confounder Many features used in radiomic studies have confounding features that can result in incorrect interpretation of results and assignment of causality. A well-known confounding factor of many features from the first-order and texture feature classes is volume. In a paper by Fave et al. , they discovered that five texture features were entirely volume-dependent, and that the dependency of other features on volume changed with pre-processing methods [39]. This points to a larger issue regarding reporting of data processing. In another study by Welch et al. [148] , a widely used radiomic signature was found to be entirely volume-dependent. It was determined that when applied to images containing simulated random noise, the signature and model had the same prognostic accuracy when used on patient images, indicating that the tumour contour was all that was required. These types of effects can result in misinterpretation of results and the overstatement of signature and model importance, and they should be thoroughly explored during feature selection and modelling building processes. There is no straight-forward way to deal with these confounding factors. Removal of confounded and correlated features can be done during feature selection using methods such as Spearman rank correlations; however, this assumes that a

confounded or correlated feature does not provide additional information, which is not always the case [88]. Alternatively, statistical methods have also been proposed to eliminate the effects of confounders for improved utilization in these types of models.

2.1.4 Building prediction models

The ultimate goal of the radiomic pipeline is to describe patterns observed in data using mathematical formulations. These formulations generate a representation of the disease as it relates to the features that have been extracted from the image. By describing the patterns in this way, a model is trained that can be used for classification of an outcome of interest. Radiomic model development involves the selection of a set of features that can be combined using statistical and machine learning methods. The combination of these features results in the ability to predict an event of interest when new observations are given. There is no single modelling methodology that will guarantee high accuracy, reliability, and efficiency; therefore, it is suggested to test various techniques. Comparison of different methods in radiomics has been performed by groups such as Parmar et al. who compared 14 different feature selection methods and 12 classification techniques in lung cancer [99]. In their work, they found that Wilcoxon-based feature selection and random forest modeling had the highest prognostic performance, with high stability against data perturbation. They also found that the classification method was the dominant source of performance variation. A similar study was performed in HN cancer by Leger et al. [77]. They compared 11 classification techniques and 12 feature selection methods and determined that random forests with maximally selected rank statistics and Spearman rank feature selection had the best performance. The different results achieved in these two studies highlight the importance of performing these types of tests on specific data sets since they can vary between sites. This section will highlight some of the key points required for model development and validation. Feature selection The average number of features that

can be extracted from a radiomic computational package varies from hundreds up to thousands (around 1000–5000) due to the large number of possible filter combinations that can be applied to the original image before feature extraction. This causes radiomics to suffer from the *curse of dimensionality*, and no classifiers are currently able to deal with an injection of such large numbers of features, particularly when using the data set sizes available for radiomic studies. In fact, even machine learning algorithms— such as random forests, which are designed to internally deal with large numbers of features—will be affected by other challenges, such as high feature correlations and confounders like volume. Techniques to deal with these problems have been developed mainly for genomic studies, which can present even higher complexity. Supervised and unsupervised feature selection can help to reduce over-fitting and increase generalizability; however, no agreed upon approach has been defined for radiomic studies. Below is a brief summary of different feature selection types. Unsupervised methods allow the learning algorithm to find structure in the input data, instead of responding to feedback from a classifier. They are called unsupervised methods since they do not make use of target information, such as patient-associated outcomes, during their procedure. The most commonly used unsupervised technique employed for feature selection involves finding similar groups of examples within the data by clustering features together. The clusters are evaluated for similarity using different metrics, such as distance between points or distance between points and a centroid [98]. It is also possible to define the number of clusters. For example, when predicting high and low risk of an event, you may define the number of clusters as 2. Alternatively, some algorithms, like hierarchical clustering, can determine the optimal number of clusters when little is known about the input data. These can help reduce feature dimensionality by defining a new set of features based on the clusters. This would involve either selecting one feature from the cluster as a cluster representative feature, or by defining a new feature based on the cluster (e.g., the centroid of the cluster). Another useful unsupervised method for feature set reduction is principal component analysis (PCA). PCA is used to reduce the

original dimensionality of the problem into orthogonal components (called principal components). The number of components is chosen as the minimum number of components that retains a certain amount of the original variance in the data (usually 95%). Principal components can then be employed in a classifier and used as a "reduced" features set. Unsupervised methods are ideal for basic exploratory analysis of features, but caution is warranted since the clusters or PCAs that are identified may have no relation to the outcome of interest. Furthermore, despite unsupervised methods being mainly used for feature set reduction, they may also be beneficial for prognostic modelling. This was demonstrated in a study by Traverso et. al. where it was shown that clusters of radiomic features could distinguish between lung cancer patients as having high or low risk of death [137]. Supervised feature selection approaches use event-of interest labels to select features that are important for the problem. They can be performed using univariate and multivariate analysis, as well as prior to or during classifier training. However, it should be noted that this approach can introduce redundant information during the training procedure. Supervised feature selection methods can largely be divided into filter, wrapper, and embedded methods. Filter methods of feature selection are performed prior to classification development, and they rank features according to a scoring criterion [50], such as a Fisher score [60]. Univariate methods only score based on the relevancy of a feature to the outcome of interest; multivariate methods score based on relevancy of the feature to the outcome of interest and redundancy of that feature in comparison to other features [29][48]. If this method is used, it should be performed on a set of features that is different than what is used for training of your classifier. This will ensure optimal generalizability in the features that are selected. Wrapper methods depend on the classifier that will be used. They act as a search method that looks at the entire feature space to find relevant and nonredundant feature subsets like recursive feature elimination [84]. A selected subset is then used with the classifier to determine how well this subset performs with the selected classifier. This method can be computationally expensive and may result in over-fitting to the type of classifier chosen. Finally,

embedded methods involve incorporation of feature selection with the classifier. Embedded methods can be similar to wrapper methods, but they tend to be more efficient as an intrinsic classification metric is used during learning. Embedded methods of feature selection are common in decision tree classification.

Classifiers

Classifiers are statistical or machine learning methods that combine pre-extracted features in ways that lead to predictive forecasts. There is often a trade-off between interpretability and complexity of classifiers [52][69]. Simple classifiers, like logistic regression, are often preferred by clinics, but they lack the ability to discover more complex feature interactions. Generalized additive models are new methods that have been applied to real healthcare problems, and they have achieved high accuracy while retaining interpretability [17]. They have not been applied to radiomic features yet, but when applied to a pneumonia risk prediction case, they recognized and allowed removal of patterns and biases in the data that had prevented complex classifiers from being used. It is up to the user to decide what aspects of a classifier are important for the utility of their radiomic models and to test various methods accordingly. Supervised classification is the most commonly used method for predictive and prognostic model development in radiomics. It involves the inference of a function from labelled training data, which consists of a set of examples that contain an input vector of features (i.e., the selected radiomic features) and a desired output (i.e., the event of interest). The function can then be iteratively learned by evaluating its performance relative to the provided input data. These types of models can be simple and intuitive, such as linear regression or Cox modelling, or more complex, like neural networks or decision trees. More complex methods often have hyper parameters that can be tuned as well [85][153]. For example, when training a random forest, the optimal number of trees, samples in leaf nodes, and samples

evaluated during splits can be tuned. It is important to tune these parameters using cross-fold validation on your training set to avoid overfitting. After finding the optimal settings, a final model can be fit on the entire training data set. Supervised learning is the most common method of classification in radiomic research. Training and testing of a radiomic classifier, it is important to ensure there is no information cross-contamination between the two data sets. This means making sure that your test set has not been seen by your classifier during training. Due to the lack of openly available data in healthcare, radiomics often relies on internal validation of models. This involves splitting a single data set into training and testing sets [12] and can measure the internal validity, or reproducibility, of the model. Internal validation can be achieved using internal holdout testing sets or k-fold cross validation. In a hold-out validation methodology pipeline, a single data set would be divided into a training and a testing data set. The training data set would be used for feature agglomeration, selection, and classifier training. This division can happen multiple times, each time evaluating the trained classifier on the hold-out testing set. The estimated error rate of the classifier can then be calculated by averaging the performance across all divisions. K-fold cross validation improves upon the hold-out validation method by preventing test set overlap. In kfold cross validation, the full data set is randomly divided into "k" subsamples. The k-1 subsamples are then used as the training data, and the let-out sample is used as the testing data; this is repeated k times for all potential subsampled combinations. The average performance of the classifier across all k-folds is taken and also gives an error estimation, but without the potential overlapping of test sets. As mentioned, internal validation of classifiers is the most commonly used method due to availability of data; however, external validation using a related, but slightly different, population of patients gives more information regarding the generalizability of the classifier. Despite the difficulty of external validation studies, it has been performed in radiomic studies for prediction of cervical cancer recurrence and identification of clinically significant portal hypertension in cirrhosis [79][81]. Most notably, Aerts et al. showed that a radiomic model trained on a lung data set

could prognosticate patients with HN cancer, as well as lung and HN cancer patients from a different institution [2]. It was ultimately discovered that the model used was a surrogate for volume, and building a complex tool that replaces a simple measure is not what the field aims to do, but this is an excellent example of external validation. Thought should also be given to the scoring metric used for your predictive classifier. Different metrics are appropriate for different outcome types and distributions. Time-dependent classifications that involve modelling techniques, like Cox regression, require statistical methods conducive to continuous variables. Log rank tests are a common hypothesis test between two survival distributions in order to determine if they are significantly different, e.g., high or low risk of event based on a time threshold. Concordance indices are another example of a metric that quantifies the quality of continuous variable ranking. It looks at the probability that, for a pair of randomly chosen samples, the sample predicted to have a higher risk will in fact experience the event before the comparing sample [102][62]. This is a commonly used metric in radiomic studies and has been used to evaluate classifiers designed for outcome prediction of stereotactic body radiation therapy lung cancer patients, comparison of classifiers for outcome prediction in HN cancer, exploration of delta radiomics for non-small cell lung cancer outcome prediction, and the study of radiomic software implementation impact on radiomic features in 18F-FDG PET HN patients. Binary classifiers are also a common event for prediction, such as patient status at three years or the presence of a malignant tumour. A confusion matrix is a simple way to visualize the number of true positive, false positive, true negative, and false negative predictions. Most statistical metrics are designed to summarize some characteristic of the confusion matrix. Accuracy is very often the first metric used for classifier evaluation. It quantifies the ratio of true predictions versus all predictions. It is a good metric when working with data sets that are nearly balanced (i.e., have the same number of events in each class); however, it should not be used for imbalanced data. For example, a data set with an event distribution of 99 to 1 might result in a classifier that learns only to predict the majority class and will have an accuracy of 99%

despite learning nothing about the minority class. Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is another metric that is best used in balanced data sets. The ROC curve shows the false positive rate vs. the true positive rate, or sensitivity vs. 1-specificity [40]. ROCs are commonly summarized using an AUC, which shows the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, similarly to how the concordance index works in time-to-event predictions. An AUC of 1 means perfect prediction, and 0.5 means no better than random. This metric has been used in radiomic studies for binary survival analysis, identification of metastases or benign lesions in breast and lung cancer, and identification of biological basis of radiomic phenotypes. However, this metric, like accuracy, is best used on balanced data sets, which are not very common in radiomic studies. Precision and recall are lesser-used metrics that are appropriate for imbalanced data sets, which are commonly used in radiomics. Precision defines the proportion of patients that are predicted positive who were actually positive, thereby providing information regarding false positives. Recall tells us the proportion of patients that are positive who were predicted to be positive, thereby giving information regarding false negatives. False negatives are a large concern in cancer care where we want to ensure we do not miss an event [89]. Precision and recall information can be presented simultaneously using an F1-score or an AUC of a precision recall curve. The F1-score takes the harmonic mean of precision and recall and allows us to learn about the false positive and false negative predictions of our classifier. These metrics are not as intuitive as accuracy, but they are more informative and often more appropriate for the data sets used in radiomics [106].

2.2 PRACTICAL CONSIDERATIONS, STANDARDS AND SAFEGUARDS

2.2.1 Clinical applications of radiomics in radiation oncology

Several recent reviews have provided qualitative overviews of the available literature regarding the applications of radiomics in radiation oncology [152][33]. The results show evidence of a nascent and rapidly growing field in which the potential prognostic/predictive power of radiomics has so far been applied to two main categories [43]: *pre-treatment tumour characterization* and *therapeutic monitoring studies*. Pre-treatment tumour characterization has included the investigation of possible correlations between imaging features and the biological and genetic properties of tumour tissues. For example, recent studies showed that imaging features helped discriminate between p16 positive and negative disease for HN cancers [99][8]. In one of the first radiomic studies, the authors used unsupervised clustering analysis to show associations between radiomic features and gene-expression patterns [2]. A Rad-TRaP framework has also been proposed to utilize radiomic classifiers to guide the generation of more focally targeted treatment plans for prostate cancer using brachytherapy and external beam radiation therapy [124]. However, the implementation of this technology into prospective clinical trials has been appropriately cautious. Therapeutic monitoring refers to the investigation of the prognostic or predictive power of radiomic features to support particular treatment decisions, or to monitor the pathological response of the patient during or after treatments. As analysed by Scalco et al. , the treatment effect following thoracic radiotherapy for NSCLC is one of the most studied subjects in this context, with a particular focus on RILI (Radiation-induced lung injuries) [118]. For example, Mattonen et al. developed a preliminary radiomic signature for the automatic classification of tumour recurrence vs. lung injuries [87]. Many publications have also been attempting to predict lung recurrences from PET images [110], while CT has been explored

for the assessment of pathological response, overall survival, and distant metastases. In HN [8][25] and prostate cancer, attempts to find radiomic features that predict for tumour response using ADC maps from DWI scans have found inconsistent results. Some studies [35][46] found prognostic or prognostic power in simple statistical features extracted in ADC maps (e.g., mean, median, 75th, 90th, and 95th percentiles), but another study [93] showed that textural features were poorly associated with biochemical recurrence. One of the reasons might be that the lower resolution of DWI imaging may have hindered consistent extraction of ADC textural information. When investigating radiomics for therapeutic monitoring, the methodology of looking at differences of feature values as a function of time has been referred to as *delta-radiomics*. Delta-radiomics is a general technique that looks at feature values as a timeline series. The main application of delta-radiomics is to compare changes in feature values between two time points, usually before the start of treatment and at treatment completion. Additional time points could be added corresponding to different in-between treatment fractions for a more detailed evaluation of treatment response. By evaluating treatment response more frequently, it is possible to detect warning signs for adverse outcomes earlier. For oesophageal cancers, Yip et al. reported that differences between pre-treatment and post-treatment radiomic features combined with maximal wall thickness changes predicted for pathological response better than morphological features alone [155]. A clinical application of delta-radiomic signatures may be earlier detection through identified signals that reflect tumour response prior to tumour volume shrinkage across time. However, feature stability across time should be carefully evaluated [95] to isolate and exclude features presenting poor time stability, as variability at each time point can compound the sources of error and variability in delta-radiomic feature measurement (e.g., image quality, ROI definition, etc.).

2.2.2 Outcome selection to address the clinical question

When designing or reviewing a radiomic study, it is critical to have a clearly defined outcome of interest that addresses the defined clinical question. One of the most popular outcomes used in radiomic studies is overall survival. In particular, the majority of the studies binarize this outcome (e.g., alive/dead after two years) for their prediction. While overall survival is an objective binary outcome, it may not be representative of the cause of patient death that is related to the clinical questions because overall survival is not strictly related to a death caused by cancer. Additionally, overall survival may not be impacted by subsequent salvage treatments following the treatment of interest, or even supportive care measures that are not related to the treatment of interest based on the primary clinical question. Other outcomes that may be used to associate radiomic features with prognosis include progression free survival (PFS) or local control (LC). Common tumour response criteria used to determine PFS and LC include the response evaluation criteria in solid tumours (RECIST), which focuses on two-dimensional tumour size changes, Cheson response criteria for malignant lymphomas that track size and metabolic changes [18], and immune-related response criteria that account for a potential increase in overall disease burden at initial treatment [151], among others [133] RECIST are the most commonly used criteria. They were written in 2000 and updated in 2009 [31] to address pitfalls and limitations within an originally drafted response criteria document by The World Health Organization in 1979 (World Health Organization 1979). RECIST contains guidelines for imaging studies performed in CT, MRI, and FDG-PET, methods of lesion measurement (e.g., all target lesions must be measured in the longest dimension), and guidance on when lesions are considered new, which is representative of disease progression. Despite its wide usage, RECIST is not without its limitations. Most notable, RECIST focuses on changes in 2-D tumour size of only five target lesions as indicators of response. This challenge is exacerbated when the reliability of tumour size measurements is considered and found to be inconsistent. Additionally, the limitations of current

methods to determine reliably these outcomes may limit the clinical interpretability and application of radiomic features found using these endpoints [36]. Most of the white papers describing these response criteria refer to the need for volumetric tracking of tumours to enable quantitative response assessment. However, the collinearity of tracking tumour volume and using these segmentations for radiomic feature extraction poses an interesting challenge. An additional factor to consider is the event rate of the outcome that is selected. By selecting an outcome that is very rare for the particular clinical question when a large number of features is extracted, there is risk of over-fitting the model. In order to avoid this problem, the empirical rule of 1 feature for 10 events has generally been accepted as a conservative estimate. Recognizing these nuances of outcome selection, a general suggestion for the radiomic community is to work closely with a clinician or clinical team to determine the outcome of interest that best addresses the clinically relevant question that can be supported by reliable availability of the appropriate feature with adequate quality. For new multicentre prospective studies, defining the standardize methodology for data collection and using established standardized criteria will simplify the consequent analysis. Finally, acquiring statistical input for the design of radiomic studies is strongly recommended.

2.2.3 Contours

An important point to be considered is that clinical contours in radiation oncology are performed by (1) looking at the particular clinical history of the patient, (2) using additional information from visual/physical inspection of the patient, and (3) embedding additional information derived from the reasoning of the clinician that can under particular contouring choices be associated with treatment decisions. For example, contours made by looking at different modalities (often "fused" together for visualization purposes) can differ from contours using just a single modality, as discussed in Riegel et al. [111] and Foroudi et al. [42]. Therefore, the goals of tumour segmenta-

tion, or source of tumour segmentation, should be noted explicitly, as the goals may influence how the tumour is segmented. For GTV segmentation, there is variability in target volume segmentation across individuals, practices, and institutions despite attempts to introduce guidelines for the homogenization of GTV contouring [13][144]. This problem becomes of interest when performing radiomic studies or validating radiomic-based models through different institutions. To guarantee that no biases are introduced when comparing or training models using data sets from different institutions, it is important to verify with clinicians the procedure used for delineations. This involves determining if the contouring procedure was driven by additional information not necessarily taken directly from visual inspection of the images. This could also involve information more related to outcomes found during the physical examination of the patient, or particular treatment decisions that might affect the original contouring strategy. Introduction of semi-automated or automated delineation algorithms could potentially reduce differences in delineations. Manual contours are prone to variability due to inter-observer variations and differences in institutional guidelines [113] [115]. This is true even when utilizing contours developed for radiation treatment plans where consistency would be expected [91]. Semi-automatic and automatic guidelines have demonstrated their effectiveness in developing clinically acceptable contours using both atlas-based [161][82] and deep learning-based methods [16]. Additionally, semi-automated contouring methods have been studied for their impact within radiomics, and they demonstrated improved radiomic feature reproducibility [101].

2.2.4 Data quality and quantity

As the quality and consistency of our images and their acquisition parameters increases, the quantity of images available tends to decrease. Therefore, when building a classifier, the data set dictates a portion of its eventual utility. This means that a large data set with low-quality images (artifacts, varied reconstruction algorithms, etc.) would gen-

erate a model with greater generalizability, but less ability to correlate features with biological processes since a feature in one patient may actually be quantifying an imaging artifact instead of the disease of interest. Alternatively, a small data set with highly curated and good-quality images ensures accurate understanding of what is being quantified in the image, but may not perform well on real-world data sets. Consideration of this trade-off is needed when building and reporting of radiomic classifiers. Scientific developments in oncology require high-quality data [10]. The *garbage-in-garbage-out* principle applies to both user-driven manual pipelines and automated pipelines. Imaging data, such as that used in radiomics, has the potential for large quality variations, which has led to standardized site specific imaging guideline development (Lewis-Jones et al. 2016). However, without an understanding of the image quality used for radiomic model development, it cannot be known whether image or patient variability is the deterministic factor in the model training and prediction. Lack of standardization in image acquisition and handling is a cause of this issue, but so is lack of image artifact consideration, as are a variety of other parameters whose impact on radiomic features have not yet been studied adequately. Radiomics is currently challenged by the lack of consistency within image acquisition parameters, such as voxel size [120] and reconstruction algorithms [68]. However, artifacts specific to certain disease sites and imaging modalities further exacerbate the data quality issue and can skew learning and validation of radiomic models. The collection of high-quality prospective data specific to radiomic analysis that supports causation instead of correlation will strengthen the findings of radiomic studies; however, until this time, data curation with respect to data quality would represent a fundamental step toward reliable and reproducible results. Artifacts come in a variety of forms that alter an image's appearance, as well as our ability to quantify the phenotype of an ROI. It is not understood to what extent quantification is impacted by the various artifacts, but it logically follows that if qualitatively the image is degraded, we cannot trust our quantitative metrics. These artifacts may present as motion artifacts in PET, CT, or MR images due to the length of scan times [109]. PET im-

age quality can also be degraded by scattering effects that impact the quantitative properties of the modality [126]. In MRI and CT images, metal artifacts caused by dental work, pacemakers, or joint replacements can impact calculated relaxation rates, diffusion metrics, and Hounsfield units (HU) [90], and current methods are only designed to salvage qualitative data, not quantitative [25]. In one study on the impact of artifacts on radiomic features, Block et al.looked at the impact of metal artifact reduction (MAR) methods on radiomic features calculated for a phantom with and without a dental artifact insert [34]. It was shown that radiomic features calculated on images after MAR were similar to those features calculated on images of the phantom without the dental artifact insert; however, regions of interest farther from the dental artifact insert had reduced stability due to the introduction of new artifacts. Additionally, Wei et al. , demonstrated the importance of DA consideration in radiomic studies by showing that the removal of DA+ patients from analysis positively impacts radiomic signature performance [146]. The lack of consensus on proper handling of artifacts often leads to patients being excluded from the data sets or included and ignored. As the availability of large retrospective data sets become more common, efficient methods of artifact identification will be needed. Recent work with deep learning achieved a precision recall AUC of 0.92 when identifying volumes containing Das [150]. Promising work in slice removal has also shown that affected portions of ROIs can be removed without significantly altering some radiomic features [44]; granted, justification would be required to explain the implications of slice removal on shape features. When developing predictive models, it is understood that more data is often preferred. Larger data sets improve statistical analysis of the model and have a higher chance of containing heterogeneities that your model may encounter during clinical usage. Additionally, small data sets have increased potential for false positive and false negative errors [6]. Smaller data sets are commonly used in radiomic studies, which could be the result of increased data quality, but is more commonly caused by data access issues. Rule-of-thumb suggestions on how many features to consider per event in the data set are often made to counter

these issues [45]; however, these are just suggestions and may not be appropriate for a specific disease site, end point, or set of features. Alternatively, gaining access to more high-quality data can increase the performance and impact of radiomic studies; however, this is not easy. Obtaining more data is a challenge that most radiomic researchers will encounter during their career. Lack of data access may be caused by a rare disease site of interest or data governance issues. This is a known area of concern and multiple things can, and are, being done to remedy it. Distributed learning is one method of accessing more data from different sites without the data needing to leave the hospital. This is achieved by enabling models to move between sites, learning at each location, and having the final results combined. Open access data is another option for increasing the size of data sets. Groups such as The Cancer Imaging Archive [20] and XNAT [86] contain openly available patient images from a variety of modalities and disease sites that are available for anyone to use. This not only improves the research of others, but it helps the publisher of the data gain greater insight into their data through collaborative and reproducibility studies. However, it is advised that feature distributions between different data sets are analysed to ensure similarity. Open access data is an important step for radiomic progression that is beginning to be recognized. In particular, the National Institutes of Health (NIH) is addressing the data sharing issue by requiring all major funding applications to have a data sharing plan in place [117].

2.2.5 Feature reproducibility and repeatability

The importance of reproducibility and repeatability of radiomic features is integral to their successful application within clinics. This parallels with quantitative imaging, where technical confidence in reproducibility and repeatability of quantitative measurements is needed for utilization in clinics and personalized cancer care [128]. *Reproducibility* refers to features that remain the same when imaged using different equipment, software, image acquisition settings, or operators

(e.g., other clinics), be that in the same subject or in different subjects [71]. It is important to understand that a feature's stability is not a sufficient condition for a feature to have certain prognostic or predictive power. However, as explained by Gudmundsson et al. for time series analysis, less robust features are likely to be found to be less important when developing a predictive or prognostic model [49]. Finding a consensus in the published literature regarding a set of "universally" reproducible features is difficult. The main reason behind this problem is that radiomic features exhibit different grades of sensitivity with respect to different settings. Furthermore, the level of sensitivity might be dependent on the particular modality or perturbation being considered. For example, a feature that showed good reproducibility for a certain anatomical site might not be reproducible for another site, even if the image modality remains the same. This challenge was highlighted via a qualitative synthesis in (Traverso et al. [138], despite a quantitative meta-analysis not being possible due to limitations of data sharing of the analysed studies or poor quality of reporting. The qualitative synthesis showed that inter-observer variability in ROI segmentations affected feature reproducibility for both PET and CT modalities. Out of all the features categories, first-order features were in general more robust than textural features, with the first-order entropy appearing to be robust in CT both for HN and lung sites. No emergent pattern regarding reproducible PET texture features was found. Digital image pre-processing prior to feature extraction was also found to affect the reproducibility of all radiomic features, with the exception of shape features. If some of the image pre-processing techniques, such as de-noising, can increase the level of signal-to-noise ratio in an image, it logically follows that other techniques, such as resampling, could worsen or improve the quality of features. *Repeatability* of radiomic features refers to features that remain the same when imaged multiple times in the same subject, be that a patient or a phantom. The most common technique that has been used to evaluate feature repeatability in human studies is called test-retest. In a test-retest scenario, patients are rescanned using the same configuration but within a short elapsed time frame from the original scan, sometimes called the "coffee-break

scan." Feature stability with respect to the two interval times is then investigated. Feature repeatability can be seen as a control group stability test. In fact, in the same short time scenario, with no changes on other settings, features are expected to not show any change in their values. Therefore, features that have poor repeatability should be discarded from further analysis [49]. Repeatability becomes fundamental when considering time-series analysis, such as the change in radiomic features between the start and finish of treatment that may indicate treatment response. When exploring the reproducibility and repeatability of radiomic features, it is often difficult to find relevant datasets that are publicly available. For the former, for example, delineations by multiple observers may be required, which represents a burden on clinicians. For the latter, test-retest human data sets would require particular approval on the patient's side, since additional exposure to radiation or prolonged scanning time is needed. Finally, we always recommend that users explicitly report the details of the entire computational pipeline used to extract features. This topic will be touched on in the next subsection.

2.2.6 Software

Software is needed during all steps of the radiomic pipeline and should be carefully considered. Many open languages, such as R (R Development Core Team 2008) and Python [139], have libraries capable of feature agglomeration, feature selection, and classifier training. However, the extraction of radiomic features often requires custom-built solutions dependent on the user's needs. It is commonly thought that the most important aspect of radiomic feature extraction software is the variety of feature classes available, but feature correctness, image and contour handling, transparency, pre-processing, and batch processing capabilities are equally, if not more, important. The variety of requirements for radiomic research often leads to building in-house solutions. Although this allows the user to customize every aspect of their pipeline, caution is warranted.

Errors in image handling and mathematical definitions of radiomic features can quickly lead to incorrect results that are difficult to trace. By using pre-built solutions, open-source libraries, or by making in-house software public, it increases accountability and produces more robust results. Additionally, by using open-source solutions and libraries, reproducibility of results increases [148]. This need is highlighted in a study by Bogowicz et al. (2017), where they demonstrated that features extracted from the same data set, but using two different software implementations, led to significant differences in 88% of the features. There are many pre-built radiomic feature extraction solutions available. MaZda [131] is a solution that was originally developed for texture analysis of mammograms in the 1990s, demonstrating the historic nature of this field. It was recently updated to handle 3-D images and is widely used and well tested [4]. It is designed to work as a stand-alone platform capable of ROI segmentation, image preprocessing, statistical analysis of features, and classifier development. Alternatively, predefined ROIs can also be imported, and generated features can be exported for statistical analysis in other software. CERR is another solution that was originally designed for importing, displaying, and analysing radiation therapy treatment plans in Matlab, but has recently been expanded to include radiomic feature extraction [24]. It can handle various imaging modalities and perform image fusion and contouring. IBEX [158] (Zhang et al. 2015) is another popular standalone solution that is compatible with various image and contour formats, has contouring capabilities, and allows for optimization of pre-processing and feature algorithm parameters that permits optimization for different modalities. If the user prefers a more custom solution, opensource libraries also exist. Pyradiomics is an opensource Python package for extraction of features from 2-D and 3-D binary masks [142]. It has a large variety of features, filtering methods, and pre-processing steps available for complete customization. Additionally, batch processing and parallel processing is easily implemented through the package or with basic Python scripting. Pyradiomics can also be implemented using 3-D Slicer [105], which provides a graphical user interface

and access to Slicer modules that include segmentation methods and DICOM handling. Pyrex [122] (Shi 2018) is another wrapper that employs original DICOM and RTStruct files of various images, and exporting of features in a variety of formats, including those compatible with radiomic oncology ontologies (Link here).

Reporting and safeguards

In emerging fields such as radiomics, the methodological details are very important to ensure productive progression of the field. Understanding the details of a study improves reproducibility and refinement of its results. In this chapter we have highlighted some of the potential areas for concern, including pre-processing of data, confounding factors, software, and model development and evaluation. This section describes some of the protocols and standards that have been published to guide radiomic researchers in proper reporting of these details, thereby ensuring good scientific developments and eventual transition of appropriate models into clinics. Broad methodological guidelines were suggested in a paper by Welch et al. to safeguard development against underlying feature dependencies and multi collinearities, while ensuring clinical engagement [148]. Through a detailed analysis they refitted a radiomic model and found vulnerabilities during a *backwards* analysis, which started at a completed model and worked toward the univariate feature selection. They ultimately discovered that the signature on a volume of noise performed equally as well as it did on patient images. To safeguard against these vulnerabilities, they suggested (1) using open-source software to increase development accountability and inter-institutional research, (2) determining added prognostic and predictive accuracy compared to clinical standards, (3) testing of underlying feature dependencies to prevent unwarranted usage of computationally expensive features, (4) testing of feature multicollinearity to improve data variance description, (5) pre-processing of data to ensure image signal quality, and (6)

describing manual contouring processes adequately to understand any prevalent signals that were used for delineation. The radiomic quality score (RQS) is a reporting guideline designed to reward and penalize methodology and encourage increased utilization of safe practices, thereby improving scientific integrity and clinical relevance [72] . It is divided into data selection (e.g., proper imaging protocols and prediction targets); medical imaging (e.g., segmentation and reporting protocols); feature extraction (e.g., feature stability and algorithm reporting); exploratory analysis (e.g., usage of clinical variables and data storage), and modelling (e.g., feature selection and validation) sections. An example of negative scoring may involve -3 points f feature reduction or adjustment for multiple testing is not performed (as this will lead to potential over-fitting of the model) or –5 points for no validation of the model. However, +15 points are awarded if validation is based on three or more data sets from distinct institutions. The RSQ is focused on data and methodology, and it does not allot scores for areas such as exploration of confounding factors and software sharing; however, it does show acceptance of the need for standards in radiomics. Interestingly, a systematic review of 41 published radiomic studies discovered that the majority had earned less than 50% of the possible points that could be awarded [116], indicating large problems in current studies and the need for usage of these guidelines. Another, more general reporting guideline for prediction models was suggested by the TRIPOD Statement [21]. TRIPOD (transparent reporting of multivariable prediction model for individual prognosis or diagnosis) is available in addition to RSQ to improve reproducibility and validity of radiomic prediction models. It is a simple checklist of 22 items that ensures adequate reporting of data sources, statistical analysis, and risk groups. It was designed as a census between methodologists, healthcare professionals, and journal editors. Furthermore, it is a general checklist for all diagnostic and prognostic models, not just radiomic-based ones. The importance of reporting guidelines has been thoroughly published [138] and will ultimately improve the field of radiomic models through greater understanding of bias risks and the potential usefulness of these

models. More recently, the American Joint Committee on Cancer (AJCC) published an article describing the criteria that must be met for endorsement of a cancer risk model which is complementary to the TRIPOD guidelines [65]. Among the criteria is the use of multiple external data sets collected over various times and locations, further necessitating the need of imaging standardization. Additionally, models can be excluded if reviewers believe there are not enough events present in the data sets. This is a challenging task for radiomics since the field is in its infancy, lending itself only to small retrospective studies that do not instil confidence in the developed models. Additionally, the lack of foresight within the healthcare domain has put us behind in the AI revolution since data management standardization is not yet the norm. Finally, the Image Biomarker Standardization Initiative (IBSI) is a comprehensive reference manual for radiomic research [164]. It exhaustively describes options for image processing, quantitative features/imaging biomarkers, imaging biomarker reporting guidelines, and data set benchmarking. The IBSI has been an international collaborative effort aimed at standardizing these processes to improve reproducibility and validation of these methods through consensus-based guidelines and definitions. It is an excellent resource to consult throughout the development of radiomic signatures.

Requirements for clinical acceptance and adoption

The number of radiomic publications has been growing exponentially; however, only a small percentage of published studies have reached the clinic for their desired utility as a decision support system. The healthcare market still offers many opportunities for investments, and commercial interests have taken an interest in transforming radiomics/quantitative imaging computer- aided diagnosis (CAD) research prototypes into FDA-approved commercial products. In July 2017, Quantitative Insights, Inc. (QI) announced that they had received regulatory clearance (via De Novo classification) from the

U.S. Food and Drug Administration (FDA) for its QuantX Advanced system (Link here). This is the industry's first CAD platform that incorporates machine learning for the evaluation of breast abnormalities. The software received approval after showing in an FDA clinical study that it could be used to enable faster and more accurate diagnoses. However, this tool is designed for application in large breast cancer screening campaigns and has radiologists as its main target users. No indications are available if this CAD will be expanded to support indications for treatment for radiation oncologists. It should be noted that the word *radiomics* is almost completely absent from the description of this software. In December 2017, the FDA premarket-approved Microsoft's *Radiomics App V1.0*. This app was designed by Inner Eye (Link here), Microsoft's research project aimed at *turning images into measurable devices using AI*. Again, this product was classified under the category of *system, image processing, radiological*, the same as QuantX however, compared to QuantX, it followed the radiomic workflow presented in this chapter more closely, including image pre-processing, multi-modality segmentation, texture analysis, etc. One of the goals of the project was to *identify the boundaries between healthy and nonhealthy cells. The boundaries can then be used for quantitative radiology and potentially for more efficient planning of radiotherapy and surgery* (Link here). If premarket approval demonstrates a promising path for the adoption of AI-driven tools in the healthcare domain, marketing of these applications will require additional effort. This is mainly related to the new FDA initiative that attempts to renew its traditional paradigm of medical device regulations to include AI and machine learning software for healthcare. This was done as an anticipatory action, which assumes many of these artificial intelligence and machine learning-driven softwares may require premarket review. In April 2019, the FDA came out with a new discussion paper entitled *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback* (Food and Drug Administration 2019) where the FDA explicitly expects commitment

about transparency and real-world performance monitoring for artificial intelligence and machine learning-based software, which is now treated at the same level as a medical device. In this view, we believe that the next generation of radiomics/AI-based prediction models should incorporate the following:

- Transparency. Full transparency regarding developed algorithms, including detailed reporting on the category of developed prediction/prognostic models. Also, benchmarking of the developed model with respect to external data sets is needed. With this in mind, a recent publication [163] introduced a framework for the benchmarking of radiomic models as TRIPOD IV (e.g., fully external validation). In addition to these efforts, it would also be beneficial to track the results of this benchmarking. This is similar to what the FDA calls *Complaint Handling*, where every user complaint is tracked, reviewed, and evaluated. Development platforms such as GitHub, GitLab, and BitBucket can offer a preliminary solution for tracing the life cycle of radiomic-based tools.

- Good practice. The FDA document cited above also references GMLP (Good Machine-Learning Practice). This incorporates three steps: (1) valid clinical association between the developed software and the target clinical condition, (2) analytical validation to test if the software correctly processes input data to generate reliable and precise output data, and (3) clinical validation to determine if the developed software output data produces the intended purpose in the target population in the context of clinical care.

- Defined life cycle. Following a well-defined life cycle includes four steps: (1) data management that specifies data collection protocols, guarantees quality assurance on training, and validates data sets; well-defined training and retraining strategy of the algorithms such that if retraining is performed, then changes to the ML architecture and parameters need to be

specified; (3) performance evaluation that clearly assesses evaluation metrics, performance targets, and involves clinicians in the "evaluation loop", and (4) planning of the update procedure that states when and how updates in verification and validation will be scheduled.

While these development steps will enable informed adoption of radiomic methods, the required processes for clinical acceptance remain unclear. The added value of radiomics and quantitative imaging in oncology depends on a number of factors and will require consideration of the methods used (including the underlying data and algorithms), as well as the ability of clinicians to appropriately interpret the results.

2.3 EMERGING TECHNOLOGIES

2.3.1 RTX-omics

Radiomics in its current state attempts to extract more quantified information from pre-treatment medical images. By using these features, radiomics has demonstrated promise in patient prognostics and outcome prediction. However, most patients will undergo treatment for their cancer that alters the natural course of their disease, resulting in a treatment-specific conditional prognosis or outcome. RTx-omics is a proposed area of research that suggests quantifying radiation therapy (RT) treatment plans to gain access to information regarding intrinsic differences in patient anatomy and dose distributions in tumours and surrounding OARs. RT lends itself well to quantification since dose plans are contained in a 3-D voxelized array, similar to an imaging volume, where a voxel represents a prescribed dose to a corresponding anatomical voxel (typically CT). Additionally, for treatment planning, contours of tumours and OARs are present due to their need in treatment plan optimization. RTx-omics has the potential to introduce new information to the automated information processing field. These

features could be combined with clinical or radiomic features for prediction of patient outcomes, recurrences, or toxicities, all of which are known to be tied to RT treatment plan quality and dose distributions [103][15][54]. In a recent study by Jiang et al. [61], machine learning methods were used to look at dose patterns in salivary glands to predict xerostomia. They looked at spatially explicit dose predictors and voxel doses in parotid glands and submandibular glands and found that dose patterns influence xerostomia at three months post-RT. This paper provides motivation for the importance of features that quantify dose and anatomy in patient plans. RTx-omics can provide more information with the inclusion of accumulated dose volumes and replans. Due to the changing anatomy of patients and variations in patient setup that occur during treatment, the planned dose volume is not what is always delivered [141][59]. By accumulating the dose across all fractions and quantifying features within that volume, we would gain a clearer picture of the dose distributions in the many ROIs of a plan [94]. Additionally, if a patient undergoes replanning, there is an opportunity for "delta" RTx-omic features to be calculated. Welch et al.vhave explored the potential of these features in combination with radiomic and clinical features for the prediction of local regional failure in HN cancer [149]. For this study, they used high-quality, low-quantity data from a single institution that did not have a sufficient number of events or planning variability to see an improvement above current clinical features. Until large data sets from multiple institutions with varied planning requirements for tumour coverage are available, other endpoints may provide more positive results. Toxicity is a promising area of study with RTx-omic features since there will be more variable dose distributions among OARs, as plans are generally optimized for tumour coverage over OAR sparing.

2.3.2 Deep learning approaches

Traditionally, radiomics has been defined by hand-engineered features—that is, features defined by a user. These types of features

are excellent for interpretation and explaining any potential ties to clinical, genomic, or phenotypic observations. However, they are also limited by the knowledge of the user that defined the feature. Deep learning—which allows a computer to generate features it finds to be important for the prediction of an event of interest—could provide new and interesting information that has previously not been conceptualized by humans. Deep learning is a supervised method of machine learning that involves training a neural network with multiple hidden layers, and it is a form of hierarchical feature learning. When working with images, a randomized filter is convolved over an image to generate a set of new, machine-learned features, and this constitutes one layer of a convolutional neural network (CNN). The resulting features can then be convolved with a new filter, with this process being repeated for many subsequent layers. A final layer predicts an event of interest based on the machine-generated features. The difference between the CNN's predicted outcome and the desired outcome supplied by the labelled training data is calculated, and the error is back-propagated through the network to update the weights between features. This is repeated multiple times until a desired performance metric is met. The potential of deep learning to be utilized with medical imaging data for predictions or prognostics has been recognized by researchers. One way in which it has garnered interest is through the extraction of "deep features" generated by a CNN. Researchers often use a pretrained deep learning network and fine tune it with their data, a process called transfer learning [123]. They are then able to extract and use the features generated by the network in combination with radiomic and clinical features. Lao et al. used "deep features" extracted in this manner in combination with hand-engineered radiomic features to train a Cox model for prediction of survival in glioblastoma multiforme patients [74]. They found that the combined features performed better than traditional risk factors, and that when combined with traditional risk factors it again improved prediction. Huynh et al. also demonstrated that deep features in combination with hand engineered features performed better than they did alone

when classifying digital mammographic images [58]. Another method by which deep neural networks can be used for the prediction of outcomes is an end-to-end process. This involves allowing the neural network to predict the outcome instead of extracting the deep features from the network. Esteva et al. demonstrated that they could achieve dermatologist-level classification of skin cancer when using transfer learning with clinical imaging [37]. They had 129,450 clinical images of skin lesions with 2032 different diseases. They trained their CNN to identify the most common cancer types (i.e., keratinoxyte carcinoma vs. benign seborrheic), as well as the deadliest (i.e., malignant melanoma vs. benign nevi) and found that it performed on par with 21 board-certified dermatologists. End-to-end CNN training has also been used for colorectal outcome prediction using images of tissue sample [64]. In this work, the CNN trained on digitized haematoxylin-eosin-stained tumour tissue microarray samples to assess the tissue microenvironment and predict prognosis directly from the histopathological images. Future work will involve prospective validation for integration into clinical workflows.

Bibliography

[1] Amy P Abernethy, Lynn M Etheredge, Patricia A Ganz, Paul Wallace, Robert R German, Chalapathy Neti, Peter B Bach, and Sharon B Murphy. Rapid-learning system for cancer care. *Journal of Clinical Oncology*, 28(27):4268, 2010.

[2] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[3] Bogdan Badic, Marie Charlotte Desseroit, Mathieu Hatt, and Dimitris Visvikis. Potential complementary value of noncontrast and contrast enhanced ct radiomics in colorectal cancers. *Academic radiology*, 26(4):469–479, 2019.

[4] Ji-In Bang, Seunggyun Ha, Sung-Bum Kang, Keun-Wook Lee, Hye-Seung Lee, Jae-Sung Kim, Heung-Kwon Oh, Ho-Young Lee, and Sang Eun Kim. Prediction of neoadjuvant radiation chemotherapy response and survival using pretreatment

[18F]FDG PET/CT scans in locally advanced rectal cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, 43(3):422–431, March 2016.

[5] Uffe Bernchou, Olfred Hansen, Tine Schytte, Anders Bertelsen, Andrew Hope, Douglas Moseley, and Carsten Brink. Prediction of lung density changes after radiotherapy by cone beam computed tomography response markers and pre-treatment factors for non-small cell lung cancer patients. *Radiotherapy and Oncology*, 117(1):17 – 22, 2015.

[6] David Jean Biau, Solen Kernéis, and Raphaël Porcher. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. *Clinical Orthopaedics and Related Research*, 466(9):2282–2288, September 2008.

[7] Ronald Boellaard, Roberto Delgado-Bolton, Wim JG Oyen, Francesco Giammarile, Klaus Tatsch, Wolfgang Eschner, Fred J Verzijlbergen, Sally F Barrington, Lucy C Pike, Wolfgang A Weber, et al. Fdg pet/ct: Eanm procedure guidelines for tumour imaging: version 2.0. *European journal of nuclear medicine and molecular imaging*, 42(2):328–354, 2015.

[8] Marta Bogowicz, Ralph T.H. Leijenaar, Stephanie Tanadini-Lang, Oliver Riesterer, Martin Pruschy, Gabriela Studer, Jan Unkelbach, Matthias Guckenberger, Ender Konukoglu, and Philippe Lambin. Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models. *Radiotherapy and Oncology*, 125(3):385–391, December 2017.

[9] Nathaniel M. Braman, Maryam Etesami, Prateek Prasanna, Christina Dubchuk, Hannah Gilmore, Pallavi Tiwari, Donna Plecha, and Anant Madabhushi. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Research*, 19(1):57, December 2017.

[10] Freddie Bray and D. Max Parkin. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *European Journal of Cancer*, 45(5):747–755, March 2009.

[11] James Brierley, Mary Gospodarowicz, and Brian O'Sullivan. The principles of cancer staging. *ecancermedicalscience*, 10, 2016.

[12] Wray Buntine and Tim Niblett. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8(1):75–85, January 1992.

[13] Neil G. Burnet. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging*, 4(2):153–161, 2004.

[14] Mathieu Burtnyk, William Apoutou N'Djin, Ilya Kobelevskiy, Michael Bronskill, and Rajiv Chopra. 3d conformal mri-controlled transurethral ultrasound prostate therapy: validation of numerical simulations and demonstration in tissue-mimicking gel phantoms. *Physics in Medicine & Biology*, 55(22):6817, 2010.

[15] Donald M. Cannon and Nancy Y. Lee. Recurrence in Region of Spared Parotid Gland After Definitive Intensity-Modulated Radiotherapy for Head and Neck Cancer. *International Journal of Radiation Oncology*Biology*Physics*, 70(3):660–665, March 2008.

[16] Michael L. Cardenas, Thomas R. Mazur, Christina I. Tsien, and Olga L. Green. A rapid, computational approach for assessing interfraction esophageal motion for use in stereotactic body radiation therapy planning. *Advances in Radiation Oncology*, 3(2):209–215, April 2018.

[17] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining - KDD '15, pages 1721–1730, Sydney, NSW, Australia, 2015. ACM Press.

[18] Bruce D. Cheson, Beate Pfistner, Malik E. Juweid, Randy D. Gascoyne, Lena Specht, Sandra J. Horning, Bertrand Coiffier, Richard I. Fisher, Anton Hagenbeek, Emanuele Zucca, Steven T. Rosen, Sigrid Stroobants, T. Andrew Lister, Richard T. Hoppe, Martin Dreyling, Kensei Tobinai, Julie M. Vose, Joseph M. Connors, Massimo Federico, and Volker Diehl. Revised Response Criteria for Malignant Lymphoma. *Journal of Clinical Oncology*, 25(5):579–586, February 2007.

[19] Jooae Choe, Sang Min Lee, Kyung-Hyun Do, Jung Bok Lee, June-Goo Lee, and Joon Beom Seo. Prognostic value of radiomic analysis of iodine overlay maps from dual-energy computed tomography in patients with resectable lung cancer. *European Radiology*, 29(2):915–923, 2019.

[20] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013.

[21] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*, 13(1):1, 2015.

[22] Gary JR Cook, Gurdip Azad, Kasia Owczarczyk, Musib Siddique, and Vicky Goh. Challenges and promises of pet radiomics. *International Journal of Radiation Oncology* Biology* Physics*, 102(4):1083–1089, 2018.

[23] Thibaud P. Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Gretchen Hermann,

Philippe Lambin, Benjamin Haibe-Kains, Raymond H. Mak, and Hugo J.W.L. Aerts. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, March 2015.

[24] Laurence E. Court, Xenia Fave, Dennis Mackin, Joonsang Lee, Jinzhong Yang, and Lifei Zhang. Computational resources for radiomics. *Translational Cancer Research*, 5(4):340–348, August 2016.

[25] Felix E Diehn, Gregory J Michalak, David R DeLone, Amy L Kotsenas, E Paul Lindell, Norbert G Campeau, Ahmed F Halaweish, Cynthia H McCollough, and Joel G Fletcher. CT Dental Artifact: Comparison of an Iterative Metal Artifact Reduction Technique with Weighted Filtered Back-Projection. *Acta Radiologica Open*, 6(11):205846011774327, November 2017.

[26] Maximilian Diehn, Christine Nardini, David S Wang, Susan McGovern, Mahesh Jayaraman, Yu Liang, Kenneth Aldape, Soonmee Cha, and Michael D Kuo. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proceedings of the National Academy of Sciences*, 105(13):5213–5218, 2008.

[27] Cuong V Dinh, Peter Steenbergen, Ghazaleh Ghobadi, Stijn WTJP Heijmink, Floris J Pos, Karin Haustermans, and Uulke A van der Heide. Magnetic resonance imaging for prostate cancer radiotherapy. *Physica Medica*, 32(3):446–451, 2016.

[28] Tai H. Dou, Thibaud P. Coroller, Joost J. M. van Griethuysen, Raymond H. Mak, and Hugo J. W. L. Aerts. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. *PLOS ONE*, 13(11):e0206108, November 2018.

[29] R.P.L. Durgabai and Ravi Bhushan Y. Feature Selection using ReliefF Algorithm. *IJARCCE*, pages 8215–8218, October 2014.

[30] F Earnest 4th, PJr Kelly, BW Scheithauer, BA Kall, TL Cascino, RL Ehman, GS Forbes, and PL Axley. Cerebral astrocytomas: histopathologic correlation of mr and ct contrast enhancement with stereotactic biopsy. *Radiology*, 166(3):823–827, 1988.

[31] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, January 2009.

[32] Issam El Naqa, PW Grigsby, A Apte, E Kidd, E Donnelly, D Khullar, S Chaudhari, Deshan Yang, M Schmitt, Richard Laforest, et al. Exploring feature-based approaches in pet images for predicting cancer treatment outcomes. *Pattern recognition*, 42(6):1162–1171, 2009.

[33] Issam El Naqa, Dan Ruan, Gilmer Valdes, Andre Dekker, Todd McNutt, Yaorong Ge, Q. Jackie Wu, Jung Hun Oh, Maria Thor, Wade Smith, Arvind Rao, Clifton Fuller, Ying Xiao, Frank Manion, Matthew Schipper, Charles Mayo, Jean M. Moran, and Randall Ten Haken. Machine learning and modeling: Data, validation, communication challenges. *Medical Physics*, 45(10):e834–e840, October 2018.

[34] Hesham Elhalawani, Timothy A. Lin, Stefania Volpe, Abdallah S. R. Mohamed, Aubrey L. White, James Zafereo, Andrew J. Wong, Joel E. Berends, Shady AboHashem, Bowman Williams, Jeremy M. Aymard, Aasheesh Kanwar, Subha Perni, Crosby D. Rock, Luke Cooksey, Shauna Campbell, Pei Yang, Khahn Nguyen, Rachel B. Ger, Carlos E. Cardenas, Xenia J. Fave, Carlo Sansone, Gabriele Piantadosi, Stefano Marrone, Rongjie Liu, Chao Huang, Kaixian Yu, Tengfei Li, Yang Yu, Youyi Zhang, Hongtu Zhu, Jeffrey S. Morris, Veerabhadran Baladandayuthapani, John W. Shumway, Alakonanda Ghosh, An-

drei Pöhlmann, Hady A. Phoulady, Vibhas Goyal, Guadalupe Canahuate, G. Elisabeta Marai, David Vock, Stephen Y. Lai, Dennis S. Mackin, Laurence E. Court, John Freymann, Keyvan Farahani, Jayashree Kaplathy-Cramer, and Clifton D. Fuller. Machine Learning Applications in Head and Neck Radiation Oncology: Lessons From Open-Source Radiomics Challenges. *Frontiers in Oncology*, 8:294, August 2018.

[35] Andrew Elson, Eric Paulson, Joseph Bovi, Malika Siker, Chris Schultz, and Peter S. Laviolette. Evaluation of pre-radiotherapy apparent diffusion coefficient (ADC): patterns of recurrence and survival outcomes analysis in patients treated for glioblastoma multiforme. *Journal of Neuro-Oncology*, 123(1):179–188, May 2015.

[36] Jeremy J. Erasmus, Gregory W. Gladish, Lyle Broemeling, Bradley S. Sabloff, Mylene T. Truong, Roy S. Herbst, and Reginald F. Munden. Interobserver and Intraobserver Variability in Measurement of Non–Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response. *Journal of Clinical Oncology*, 21(13):2574–2582, July 2003.

[37] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.

[38] Ming Fan, Hui Li, Shijian Wang, Bin Zheng, Juan Zhang, and Lihua Li. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast cancer. *PLOS ONE*, 12(2):e0171683, February 2017.

[39] Xenia Fave, Dennis Mackin, Jinzhong Yang, Joy Zhang, David Fried, Peter Balter, David Followill, Daniel Gomez, A Kyle Jones, Francesco Stingo, et al. Can radiomics features be reproducibly measured from cbct images for patients with non-small cell lung cancer? *Medical physics*, 42(12):6784–6797, 2015.

[40] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[41] Sandra Fiset, Mattea L. Welch, Jessica Weiss, Melania Pintilie, Jessica L. Conway, Michael Milosevic, Anthony Fyles, Alberto Traverso, David Jaffray, Ur Metser, Jason Xie, and Kathy Han. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and Oncology*, 135:107–114, June 2019.

[42] F Foroudi, A Haworth, A Pangehel, J Wong, P Roxby, G Duchesne, S Williams, and Kh Tai. Inter-observer variability of clinical target volume delineation for bladder cancer using CT and cone beam CT. *Journal of Medical Imaging and Radiation Oncology*, 53(1):100–106, February 2009.

[43] I. Gardin, V. Grégoire, D. Gibon, H. Kirisli, D. Pasquier, J. Thariat, and P. Vera. Radiomics: Principles and radiotherapy applications. *Critical Reviews in Oncology/Hematology*, 138:44–50, June 2019.

[44] Rachel B. Ger, Shouhao Zhou, Pai-Chun Melinda Chi, Hannah J. Lee, Rick R. Layman, A. Kyle Jones, David L. Goff, Clifton D. Fuller, Rebecca M. Howell, Heng Li, R. Jason Stafford, Laurence E. Court, and Dennis S. Mackin. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Scientific Reports*, 8(1):13047, December 2018.

[45] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[46] Adam Gladwish, Michael Milosevic, Anthony Fyles, Jason Xie, Jaydeep Halankar, Ur Metser, Haiyan Jiang, Nathan Becker, Wilfred Levin, Lee Manchul, Warren Foltz, and Kathy Han. Association of Apparent Diffusion Coefficient with Disease Recurrence in Patients with Locally Advanced Cervical Cancer Treated

with Radical Chemotherapy and Radiation Therapy. *Radiology*, 279(1):158–166, April 2016.

[47] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[48] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized Fisher Score for Feature Selection. *arXiv:1202.3725 [cs, stat]*, February 2012. arXiv: 1202.3725.

[49] Steinn Gudmundsson, Thomas Philip Runarsson, and Sven Sigurdsson. Test–retest reliability and feature selection in physiological time series classification. *Computer Methods and Programs in Biomedicine*, 105(1):50–60, January 2012.

[50] Ernest L Hall, Richard P Kruger, Samuel J Dwyer, David L Hall, Robert W Mclaren, and Gwilyu S Lodwick. A survey of preprocessing and feature extraction techniques for radiographic images. *IEEE Transactions on Computers*, 100(9):1032–1044, 1971.

[51] Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and AndrA Vanhoutte. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing & Management*, 25(3):315 – 318, 1989.

[52] Satoshi Hara and Kohei Hayashi. Making Tree Ensembles Interpretable. *arXiv:1606.05390 [stat]*, June 2016. arXiv: 1606.05390.

[53] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, November 1973.

[54] Paul M. Harari. Beware the Swing and a Miss: Baseball Precautions for Conformal Radiotherapy. *International Journal of Radiation Oncology*Biology*Physics*, 70(3):657–659, March 2008.

[55] Charles A Harlow and Sharon A Eisenbeis. The analysis of radiographic images. *IEEE Transactions on Computers*, 100(7):678–689, 1973.

[56] Mathieu Hatt, Mohamed Majdoub, Martin Vallières, Florent Tixier, Catherine Cheze Le Rest, David Groheux, Elif Hindié, Antoine Martineau, Olivier Pradier, Roland Hustinx, 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*, 56(1):38–44, 2015.

[57] Mathieu Hatt, Mohamed Majdoub, Martin Vallières, Florent Tixier, Catherine Cheze Le Rest, David Groheux, Elif Hindié, Antoine Martineau, Olivier Pradier, Roland Hustinx, Remy Perdrisot, Remy Guillevin, Issam El Naqa, and Dimitris Visvikis. 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: Official Publication, Society of Nuclear Medicine*, 56(1):38–44, January 2015.

[58] Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, August 2016.

[59] David Jaffray, Patrick Kupelian, Toufik Djemil, and Roger M Macklis. Review of image-guided radiation therapy. *Expert Review of Anticancer Therapy*, 7(1):89–103, January 2007.

[60] R. I. Jennrich and P. F. Sampson. Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18(1):11–17, 1976.

[61] Wei Jiang, Pranav Lakshminarayanan, Xuan Hui, Peijin Han, Zhi Cheng, Michael Bowers, Ilya Shpitser, Sauleh Siddiqui, Rus-

sell H. Taylor, Harry Quon, and Todd McNutt. Machine Learning Methods Uncover Radiomorphologic Dose Patterns in Salivary Glands that Predict Xerostomia in Patients with Head and Neck Cancer. *Advances in Radiation Oncology*, 4(2):401–412, April 2019.

[62] Le Kang, Weijie Chen, Nicholas A. Petrick, and Brandon D. Gallas. Comparing two correlated *C* indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in Medicine*, 34(4):685–703, February 2015.

[63] A. Kassner and R.E. Thornhill. Texture Analysis: A Review of Neurologic MR Imaging Applications. *American Journal of Neuroradiology*, 31(5):809–816, May 2010.

[64] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):e1002730, January 2019.

[65] Michael W. Kattan, Kenneth R. Hess, Mahul B. Amin, Ying Lu, Karl G.M. Moons, Jeffrey E. Gershenwald, Phyllis A. Gimotty, Justin H. Guinney, Susan Halabi, Alexander J. Lazar, Alyson L. Mahar, Tushar Patel, Daniel J. Sargent, Martin R. Weiser, Carolyn Compton, and members of the AJCC Precision Medicine Core. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine: AJCC Criteria for Risk Models in Precision Medicine. *CA: A Cancer Journal for Clinicians*, 66(5):370–374, September 2016.

[66] Larry G Kessler, Huiman X Barnhart, Andrew J Buckler, Kingshuk Roy Choudhury, Marina V Kondratovich, Alicia Toledano,

Alexander R Guimaraes, Ross Filice, Zheng Zhang, Daniel C Sullivan, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Statistical methods in medical research*, 24(1):9–26, 2015.

[67] Satoshi Kida, Shizuo Kaji, Kanabu Nawa, Toshikazu Imae, Takahiro Nakamoto, Sho Ozaki, Takeshi Ohta, Yuki Nozawa, and Keiichi Nakagawa. Visual enhancement of cone-beam ct by use of cyclegan. *Medical physics*, 47(3):998–1010, 2020.

[68] Hyungjin Kim, Chang Min Park, Myunghee Lee, Sang Joon Park, Yong Sub Song, Jong Hyuk Lee, Eui Jin Hwang, and Jin Mo Goo. Impact of Reconstruction Algorithms on CT Radiomic Features of Pulmonary Tumors: Analysis of Intra- and Inter-Reader Variability and Inter-Reconstruction Algorithm Variability. *PLOS ONE*, 11(10):e0164924, October 2016.

[69] Min-Joo Kim, Seu-Ran Lee, Min-Young Lee, Jason W. Sohn, Hyong Geon Yun, Joon Yong Choi, Sang Won Jeon, and Tae Suk Suh. Characterization of 3D printing techniques: Toward patient specific quality assurance spine-shaped phantom for stereotactic body radiation therapy. *PloS One*, 12(5):e0176227, 2017.

[70] Remy Klaassen, Ruben THM Larue, Banafsche Mearadji, Stephanie O van der Woude, Jaap Stoker, Philippe Lambin, and Hanneke WM van Laarhoven. Feasibility of ct radiomics to predict treatment response of individual liver metastases in esophagogastric cancer patients. *PloS one*, 13(11):e0207362, 2018.

[71] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, November 2012.

[72] Philippe Lambin, Ralph T. H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E. C. de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T. H. M. Larue, Aniek J. G. Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews. Clinical Oncology*, 14(12):749–762, December 2017.

[73] Philippe Lambin, Erik Roelofs, Bart Reymen, Emmanuel Rios Velazquez, Jeroen Buijsen, Catharina M.L. Zegers, Sara Carvalho, Ralph T.H. Leijenaar, Georgi Nalbantov, Cary Oberije, M. Scott Marshall, Frank Hoebers, Esther G.C. Troost, Ruud G.P.M. van Stiphout, Wouter van Elmpt, Trudy van der Weijden, Liesbeth Boersma, Vincenzo Valentini, and Andre Dekker. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiotherapy and Oncology*, 109(1):159–164, October 2013.

[74] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports*, 7(1):10353, December 2017.

[75] Ruben THM Larue, Remy Klaassen, Arthur Jochems, Ralph TH Leijenaar, Maarten CCM Hulshof, Mark I van Berge Henegouwen, Wendy MJ Schreurs, Meindert N Sosef, Wouter van Elmpt, Hanneke WM van Laarhoven, et al. Pre-treatment ct radiomics to predict 3-year overall survival following chemoradiotherapy of esophageal cancer. *Acta Oncologica*, 57(11):1475–1481, 2018.

[76] Sangjune Laurence Lee, Jenny Lee, Tim Craig, Alejandro Berlin, Peter Chung, Cynthia Ménard, and Warren D. Foltz. Changes

in apparent diffusion coefficient radiomics features during dose-painted radiotherapy and high dose rate brachytherapy for prostate cancer. *Physics and Imaging in Radiation Oncology*, 9:1–6, January 2019.

[77] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber, Inge Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Mechthild Krause, Michael Baumann, Esther G. C. Troost, Steffen Löck, and Christian Richter. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):13206, December 2017.

[78] Ralph T.H. Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter J.C. van Elmpt, Esther G.C. Troost, Ronald Boellaard, Hugo J.W.L Aerts, Robert J. Gillies, and Philippe Lambin. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Scientific Reports*, 5(1), September 2015.

[79] Fuquan Liu, Zhenyuan Ning, Yanna Liu, Dengxiang Liu, Jie Tian, Hongwu Luo, Weimin An, Yifei Huang, Jialiang Zou, Chuan Liu, Changchun Liu, Lei Wang, Zaiyi Liu, Ruizhao Qi, Changzeng Zuo, Qingge Zhang, Jitao Wang, Dawei Zhao, Yongli Duan, Baogang Peng, Xingshun Qi, Yuening Zhang, Yongping Yang, Jinlin Hou, Jiahong Dong, Zhiwei Li, Huiguo Ding, Yu Zhang, and Xiaolong Qi. Development and validation of a radiomics signature for clinically significant portal hypertension in cirrhosis (CHESS1701): a prospective multicenter study. *EBioMedicine*, 36:151–158, October 2018.

[80] Eitan Lovat, Musib Siddique, Vicky Goh, Rosalie E Ferner, Gary JR Cook, and Victoria S Warbey. The effect of post-injection 18 f-fdg pet scanning time on texture analysis of peripheral nerve

sheath tumours in neurofibromatosis-1. *Ejnmmi Research*, 7(1):35, 2017.

[81] François Lucia, Dimitris Visvikis, Marie-Charlotte Desseroit, Omar Miranda, Jean-Pierre Malhaire, Philippe Robin, Olivier Pradier, Mathieu Hatt, and Ulrike Schick. Prediction of outcome using pretreatment 18F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. *European Journal of Nuclear Medicine and Molecular Imaging*, 45(5):768–786, 2018.

[82] Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, and Andre Dekker. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*, 126(2):312–317, February 2018.

[83] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, and Laurence Court. Measuring Computed Tomography Scanner Variability of Radiomics Features:. *Investigative Radiology*, 50(11):757–765, November 2015.

[84] Sebastián Maldonado and Richard Weber. A wrapper method for feature selection using Support Vector Machines. *Information Sciences*, 179(13):2208–2217, June 2009.

[85] Rafael G. Mantovani, Tomas Horvath, Ricardo Cerri, Joaquin Vanschoren, and Andre C.P.L.F. de Carvalho. Hyper-Parameter Tuning of a Decision Tree Induction Algorithm. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 37–42, Recife, October 2016. IEEE.

[86] Daniel S. Marcus, Timothy R. Olsen, Mohana Ramaratnam, and Randy L. Buckner. The extensible neuroimaging archive toolkit: An informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*, 5(1):11–33, March 2007.

[87] Sarah A. Mattonen, David A. Palma, Carol Johnson, Alexander V. Louie, Mark Landis, George Rodrigues, Ian Chan, Roya Etemad-Rezai, Timothy P.C. Yeung, Suresh Senan, and Aaron D. Ward. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. *International Journal of Radiation Oncology*Biology*Physics*, 94(5):1121–1128, April 2016.

[88] E. Matzner-Lober, C.M. Suehs, A. Dohan, and N. Molinari. Thoughts on entering correlated imaging variables into a multivariable model: Application to radiomics and texture analysis. *Diagnostic and Interventional Imaging*, 99(5):269–270, May 2018.

[89] L. Daniel Maxim, Ron Niebo, and Mark J. Utell. Screening tests: a review with examples. *Inhalation Toxicology*, 26(13):811–828, November 2014.

[90] Fabian Morsbach, Sebastian Bickelhaupt, Guido A. Wanner, Andreas Krauss, Bernhard Schmidt, and Hatem Alkadhi. Reduction of Metal Artifacts from Hip Prostheses on CT Images of the Pelvis: Value of Iterative Reconstructions. *Radiology*, 268(1):237–244, July 2013.

[91] Benjamin E. Nelms, Wolfgang A. Tomé, Greg Robinson, and James Wheeler. Variations in the Contouring of Organs at Risk: Test Case From a Patient With Oropharyngeal Cancer. *International Journal of Radiation Oncology*Biology*Physics*, 82(1):368–378, January 2012.

[92] Ke Nie, Liming Shi, Qin Chen, Xi Hu, Salma K Jabbour, Ning Yue, Tianye Niu, and Xiaonan Sun. Rectal cancer: assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric mri. *Clinical cancer research*, 22(21):5256–5264, 2016.

[93] Line Nilsen, Anne Fangberget, Oliver Geier, Dag Rune Olsen, and Therese Seierstad. Diffusion-weighted magnetic resonance

imaging for pretreatment prediction and monitoring of treatment response of patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. *Acta Oncologica*, 49(3):354–360, January 2010.

[94] Carolyn J. Niu, Warren D. Foltz, Michael Velec, Joanne L. Moseley, Adil Al-Mayah, and Kristy K. Brock. A novel technique to enable experimental validation of deformable dose accumulation: Experimental validation of deformable dose accumulation. *Medical Physics*, 39(2):765–776, January 2012.

[95] Nancy A Obuchowski, Huiman X Barnhart, Andrew J Buckler, Gene Pennello, Xiao-Feng Wang, Jayashree Kalpathy-Cramer, Hyun J (Grace) Kim, Anthony P Reeves, and for the Case Example Working Group. Statistical issues in the comparison of quantitative imaging biomarker algorithms using pulmonary nodule volume as an example. *Statistical Methods in Medical Research*, 24(1):107–140, February 2015.

[96] Fanny Orlhac, Sarah Boughdad, Cathy Philippe, Hugo Stalla-Bourdillon, Christophe Nioche, Laurence Champion, Michaël Soussan, Frédérique Frouin, Vincent Frouin, and Irène Buvat. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *Journal of Nuclear Medicine*, 59(8):1321–1328, August 2018.

[97] Brian O'Sullivan, Shao Hui Huang, Jie Su, Adam S Garden, Erich M Sturgis, Kristina Dahlstrom, Nancy Lee, Nadeem Riaz, Xin Pei, Shlomo A Koyfman, et al. Development and validation of a staging system for hpv-related oropharyngeal cancer by the international collaboration on oropharyngeal cancer network for staging (icon-s): a multicentre cohort study. *The Lancet Oncology*, 17(4):440–451, 2016.

[98] Shraddha Pandit and Suchita Gupta. A Comparative Study on Distance Measuring Approaches for Clustering. *International*

Journal of Research in Computer Science, 2(1):29–31, December 2011.

[99] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo J. W. L. Aerts. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports*, 5(1):13087, October 2015.

[100] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains, Philippe Lambin, and Hugo J. W. L. Aerts. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLOS ONE*, 9(7):1–8, 07 2014.

[101] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains, Philippe Lambin, and Hugo J. W. L. Aerts. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. *PLoS ONE*, 9(7):e102107, July 2014.

[102] Michael J. Pencina and Ralph B. D'Agostino. OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, July 2004.

[103] Lester J. Peters, Brian O'Sullivan, Jordi Giralt, Thomas J. Fitzgerald, Andy Trotti, Jacques Bernier, Jean Bourhis, Kally Yuen, Richard Fisher, and Danny Rischin. Critical Impact of Radiotherapy Protocol Compliance and Quality in the Treatment of Advanced Head and Neck Cancer: Results From TROG 02.02. *Journal of Clinical Oncology*, 28(18):2996–3001, June 2010.

[104] Elisabeth Pfaehler, Roelof J. Beukinga, Johan R. de Jong, Riemer H. J. A. Slart, Cornelis H. Slump, Rudi A. J. O. Dierckx, and Ronald Boellaard. Repeatability of 18 F-FDG PET radiomic

features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Medical Physics*, 46(2):665–678, February 2019.

[105] Csaba Pinter, Andras Lasso, An Wang, David Jaffray, and Gabor Fichtinger. SlicerRT: Radiation therapy research toolkit for 3D Slicer: SlicerRT: Radiation therapy research toolkit for 3D Slicer. *Medical Physics*, 39(10):6332–6338, September 2012.

[106] David Martin Ward Powers. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2011.

[107] Prateek Prasanna, Jay Patel, Sasan Partovi, Anant Madabhushi, and Pallavi Tiwari. Radiomic features from the peritumoral brain parenchyma on treatment-naïve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings. *European Radiology*, 27(10):4188–4197, October 2017.

[108] James M Provenzale, Srinivasan Mukundan, and Daniel P Barboriak. Diffusion-weighted and perfusion mr imaging for brain tumor characterization and assessment of treatment response. *Radiology*, 239(3):632–649, 2006.

[109] Audrey Pépin, Joël Daouk, Pascal Bailly, Sébastien Hapdey, and Marc-Etienne Meyer. Management of respiratory motion in PET/computed tomography: the state of the art. *Nuclear Medicine Communications*, 35(2):113–122, February 2014.

[110] Sylvain Reuzé, Antoine Schernberg, Fanny Orlhac, Roger Sun, Cyrus Chargari, Laurent Dercle, Eric Deutsch, Irène Buvat, and Charlotte Robert. Radiomics in Nuclear Medicine Applied to Radiation Therapy: Methods, Pitfalls, and Challenges. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1117–1142, November 2018.

[111] Adam C. Riegel, Anthony M. Berson, Sylvie Destian, Tracy Ng, Lawrence B. Tena, Robin J. Mitnick, and Ping S. Wong. Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. *International Journal of Radiation Oncology*Biology*Physics*, 65(3):726–732, July 2006.

[112] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Cristiana Fanciullo, Alessio Giuseppe Morganti, and Massimo Bellomi. Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental*, 2(1):36, December 2018.

[113] Dale Roach, Lois C Holloway, Michael G Jameson, Jason A Dowling, Angel Kennedy, Peter B Greer, Michele Krawiec, Robba Rai, Jim Denham, Jeremiah De Leon, Karen Lim, Megan E Berry, Rohen T White, Sean A Bydder, Hendrick T Tan, Jeremy D Croker, Alycea McGrath, John Matthews, Robert J Smeenk, and Martin A Ebert. Multi-observer contouring of male pelvic anatomy: Highly variable agreement across conventional and emerging structures of interest. *Journal of Medical Imaging and Radiation Oncology*, 63(2):264–271, April 2019.

[114] Bert-Ram Sah, Kasia Owczarczyk, Musib Siddique, Gary JR Cook, and Vicky Goh. Radiomics in esophageal and gastric cancer. *Abdominal radiology*, 44(6):2048–2058, 2019.

[115] Helena Sandström, Hidefumi Jokura, Caroline Chung, and Iuliana Toma-Dasu. Multi-institutional study of the variability in target delineation for six targets commonly treated with radiosurgery. *Acta Oncologica*, 57(11):1515–1520, November 2018.

[116] Sebastian Sanduleanu, Henry C. Woodruff, Evelyn E.C. de Jong, Janna E. van Timmeren, Arthur Jochems, Ludwig Dubois, and Philippe Lambin. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiotherapy and Oncology*, 127(3):349–360, June 2018.

[117] Francesco Sardanelli, Marco Alì, Myriam G. Hunink, Nehmat Houssami, Luca M. Sconfienza, and Giovanni Di Leo. To share or not to share? Expected pros and cons of data sharing in radiological research. *European Radiology*, 28(6):2328–2335, June 2018.

[118] E. Scalco, S. Moriconi, and G. Rizzo. Texture analysis to assess structural modifications induced by radiotherapy. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5219–5222, Milan, August 2015. IEEE.

[119] Matthew Seidler, Behzad Forghani, Caroline Reinhold, Almudena Pérez-Lara, Griselda Romero-Sanchez, Nikesh Muthukrishnan, Julian L Wichmann, Gabriel Melki, Eugene Yu, and Reza Forghani. Dual-energy ct texture analysis with machine learning for the evaluation and characterization of cervical lymphadenopathy. *Computational and Structural Biotechnology Journal*, 17:1009–1015, 2019.

[120] Muhammad Shafiq-ul Hassan, Geoffrey G. Zhang, Kujtim Latifi, Ghanim Ullah, Dylan C. Hunt, Yoganand Balagurunathan, Mahmoud Abrahem Abdalah, Matthew B. Schabath, Dmitry G. Goldgof, Dennis Mackin, Laurence Edward Court, Robert James Gillies, and Eduardo Gerardo Moros. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Medical Physics*, 44(3):1050–1062, mar 2017.

[121] Lalitha K Shankar, John M Hoffman, Steve Bacharach, Michael M Graham, Joel Karp, Adriaan A Lammertsma, Steven Larson, David A Mankoff, Barry A Siegel, Annick Van den Abbeele, et al. Consensus recommendations for the use of 18f-fdg pet as an indicator of therapeutic response in patients in national cancer institute trials. *Journal of Nuclear Medicine*, 47(6):1059–1066, 2006.

[122] Zhenwei Shi, Alberto Traverso, Johan Soest, Andre Dekker, and Leonard Wee. Technical Note: Ontology-guided radiomics anal-

ysis workflow (O-RAW). *Medical Physics*, 46(12):5677–5684, December 2019.

[123] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, May 2016.

[124] Rakesh Shiradkar, Tarun K Podder, Ahmad Algohary, Satish Viswanath, Rodney J. Ellis, and Anant Madabhushi. Radiomics based targeted radiotherapy planning (Rad-TRaP): a computational framework for prostate cancer treatment planning with MRI. *Radiation Oncology*, 11(1):148, December 2016.

[125] Isaac Shiri, Arman Rahmim, Pardis Ghaffarian, Parham Geramifar, Hamid Abdollahi, and Ahmad Bitarafan-Rajabi. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. *European Radiology*, 27(11):4498–4509, November 2017.

[126] Paul Shreve and David W. Townsend. *Clinical PET-CT in Radiology: Integrated Imaging in Oncology*. Springer New York, New York, NY, 2011.

[127] Amita Shukla-Dave, Nancy A Obuchowski, Thomas L Chenevert, Sachin Jambawalikar, Lawrence H Schwartz, Dariya Malyarenko, Wei Huang, Susan M Noworolski, Robert J Young, Mark S Shiroishi, et al. Quantitative imaging biomarkers alliance (qiba) recommendations for improved precision of dwi and dce-mri derived biomarkers in multicenter oncology trials. *Journal of Magnetic Resonance Imaging*, 49(7):e101–e121, 2019.

[128] Amita Shukla-Dave, Nancy A. Obuchowski, Thomas L. Chenevert, Sachin Jambawalikar, Lawrence H. Schwartz, Dariya Malyarenko, Wei Huang, Susan M. Noworolski, Robert J. Young,

Mark S. Shiroishi, Harrison Kim, Catherine Coolens, Hendrik Laue, Caroline Chung, Mark Rosen, Michael Boss, and Edward F. Jackson. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *Journal of Magnetic Resonance Imaging*, 49(7):e101–e121, June 2019.

[129] Roel J.H.M. Steenbakkers, Joop C. Duppen, Isabelle Fitton, Kirsten E.I. Deurloo, Lambert Zijp, Apollonia L.J. Uitterhoeve, Patrick T.R. Rodrigus, Gijsbert W.P. Kramer, Johan Bussink, Katrien De Jaeger, José S.A. Belderbos, Augustinus A.M. Hart, Peter J.C.M. Nowak, Marcel van Herk, and Coen R.N. Rasch. Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: A 'Big Brother' evaluation. *Radiotherapy and Oncology*, 77(2):182–190, November 2005.

[130] Radka Stoyanova, Mandeep Takhar, Yohann Tschudi, John C. Ford, Gabriel Solórzano, Nicholas Erho, Yoganand Balagurunathan, Sanoj Punnen, Elai Davicioni, Robert J. Gillies, and Alan Pollack. Prostate cancer radiomics and the promise of radiogenomics. *Translational Cancer Research*, 5(4):432–447, August 2016.

[131] Michal Strzelecki, Piotr Szczypinski, Andrzej Materka, and Artur Klepaczko. A software tool for automatic classification and segmentation of 2D/3D medical images. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 702:137–140, February 2013.

[132] Yiqun Sun, Panpan Hu, Jiazhou Wang, Lijun Shen, Fan Xia, Gan Qing, Weigang Hu, Zhen Zhang, Chao Xin, Weijun Peng, et al. Radiomic features of pretreatment mri could identify t stage in patients with rectal cancer: Preliminary findings. *Journal of Magnetic Resonance Imaging*, 48(3):615–621, 2018.

[133] Temel Tirkes, Margaret A. Hollar, Mark Tann, Marc D. Kohli, Fatih Akisik, and Kumaresan Sandrasegaran. Response Criteria in Oncologic Imaging: Review of Traditional and New Criteria. *RadioGraphics*, 33(5):1323–1341, September 2013.

[134] Georgia D. Tourassi. Journey toward Computer-aided Diagnosis: Role of Image Texture Analysis. *Radiology*, 213(2):317–320, November 1999.

[135] Alberto Traverso, Michal Kazmierski, Zhenwei Shi, Petros Kalendralis, Mattea Welch, Henrik Dahl Nissen, David Jaffray, Andre Dekker, and Leonard Wee. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Physica Medica*, 61:44–51, May 2019.

[136] Alberto Traverso, Michal Kazmierski, Mattea L. Welch, Jessica Weiss, Sandra Fiset, Warren D. Foltz, Adam Gladwish, Andre Dekker, David Jaffray, Leonard Wee, and Kathy Han. Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients. *Radiotherapy and Oncology*, page S0167814019330476, August 2019.

[137] 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*, 71:24–30, March 2020.

[138] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[139] José Unpingco. *Python for Probability, Statistics, and Machine Learning*. Springer International Publishing, Cham, 2019.

[140] Martin Vallières, Alex Zwanenburg, Bodgan Badic, Catherine Cheze Le Rest, Dimitris Visvikis, and Mathieu Hatt. Responsible Radiomics Research for Faster Clinical Translation. *Journal of Nuclear Medicine*, 59(2):189–193, February 2018.

[141] Astrid van der Horst, Antonetta C. Houweling, Geertjan van Tienhoven, Jorrit Visser, and Arjan Bel. Dosimetric effects of anatomical changes during fractionated photon radiation therapy in pancreatic cancer patients. *Journal of Applied Clinical Medical Physics*, 18(6):142–151, November 2017.

[142] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[143] Janna E. van Timmeren, Ralph T.H. Leijenaar, Wouter van Elmpt, Bart Reymen, Cary Oberije, René Monshouwer, Johan Bussink, Carsten Brink, Olfred Hansen, and Philippe Lambin. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam ct images. *Radiotherapy and Oncology*, 123(3):363 – 369, 2017.

[144] Shalini K. Vinod, Michael G. Jameson, Myo Min, and Lois C. Holloway. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiotherapy and Oncology*, 121(2):169–179, November 2016.

[145] Stefan Wachter, Natascha Wachter-Gerstner, Thomas Bock, Gregor Goldner, György Kovacs, Annette Fransson, and Richard Pötter. Interobserver comparison of ct and mri-based prostate apex definition clinical relevance for conformal radiotherapy treatment planning. *Strahlentherapie und Onkologie*, 178(5):263–268, 2002.

[146] Lise Wei, Benjamin Rosen, Martin Vallières, Thong Chotchutipan, Michelle Mierzwa, Avraham Eisbruch, and Issam El Naqa. Automatic recognition and analysis of metal streak artifacts in head and neck computed tomography for radiomics modeling. *Physics and Imaging in Radiation Oncology*, 10:49–54, April 2019.

[147] Mattea L Welch and David A Jaffray. radiomics: the new world or another road to el dorado?, 2017.

[148] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[149] Mattea L. Welch, Chris McIntosh, Andrea McNiven, Shao Hui Huang, Bei-Bei Zhang, Leonard Wee, Alberto Traverso, Brian O'Sullivan, Frank Hoebers, Andre Dekker, and David A. Jaffray. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. *Physica Medica*, 70:145–152, February 2020.

[150] Mattea L Welch, Chris McIntosh, Tom G Purdie, Leonard Wee, Alberto Traverso, Andre Dekker, Benjamin Haibe-Kains, and David A Jaffray. Automatic classification of dental artifact status for efficient image veracity checks: effects of image resolution and convolutional neural network depth. *Physics in Medicine & Biology*, 65(1):015005, January 2020.

[151] J. D. Wolchok, A. Hoos, S. O'Day, J. S. Weber, O. Hamid, C. Lebbe, M. Maio, M. Binder, O. Bohnsack, G. Nichol, R. Humphrey, and F. S. Hodi. Guidelines for the Evaluation of Immune Therapy Activity in Solid Tumors: Immune-Related Response Criteria. *Clinical Cancer Research*, 15(23):7412–7420, December 2009.

[152] Jia Wu, Khin Khin Tha, Lei Xing, and Ruijiang Li. Radiomics and radiogenomics for precision radiotherapy. *Journal of Radiation Research*, 59(suppl_1):i25–i31, March 2018.

[153] Wei Xia, Ying Chen, Rui Zhang, Zhuangzhi Yan, Xiaobo Zhou, Bo Zhang, and Xin Gao. Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data—a preliminary study. *Physics in Medicine & Biology*, 63(3):035044, February 2018.

[154] X Xu, X Sui, W Zhong, Y Xu, Z Wang, J Jiang, Y Ge, L Song, Q Du, X Wang, et al. Clinical utility of quantitative dual-energy ct iodine maps and ct morphological features in distinguishing small-cell from non-small-cell lung cancer. *Clinical radiology*, 74(4):268–277, 2019.

[155] C. Yip, F. Davnall, R. Kozarski, D. B. Landau, G. J. R. Cook, P. Ross, R. Mason, and V. Goh. Assessment of changes in tumor heterogeneity following neoadjuvant chemotherapy in primary esophageal cancer: Tumor heterogeneity in esophageal cancer. *Diseases of the Esophagus*, 28(2):172–179, February 2015.

[156] Stephen S F Yip and Hugo J W L Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, July 2016.

[157] Evangelia I Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R Melhem, and Christos Davatzikos. Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(6):1609–1618, 2009.

[158] Lifei Zhang, David V. Fried, Xenia J. Fave, Luke A. Hunter, Jinzhong Yang, and Laurence E. Court. An open infrastructure software platform to facilitate collaborative work in radiomics:

Open infrastructure platform for radiomics. *Medical Physics*, 42(3):1341–1353, February 2015.

[159] Yucheng Zhang, Anastasia Oikonomou, Alexander Wong, Masoom A. Haider, and Farzad Khalvati. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. *Scientific Reports*, 7:46349, 2017.

[160] Yupeng Zhang, Baorui Zhang, Fei Liang, Shikai Liang, Yuxiang Zhang, Peng Yan, Chao Ma, Aihua Liu, Feng Guo, and Chuhan Jiang. Radiomics features on non-contrast-enhanced ct scan can precisely classify avm-related hematomas from other spontaneous intraparenchymal hematoma types. *European radiology*, 29(4):2157–2165, 2019.

[161] Rongrong Zhou, Zhongxing Liao, Tinsu Pan, Sarah A. Milgrom, Chelsea C. Pinnix, Anhui Shi, Linglong Tang, Ju Yang, Ying Liu, Daniel Gomez, Quynh-Nhu Nguyen, Bouthaina S. Dabaja, Laurence Court, and Jinzhong Yang. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. *Radiotherapy and Oncology*, 122(1):66–71, January 2017.

[162] Ivan Zhovannik, 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*, 19:33–38, November 2019.

[163] Alex Zwanenburg. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(13):2638–2655, December 2019.

[164] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

3

Repeatability and reproducibility of radiomic features: a systematic review

Adapted from: **"Repeatability and reproducibility of radiomic features: a systematic review"**. A Traverso, L Wee, A Dekker, R Gillies. International Journal of Radiation Oncology* Biology* Physics 102 (4), 1143-1158. (2018).

Abstract

An ever-growing number of predictive models used to inform clinical decision making have included quantitative, computer-extracted imaging biomarkers, or *radiomic features*. Broadly generalizable validity of radiomics-assisted models may be impeded by concerns about reproducibility. We offer a qualitative synthesis of 41 studies that specifically investigated the repeatability and reproducibility of radiomic features, derived from a systematic review of published peer-reviewed literature. The PubMed electronic database was searched using combinations of the broad Haynes and Ingui filters along with a set of text words specific to cancer, radiomics (including texture analyses), reproducibility, and repeatability. This review has been reported in compliance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. From each full-text article, information was extracted regarding cancer type, class of radiomic feature examined, reporting quality of key processing steps, and statistical metric used to segregate stable features. Among 624 unique records, 41 full-text articles were subjected to review. The studies primarily addressed non-small cell lung cancer and oropharyngeal cancer. Only 7 studies addressed in detail every methodologic aspect related to image acquisition, pre-processing, and feature extraction. The repeatability and reproducibility of radiomic features are sensitive at various degrees to processing details such as image acquisition settings, image reconstruction algorithm, digital image pre-processing, and software used to extract radiomic features. First-order features were overall more reproducible than shape metrics and textural features. Entropy was consistently reported as one of the most stable first-order features. There was no emergent consensus regarding either shape metrics or textural features; however, coarseness and contrast appeared among the least reproducible. Investigations of feature repeatability and reproducibility are currently limited to a small number of cancer types. Reporting quality could be improved regarding details of feature extraction software, digital image manipulation

(pre-processing), and the cut-off value used to distinguish stable features.

3.1 INTRODUCTION

Medical imaging is widespread, and its value is firmly established in routine oncologic practice. Image-based biomarkers are used during screening, staging, stratifying, and intervention planning (surgery and radiation therapy)[20][57][59] and for predicting treatment outcomes [1]. In current practice, a radiologist semantically annotates only a small number of radiologic features as having some clinical significance during manual assessment of the images (i.e. with the unaided human eye). These few features may include Response Evaluation Criteria in Solid Tumours [19] and World Health Organization criteria [62] for treatment response; a change in the mean apparent diffusion coefficient [10]; morphologic descriptors (e.g. spiculation) [74]; or the number of voxels exceeding a threshold for selective uptake of a radioactive tracer [63]. Tumour phenotypes, as they are manifest in medical images, may contain more information than can be readily processed by the unaided human eye. Recent studies have suggested that complex shape metrics and the nonuniform appearance of the tumour mass in images (i.e. texture) also provide information about the likely outcome of the disease [58] [13][2][49]. Radiomics is the computerized extraction of quantitative features from medical images, beyond the level of detail accessible to an unaided human eye, with the intent to automatically label clinically significant tumour phenotypes [41] . Radiomics entails large-scale batch processing [40] via high-throughput computational "pipelines" that integrate some or all of the following steps: image pre-processing, tumour segmentation, feature extraction, feature selection, machine learning–based predictive modelling, and model validation [42]. A systematic review of false discovery rates in textural analysis of medical images [11] clearly showed that optimal cut-off selection for tuning machine-learning predictive models [3] in combination with a large number of candidate image features (approximately 100) leads to an increased risk of type I error. Furthermore, related sets of image-derived features tend to be strongly correlated with each other, and this increases the risk of falsely significant associations. There are strong interclass correlations for features derived from sim-

ilar matrix operations [4], and there may be correlations to the absolute tumour volume [55]. There are ways to reduce the risk of a false-positive association. First, only features with high repeatability and high reproducibility should be used for training the predictive models. "Repeatability" refers to features that remain the same when imaged multiple times in the same subject, be that a human person or a suitable phantom [40] [53]. "Reproducibility" refers to features that remain the same when imaged using different equipment, different software, different image acquisition settings, or different operators (e.g. other clinics), be that in the same subject or in different subjects [40][53]. Second, estimates of predictive performance in single-institution cohorts should include multiple-folded repeated cross validation to minimize the risk of overfitting [7]. Last, assessment of predictive models based on radiomic features should be based on independent external validation in multi-institutional settings [14]. The purpose of this systematic review was to determine which broadly generic type of radiomic features has been shown to be repeatable and/or reproducible in peer-reviewed studies and, if applicable, what degree of repeatability and reproducibility might be achievable. It was out of the scope of the reviewers performing this study to provide any subjective evaluation concerning the goodness of a study. For example, when evaluating the quality of reporting of a study, we only retrieved objective information from the article, without assigning a score aiming at judging the overall quality of the article.

3.2 METHODS AND MATERIAL

3.2.1 Eligibility criteria

We conducted this systematic review during March and April 2017. Reporting of this review complies with the PRISMA-P Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement [50]. The included articles met all of the eligibility criteria given in the subsequent paragraphs.

3.2.2 Report design

We included only peer-reviewed full-text reports published in journals that presented full results of repeatability and/or reproducibility tests on radiomic features. With full results, we intend all the articles that matching the inclusion criteria defined in material and methods, presented a statistical analysis of radiomics reproducibility and repeatability. Only articles that included (in their titles or abstracts) at least 1 of the search words specified in the search string were identified.

3.2.3 Population

With regard to the population reported in the study, we included either (1) studies of human persons diagnosed with 1 (or more) known and clearly stated primary solid tumour where medical imaging in the form of computed tomography (CT), positron emission tomography (PET), and/or magnetic resonance imaging (MRI) was used or (2) studies consisting of radiologic phantoms where medical imaging in the form of CT, PET, and/or MRI was used. We excluded studies consisting of animal subjects, studies using biological samples taken from the human body, nonclinical imaging studies, or studies in which the type of primary tumour was not objectively known.

3.2.4 Outcomes

The primary criterion for inclusion was an assessment of the repeatability and/or reproducibility of any number of radiomic features with respect to any equipment-, scan-, subject-, or observer-related cause. Included studies also had to report at least 1 of the following quantitative outcomes of interest: variability of radiomic features with respect to image acquisition parameters, imaging modalities examined, or effect of pre-processing steps applied to the images from which features were extracted.

3.2.5 Language

Only full-text reports in the English language were included in this review.

3.2.6 Information sources

The Cochrane Database of Systematic Reviews was screened for any previous systematic reviews addressing repeatability and/or reproducibility of radiomic features. An electronic search was conducted in PubMed (MEDLINE citations had been previously merged into the PubMed repository). For all articles for which the full text was obtained for data extraction, the bibliographic references within them were also screened for potentially eligible studies. No search was made in gray-literature sources for unpublished material or conference proceedings.

3.2.7 Search strategy

A search of PubMed citations was performed using the broad Haynes [31] and Ingui [35] filters in combination with the modifications proposed by Geersing et al [27] (each combined using "OR"). For the final database search, additional criteria of "cancer" (Medical Subject Headings major topic) and text terms that were each related to reproducibility, repeatability (fundamental to include test-retest studies), variability, and radiomics (including textural analyses) were also included. All PubMed search results were admitted up to and including the second week of April 2017.

3.2.8 Study records

Data management

Electronic full-text articles were downloaded using university library subscriptions. A review-specific SharePoint (Microsoft Corporation, USA) page was set up to handle document collection, data extraction forms, and dissemination of reviewer findings.

Selection process

Two reviewers worked independently throughout all phases of the study selection process (abstract screening, eligibility, and inclusion for full-text evaluation). They compared the titles and abstracts against the inclusion criteria. Each reported whether an abstract was eligible for evaluation in full. Disagreements were resolved by consensus. All of the articles deemed eligible were successfully downloaded. Two reviewers independently evaluated whether the full-text reports were suitable for inclusion and synthesis. Disagreements were again resolved by consensus. A third reviewer was available if disagreements could not be resolved, but this option was not exercised. Reasons for excluding a specific full-text article were documented.

3.2.9 Data extraction: data items

We extracted information about the population used in the studies, including the sample size and type of primary tumour for human studies and the phantom details for phantom studies. The inclusion of metastatic, secondary, or synchronous tumours was noted, as was any case in which the nonprimary tumour was used in the derivation of radiomic features. The study design and image modality used were noted, including any image acquisition parameters explicitly stated in the text. We noted the total number of radiomic features tested and

grouped these features according to (1) shape features (defining the 2-dimensional or 3-dimensional [3D] properties of the tumour, e.g. volume or surface area); (2) first-order statistics (derived from statistical moments of the image intensity histogram); and (3) higher-order textural features (describing spatial patterns of voxel intensities) [6]. In addition, we noted the names and versions of the software used to quantitatively extract radiomic features, including whether any particular pre-processing steps were applied to the images before feature extraction. Finally, we noted the statistical methods used as a metric of repeatability and/or reproducibility of the studied features.

3.2.10 Outcomes and prioritizations

Primary outcome

The primary outcome of interest in this review synthesis was the degree of repeatability or reproducibility of radiomic features, along with any independent validation used to test repeatability and reproducibility at an external institution.

Secondary outcomes

The secondary outcomes were the impact of image acquisition settings on the reproducibly of features and the effect of pre-processing imaging filters applied before feature extraction.

Additional outcomes

Additional outcomes were the statistics and metrics used for reporting robustness and reliability and the investigation of the impact of different segmentation methods used to define the regions of interest (ROIs).

Risk of bias in individual studies

To assess the risk of bias in each study, 2 reviewers independently extracted detailed information from the reports in the following specific domains:

- Characteristics of the cohort used to perform the study in the case of human studies or characteristics of the phantom used to perform the study in the case of phantom studies.
- Description of the software used to compute the features
- Image acquisition parameters reported in the study
- Filtering and/or image pre-processing operation(s) performed on the original scan before the radiomic features underwent computation
- Segmentation method(s) used to derive an ROI
- Use of either cross validation or independent validation to show that features are repeatable and/or reproducible after folding or in separate data sets
- Threshold (cut-off) values used in repeatability and/or reproducibility metric(s) to segregate features

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis classifications were not applicable here because tests of repeatability and reproducibility of radiomic features do not strictly link to diagnostic verification or predictive performance. The impact of undocumented (or inadequately reported) steps on the potential repeatability or reproducibility of features was noted. Discrepancies in data extraction between the 2 reviewers were resolved by consensus after discussion. A third reviewer was available to resolve a deadlock, but this option was not needed.

3.2.11 Data synthesis

The included studies were not uniform by way of reported metrics, and we could not attempt a quantitative meta-analysis of pooled metrics. A systematic qualitative synthesis is given in this publication, with details presented in text and tables to summarize our findings about the included studies.

Subgroup analyses

The summary findings on repeatability and reproducibility of features were grouped by the following: disease type (lung cancer, head and neck cancer, and other anatomic sites) or phantom study and type of imaging modality (CT, PET, and MRI).

3.3 RESULTS

3.3.1 Literature search results

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram is shown in Figure 3.1. The PubMed search yielded 624 abstracts for screening against our selection criteria, reduced to 623 after elimination of duplicates. The full text was retrieved for 52 abstracts deemed suitable for in-depth evaluation, including 5 that were located in the references of retrieved studies and 2 previously known studies. After full-text evaluation, 11 studies were excluded because they did not meet the aforementioned eligibility criteria. A qualitative synthesis was derived from 41 studies, of which 35 were performed in human subjects and 6 were exclusively performed on radiologic phantoms. A detailed checklist is provided in Appendix E1 (available online at the following link).

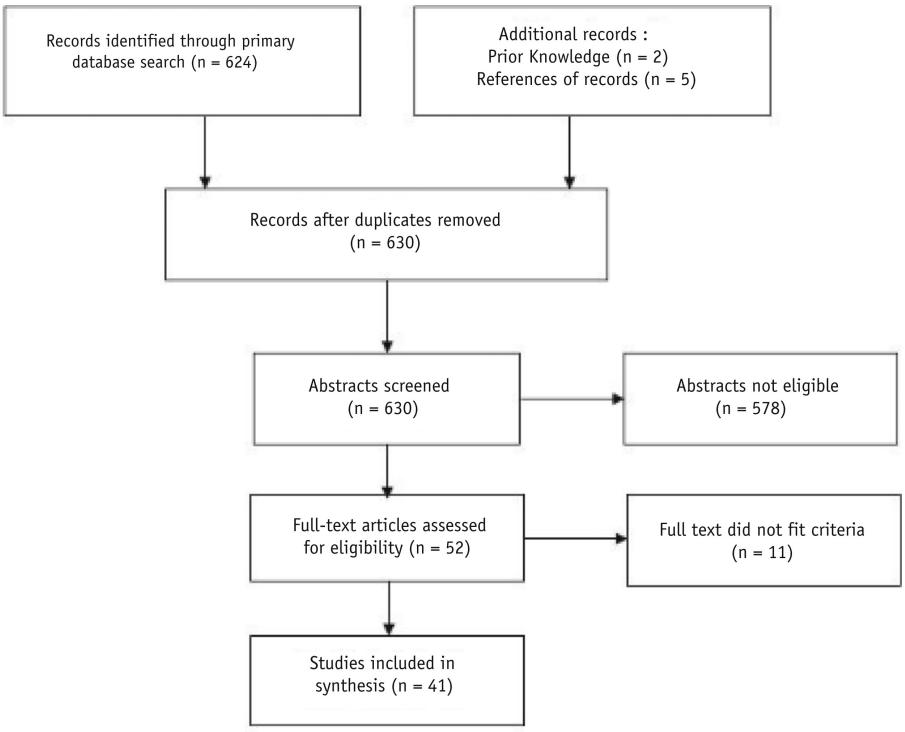


Figure 3.1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram. The primary PubMed search returned 624 records. A further 5 records were added from references in full-text articles. Two records were added owing to prior knowledge. After screening and full-text assessment, a total of 41 studies were included in the qualitative synthesis.

3.3.2 Human study characteristics

Table 3.3.2 summarizes the general characteristics of the human studies. The vast majority of studies addressed lung cancer (25 of 35 studies), of which 21 specifically addressed non-small cell lung cancer (NSCLC). There were 3 studies each on head and neck cancer, oesophageal cancer, and rectal cancer. There was only 1 study each on breast cancer and cervical cancer. In 2 studies, multiple cancer types were combined, but details within the subgroups of cancer types were not specified [26][30]. Two studies combined features from multiple specified cancers [55][65]. The number of patients reported in the retrieved studies ranged from 10 (33) to 555 (54), and only 1 study was prospective [32]. Few studies (7 of 41) specifically referred to a publicly available image set. The imaging modalities mentioned in the human lung studies were PET (17 of 35), CT (17 of 35), MRI (1 of 35), and cone beam CT (CBCT) (2 of 35). Two studies used multiple imaging modalities. Six studies exclusively investigated feature repeatability; all others examined either reproducibility alone or both reproducibility and repeatability. The number of investigated radiomic features in the studies ranged from 4 [69] to 830 [36]. The latter was a multi-institutional study, so it was unclear whether the number included repeated instances of some of the features. All the studies included textural analysis; the majority (28 of 35) also evaluated first-order features, but less than half (15 of 35) evaluated shape metrics. Fourteen studies investigated all categories of features.

3.3.3 Phantom study characteristics

Table 3.3 shows the main characteristics of 6 studies exclusively concerning radiologic phantoms. Among these, CT was the most common image modality (5 of 6), and PET was investigated in only 1 study. We did not locate any phantom study of repeatable and/or reproducible features from MRI. All of these studies investigated feature reproducibility, and only 1 phantom study was prospective

Table 1 Summary of human studies included in analysis

Reference	Disease	Modality investigated	No. of patients	Primary set made public	Total NO. Of features	Class of features	Statistical analysis	Type of study
Aerts et al (4), 2014	NSCLC	CT	52	Yes	440	FO, SM, TA	ICC	Repeatability, reproducibility
Balagurunathan et al (28), 2014	NSCLC	CT	32	Yes	330	FO, SM, TA	CCC	Reproducibility
Cheng et al (29), 2016	NSCLC	PET	56	No	12	TA	ICC	Reproducibility
Coroller et al (30), 2016	NSCLC	CT	32	Yes	1603	FO, SM, TA	ICC	Repeatability
Desseroit et al (31), 2017	NSCLC	PET, CT	73	No	49	FO, TA	ICC	Repeatability
Desseroit et al (32), 2016	NSCLC	CT	32	Yes	34	FO, SM, TA	Mean standard deviation	Repeatability
Fave et al (33), 2015	NSCLC	CBCT	10	No	68	FO, TA	CCC	Reproducibility
Fave et al (34), 2015	NSCLC	CT	40	No	20	FO, SM, TA	Spearman correlation	Reproducibility
Fave et al (35), 2016	NSCLC	CT	134	No	55	FO, TA	Spearman correlation	Reproducibility
Fried et al (36), 2014	NSCLC	CT	91	No	30	FO, TA	CCC	Reproducibility
Huynh et al (37), 2017	NSCLC	CT	32	Yes	644	FO, SM, TA	ICC	Repeatability
Kalpathy-Cramer et al (38), 2016	NSCLC	CT	40	Yes	830	FO, SM, TA	CCC	Reproducibility
Leijenaar et al (39), 2013	NSCLC	PET	34	No	108	FO, SM, TA	ICC	Repeatability
Mackin et al (40), 2015	NSCLC	CT	20	No	10	FO, TA	Mean standard deviation	Reproducibility
Oliver et al (41), 2015	NSCLC	PET	23	No	56	FO, SM, TA	Average percentage difference	Reproducibility
Parmar et al (42), 2014	NSCLC	PET	20	Yes	56	FO, SM, TA	ICC	Reproducibility
Van Velden et al (43), 2016	NSCLC	PET	11	No	105	FO, SM, TA	ICC	Reproducibility
Yan et al (44), 2015	NSCLC	PET	17	No	64	FO, TA	Mean standard deviation	Reproducibility
Yip et al (45), 2014	NSCLC	PET	26	No	4	TA	Relative difference	Reproducibility
Zhao et al (46), 2016	NSCLC	CT	32	No	89	FO, SM, TA	CCC	Repeatability, reproducibility
Grootjans et al (47), 2016	Lung cancer	PET	60	No	4	TA	Mean standard deviation	Reproducibility
Koo et al (48), 2017	Lung cancer	CT	194	No	9	FO, SM	ICC	Reproducibility
Lasnon et al (49), 2016	Lung cancer	PET	60	No	5	TA	Mean standard deviation	Reproducibility
Bagher-Ebadian et al (50), 2017	Oropharyngeal cancer	CT, CBCT	18	No	165	FO, TA	Mean absolute percentage change	Reproducibility
Bogowicz et al (51), 2016	Oropharyngeal cancer	CT	22	No	315	FO, TA	ICC	Reproducibility

(*continued on next page*)

Table 1 (*continued*)

Reference	Disease	Modality investigated	No. of patients	Primary set made public	Total NO. Of features	Class of features	Statistical analysis	Type of study
Lu et al (52), 2016	Oropharyngeal cancer	PET	40	No	88	FO, SM, TA	ICC	Reproducibility
Doumou et al (53), 2015	Esophageal cancer	PET	64	No	57	TA	Spearman correlation	Reproducibility
Hatt et al (54), 2013	Esophageal cancer	PET	50	No	10	FO, TA	Mean standard deviation	Reproducibility
Tixier et al (55), 2012	Esophageal cancer	PET	16	No	25	FO, TA	ICC	Reproducibility
Hu et al (56), 2016	Rectal cancer	CT	40	No	775	TA	ICC	Repeatability
Orlhac et al (19), 2014	Rectal cancer, NSCLC, breast cancer	PET	28, 24, 54	No	41	FO, TA	Spearman correlation	Reproducibility
Van Timmeren et al (57), 2016	Rectal cancer, lung cancer	CT	40, 40	No	542	FO, SM, TA	CCC	Repeatability
Guan et al (58), 2016	Cervical cancer	MR	51	No	8	FO, TA	ICC	Reproducibility
Galavis et al (59), 2010	Various solid tumors	PET	20	No	50	FO, TA	Average percentage difference	Reproducibility
Hatt et al (60), 2015	Various solid tumors	PET	555	No		TA	Spearman correlation	Reproducibility

Abbreviations: CBCT = cone beam computed tomography; CCC = concordance correlation coefficient; CT = computed tomography; FO = first order; ICC = intraclass correlation coefficient; MR = magnetic resonance; NSCLC = non-small cell lung cancer; PET = positron emission tomography; SM = shape metric; TA = textural analysis.

Figure 3.2: Summary of human studies included in analysis. Abbreviations: CBCT = cone beam computed tomography; CCC = concordance correlation coefficient; CT = computed tomography; FO = first order; ICC = intraclass correlation coefficient; MR = magnetic resonance; NSCLC = non-small cell lung cancer; PET = positron emission tomography; SM = shape metric; TA = textural analysis.

Table 2 Summary of pure phantom studies included in analysis

Reference	Phantom used	Modality investigated	Primary set made public	Total no. of features	Class of features	Statistical analysis	Type of study
Buch et al (61), 2017	Nonanatomic in-house phantom	CT	No	42	FO, TA	Student *t* test	Reproducibility
Forgacs et al (62), 2016	NEMA-IQ phantom	PET	No	26	TA	Coefficient of variation	Reproducibility
Kim et al (63), 2015	AAPM CT performance phantom model 76-410-4130	CT	No	5	TA	Unclear	Reproducibility
Lo et al (64), 2016	Water phantom	CT	No	26	FO, TA	Mean standard deviation	Reproducibility
Shafiq-Ul-Hassan et al (65), 2017	Credence Cartridge Radiomics phantom	CT	No	213	FO, SM, TA	Coefficient of variation	Reproducibility
Zhao et al (66), 2014	Anthropomorphic thorax phantom	CT	No	15	SM, TA	Multilinear regression	Reproducibility

Abbreviations: AAPM = American Association of Physicists in Medicine; CT = computed tomography; FO = first order; NEMA-IQ = National Electrical Manufacturers Association–Image Quality; PET = positron emission tomography; SM = shape metric; TA = textural analysis.

Figure 3.3: Summary of pure phantom studies included in analysis. Abbreviations: AAPM = American Association of Physicists in Medicine; CT = computed tomography; FO = first order; NEMA-IQ = National Electrical Manufacturers Association–Image Quality; PET = positron emission tomography; SM = shape metric; TA = textural analysis.

[60]. The number of investigated radiomic features in the phantom studies ranged from 5 [38] to 213 [60]. All studies included textural analysis, 3 evaluated first-order features, and 2 evaluated shape metrics. Only 1 study investigated all categories of features [60].

3.3.4 Quality of reporting in included studies

Human studies

Table 3.4 gives a summary of methodology and reporting quality for human studies. In general, methodologic aspects were adequately documented. However, only 7 of 35 studies reported detailed information in every one of the aforementioned quality domains [6][22][56][8][47][32][18]. In 3 quality aspects, the overall standard of reporting was lower: (1) providing details of the software implementation to extract radiomic features, (2) providing details pertaining to image pre-processing before extracting radiomic

features, and (3) stating the cut-off value for discriminating a subset of repeatable and/or reproducible features. One study did not provide detailed information about the disease groups used in the analysis, apart from stating that different types of solid cancers were included [26]. The practice of pooling heterogeneous tumours is questionable because there is no a priori reason to assume that any arbitrary feature that may be stable in one disease site will also prove to be stable in others. Software details (application framework used for analysis, programming language, and version) were not reported in detail in 16 studies. Standards for radiomic features have not yet been universally adopted; therefore, software should be described because differences due to feature extraction are likely to influence the apparent stability of features. All but 2 studies [17][16] provided detailed tables describing image acquisition settings including information about scanners (manufacturer, model, reconstruction package, and software version) and scan protocols. Eight studies lacked detailed descriptions regarding pre-processing steps (if any) applied to the original images [1][12][34][37][44][54][71]. Digital image manipulations (e.g. voxel size resampling, de-noising, and sharpening) are known to drastically alter the extracted values, and this is likely to hamper reproducibility across data sets. Six studies did not provide sufficient information regarding the segmentation procedures to define an ROI [17][22][34][48][69][28] [65]. Differences in segmentation methods are likely to bias the stability of shape metrics and perhaps textural and first-order features. Fifteen studies did not document the cut-off value used in their statistical metrics to discriminate between reproducible and irreproducible features. One of these selected only the top-ranking feature from each of 4 feature groups (first order, shape metric, texture, and wavelet filtered) [1]. The subset of stable radiomic features selected in a given study obviously depends on arbitrary threshold values of the repeatability or reproducibility metric; therefore, the cut-off criterion should be clearly stated. Two studies validated feature stability in texture phantoms combined with publicly available clinical images [22][23]. One study assessed feature reproducibility across 7 institutions [37].

Seven studies made their primary data set of clinical images publicly available to other researchers [1][6][15][17][33][37][56], and some of these studies used the same publicly available data set.

Phantom studies

Table 3.5 provides an overview of methodologic aspects and reporting quality for the phantom studies. All studies reported information about the phantom used in the analysis. All provided detailed tables describing image acquisition settings, including information about scanners (manufacturer, model, reconstruction package, and software version) and scan protocols. However, no studies reported detailed information in every one of the aforementioned quality domains. In 2 quality aspects, the overall standard of reporting was lower: (1) providing details of the software implementation to extract radiomic features and (2) stating the cut-off value for discriminating within a subset of reproducible features. Four studies made use of commercially available phantoms originally designed for scanner calibration or image quality checks, whereas 2 studies used an in-house texture phantom [9][60]. Software details (application framework used for analysis, programming language, and version) were not reported in the majority of studies. Five studies described their in-house software as based on MATLAB (The MathWorks) [9][38][24][46][60][71] and Kim et al [38] developed a plug-in for the open-source software ImageJ (National Institutes of Health), but none of these studies made their code accessible. Only Forgacs et al [24] lacked a detailed description regarding pre-processing steps (if any) that were applied to the original images. One study did not provide sufficient information regarding the segmentation procedures to define an ROI [24]. Three studies relied on manual segmentations [9][38][60], and the remainder used semiautomated segmentations while also specifying the algorithms used. Five studies did not document the cut-off values used in their statistical metrics to discriminate between reproducible and irreproducible features

Table 3 Quality of reporting analysis for human studies

Reference	Cohort details stated	Software details stated	Image acquisition settings provided	Preprocessing details provided	Segmentation details provided	Metric thresholds stated
Aerts et al (4), 2014	Yes	Yes	Yes	No	Yes	No
Bagher-Ebadian et al (50), 2017	Yes	Yes	Yes	Yes	Yes	No
Balagurunathan et al (28), 2014	Yes	Yes	Yes	Yes	Yes	Yes
Bogowicz et al (51), 2016	Yes	Yes	Yes	Yes	Yes	Yes
Cheng et al (29), 2016	Yes	Yes	Yes	No	Yes	No
Coroller et al (30), 2016	Yes	No	Yes	Yes	Yes	Yes
Desseroit et al (31), 2017	Yes	No	No	Yes	Yes	No
Desseroit et al (32), 2016	Yes	No	Yes	No	No	No
Doumou et al (53), 2015	Yes	Yes	Yes	Yes	Yes	Yes
Fave et al (33), 2015	Yes	Yes	Yes	Yes	Yes	Yes
Fave et al (34), 2015	Yes	Yes	Yes	Yes	Yes	No
Fave et al (35), 2016	Yes	Yes	Yes	Yes	No	Yes
Fried et al (36), 2014	Yes	Yes	Yes	Yes	Yes	Yes
Galavis et al (59), 2010	No	Yes	Yes	Yes	Yes	Yes
Guan et al (58), 2016	Yes	No	Yes	Yes	Yes	Yes
Grootjans et al (47), 2016	Yes	No	Yes	Yes	Yes	No
Hatt et al (54), 2013	Yes	No	Yes	Yes	Yes	No
Hatt et al (60), 2015	Yes	No	Yes	Yes	Yes	No
Hu et al (56), 2016	Yes	Yes	Yes	Yes	Yes	Yes
Huynh et al (37), 2017	Yes	Yes	Yes	No	No	Yes
Kalpathy-Cramer et al (38), 2016	Yes	Yes	Yes	No	Yes	Yes
Koo et al (48), 2017	Yes	No	Yes	Yes	Yes	No
Lasnon et al (49), 2016	Yes	No	Yes	Yes	Yes	No
Leijenaar et al (39), 2013	Yes	Yes	Yes	No	Yes	Yes
Lu et al (52), 2016	Yes	Yes	Yes	Yes	Yes	Yes
Mackin et al (40), 2015	Yes	Yes	Yes	Yes	No	No
Oliver et al (41), 2015	Yes	No	Yes	No	Yes	Yes
Orlhac et al (19), 2014	Yes	No	Yes	Yes	Yes	No
Parmar et al (42), 2014	Yes	Yes	Yes	Yes	Yes	Yes
Tixier et al (55), 2012	Yes	No	Yes	Yes	Yes	No
Van Velden et al (43), 2016	Yes	No	Yes	Yes	Yes	Yes
Van Timmeren et al (57), 2016	Yes	No	Yes	Yes	No	Yes
Yan et al (44), 2015	Yes	No	Yes	Yes	Yes	Yes
Yip et al (45), 2014	Yes	Yes	Yes	Yes	No	No
Zhao et al (46), 2016	Yes	No	Yes	No	Yes	Yes

Each cell indicates whether it was possible to evince sufficient information from the text to re-create the experiment (yes) or not (no).

Figure 3.4: Quality of reporting analysis for human studies. Each cell indicates whether it was possible to evince sufficient information from the text to re-create the experiment (yes) or not (no).

Table 4 Quality of reporting analysis for phantom studies

Reference	Phantom details stated	Software details stated	Image acquisition settings provided	Preprocessing details provided	Segmentation details provided	Metric thresholds stated
Buch et al (61), 2017	Yes	Yes	Yes	Yes	Yes	No
Forgacs et al (62), 2016	Yes	No	Yes	No	No	Yes
Kim et al (63), 2015	Yes	Yes	Yes	Yes	Yes	No
Lo et al (64), 2016	Yes	No	Yes	Yes	Yes	No
Shafiq-ul-Hassan et al (65), 2017	Yes	No	Yes	Yes	Yes	No
Zhao et al (66), 2014	Yes	No	Yes	Yes	Yes	No

Each cell indicates whether it was possible to evince sufficient information from the text to re-create the experiment (yes) or not (no).

Figure 3.5: Quality of reporting analysis for phantom studies. Each cell indicates whether it was possible to evince sufficient information from the text to re-create the experiment (yes) or not (no).

[9][38][46][60][72]. The statistical analysis was unclear in the text in the study by Kim et al [38].

3.3.5 Radiomic features according to cancer diagnosis

Lung cancers

The imaging modalities among lung cancer studies were CT (14 studies), PET (11 studies), and CBCT (1 study). Desseroit et al [17] investigated PET and CT at the same time. All of the studies investigating PET acquired these images using a combined PET-CT scanner. Twenty-one studies used NSCLC data sets, and 4 studies used a combination of different lung cancers [28][39][43][65]. Three studies evaluated the reproducibility of radiomic features with respect to multiple manual segmentations in the same patient (interobserver sensitivity) by use of PET [1][44][66]. Each showed that interobserver differences in delineations affected feature reproducibility to some degree. Interobserver differences were amplified in textural features. Leijenaar et al [44] found that features with high test-retest repeatability were also less affected by interobserver differences. Parmar et al [56] studied interobserver variability with respect to manual versus

semiautomated segmentation (3D Slicer) and concluded that semiautomated methods improved feature reproducibility in PET. Orlhac et al [55] studied textural feature reproducibility with respect to 2 different semiautomated segmentation algorithms: an adaptive method [52] and a more conventional thresholding method based on 40% of the maximum standardized uptake value. Homogeneity, contrast, dissimilarity, and coarseness were found to be the most reproducible features. The only multi-centre CT study found that among shape metrics, 3 variants—local shape descriptors, global shape descriptors, and textural features—had the highest variation with respect to segmentations, whereas size measures had the least variation resulting from segmentation [37]. First-order features were highly reproducible across participating centres, and there were strong internal correlations within each class of features. Four studies examined the impact of different PET image reconstruction algorithms or image processing filters [55][66][68][43]. Grid size had a larger impact on feature reproducibility than did simple Gaussian filters applied inside the image reconstruction algorithms; the latter affected reproducibility in only shape and textural features. Gray-level resampling sensitively affects textural feature reproducibility, whereas first-order features are less affected. Differences in reconstruction algorithms strongly affect feature reproducibility, with the exception of first-order entropy. Entropy was reproducible for both image pre-processing and several reconstruction algorithms. Three studies compared features using free-breathing PET versus respiratory-gated PET to evaluate the impact of motion on reproducibility but showed conflicting results. In the study by Oliver et al [54], spatial blurring effects due to respiratory motion and intrinsic noise during acquisition were major factors leading to irreproducibility. Similar results were seen when textural features on 3D versus 4-dimensional (4D) PET were compared [69]. The latter study concluded that 4D imaging reduced motion artifacts, producing less blurred images and potentially more reproducible textural features. However, in the study by Grootjans et al [28], the differences between features derived from 3D and 4D imaging were not statistically significant. Zhao et al [71] evaluated the combined effect of 3 different CT slice thick-

nesses and 2 different CT reconstruction algorithms. Many features that had been repeatable under test-retest conditions became irreproducible with respect to altered slice thickness and image reconstruction settings, with first-order features and shape metrics being less sensitive than textural features. Fave et al [23] tested the effect of different image pre-processing filters, such as bit-depth resampling and smoothing filters, on CT radiomic features. Correlation to tumour volume and use of pre-processing filters increased the chance of features being significant on univariate analysis against outcome, but whether these retain their predictive value in an independent validation set remains unclear. Three studies focused exclusively on repeatability using the same test-retest images [34][17][65] (31, 37, 57). They consistently found shape metrics and first-order features to be highly repeatable, but there was no consensus on repeatable textural features. Two studies investigated feature reproducibility across different scanning equipment using a specially constructed texture phantom: Fave et al [22] using CBCT and Mackin et al [48] using CT. Both investigations went on to validate their phantom results using clinical images. Feature reproducibility using CBCT was adversely affected by motion and scattered radiation, whereas inter-scanner CT differences were found to be of the same magnitude as interpatient feature differences.

Head and neck cancers

Three studies were concerned with head and neck cancers; all primary tumours were located in the oropharynx. Modalities investigated were CT [8], CT and CBCT [5], and PET [47]. Each of these investigated some aspect of image resampling filters or other pre-processing regarding feature reproducibility. Bagher-Ebadian et al [5] applied different smoothing, sharpening, and noise filters to CBCT and CT images and found that feature reproducibility in both modalities was most strongly affected by high-pass filters and logarithmic filters. Smoothing filters and Gaussian noise kernels had a similar but smaller impact on reproducibility. The authors found

no major differences in reproducibility between CT and CBCT. The effect of gray-level discretization was discussed in 2 studies: those by Bogowicz et al [8] using CT and Lu et al [47] using PET. The first found that bin size strongly affected reproducibility on perfusion CT, but the second found a qualitatively similar though less severe impact on reproducibility in PET. Lu et al [47] also compared different PET segmentation methods (manual, semiautomated, and fully automated) and found that more than half of the radiomic features were reproducible. In addition, the sensitivity of features due to segmentation differences was less than that due to voxel dimension resampling.

Oesophageal cancers

Three studies were concerned with oesophageal cancer. All 3 investigated PET modalities. Tixier et al [64] investigated the effect of different PET reconstruction algorithms in oesophageal cancer. The most reproducible tumour heterogeneity markers were entropy, homogeneity, and dissimilarity (for local characterization) and variability in the size and intensity of homogeneous tumour regions (for regional characterization). The other 2 studies investigated how different thresholding-based semiautomated segmentation algorithms affect feature reproducibility. The impact was less marked with first-order features than with textural features. Entropy was the most reproducible first-order feature, and homogeneity was the most reproducible textural feature. Segmentation affected reproducibility more than either smoothing or filtering.

Rectal cancers

Three studies were concerned with rectal cancers. Modalities investigated included CT (2 studies) and PET (1 study). Orlhac et al [55] studied textural feature reproducibility using PET with respect

to different segmentation algorithms and gray-level resampling. Only a few features (homogeneity, contrast, dissimilarity, and coarseness) were found to be highly reproducible with respect to segmentation and resampling. Two studies investigated feature repeatability using CT [32][65]. Shape features were again found to be the most repeatable, and higher-order textural features were the least reproducible. Normalizing the extracted values by ROI volume generally improved the overall repeatability of features.

Other cancers

Studies were limited for other cancers. Guan et al [29] investigated feature reproducibility in the apparent diffusion coefficient from MRI of cervical cancer with respect to interobserver and intra-observer variability. All entropy measures were highly reproducible independent of observer effects. Orlhac et al [55] investigated textural feature reproducibility using PET in breast cancer. Only a few features (contrast, coarseness, and high gray-level run emphasis) were reproducible with respect to the number of gray levels used for resampling.

3.3.6 Radiomic features according to imaging modality

Positron emission tomography

PET was the second most common imaging modality overall and the most common in lung cancer. First-order statistics derived from a standard uptake value (SUV) histogram, such as mean SUV and maximum SUV, were consistently among the most repeatable and reproducible. Interclass correlation coefficients of these features were consistently higher than 0.95. First-order PET features were generally robust with respect to segmentation, but textural features consistently showed greater sensitivity to segmentation differences [1][55][44][56][66][18]. The choice of image reconstruction algorithm

had a greater effect on reproducibility of shape metrics and textural features relative to first-order features [66][68][43].

Computed tomography

Studies using CT were most common in head and neck cancer (2 studies) [5][8], and CT was the second most common modality in lung cancer (14 studies). First-order and shape CT features were generally more repeatable than textural features [17][32][65]. Slice thickness resampling and different reconstruction algorithms strongly degraded feature reproducibility [23][71][5][8]. The magnitude of degradation was greater for textural features than for first-order features.

Cone beam CT

CBCT was used in 1 NSCLC study [22] and 1 oropharyngeal cancer study [5]. Radiomic feature reproducibility on CBCT appeared to be adversely affected by scattered X rays and specifics of the imaging device. Low-amplitude noise and smoothing did not appear to affect the correlation of CBCT features to planning CT.

Radiomic features according to phantom studies

All the studies that investigated reproducibility on CT agreed that voxel size resampling strongly affected feature reproducibility. Zhao et al [72] demonstrated substantial differences in reproducibility when comparing 1.25- and 5-mm slices. Volume, homogeneity, and energy (gray-level co-occurrence matrix) were more reproducible for the finer slice thickness. This study recommended using images with a slice thickness between 1 and 2.5 mm for radiomic analysis. Studies confirmed that other CT acquisition parameters, such as tube voltage or tube current, had no influence on feature reproducibility [9][24].

3.3.7 Predictive or prognostic power of reproducible or repeatable features

Of the 35 articles investigating feature reproducibility and repeatability in human studies, only 11 also addressed the prognostic or predictive value of computed features (Table E1, available online at the following link). Ten studies used NSCLC data sets, and only 1 study was based on oesophageal cancers. Five studies investigated clinical outcome and patients' overall survival, 2 investigated the role of features in stratifying patients according to poor or good prognosis (based on mean overall survival), 3 investigated the pathologic response to treatment, and 1 investigated tumour recurrence. All studies agreed that models including quantitative imaging features have better performance than models including only clinical features. The majority of the studies found textural analysis features to be predictive or prognostic. Unfortunately, there is no consensus on most predictive textural analysis features. In addition, some studies found some first-order features to be predictive, but unfortunately, it was not possible to find a consensus. We suggest that authors clearly document the procedure adopted for feature selection for their models, possibly by making use of workflow figures.

3.3.8 Methodologic issues identified in review

Accessibility of software for feature extraction and of image collections

The included studies used a wide range of software to process images and extract features. Fourteen studies specifically identified MATLAB as the framework for their feature extraction algorithms. Software in the studies by Bogowicz et al [8] and Kim et al [38] were based on in-house code written for Python and ImageJ, respectively. Twenty studies did not report any details about the software used. Only 1 of the aforementioned studies has made its source code available in a GitHub repository [47]. Among MATLAB users, only Aerts et al [1]

and Balagurunathan et al [6] have made their image sets publicly accessible online. Kalpathy-Cramer et al [37] and Oliver et al [54] also used in-house created software, but neither provided additional details or made the software publicly accessible. Kalpathy-Cramer et al [37] provided open access to images and structure sets (for 40 patients and 1 phantom) via The Cancer Imaging Archive (TCIA). Four studies used the IBEX open-source radiomics package [70], developed by MD Anderson Cancer Center, but their image collections are not publicly accessible online [23][21][23][48]. Among phantom studies, only 2 reported the software used to extract radiomic features. Buch et al [9] used MATLAB as the framework application for their feature extraction algorithms, and Kim et al [38] developed a dedicated plug-in for ImageJ. However, none of the phantom studies had publicly released their image sets. It would be difficult to compare, for consistency and standardization, the radiomic features extracted by different software implementations if values for a canonical set of features were not openly accessible. Furthermore, feature stability and predictive performance of radiomic features cannot easily be externally validated unless other researchers have access to either the extraction software or the medical images (or both).

Heterogeneity in statistical metric and cut-off values

The human subject studies in this synthesis were highly heterogeneous regarding statistical metrics for repeatability and/or reproducibility. The metrics encountered were the intraclass correlation coefficient (ICC) in 14 studies, concordance correlation coefficient (CCC) in 7 studies, Spearman rank correlation in 5 studies, and various descriptive measures of difference among the remaining 9 studies. Some studies reported more than a single metric. However, the specific cut-off values used to segregate stable from unstable features were not always stated. When stated, the threshold values were highly study dependent. This led to differences in the individual features that were deemed repeatable or reproducible, and there

was no universal consensus. The ICC metric [67] is appropriate where one expects strong correlation within a given class but weak correlation between classes, and it was most commonly reported in reproducibility experiments. Five of these ICC-based studies failed to report the threshold value used to consider a feature as reproducible. The others defined a feature as highly reproducible if ICC was >0.9 [15][66][8]; if ICC was >0.81 [29]; if ICC was >0.8 [34][44]; and finally, if ICC was ≥ 0.8 [56][47]. The CCC metric [45] assumed each observation was independent and was commonly reported in both repeatability and reproducibility studies. In the studies by Kalpathy-Cramer et al [37] and Zhao et al [71], the cut-off was set at CCC ≥ 0.75. Other reported cut-offs were CCC ≥ 0.8 [22][32], CCC >0.85 [65] and CCC >0.9 [6][25]. Spearman rank correlation [51] measures the ordinal correlation between features in 2 experiments and was reported in 5 studies. Human studies also tended to stratify features into ordinal groups (e.g. poor, medium, or high reproducibility or repeatability) according to the statistical metric. No study made available its calculated metrics at the level of the individual feature. We did not attempt a meta-analysis of summary statistics in this review. The phantom studies also used diverging statistical metrics. In the study by Kim et al [38], the metric was ambiguous. Two studies used the coefficient of variation 62, 65; one study used the mean standard deviation [46]; one study used a multilinear regression method [72]; and Buch et al [9] used a t test. Only Forgacs et al [24] reported the cut-off used to select reproducible features. For phantom studies, lack of consensus also excluded quantitative meta-analysis of the results.

Reporting of digital image manipulations before feature extraction

Radiomic feature values appeared to be sensitive to pre-processing filters applied to the original image. There was some consensus that first-order features were not as sensitive to image pre-processing as were textural features. Because the latter class of features is

highly sensitive to perturbations in local intensity distribution and short-range correlations, the use of prefilters might have enhanced certain details and eliminated information from others. However, the aforementioned pre-processing steps (if any) were embedded within each software implementation and were seldom explicitly documented. Discrepancies between studies using the same image modalities and the same software may be partly due to undocumented differences in pre-processing, but it would be impossible to rule out differences resulting from image reconstruction algorithm or image acquisition settings.

3.3.9 Qualitative synthesis

Lung studies generally agreed that ROI segmentation affects the reproducibility of radiomic features for both PET and CT modalities, especially among the shape metrics and textural features. Image reconstruction algorithms revealed a difference between filtration and voxel sampling. The former had more impact on reproducibility of textural features, but the latter reduced the reproducibility of all features. Respiratory motion appears to have had a significant adverse impact on reproducibility of PET and CBCT features. Feature values were correlated to ROI volume in some software implementations, which may lead to a confounding association with certain outcomes. In general, the head and neck cancer studies agreed that either modifying voxel size or applying intensity discretization influenced feature reproducibility for both CT and PET, but PET seemed overall less sensitive with respect to differences in segmentation. This review did not find any studies addressing differences in image acquisition settings, reconstruction algorithms, or scanners for head and neck cancers. In PET human subject studies, first-order entropy was one of the most stable features across multiple settings. There were mixed findings for reproducibility of skewness and kurtosis. Shape metrics were also reproducible using PET but were less reproducible using CT, likely because of the manual delineation

sensitivity of the latter. Coarseness and contrast appeared to be least stable among the textural features. There was no overall pattern for stable textural features, nor were any significant differences noted due to different isotopes [47] (i.e. 18F and 11C). First-order entropy emerged as among the most consistently reproducible features on CT for both oropharyngeal and lung cancers. Single-institution studies concluded that CT shape metrics were highly reproducible, but the only multi-institutional study concluded that shape descriptors (i.e. flatness and sphericity) and textural features were the least reproducible [37]. Kalpathy-Cramer et al [37] also showed that first-order CT features were highly reproducible across participating centres, even for skewness and kurtosis. There was consensus that certain texture features, such as coarseness and contrast, were poorly reproducible. No emergent pattern regarding reproducible PET texture features was found. No overall trend emerged regarding repeatable and reproducible CBCT textural features. Among first-order features, entropy was one of the most stable features, whereas kurtosis was the least stable. In general, all phantom studies on CT consistently reported that first-order features such as histogram mean and entropy were the most reproducible features. Similar results for entropy were observed on PET radiomic analysis [24] when examining reproducibility with respect to different acquisition time intervals and reconstruction settings. The aforementioned qualitative synthesis across all included studies has been summarized in Figure 3.6, indicating which process steps are most likely, probable, or least likely to affect the repeatability and reproducibility of radiomic features.

	FIRST ORDER	SHAPE METRICS	TEXTURE ANALYSIS	COMMENTS
ROI SEGMENTATION				
MANUAL DELINEATION	♦	♦♦♦	♦♦♦	Mainly PET studies and one multi-center CT study.Shape metrics from PET may be less subject to inter-observer differences. Semi-automated methods generally improve reproducibility.
SEMI-AUTO / AUTO	♦	♦♦	♦♦	
IMAGE RECONSTRUCTION				
RECONSTRUCTION FILTER	♦	♦♦	♦♦♦	Consistent in a few CT and PET studies of NSCLC.
VOXEL SAMPLING	♦♦	♦♦	♦♦♦	
IMAGE ACQUISITION SETTINGS				
RESPIRATORY MOTION	♦♦	♦♦	♦♦	Consistent over single-institution PET and CBCT studies of NSCLC.
SCATTERED RADIATION	♦♦	?	♦♦	In one CBCT study of NSCLC, but did not evaluate shape metrics.
CT SCANNER	♦♦	♦♦	♦♦	In one multi-institutional CT study in NSCLC , effects were similar in magnitude to inter-patient differences.
DIGITAL IMAGE PRE-PROCESSING				
NOISE AND SMOOTHING	♦♦	?	♦♦	Single-center CBCT and planning CT study in H&N; smoothing and noise have less effect than high-pass and logarithmic filters.
INTENSITY DISCRETIZATION	♦♦	♦♦	♦♦	Consistent in H&N studies of perfusion CT and PET, bin size may have less impact in PET.
CONSENSUS ABOUT MOST STABLE OR LEAST STABLE RADIOMIC FEATURES	Entropy was consistently among the most repeatable/reproducible first-order features. There were inconsistent findings for skewness and kurtosis. Certain shape metrics may be reproducible in PET, and slightly less reproducible in CT, though it is unclear which individual features prove to be stable. No emergent pattern or consensus for highly reproducible textural features. Coarseness and contrast were among the least reproducible.			

Figure 3.6: Qualitative synthesis of radiomic feature classes, indicating processing steps that are either highly likely (3 diamonds), probable (2 diamonds), or less likely (1 diamond) to exert an adverse effect on repeatability and reproducibility for each class of radiomic features. Feature classes for which no information was available are marked as unknown (question mark). Abbreviations: CBCT = cone beam computed tomography; CT = computed tomography; HN = head and neck cancer; NSCLC = non-small cell lung cancer; PET = positron emission tomography; ROI = region of interest.

3.4 DISCUSSION

The total number of published predictive modelling studies using image-based quantitative features has been rapidly rising, but global consensus about features that are repeatable and reproducible has not yet emerged. Lack of unified synthesis could potentially undermine future discussions about clinical applicability and prospective multi-institutional external-validation trials. The primary objective of this review was to identify radiomic features that were shown to be repeatable and reproducible through an electronic search of peer-reviewed journal publications. We also evaluated the methodologic details provided in each of the studies. Summaries of our findings have been presented in tables. We located a number of general reviews focusing on the process and challenges of radiomic studies [42][61]. The previous work has drawn particular attention to the lack of standardization [73] and need for calibration of imaging settings. At the time of this writing, there has been no systematic review focusing on repeatability and/or reproducibility studies of radiomic features. General recommendations for radiomic research To homogenize radiomic reproducibility and repeatability studies, we suggest that the community perform benchmarking studies on common, shared, and publicly available data sets. In particular, this concept has already been proposed within the Image Biomarker Standardization Initiative, where different institutions computed features on a common data set. However, to expand this effort, we have been working on (1) providing users with a common repository with shared data sets for feature benchmarking; (2) providing a computational infrastructure, which directly connects to the repository; and (3) suggesting a standardized way of reporting and collecting computational results. With regard to point 1, we believe that common data sets should include both phantom and human studies. Because features could be dependent on several acquisition parameters (e.g. slice thicknesses and different scanning protocols and/or scanner manufacturers), our recommendations are as follows:

- Include benchmarking data sets collected by different institutions to guarantee the maximum heterogeneity in terms of the aforementioned parameters.
- Include different data sets for most common modalities and diseases because, as we have shown in the review, feature reproducibility results can be different when considering different diseases and/or different modalities.

With regard to point 2, we are currently working on developing an infrastructure, based on workflow programming language, that allows users to connect to the mentioned repository and run their feature extraction software. This infrastructure can be expanded by introducing in the workflow a *benchmarking module*, where users can easily *test* their software on common data sets, directly choosing them according to the modality and/or disease population of interest. In such a module, computational results can then automatically be uploaded, and a "sanity" report of feature reproducibility, compared with values already obtained by other institutions (benchmarking), is returned to the user. We believe that such an infrastructure not only will stimulate users to perform benchmarking calculations but also will help them in terms of debugging their software in case of possible errors. Point 3 is strictly related to point 2. In fact, benchmarking intrinsically brings the concept of comparisons. For this reason, a standardized way of reporting should be preferred. As already pointed out in this article, users should report not only the obtained features' raw values but also the configurations and/or parameters used to perform computations, together with details of the software used for computations. To facilitate this process, we are working on providing users with standard template tables that need to be filled in by the users and that include all the information mentioned earlier. In our view, this represents the first step toward homogenizing and increasing the quality of reporting. However, to facilitate feature comparison, we recommend using ontology techniques combined with Semantic Web to transform template tables into semantically linked data that can easily be queried by means of universal concepts defined by the ontology. Finally, to in-

crease the general validity of a radiomics-based model, we believe that external validation of the developed model should be performed, and only reproducible and repeatable features should be included in the model. To achieve this goal, we suggest using a distributed learning approach. In fact, in a distributed learning environment, models are "learned" and validated in different centres to increase their general validity. In addition, we recommend that authors describe in detail the procedure adopted for selecting features in the model. We suggest the following possible workflow:

- Perform a reproducibility experiment and rank features from most reproducible to least reproducible.
- Start scanning the list, picking up the most reproducible features; train the model; and validate the model externally to investigate the predictive or prognostic power of the features.

As a final point, we recognize the need to create a community of radiomics users, sharing common methodology in terms of both feature computation and methodology. In addition, we believe that this community should be guided by findable, accessible, interoperable, and reusable (FAIR) principles for a standardized, reliable, and reproducible use of radiomics.

3.4.1 Limitations of review

We are not able to rule out possible publication bias toward favourable results among the included studies. We did not generate a funnel plot because of the relatively low number of eligible studies and because of the specific exclusion of unpublished reports and conference proceedings. Every published study included in our review identified at least 1 radiomic feature that was repeatable or reproducible. This systematic review was limited to only 2 reviewers; though a third reviewer was available to resolve disagreements, this option was not exercised. No disagreements were found after discussion, when comparing the results of the quality of reporting and the qualitative synthe-

sis. Furthermore, our search was limited to only 1 literature repository (PubMed) after its incorporation of MEDLINE and Embase citations. We did not permit conference proceedings, non–peer-reviewed publications, or other sources of gray literature in this review, which may have limited the number of studies located. As a fundamental final point, this review would have benefited from a quantitative synthesis of the analysed articles. Unfortunately, as already mentioned, the reviewed studies applied arbitrary cut-offs during the statistical analysis of reproducible and repeatable features; in some cases, as we documented in our article, thresholds used to define a feature as "reproducible" were not reported. We have performed the analysis that was amenable to us at this time. However, we strongly support consensus toward a standardized metric to quantitatively evaluate reporting of radiomic studies that will be useful for the community. No such consensus presently exists, and our review was the first attempt to describe what has been reported in the reviewed literature. We did not specifically propose a metric to evaluate the quality of reporting because this needs to be a consensus effort by our community. This can be seen as one of our study's limitations. An anticipated update of the current review is proposed for April 2019. We hope by that date to have agreed on a common quantitative evaluation metric within the research community so that we will be able to update the review, in terms of not only up-to-date publications but also the inclusion of a quantitative analysis.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] S Alobaidli, S McQuaid, C South, V Prakash, P Evans, and A Nisbet. The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning. *The British Journal of Radiology*, 87(1042):20140369, October 2014.

[3] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *JNCI Journal of the National Cancer Institute*, 86(11):829–835, June 1994.

[4] Ulas Bagci, Jianhua Yao, Kirsten Miller-Jaster, Xinjian Chen, and Daniel J. Mollura. Predicting Future Morphological Changes of Lesions from Radiotracer Uptake in 18F-FDG-PET Images. *PLoS ONE*, 8(2):e57105, February 2013.

[5] Hassan Bagher-Ebadian, Farzan Siddiqui, Chang Liu, Benjamin Movsas, and Indrin J. Chetty. On the impact of smoothing and noise on robustness of CT and CBCT radiomics features for patients with head and neck cancers. *Medical Physics*, 44(5):1755–1770, May 2017.

[6] Yoganand Balagurunathan, Yuhua Gu, Hua Wang, Virendra Kumar, Olya Grove, Sam Hawkins, Jongphil Kim, Dmitry B. Goldgof, Lawrence O. Hall, Robert A. Gatenby, and Robert J. Gillies. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Translational Oncology*, 7(1):72–87, February 2014.

[7] Knut Baumann. Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, 22(6):395–406, June 2003.

[8] M Bogowicz, O Riesterer, R A Bundschuh, P Veit-Haibach, M Hüllner, G Studer, S Stieb, S Glatz, M Pruschy, M Guckenberger, and S Tanadini-Lang. Stability of radiomic features in CT perfusion maps. *Physics in Medicine and Biology*, 61(24):8736–8749, December 2016.

[9] K Buch, B Li, MM Qureshi, H Kuno, SW Anderson, and O Sakai. Quantitative assessment of variation in ct parameters on texture features: pilot study using a nonanatomic phantom. *American Journal of Neuroradiology*, 38(5):981–985, 2017.

[10] Nail Bulakbasi, Inanc Guvenc, Onder Onguru, Ersin Erdogan, Cem Tayfun, and Taner Ucoz. The Added Value of the Apparent Diffusion Coefficient Calculation to Magnetic Resonance Imaging in the Differentiation and Grading of Malignant Brain Tumors:. *Journal of Computer Assisted Tomography*, 28(6):735–746, November 2004.

[11] Anastasia Chalkidou, Michael J. O'Doherty, and Paul K. Marsden. False Discovery Rates in PET and CT Studies with Texture

Features: A Systematic Review. *PLOS ONE*, 10(5):e0124165, May 2015.

[12] Nai-Ming Cheng, Yu-Hua Dean Fang, Din-Li Tsan, Ching-Han Hsu, and Tzu-Chen Yen. Respiration-averaged ct for attenuation correction of pet images–impact on pet texture features in non-small cell lung cancer patients. *PloS one*, 11(3):e0150509, 2016.

[13] Sugama Chicklore, Vicky Goh, Musib Siddique, Arunabha Roy, Paul K. Marsden, and Gary J. R. Cook. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(1):133–140, January 2013.

[14] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*, 13(1):1, 2015.

[15] Thibaud P Coroller, Vishesh Agrawal, Vivek Narayan, Ying Hou, Patrick Grossmann, Stephanie W Lee, Raymond H Mak, and Hugo JWL Aerts. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiotherapy and oncology*, 119(3):480–486, 2016.

[16] Marie-Charlotte Desseroit, Florent Tixier, Wolfgang A Weber, Barry A Siegel, Catherine Cheze Le Rest, Dimitris Visvikis, and Mathieu Hatt. 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*, 58(3):406–411, 2017.

[17] Marie-Charlotte Desseroit, Dimitris Visvikis, Florent Tixier, Mohamed Majdoub, Rémy Perdrisot, Rémy Guillevin, Catherine Cheze Le Rest, and Mathieu Hatt. Development of a nomogram combining clinical staging with 18 f-fdg pet/ct image fea-

tures in non-small-cell lung cancer stage i–iii. *European journal of nuclear medicine and molecular imaging*, 43(8):1477–1485, 2016.

[18] Georgia Doumou, Musib Siddique, Charalampos Tsoumpas, Vicky Goh, and Gary J. Cook. The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer. *European Radiology*, 25(9):2805–2812, September 2015.

[19] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, January 2009.

[20] Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zheng. Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients. *IEEE Transactions on Biomedical Engineering*, 63(5):1034–1043, May 2016.

[21] Xenia Fave, Molly Cook, Amy Frederick, Lifei Zhang, Jinzhong Yang, David Fried, Francesco Stingo, and Laurence Court. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Computerized Medical Imaging and Graphics*, 44:54–61, 2015.

[22] Xenia Fave, Dennis Mackin, Jinzhong Yang, Joy Zhang, David Fried, Peter Balter, David Followill, Daniel Gomez, A Kyle Jones, Francesco Stingo, et al. Can radiomics features be reproducibly measured from cbct images for patients with non-small cell lung cancer? *Medical physics*, 42(12):6784–6797, 2015.

[23] Xenia Fave, Lifei Zhang, Jinzhong Yang, Dennis Mackin, Peter Balter, Daniel Gomez, David Followill, A Kyle Jones, Francesco Stingo, et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-

small cell lung cancer. *Translational Cancer Research*, 5(4):349–363, 2016.

[24] Attila Forgacs, Hermann Pall Jonsson, Magnus Dahlbom, Freddie Daver, Matthew D. DiFranco, Gabor Opposits, Aron K. Krizsan, Ildiko Garai, Johannes Czernin, Jozsef Varga, et al. A study on the basic criteria for selecting heterogeneity parameters of f18-fdg pet images. *PloS one*, 11(10):e0164113, 2016.

[25] David V Fried, Susan L Tucker, Shouhao Zhou, Zhongxing Liao, Osama Mawlawi, Geoffrey Ibbott, and Laurence E Court. Prognostic value and reproducibility of pretreatment ct texture features in stage iii non-small cell lung cancer. *International Journal of Radiation Oncology* Biology* Physics*, 90(4):834–842, 2014.

[26] Paulina E. Galavis, Christian Hollensen, Ngoneh Jallow, Bhudatt Paliwal, and Robert Jeraj. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncologica*, 49(7):1012–1016, October 2010.

[27] Geert-Jan Geersing, Walter Bouwmeester, Peter Zuithoff, Rene Spijker, Mariska Leeflang, and Karel Moons. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. *PLoS ONE*, 7(2):e32844, February 2012.

[28] Willem Grootjans, Florent Tixier, Charlotte S van der Vos, Dennis Vriens, Catherine C Le Rest, Johan Bussink, Wim JG Oyen, Lioe-Fee de Geus-Oei, Dimitris Visvikis, and Eric P Visser. The impact of optimal respiratory gating and image noise on evaluation of intratumor heterogeneity on 18f-fdg pet imaging of lung cancer. *Journal of nuclear medicine*, 57(11):1692–1698, 2016.

[29] Y. Guan, W. Li, Z. Jiang, B. Zhang, Y. Chen, X. Huang, J. Zhang, S. Liu, J. He, Z. Zhou, and Y. Ge. Value of whole-lesion apparent diffusion coefficient (ADC) first-order statistics and texture

features in clinical staging of cervical cancers. *Clinical Radiology*, 72(11):951–958, November 2017.

[30] M. Hatt, M. Majdoub, M. Vallieres, F. Tixier, C. C. Le Rest, D. Groheux, E. Hindie, A. Martineau, O. Pradier, R. Hustinx, R. Perdrisot, R. Guillevin, I. El Naqa, and D. Visvikis. 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*, 56(1):38–44, January 2015.

[31] R Brian Haynes, K Ann McKibbon, Nancy L Wilczynski, Stephen D Walter, and Stephen R Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*, 330(7501):1179, May 2005.

[32] Panpan Hu, Jiazhou Wang, Haoyu Zhong, Zhen Zhou, Lijun Shen, Weigang Hu, and Zhen Zhang. Reproducibility with repeat CT in radiomics study for rectal cancer. *Oncotarget*, 7(44):71440–71446, November 2016.

[33] Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, August 2016.

[34] Elizabeth Huynh, Thibaud P Coroller, Vivek Narayan, Vishesh Agrawal, John Romano, Idalid Franco, Chintan Parmar, Ying Hou, Raymond H Mak, and Hugo JWL Aerts. Associations of radiomic data extracted from static and respiratory-gated ct scans with disease recurrence in lung cancer patients treated with sbrt. *PloS one*, 12(1):e0169172, 2017.

[35] B. J. Ingui and M. A. M. Rogers. Searching for Clinical Prediction Rules in MEDLINE. *Journal of the American Medical Informatics Association*, 8(4):391–397, July 2001.

[36] Jayashree Kalpathy-Cramer, Artem Mamomov, Binsheng Zhao, Lin Lu, Dmitry Cherezov, Sandy Napel, Sebastian Echegaray, Daniel Rubin, Michael McNitt-Gray, Pechin Lo, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography*, 2(4):430–437, December 2016.

[37] Jayashree Kalpathy-Cramer, Artem Mamomov, Binsheng Zhao, Lin Lu, Dmitry Cherezov, Sandy Napel, Sebastian Echegaray, Daniel Rubin, Michael McNitt-Gray, Pechin Lo, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography*, 2(4):430, 2016.

[38] Hyun Gi Kim, Yong Eun Chung, Young Han Lee, Jin Young Choi, Mi Suk Park, Myeong Jin Kim, and Ki Whang Kim. Quantitative analysis of the effect of iterative reconstruction using a phantom: determining the appropriate blending percentage. *Yonsei Medical Journal*, 56(1):253–261, January 2015.

[39] Hyun Jung Koo, Yu Sub Sung, Woo Hyun Shim, Hai Xu, Chang-Min Choi, Hyeong Ryul Kim, Jung Bok Lee, and Mi Young Kim. Quantitative computed tomography features for predicting tumor recurrence in patients with surgically resected adenocarcinoma of the lung. *PLoS One*, 12(1):e0167955, 2017.

[40] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, November 2012.

[41] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, AndrÃ©

Dekker, and Hugo J.W.L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, March 2012.

[42] Ruben T H M Larue, Gilles Defraene, Dirk De Ruysscher, Philippe Lambin, and Wouter van Elmpt. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *The British Journal of Radiology*, 90(1070):20160665, February 2017.

[43] Charline Lasnon, Mohamed Majdoub, Brice Lavigne, Pascal Do, Jeannick Madelaine, Dimitris Visvikis, Mathieu Hatt, and Nicolas Aide. 18 f-fdg pet/ct heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer. *European journal of nuclear medicine and molecular imaging*, 43(13):2324–2335, 2016.

[44] Ralph TH Leijenaar, Sara Carvalho, Emmanuel Rios Velazquez, Wouter JC Van Elmpt, Chintan Parmar, Otto S Hoekstra, Corneline J Hoekstra, Ronald Boellaard, André LAJ Dekker, Robert J Gillies, et al. Stability of fdg-pet radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica*, 52(7):1391–1397, 2013.

[45] L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989.

[46] P Lo, S Young, HJ Kim, MS Brown, and MF McNitt-Gray. Variability in ct lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. *Medical physics*, 43(8Part1):4854–4865, 2016.

[47] Lijun Lu, Wenbing Lv, Jun Jiang, Jianhua Ma, Qianjin Feng, Arman Rahmim, and Wufan Chen. Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. *Molecular Imaging and Biology*, 18(6):935–945, December 2016.

[48] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, and Laurence Court. Measuring Computed Tomography Scanner Variability of Radiomics Features:. *Investigative Radiology*, 50(11):757–765, November 2015.

[49] Kenneth A. Miles, Balaji Ganeshan, and Michael P. Hayball. CT texture analysis using the filtration-histogram method: what do the measurements mean? *Cancer Imaging*, 13(3):400–406, 2013.

[50] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7):e1000097, July 2009.

[51] Leann Myers and Maria J. Sirois. Spearman Correlation Coefficients, Differences between. In Samuel Kotz, Campbell B. Read, N. Balakrishnan, Brani Vidakovic, and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, page ess5050.pub2. John Wiley & Sons, Inc., Hoboken, NJ, USA, August 2006.

[52] Ursula Nestle, Stephanie Kremp, Andrea Schaefer-Schuler, Christiane Sebastian-Welsch, Dirk Hellwig, Christian Rübe, and Carl-Martin Kirsch. Comparison of different methods for delineation of 18f-fdg pet–positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. *Journal of Nuclear Medicine*, 46(8):1342–1348, 2005.

[53] James P. B. O'Connor, Eric O. Aboagye, Judith E. Adams, Hugo J. W. L. Aerts, Sally F. Barrington, Ambros J. Beer, Ronald Boellaard, Sarah E. Bohndiek, Michael Brady, Gina Brown, David L. Buckley, Thomas L. Chenevert, Laurence P. Clarke, Sandra Collette, Gary J. Cook, Nandita M. deSouza, John C. Dickson, Caroline Dive, Jeffrey L. Evelhoch, Corinne Faivre-Finn, Ferdia A. Gallagher, Fiona J. Gilbert, Robert J. Gillies, Vicky Goh, John R. Griffiths, Ashley M. Groves, Steve Halligan, Adrian L. Harris,

David J. Hawkes, Otto S. Hoekstra, Erich P. Huang, Brian F. Hutton, Edward F. Jackson, Gordon C. Jayson, Andrew Jones, Dow-Mu Koh, Denis Lacombe, Philippe Lambin, Nathalie Lassau, Martin O. Leach, Ting-Yim Lee, Edward L. Leen, Jason S. Lewis, Yan Liu, Mark F. Lythgoe, Prakash Manoharan, Ross J. Maxwell, Kenneth A. Miles, Bruno Morgan, Steve Morris, Tony Ng, Anwar R. Padhani, Geoff J. M. Parker, Mike Partridge, Arvind P. Pathak, Andrew C. Peet, Shonit Punwani, Andrew R. Reynolds, Simon P. Robinson, Lalitha K. Shankar, Ricky A. Sharma, Dmitry Soloviev, Sigrid Stroobants, Daniel C. Sullivan, Stuart A. Taylor, Paul S. Tofts, Gillian M. Tozer, Marcel van Herk, Simon Walker-Samuel, James Wason, Kaye J. Williams, Paul Workman, Thomas E. Yankeelov, Kevin M. Brindle, Lisa M. McShane, Alan Jackson, and John C. Waterton. Imaging biomarker roadmap for cancer studies. *Nature Reviews Clinical Oncology*, 14(3):169–186, March 2017.

[54] Jasmine A Oliver, Mikalai Budzevich, Geoffrey G Zhang, Thomas J Dilling, Kujtim Latifi, and Eduardo G Moros. Variability of image features computed from conventional and respiratory-gated pet/ct images of lung cancer. *Translational oncology*, 8(6):524–534, 2015.

[55] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat. Tumor Texture Analysis in 18F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. *Journal of Nuclear Medicine*, 55(3):414–422, March 2014.

[56] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains, Philippe Lambin, and Hugo J. W. L. Aerts. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. *PLoS ONE*, 9(7):e102107, July 2014.

[57] Vlad Popovici, Eva Budinská, Ladislav Dušek, Michal Kozubek, and Fred Bosman. Image-based surrogate biomarkers for molecular subtypes of colorectal cancer. *Bioinformatics*, 33(13):2002–2009, 01 2017.

[58] E. Scalco, S. Moriconi, and G. Rizzo. Texture analysis to assess structural modifications induced by radiotherapy. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5219–5222, Milan, August 2015. IEEE.

[59] Elisa Scalco and Giovanna Rizzo. Texture analysis of medical images for radiotherapy applications. *The British Journal of Radiology*, 90(1070):20160642, February 2017.

[60] Muhammad Shafiq-ul Hassan, Geoffrey G. Zhang, Kujtim Latifi, Ghanim Ullah, Dylan C. Hunt, Yoganand Balagurunathan, Mahmoud Abrahem Abdalah, Matthew B. Schabath, Dmitry G. Goldgof, Dennis Mackin, Laurence Edward Court, Robert James Gillies, and Eduardo Gerardo Moros. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Medical Physics*, 44(3):1050–1062, mar 2017.

[61] M. Sollini, L. Cozzi, L. Antunovic, A. Chiti, and M. Kirienko. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Scientific Reports*, 7(1):358, December 2017.

[62] Patrick Therasse, Susan G. Arbuck, Elizabeth A. Eisenhauer, Jantien Wanders, Richard S. Kaplan, Larry Rubinstein, Jaap Verweij, Martine Van Glabbeke, Allan T. van Oosterom, Michaele C. Christian, and Steve G. Gwyther. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *JNCI: Journal of the National Cancer Institute*, 92(3):205–216, February 2000.

[63] Joseph A. Thie. Understanding the standardized uptake value, its methods, and implications for usage. *Journal of Nuclear Medicine:*

Official Publication, Society of Nuclear Medicine, 45(9):1431–1434, September 2004.

[64] Florent Tixier, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 53(5):693–700, May 2012.

[65] Janna E. van Timmeren, Ralph T. H. Leijenaar, Wouter van Elmpt, Jiazhou Wang, Zhen Zhang, André Dekker, and Philippe Lambin. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*, 2(4):361–365, December 2016.

[66] Floris HP van Velden, Gerbrand M Kramer, Virginie Frings, Ida A Nissen, Emma R Mulder, Adrianus J de Langen, Otto S Hoekstra, Egbert F Smit, and Ronald Boellaard. Repeatability of radiomic features in non-small-cell lung cancer [18 f] fdg-pet/ct studies: impact of reconstruction and delineation. *Molecular imaging and biology*, 18(5):788–795, 2016.

[67] Joseph P Weir. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1):231–240, 2005.

[68] Jianhua Yan, Jason Lim Chu-Shern, Hoi Yin Loi, Lih Kin Khor, Arvind K Sinha, Swee Tian Quek, Ivan WK Tham, and David Townsend. Impact of image reconstruction settings on texture features in 18f-fdg pet. *Journal of nuclear medicine*, 56(11):1667–1673, 2015.

[69] Stephen Yip, Keisha McCall, Michalis Aristophanous, Aileen B. Chen, Hugo J. W. L. Aerts, and Ross Berbeco. Comparison of Texture Features Derived from Static and Respiratory-Gated PET

Images in Non-Small Cell Lung Cancer. *PLoS ONE*, 9(12):e115510, December 2014.

[70] Lifei Zhang, David V Fried, Xenia J Fave, Luke A Hunter, Jinzhong Yang, and Laurence E Court. Ibex: an open infrastructure software platform to facilitate collaborative work in radiomics. *Medical physics*, 42(3):1341–1353, 2015.

[71] Binsheng Zhao, Yongqiang Tan, Wei-Yann Tsai, Jing Qi, Chuanmiao Xie, Lin Lu, and Lawrence H. Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports*, 6(1), September 2016.

[72] Binsheng Zhao, Yongqiang Tan, Wei Yann Tsai, Lawrence H Schwartz, and Lin Lu. Exploring variability in ct characterization of tumors: a preliminary phantom study. *Translational oncology*, 7(1):88–93, 2014.

[73] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

[74] C V Zwirewich, S Vedal, R R Miller, and N L Müller. Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation. *Radiology*, 179(2):469–476, May 1991.

4

Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing

Adapted from: **"Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing"**. A Traverso, M Kazmierski, Z Shi, P Kalendralis, M Welch, HD Nissen, D Jaffray, A Dekker, L Wee. Physica Medica 61, 44-51. (2019).

Abstract

Quantitative imaging features (radiomics) extracted from apparent diffusion coefficient (ADC) maps of rectal cancer patients can provide additional information to support treatment decision. Most available radiomic computational packages allow extraction of hundreds to thousands of features. However, two major factors can influence the reproducibility of radiomic features: interobserver variability, and imaging filtering applied prior to features extraction. In this exploratory study we seek to determine to what extent various commonly-used features are reproducible with regards to the mentioned factors using ADC maps from two different clinics (56 patients). Features derived from intensity distribution histograms are less sensitive to manual tumour delineation differences, noise in ADC images, pixel size resampling and intensity discretization. Shape features appear to be strongly affected by delineation quality. On the whole, textural features appear to be poorly or moderately reproducible with respect to the image preprocessing perturbations we reproduced.

4.1 INTRODUCTION

Neo-adjuvant chemoradiotherapy (NACRT) followed by total mesorectal excision (TME) is the accepted standard of care for locally advanced rectal cancer (LARC) due to conclusive evidence of superior clinical outcome [20] [24] [9] [6][19]. However, TME is a highly invasive procedure leading to bowel and bladder complications, and its added value for *good responders* is currently being debated [23] [18]. Magnetic Resonance Imaging (MRI) has flexibility for imaging anatomy, physiological parameters and biochemical function, through appropriate choice of pulse sequences. Diffusion-weighted imaging in MRI allows construction of 3D maps of apparent diffusion coefficient (ADC) of water molecules, that are promising markers of internal tumour pores and cellular interstices wherein water molecules can migrate [5]. A change in mean value of ADC has been shown to be associated with tumour response in a number of different cancers, including LARC [12], [2]. Joye et al. showed that combined PET (Positron Emission Tomography) and MRI imaging parameters were strongly associated with pCR or near-pCR [14]. The above-mentioned results led to an active search for quantitative imaging biomarkers (radiomic features) that could have prognostic/predictive power to support indication for treatment. Radiomics refers to computerized extraction of a large number of quantitative image metrics from medical images, that may reveal a deeper level of detail than is accessible to an unaided human eye, with the intent of defining tumour sub-types [1]. While radiomics has been successfully applied for clinical outcome predictions in Computed Tomography (CT) and Positron Emission Tomography (PET), its application to MRI is less advanced. Despite the chosen modality, recent publications showed the importance of evaluating radiomic features sensitivity with respect to several scenarios: different acquisition settings, inter-observer variability in tumour's delineations, choice of particular computational settings prior to features extraction (i.e. image pre-processing). The results affirm that different categories of radiomic features are, in different

forms, affected by the abovementioned scenarios. For example, textural metrics have been shown to sensitively change their values when computed using different quantization. Trying to isolate a set of features which appear to be robust to all these factors is of interest. Again, most of the available work on this topic was carried on lung and head and neck cancers using CT or PET. However, Hu et al. [13] did demonstrate that volume-normalized features were more stable than not normalized features extracted from CT; while for MRI global textural descriptors showed more temporal stability than local-regional texture parameters [10]. In this exploratory study of ADC radiomic features, we seek to determine to what extent are various commonly-used features sensitive to inter-observer disagreements in tumour delineation and the application of digital image filter prior to radiomic feature extraction, which is an adopted procedure used in most radiomic studies.

4.2 MATERIAL AND METHODS

4.2.1 Images

Ethical clearance was obtained for re-analysis of pre-radiotherapy LARC images collected between 2009 and 2012 by a Dutch radiotherapy clinic for inclusion in the THeragnostic Utilities for Neoplastic Diseases of the Rectum (THUNDER) clinical trial (NCT00969657, dataset described in [28]). A subset of 23 patients was retrospectively extracted from the THUNDER set having a pre-treatment diffusion-weighted imaging (DWI) examination at gradients of 0, 300 s/mm2 and 1100 s/mm2. ADC maps were constructed directly from the above field gradients in the Siemens (Erlangen, Germany) MR scanner console. A retrospective set of 33 LARC patients undergoing routine care were extracted with review board permission at a Danish radiotherapy clinic (population details for all the cohorts available in the Supplementary material). Images with the same DWI field gradients had been obtained using

a Philips (Eindhoven, The Netherlands) MR scanner. ADC maps were then constructed using an in-house Matlab script (MathWorks, Natick, USA). The ADC maps were calculated on a voxel-by-voxel basis using axial slices for all the cohorts. The above datasets are hereafter referred to as the "THUNDER" and "CLINIC" cohorts, respectively. Key elements of the image acquisition settings are given in Table 1 for each cohort. Other than reconstructed slice thickness and pulse sequence repetition time, the imaging parameters were nominally closely matched across the two devices. Gross tumour volume (GTV) delineation GTVs were manually delineated via a standardized consensus method between the operators. Specifically, the ADC was overlaid with a constant false-colour lookup table over the 1100s/mm2 image. Some anatomical details were visible in the latter, and using these as a guide, an outline of the hyper-intense ADC region inside and adjoining the rectum was then drawn in by hand. On the THUNDER cohort, three observers, working independently, delineated the tumour on a Mirada (Mirada Medical, Oxford, UK) workstation. In the CLINIC cohort, two observers, working independently, delineated the tumour on an Oncentra External Beam (Elekta AB, Stockholm, Sweden) workstation. One common observer (author AT) delineated on both THUNDER and CLINIC cohorts. Observers (median experience, 4 years; range 1–10) were trained by a resident radiation oncologist to identify relevant normal and abnormal anatomical structures within the ADC maps. In addition, original CT scans with annotated lesions for all the patients were available to the observers, so that they could be guided in the delineations in the ADC maps. The median DICE for both the cohorts was 0.75 (range 0.6–0.90). At the end, delineations were exported into a single DICOM RT Structure Set file per patient. Each patient's ADC map was also exported in standard DICOM format.

4.2.2 Image pre-processing

Digital pre-processing on ADC maps was applied prior to extracting features. This was intended to test the sensitivity of histogram and textural features, since shape features in PyRadiomics are entirely independent of pre-processing. For each patient, a baseline radiomic feature value was calculated on the native (unprocessed) map. Subsequently, the native ADC map was altered using one digital image pre-processing operations at a time – (i) filtering, (ii) pixel dimension resampling and (iii) intensity value discretization. All pre-processing was performed using only the functions embedded within the open-source PyRadiomics library [26], which were themselves based on SimpleITK functions [17], [31]. All the filters were applied in 3D. Mathematical details of the image pre-processing operations are provided in the Supplementary Materials.

4.2.3 Features extraction

Radiomic feature extraction was performed with PyRadiomics. The open-source PyRex extensions (Link here) were used to manage the conversions of DICOM and DICOM RT Structure files to binary masks. A total of 70 radiomic features were extracted from each subject; 18 first-order (FO) features based on the intensity histogram, 13 shape metrics (SM), 23 features based on gray-level co-occurrence matrices (GLCM) and 16 features based on gray-level size-zone matrices (GLSZM). Mathematical definitions of these features are given on the PyRadiomics feature documentation page (https://pyradiomics.readthedocs.io/en/latest/features.html).
Details used for computations, are specified in the Supplementary Materials. It is important to note that out of the 70 features, 6 features available in PyRadiomics are not defined in the IBSI (Image Biomarker Standardization Initiative), namely: Maximum 3D Diameter Column. Slice, and Row (SM); Total Energy (FO); GLCM Homogenity1/2. All the remaining features correspond to the definitions provided by IBSI.

4.2.4 Statistical analysis

Statistical analysis was performed in R Studio (v1.1.383), R (v 3.5.1) and Python (v3.6.4). A Concordance Correlation Coefficient (CCC) [16] was chosen as the reproducibility metric to evaluate the agreement of radiomic feature values in the perturbed image (with pre-processing filters, re-binning and resampling) with respect to the baseline feature values in the native ADC map. For each possible image pre-processing function, a CCC was computed for the feature value in the perturbed ADC relative to the native ADC map. The reported stability metric is the mean value of CCC over all observers in the combined THUNDER and CLINIC sets. For inter-observer dependence, we computed the Intraclass Correlation Coefficient (ICC) [3] across patients for each feature We proposed that a feature was reproducible if CCC$\geq$0.85, in keeping with one of the most commonly used thresholds reported in the literature [25]. Moderately reproducible features were arbitrarily defined as 0.65$<$CCC$<$0.85. However, features with CCC$\leq$0.65 were deemed poorly reproducible. A threshold value of 0.85 was used also for the ICC values to define reproducibility. To quantify the degree of reproducibility of features between the two datasets, the features were ordered by descending mean CCC and compared using the Spearman Rank correlation coefficient [32]. Results were considered statistically significant if p-value $<$0.05. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation at least as extreme as the one computed from these datasets. Fig. 4.1 proposes a sketch representation of the workflow used for the experiments.

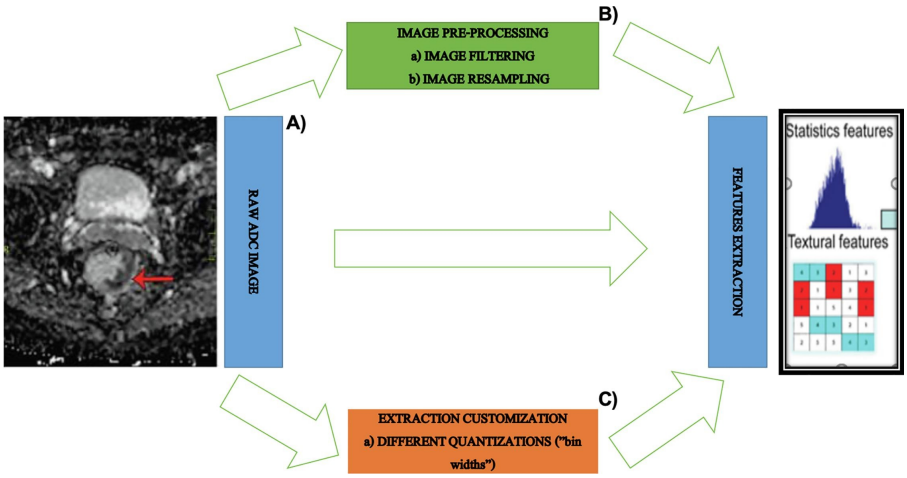


Figure 4.1: Schematic representation of the workflow used in the analysis. Radiomic features are extracted from the GTVs annotated in the ADC maps. Different configurations were considered: (a) direct extraction from raw image, using default PyRadiomics settings; (b) customization of the extraction introducing different image pre-processing steps such as filtering or image resampling; (c) customization of the extraction, without modifying the original images, but considering different quantizations when computing features. The effects of (b) and (c) are then evaluating comparing differences in feature values with respect to (a) using concordance correlation coefficients.

4.3 RESULTS

4.3.1 Inter-observer dependence

The overall sensitivity of feature types with respect to differences in manual GTV delineations were compared using the ICC metric. A box and whisker boxplot summarising the median ICC and its distribution for four feature types is given ins Fig. 4.2. Among the FO and GLCM feature types from the native ADC maps, the median ICC was consistently high in both THUNDER and CLINIC datasets. Major divergences appear for the GLSZM and SM feature types, with respect to the persons performing the delineations in the THUNDER and CLINIC datasets, respectively, such that GLSZM and SM features appeared more reproducible in the latter. There was significant spread in ICC for every feature type, so even within a related group of features certain individual features are much less sensitive to delineation differences than others.

4.3.2 Effect of resampling with interpolation

A heatmap of CCC ranges with respect to axial pixel dimension resampling is given as Fig. 4.3. At a glance, it is clear to see that the FO features are generally reproducible with respect to scale changes in pixel dimensions. The single FO feature that falls below CCC of 0.65 relative to the native ADC, after perturbation, happens to be *Energy* in this study. The GLCM features are moderately reproducible with resampling, since many features retain good or moderate reproducibility over a wide range of resampling. The majority of GLSZM features are, on the whole, poorly reproducible.

4.3.3 Effect of intensity value discretization

A heatmap of CCC ranges with respect to changes in the width of discrete intensity *bins* is given as Fig. 4.4. The overall trend here, once

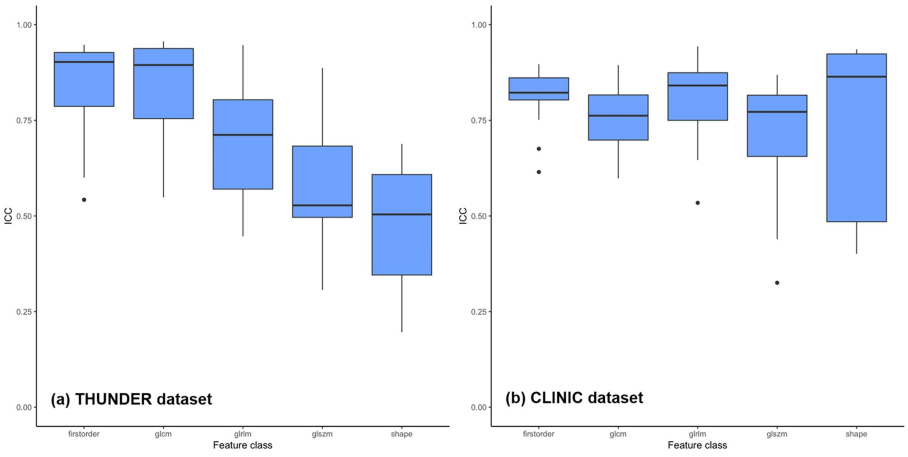


Figure 4.2: Box and whisker plots of intraclass correlation coefficient (ICC) grouped by type of feature – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM) and shape metrics – for the (a) THUNDER and (b) Danish CLINIC datasets. The solid bars represent the median. The upper and lower edges of the boxes represent the upper and lower quartiles of the ICC distribution, respectively

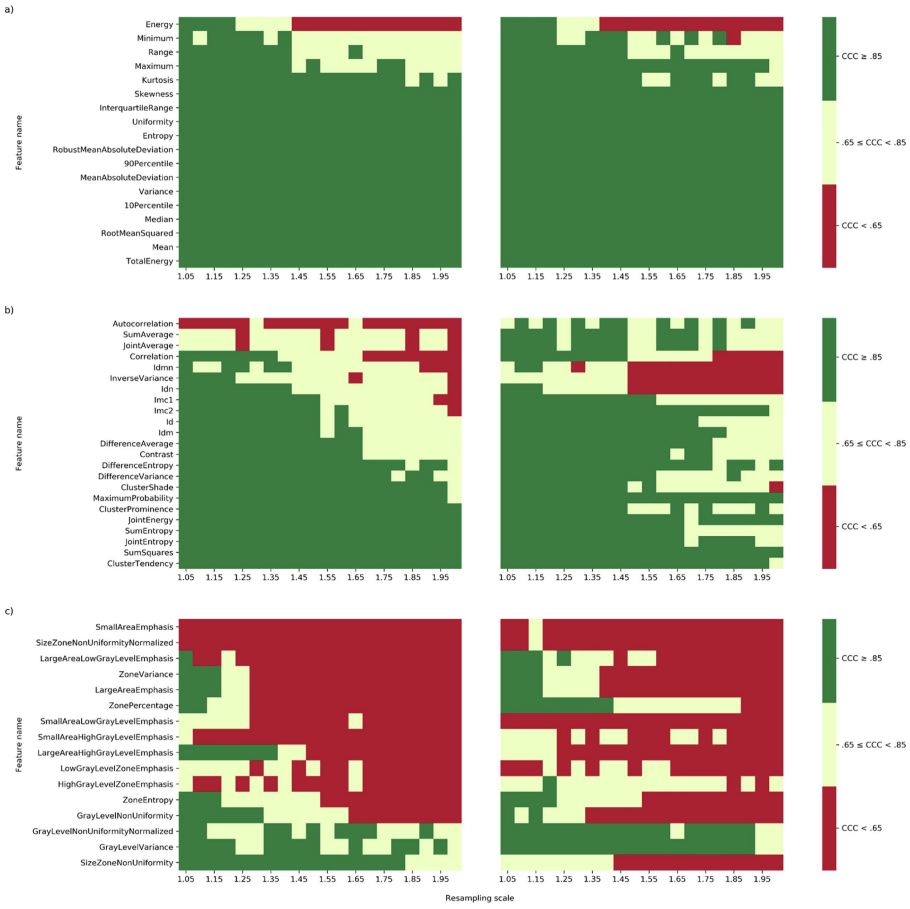


Figure 4.3: Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbation introduced is resampling of the pixel dimensions in the axial plane with interpolation, the magnitude of which is shown along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of resampling of the image pixel in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

again, is that FO features (except for Kurtosis and Skewness) are generally reproducible over a wide range of intensity discretization bin widths; however, nearly all of the GLCM and GLSZM features are poorly reproducible. One texture feature – *GLSZM Gray Level Non-Uniformity* – could be a potential feature with good to moderate reproducibility with respect to image intensity discretization. However, previous studies pointed out the strong correlation between this feature and tumour volume. In the Supplementary material a list of most reproducible features is supplied.

4.3.4 Effect of applying digital image filters

A heatmap of CCC ranges with respect to application of different types of digital image filters is given as Fig. 4.5. In regard to feature types, FO features seem to be largely reproducible after additive Gaussian noise. The overall picture is more mixed with curvature flow, Laplacian and Gaussian smoothing filters, but it generally holds that GLCM and GLSZM feature types are poorly reproducible.

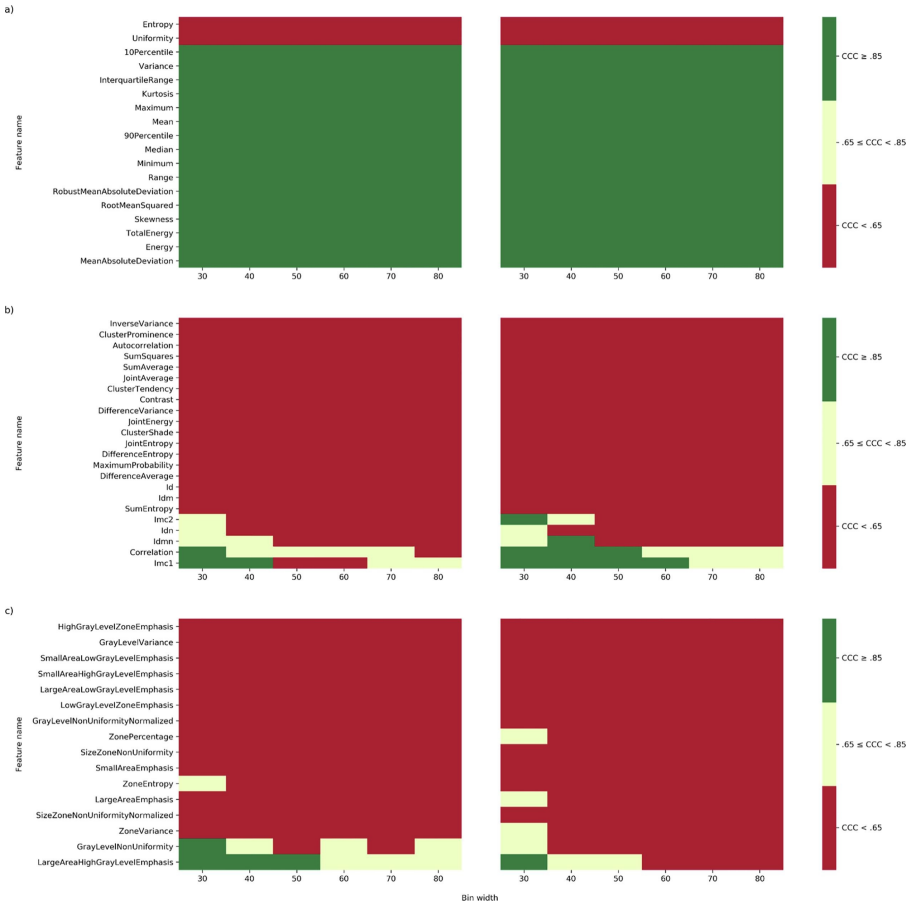


Figure 4.4: Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbation introduced is the discretization bin width for the image intensity values, the magnitude of which is shown along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of intensity discretization in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

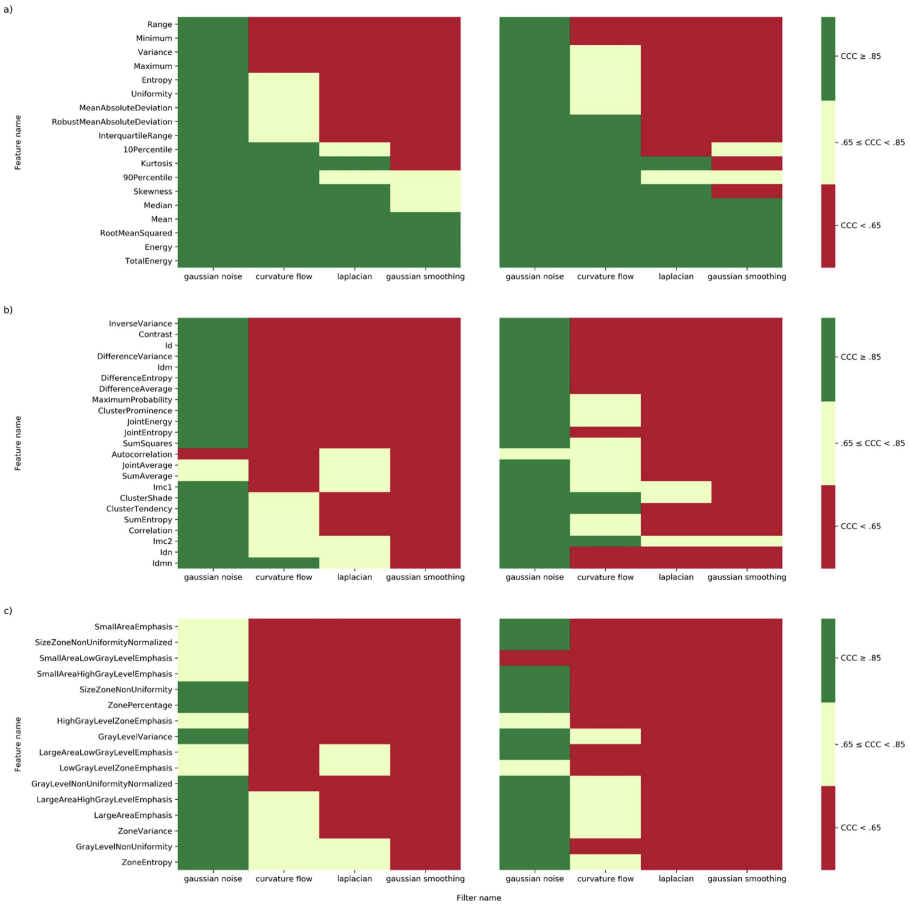


Figure 4.5: Heatmap defining ranges of concordance correlation coefficient (CCC). The perturbations introduced are four different digital image manipulation filters as described in the main text, which is denoted along the horizontal axis. Each row in the image corresponds to a particular feature within one of the feature types – first order (FO), gray-level co-occurrence matrices (GLCM), gray-level size-zone matrices (GLSZM). Shape metrics are not evaluated, because they are independent of digital filtering in the pyradiomics implementation. Results are shown on the left side for the CLINICAL cohort and on the right side for the THUNDER dataset.

4.4 DISCUSSION

We have used an ICC metric to examine the overall reproducibility of radiomic features with respect to tumour (GTV) delineation differences between groups of observers. Overall, we find that FO and GLCM feature types are less sensitive to manual delineation differences, but within each feature type a wide spread of ICCs are observed. We hypothesize that the experience level of an observer plays a key role in feature reproducibility, since we observe the median and distribution of feature ICC values appear more consistent across all feature types in the CLINIC set than compared to the THUNDER set. However, it is well known that interobserver variability increases as the quality of the image decreases. In fact, agreement between clinicians delineating on Computed Tomography (CT) is usually larger than on MRI or ADC, due to higher signal to noise ratio [30][7]. To improve agreement, studies suggest defining strong protocols for delineations, and training of the observers to strictly follow the mentioned protocols. Qualitatively, taking both datasets into account, the overall trend for increasing risk of feature group irreproducibility appears to be – FO (least risk), GLCM, GLSZM and SM (greatest risk). This hypothesis finds some support in recent literature, where a recent study by van Heeswijk et al. [27] was able to identify a FO feature that was reproducible when a fast approximate delineation was used in place of a time-consuming precise delineation of a rectal tumour on ADC maps. However, that specific study did not consider other feature groups, except for a subset of FO features. We also examined the reproducibility of types of radiomic features when a range of different image pre-processing operations were applied prior to feature extraction. We used a CCC metric to compare the feature value in the processed ADC map versus the same feature value in the native (unperturbed) ADC map. We have detected the overall trend that FO feature types were robust with respect to many of these perturbations, but GLCM and GLSZM were in general sensitive to such pre-processing. It is not surprising that FO features were less impacted by image pre-processing than textural features. In fact, FO features can be considered as global

statistical descriptors, while textural features provide a local measurement by looking at particular patterns inside gray values. Being a local measurement, any image pre-processing that alters the local values, or the matrixes used for computation of TA, can produce values that are much different than the features computed on the original image. For example, a study [15] performed on a dedicated texture phantom for radiomics studies, showed TA features to be very sensitive to the bin width chosen for computation. In totality, our results suggest that overall global intensity-based descriptors (such as the FO type) may be more tolerant to differences in GTV delineation accuracy, pixel dimensions, noise level and image-enhancing digital filters compared to textural features such as GLCM and GLSZM. These results were found to be consistent across the two different cohorts ($p<0.01$). The found results are in line with a recent study [25] proposing a qualitative synthesis of 41 studies investigating the repeatability and reproducibility of radiomic features. From the analysis, textural features were found to be more sensitive than FO features with respect to inter-observer variability and image processing. However, the analysis also revealed the lack of a consensus. Furthermore, it shows that results could depend on the modality or the anatomical site considered. Unfortunately, due to the lack of literature investigating this topic for rectal cancers in MRI, it is not possible to have a quantitate meta-analysis. Nevertheless, as this study also shows, it becomes fundamental to report the exact details used for the computations prior to features extraction.

It is important to note that we do not make a claim about the potential predictive power of feature types, nor is it in the scope of this study to identify any set of features as more preferable than others. The CCC metrics show that image pre-processing has the potential to strongly change the value of some radiomic features relative to the same feature value in the unperturbed native image, but the data cannot substantiate whether this change is leading to better or worse predictive performance in the final model. Also, as pointed out in [29], when considering the prognostic/predictive power of radiomic features, their correlation with accepted clinical factors (such as for example tumour ex-

tension) should be considered. This is to avoid redundant information that might increase the risk of overfitting, while only features that provide additional information besides other predictors should be kept [11]. The purpose of this study is to emphasize that differences in the steps leading up to feature extraction could negatively affect the wider generalizability of any given model developed using radiomic features as a signature. For instance, if it is known in advance that the intended application of a radiomic signature might include a wide range of pixel dimensions, it may be preferable to prioritize features whose values do not change greatly as a function of pixel size. Alternatively, if a particular radiomics signature uses a specific image-enhancing digital filter, it is almost certain that the exact same digital filter will be needed to obtain reasonable validation results, particularly if complex textural features are part of the radiomic signature being validated. Finally, it is important to verify the correlation between features. Our cross-institutional dataset used in this investigation did not allow us to investigate additional aspects of ADC reproducibility. For example, it is well known that intensity in MR images may drift significantly over time. We did not have the data to examine temporal stability of ADC feature values, though this has been investigated by Newitt et al. [22] for breast tumours. Here, we used only THUNDER imaging series that were a nominal match of the field gradients available from the Danish CLINIC dataset. As pointed out by others, radiomic features may also depend on the number of unique DWI gradients and the magnitude of those gradients used when generating an ADC map [21], [8]. As an additional limitation, we considered in our experiment the preliminary example of radiomic feature sensitivity with respect to the introduction of gaussian noise and the application of gaussian blurring to possibly reduce the noise. This was meant a) to test features' behaviour in an 'extreme situation', considering that ADC maps already present a relevant intrinsic level of noise; b) verifying the sensitivity of features with respect to one possible de-noising technique. Further studies are needed to investigate the impact of noise in features' reproducibility. Bologna et al. [4] further suggests that ADC feature reproducibility will depend on the region of the body being examined. This suggests

that repeatability and reproducibility should be considered early in the radiomic model development process, by way of a priori feature selection. This would lead to better generalizability in external validation. Our future plans include the extension of our study to additional MRI sequences, such as T1 or T2 weighted imaging, which are often used as standard imaging for pelvic malignancies. In particular, we would like to verify if our results can be validated on different modalities, but within the same anatomical site. This will provide us with an initial evaluation of the sensitivity of radiomics features and related imaging pre-processing as a function of different modalities.

4.5 CONCLUSIONS

Evidence in literature clearly points towards the need to evaluate reproducibility of radiomic features derived on ADC maps. In this work, we demonstrated that – generally speaking – the mathematically simpler features, such as those derived from intensity distribution histograms, are less sensitive to manual tumour delineation differences, noise in ADC images, pixel size resampling and intensity discretization. Shape features appear to be strongly affected by delineation quality, and the expertise among groups of observer plays a role. On the whole, GLCM and GLSZM features appear to be poorly or moderately reproducible with respect to the image pre-processing perturbations we reproduced. Further studies are required to elucidate the role of diffusion gradients and temporal stability of DWI scans in order to develop the role of radiomic analysis in supporting treatment response monitoring in locally advanced rectal cancer.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] Salvatore Amodeo, Alan S. Rosman, Vincenzo Desiato, Nicole M. Hindman, Elliot Newman, Russell Berman, H. Leon Pachter, and Marcovalerio Melis. MRI-Based Apparent Diffusion Coefficient for Predicting Pathologic Response of Rectal Cancer After Neoadjuvant Therapy: Systematic Review and Meta-Analysis. *American journal of roentgenology*, 211(5):W205–W216, 2018.

[3] John J. Bartko. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, 19(1):3–11, August 1966.

[4] Marco Bologna, Valentina D. A. Corino, Eros Montin, Antonella Messina, Giuseppina Calareso, Francesca G. Greco, Silvana Sdao, and Luca T. Mainardi. Assessment of Stability and Discrimination Capacity of Radiomic Features on Apparent Diffusion Coefficient Images. *Journal of Digital Imaging*, 31(6):879–894, December 2018.

[5] Susanne Bonekamp, Celia P. Corona-Villalobos, and Ihab R. Kamel. Oncologic applications of diffusion-weighted MRI in the body. *Journal of magnetic resonance imaging: JMRI*, 35(2):257–279, February 2012.

[6] Jean-FranĂ§ois Bosset, Laurence Collette, Gilles Calais, Laurent Mineur, Philippe Maingon, Ljiljana Radosevic-Jelic, Alain Daban, Etienne Bardet, Alexander Beny, Jean-Claude Ollier, and EORTC Radiotherapy Group Trial 22921. Chemotherapy with preoperative radiotherapy in rectal cancer. *The New England Journal of Medicine*, 355(11):1114–1123, September 2006.

[7] Jeroen Buijsen, Jørgen van den Bogaard, Hiska van der Weide, Stephanie Engelsman, Ruud van Stiphout, Marco Janssen, Geerard Beets, Regina Beets-Tan, Philippe Lambin, and Guido Lammering. FDG-PET-CT reduces the interobserver variability in rectal tumor delineation. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 102(3):371–376, March 2012.

[8] Luguang Chen, Fu Shen, Zhihui Li, Haidi Lu, Yukun Chen, Zhen Wang, and Jianping Lu. Diffusion-weighted imaging of rectal cancer on repeatability and cancer characterization: an effect of b-value distribution study. *Cancer Imaging: The Official Publication of the International Cancer Imaging Society*, 18(1):43, November 2018.

[9] Jean-Pierre Gérard, Thierry Conroy, Franck Bonnetain, Olivier Bouché, Olivier Chapet, Marie-Thérèse Closon-Dejardin, Michel Untereiner, Bernard Leduc, Éric Francois, Jean Maurel, et al. Preoperative radiotherapy with or without concurrent fluorouracil and leucovorin in t3-4 rectal cancers: results of ffcd 9203. *Journal of clinical oncology*, 24(28):4620–4625, 2006.

[10] Sofia Gourtsoyianni, Georgia Doumou, Davide Prezzi, Benjamin Taylor, J. James Stirling, N. Jane Taylor, Musib Siddique, Gary J. R. Cook, Robert Glynne-Jones, and Vicky Goh. Primary Rectal Can-

cer: Repeatability of Global and Local-Regional MR Imaging Texture Features. *Radiology*, 284(2):552–561, 2017.

[11] M. Hatt, M. Majdoub, M. Vallieres, F. Tixier, C. C. Le Rest, D. Groheux, E. Hindie, A. Martineau, O. Pradier, R. Hustinx, R. Perdrisot, R. Guillevin, I. El Naqa, and D. Visvikis. 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*, 56(1):38–44, January 2015.

[12] Linda Heijmen, Maartje C. H. M. Verstappen, Edwin E. G. W. Ter Voert, Cornelis J. A. Punt, Wim J. G. Oyen, Lioe-Fee de Geus-Oei, John J. Hermans, Arend Heerschap, and Hanneke W. M. van Laarhoven. Tumour response prediction by diffusion-weighted MR imaging: ready for clinical use? *Critical Reviews in Oncology/Hematology*, 83(2):194–207, August 2012.

[13] Panpan Hu, Jiazhou Wang, Haoyu Zhong, Zhen Zhou, Lijun Shen, Weigang Hu, and Zhen Zhang. Reproducibility with repeat CT in radiomics study for rectal cancer. *Oncotarget*, 7(44):71440–71446, November 2016.

[14] Ines Joye, Annelies Debucquoy, Christophe M. Deroose, Vincent Vandecaveye, Eric Van Cutsem, Albert Wolthuis, André D'Hoore, Xavier Sagaert, Mu Zhou, Olivier Gevaert, and Karin Haustermans. Quantitative imaging outperforms molecular markers when predicting response to chemoradiotherapy for rectal cancer. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 124(1):104–109, 2017.

[15] Ruben T. H. M. Larue, Janna E. van Timmeren, Evelyn E. C. de Jong, Giacomo Feliciani, Ralph T. H. Leijenaar, Wendy M. J. Schreurs, Meindert N. Sosef, Frank H. P. J. Raat, Frans H. R. van der Zande, Marco Das, Wouter van Elmpt, and Philippe Lambin. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice

thicknesses: a comprehensive phantom study. *Acta Oncologica*, 56(11):1544–1553, November 2017.

[16] L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989.

[17] Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. The Design of SimpleITK. *Frontiers in Neuroinformatics*, 7, 2013.

[18] Bin Ma, Qingzhou Xu, Yongxi Song, Peng Gao, and Zhenning Wang. Current issues of preoperative radio(chemo)therapy and its future evolution in locally advanced rectal cancer. *Future Oncology (London, England)*, 13(27):2489–2501, November 2017.

[19] Monique Maas, Patty J Nelemans, Vincenzo Valentini, Prajnan Das, Claus Rödel, Li-Jen Kuo, Felipe A Calvo, Julio García-Aguilar, Rob Glynne-Jones, Karin Haustermans, Mohammed Mohiuddin, Salvatore Pucciarelli, William Small, Javier Suárez, George Theodoropoulos, Sebastiano Biondo, Regina GH Beets-Tan, and Geerard L Beets. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *The Lancet Oncology*, 11(9):835–844, September 2010.

[20] JK MacFarlane, RDH Ryall, and RJ Heald. Mesorectal excision for rectal cancer. *The lancet*, 341(8843):457–460, 1993.

[21] David C. Newitt, Dariya Malyarenko, Thomas L. Chenevert, C. Chad Quarles, Laura Bell, Andriy Fedorov, Fiona Fennessy, Michael A. Jacobs, Meiyappan Solaiyappan, Stefanie Hectors, Bachir Taouli, Mark Muzi, Paul E. Kinahan, Kathleen M. Schmainda, Melissa A. Prah, Erin N. Taber, Christopher Kroenke, Wei Huang, Lori R. Arlinghaus, Thomas E. Yankeelov, Yue Cao, Madhava Aryal, Yi-Fen Yen, Jayashree Kalpathy-Cramer, Amita Shukla-Dave, Maggie Fung, Jiachao Liang, Michael Boss, and

Nola Hylton. Multisite concordance of apparent diffusion coefficient measurements across the NCI Quantitative Imaging Network. *Journal of Medical Imaging*, 5(1):011003, January 2018.

[22] David C. Newitt, Zheng Zhang, Jessica E. Gibbs, Savannah C. Partridge, Thomas L. Chenevert, Mark A. Rosen, Patrick J. Bolan, Helga S. Marques, Sheye Aliu, Wen Li, Lisa Cimino, Bonnie N. Joe, Heidi Umphrey, Haydee Ojeda-Fournier, Basak Dogan, Karen Oh, Hiroyuki Abe, Jennifer Drukteinis, Laura J. Esserman, Nola M. Hylton, and ACRIN Trial Team and I-SPY 2 TRIAL Investigators. Test-retest repeatability and reproducibility of ADC measures by breast DWI: Results from the ACRIN 6698 trial. *Journal of magnetic resonance imaging: JMRI*, 49(6):1617–1628, 2019.

[23] Joseph M. Plummer, Pierre-Anthony Leake, and Matthew R. Albert. Recent advances in the management of rectal cancer: No surgery, minimal surgery or minimally invasive surgery. *World Journal of Gastrointestinal Surgery*, 9(6):139–148, June 2017.

[24] Rolf Sauer, Heinz Becker, Werner Hohenberger, Claus Rödel, Christian Wittekind, Rainer Fietkau, Peter Martus, Jörg Tschmelitsch, Eva Hager, Clemens F Hess, et al. Preoperative versus postoperative chemoradiotherapy for rectal cancer. *New England Journal of Medicine*, 351(17):1731–1740, 2004.

[25] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[26] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[27] Miriam M. van Heeswijk, Doenja M. J. Lambregts, Monique Maas, Max J. Lahaye, Z. Ayas, Jos M. G. M. Slenter, Geerard L. Beets, Frans C. H. Bakers, and Regina G. H. Beets-Tan. Measuring the apparent diffusion coefficient in primary rectal tumors: is there a benefit in performing histogram analyses? *Abdominal Radiology (New York)*, 42(6):1627–1636, 2017.

[28] Ruud G. P. M. van Stiphout, Vincenzo Valentini, Jeroen Buijsen, Guido Lammering, Elisa Meldolesi, Johan van Soest, Lucia Leccisotti, Alessandro Giordano, Maria A. Gambacorta, Andre Dekker, and Philippe Lambin. Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: a multicentric prospective study with external validation. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 113(2):215–222, November 2014.

[29] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[30] C. Weltens, J. Menten, M. Feron, E. Bellon, P. Demaerel, F. Maes, W. Van den Bogaert, and E. van der Schueren. Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 60(1):49–59, July 2001.

[31] Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3):290–303, 2018.

[32] Jerrold H. Zar. Significance Testing of the Spearman Rank Correlation Coefficient. *Journal of the American Statistical Association*, 67(339):578–580, September 1972. Publisher: Taylor & Francis.

5

Repeatability and reproducibility of MRI-based radiomic features in cervical cancer

Adapted from: **"Repeatability and reproducibility of MRI-based radiomic features in cervical cancer"**. S Fiset, ML Welch, J Weiss, M Pintilie, JL Conway, M Milosevic, A Fyles, A Traverso, D Jaffray, U Metser, J Xie, K Han. Radiotherapy and Oncology 135, 107-114. (2019). Contribution: radiomic analysis, statistical analysis, manuscript writing.

Abstract

The aims of this study are to evaluate the stability of radiomic features from T2-weighted MRI of cervical cancer in three ways: (1) repeatability via test–retest; (2) reproducibility between diagnostic MRI and simulation MRI; (3) reproducibility in inter-observer setting. This retrospective cohort study included FIGO stage IB-IVA cervical cancer patients treated with chemoradiation between 2005 and 2014. There were three cohorts of women corresponding to each aim of the study: (1) 8 women who underwent test–retest MRI; (2) 20 women who underwent MRI on different scanners (diagnostic and simulation MRI); (3) 34 women whose diagnostic MRIs were contoured by three observers. Radiomic features based on first-order statistics, shape features and texture features were extracted from the original, Laplacian of Gaussian (LoG)-filtered and wavelet-filtered images, for a total of 1761 features. Stability of radiomic features was assessed using intraclass correlation coefficient (ICC). The inter-observer cohort had the most reproducible features (95.2% with ICC $\geq$0.75) whereas the diagnostic–simulation cohort had the fewest (14.1% with ICC $\geq$0.75). Overall, 229 features had ICC $\geq$0.75 in all three tests. Shape features emerged as the most stable features in all cohorts. The diagnostic–simulation test resulted in the fewest reproducible features. Further research in MRI-based radiomics is required to validate the use of reproducible features in prognostic models.

5.1 INTRODUCTION

Radiomics, the automated high-throughput extraction of quantitative imaging features, is hypothesized to capture the histological heterogeneity inherent to solid tumours [14][26][13]. The potential of radiomics has instigated a multitude of modality- and site-specific investigations to provide robust diagnostic and prognostic models. Generally, computed tomography (CT)-based radiomics have dominated the literature; however, magnetic resonance imaging (MRI) is gaining popularity owing to its superior soft tissue contrast [19]. While radiomics is changing the landscape of cancer imaging research, the lack of consistency in analysis and feature reporting make comparison and repetition of studies difficult [18]. Consequently, studies looking at the repeatability (comparison under constant condition) and reproducibility (comparison under varying conditions) of radiomic features have become increasingly common [31]. Identification of reproducible and repeatable features, and their inclusion in predictive models, are key to ensuring model generalizability. An important indicator of feature repeatability is test–retest, a comparison of radiomic features from two images of the same patient acquired within a short timeframe. Studies looking at two sets of CT images acquired within 15 minutes to 2 weeks found that 29%–98% of calculated features were not repeatable, thus confirming the need for robust feature selection [10][33][23][29][2][28][4][1]. There have been no conclusive studies regarding the test–retest robustness of MR-based radiomic features. In addition to a diagnostic MRI, patients planned for radiotherapy often undergo a simulation MRI in treatment position for radiation treatment planning using a different MRI scanner and image acquisition protocol. Clinical applicability of radiomics will be dependent on its widespread external generalizability. It is therefore essential to identify radiomic features that are able to transcend such differences between image acquisition parameters. Additionally, tumour delineation uncertainty can translate into significant variability in radiomic feature accuracy [8][11]. The need for assessment of inter-observer variability in MRI radiomics is further substantiated by two published studies

which have shown better reproducibility than CT [7][25]. Ultimately, there is a need to identify MRI-based radiomic features that are robust and stable against inevitable variation in clinical data. We hypothesize that we will identify MRI-based radiomic features that are robust to tests of repeatability and reproducibility, which can be utilized in predictive radiomics models. Accordingly, the aims of this study are to evaluate the stability of radiomic features from T2-weighted MRI of cervical cancer in three ways: (1) repeatability via test–retest; (2) reproducibility between diagnostic MRI and simulation MRI; (3) reproducibility in inter-observer setting.

5.2 MATERIAL AND METHODS

5.2.1 Study population

This retrospective cohort study was approved by the institutional research board, with waiver of informed consent. We retrospectively identified all patients with stage IB-IVA cervical cancer who were treated at our centre with chemoradiation between 2005 and 2014. Those who did not undergo diagnostic MRI at our centre prior to treatment were excluded. There were three cohorts of women: (1) 8 women who underwent test–retest simulation MRIs (within 14–47 min); (2) 20 women who underwent a diagnostic MRI and a simulation MRI within an average timeframe of 8 days; (3) 34 patients whose diagnostic MRIs were contoured by three observers (Fig. 5.1). There was overlap between the three patient cohorts. Table 5.2 outlines patient demographics.

5.2.2 Image acquisition

All images were acquired on clinical MR scanners with axial T2-weighted turbo spin-echo (TSE) sequence. Scanner and imaging parameters are listed in Table 5.2. All imaging parameters were the same between images from a single patient in the test–retest cohort

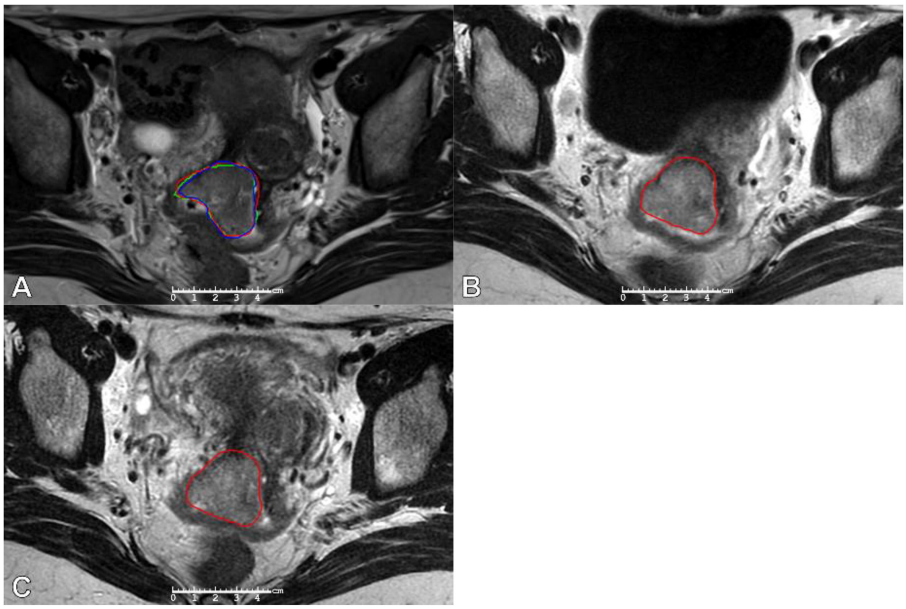


Figure 5.1: Axial T2-weighted MR images of a patient with FIGO stage 2B cervical cancer. (A) Diagnostic MR images with contours by three observers, (B) Radiotherapy simulation MR images acquired on a different scanner approximately 2 hours following A and (C) Radiotherapy simulation MR images acquired 20 minutes following B on the same scanner, after a bathroom break.

Table 1
Patient demographics and image parameters for the three patient cohorts in this study (test–retest, diagnostic–simulation, inter-observer).

Characteristics	Patient cohorts Test–retest	Diagnostic–simulation		Inter-observer
Number of patients	8	20		34
Age, mean ± SD (*yrs*)	54 (9.4)	51 (8.8)		49 (26 – 70)
Tumor characteristics				
FIGO Stage (*n*)				
1B	4	8		13
2A	2	2		5
2B	1	4		10
3A	0	1		0
3B	1	5		6
Max dimension on imaging, mean ± SD (*cm*)	4.1 ± 1.4	4.8 ± 1.6		5.0 ± 1.5
Volume, mean ± SD (cm^3)	29.2 ± 21.0	49.5 ± 44.1		46.0 ± 33.3
Time between images, median (range)	23 min (14–47 min)	8 days (0–14 days)		N/A
Image Parameters	**Test–Retest***	**Diagnostic**	**Simulation**	**Inter-observer**
Number of images (*n*)	8	20	20	34
MR Scanner (*n*)				
GE Signa Excite	0	12	2	20
GE Signa HDx	7	1	14	0
Siemens Verio	1	3	2	9
Siemens Avanto	0	4	2	5
Magnetic Field (*n*)				
1.5 T	7	17	18	25
3.0 T	1	3	2	9
Sequence, median (range)				
Slice Thickness (*mm*)	4	4	4	4
Axial resolution (*mm*)	0.43 (0.43–0.69)	0.43 (0.43–0.78)	0.45 (0.39–0.78)	0.43 (0.43–0.69)
TE (*ms*)	102 (92–102)	98 (88–104)	104 (92–106)	98 (88–106)
TR (*ms*)	6100 (3850–6500)	3820 (3050–6340)	5842 (3500–6500)	3790 (3100–7667)

SD = standard deviation.
* All imaging parameters were the same between images from a single patient in the test–retest cohort.

Figure 5.2: Patient demographics and image parameters for the three patient cohorts in this study (test–retest, diagnostic–simulation, inter-observer).

and the inter-observer cohort. The diagnostic–simulation cohort had differences in imaging parameters for a given patient including scanner model, magnetic field, echo time (TE) and repetition time (TR), as is expected in real clinical scenarios. The cervical tumour was manually delineated on all images by one gynaecologic radiation oncologist (KH) with 5 years of experience using Raystation 6 (RaySearch Laboratories). To minimize intra-observer contouring variability between diagnostic and simulation MRIs in the test–retest cohort, each patient's images were soft-tissue co-registered. Co-registration involved contouring the diagnostic T2 MRI, propagating the contour onto the simulation MRI, and modifying as needed. Additionally, the inter-observer cohort was subsequently contoured by two other gynaecologic radiation oncology observers (JC and JX with 1 and 10 years of experience, respectively).

5.2.3 Feature extraction

After contouring, all DICOM images and associated contours were exported and resampled to 0.6×0.6×4mm to exclude potential confounding by variable in-plane resolutions. Resampling was performed using B-spline interpolation which has been shown to retain tissue contrast differences and has good reproducibility [16][15]. Resampling and subsequent feature extraction were performed using the open-source PyRadiomics (v.1.3.0) package for Python (v. 3.6.5) [32]. The custom script which was used to run PyRadiomics is included in this paper as Supplementary materials. The PyRadiomics platform was selected for radiomic feature extraction to increase accountability and refinement of methodologies. Additionally, this platform was validated against the Image Biomarker Standardization Initiative benchmark values [36]. MRI gray values (signal intensity) are generally relative and cannot be compared between images. To ensure better comparability of gray values, normalization was performed on the images by cantering at the mean and dividing by standard deviation of the gray values in the image as per PyRadiomics standard. In both the

literature and the PyRadiomics documentation, a fixed bin width is recommended as opposed to a fixed bin count [27]. An analysis was performed to determine a suitable bin width value. Due to the normalization, smaller bin widths rather than the default value of 25 in PyRadiomics were required to achieve a sufficient number of bin counts for each patient. A fixed bin width of 0.05 was deemed suitable as it resulted in an average of 54 bins (minimum 17, maximum 95) in the original images. While the sources mentioned above recommend the fixed bin width method, the Image Biomarker Standardization Initiative recommends the use of fixed bin count for T2-weighted MR [22]. To further evaluate the differences between the two methods, the analysis was repeated using a fixed bin count of 64, which has been commonly used in the literature with good reproducibility in PET studies [24][30][20]. A total of 1761 features were computed for each image. The main groupings of texture analysis features were (1) First-order statistics based on pixel gray-level histograms, 18 features; (2) Shape metrics, 13 features; (3) Statistical features derived from texture matrices including gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), neighbouring gray tone difference matrix (NGTDM), 74 features; (4) Statistical features derived from texture matrices in Laplacian-of-Gaussian (LoG) filtered domain (0.5–5.0 mm kernels), 920 features; and (5) Statistical features derived from texture matrices in wavelet filtered domains, 736 features. Texture matrices were calculated in 3 dimensions, resulting in 2 neighbors for each of 13 angles. As per PyRadiomics default, feature values are calculated in all directions and the mean was recorded. No weighting to distance was applied to the GLCM matrix.

5.2.4 Statistics

Feature stability was evaluated using the Intraclass correlation coefficient (ICC). ICC(1,1) was used for the test–retest and diagnostic–simulation cohorts, whereas ICC(2,1) was used for the

inter-observer cohort. Here, an ICC of $\geq$0.75–0.89 was considered good reproducibility and an ICC$\geq$0.90 was considered excellent reproducibility as recommended by Koo et al. [12]. The Dice coefficient was used as a metric for the spatial overlap accuracy of the three manual contours for the images in the inter-observer cohort. A Dice coefficient of 0 indicated no overlap and a value of 1 corresponded to exact overlap [35]. Features which were highly correlated were grouped together in clusters to avoid skewing results if many features show high ICC but, in fact, are all highly correlated and would not add additional value to a radiomics model. Cluster sizes of 10, 100, 200, 300, 400, 500, and 600 were examined. The optimal cluster size was decided to be the one which 75% of pairs of features in a cluster are correlated with a Pearson correlation coefficient above 0.9. This ensures that clusters are highly correlated within themselves but still reduces the number of features. The representative feature from each cluster was selected as the feature with the highest median correlation with the other members of the cluster. The Pearson correlation coefficient was used to evaluate the relationship between features and tumour volume. Volume is a known prognostic indicator; therefore, features which are highly correlated with volume do not add meaningful information to a radiomics model and volume dependency can artificially increase a feature's repeatability [6]. In order to determine whether a specific LoG filter or wavelet decomposition offered superior feature stability, the first-order and texture features calculated on the original image versus the same features calculated in 19 image domains were compared. Therefore, the original image was compared with 10 images from LoG kernel sizes ranging between 0 and 5mm, and 8 images from the wavelet decompositions. For this analysis, the difference between the ICC of the original image and each filtered image was calculated. Only the features which exhibited an ICC$\geq$0.5 in one of the image domains were included in the analysis to reduce artificially high differences between very low ICCs which are not of interest for potential inclusion in radiomics models. An alternative measure of agreement, Krippendorff's alpha, was calculated to assess

for reliability in the three cohorts. Krippendorff's alpha ranges from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement [9]. All statistics were performed with R package v 3.4.2, 2017.

5.3 RESULTS

The Dice coefficients were calculated for the contours on each patient. The mean±standard deviation Dice coefficients were 0.92±0.03, 0.90±0.06 and 0.91±0.06 between observers 1 and 2, 1 and 3, and 2 and 3, respectively. ICC values for the fixed bin width and fixed bin count methods for the original image domain are provided in Table E1 (online). The fixed bin width method produced higher ICCs for the inter-observer cohort whereas the fixed bin count method resulted in higher ICCs for the test–retest cohort. The two methods were approximately equal for the diagnostic–simulation cohort. Therefore, neither method emerged as superior for this study. Only results from the fixed bin width method are reported for the remainder of this text. The number of features that fell within either the "good" ($\geq$0.75–0.89) or "excellent" ($\geq$0.9) ICC category for each cohort is presented in Table 5.3. The shape metrics have the highest percentage of features in the "excellent" ICC group in all three cohorts. Overall, the diagnostic–simulation cohort showed the fewest features with "good" or "excellent" ICC, 14.1% of all features. This contrasts with the test–retest and inter-observer cohorts from which 52.1% and 95.2% of features had *good* or *excellent* reproducibility. In addition to analysing all the features separately, features which were highly correlated were clustered. The optimal number of clusters was 300 where 75% of pairs within each cluster have a Pearson correlation of above 0.90 or below -0.90. When analysing only 1 representative feature from each cluster, the percentage of features which demonstrated *excellent* or *good* reproducibility in the three cohorts remained largely unchanged as listed in Table E2 (online). This confirms that there is no skewing of results from highly correlated features. The diagnostic–simulation cohort again demonstrated the fewest reproducible features with

Table 2
Number of features (n) and percentage of their groups (%) which fall into excellent ICC category (ICC≥0.9), good category (ICC≥0.75–0.89) and other (ICC < 0.75) for all features and distinct feature types (first-order, shape, texture, LoG filtered and wavelet filtered).

	Test–Retest		Diagnostic–Simulation		Inter-observer	
Feature Category (n)	n	%	n	%	n	%
All features (1761)						
ICC ≥ 0.9	398	22.6	109	6.2	1310	74.4
ICC ≥ 0.75–0.89	519	29.5	139	7.9	366	20.8
ICC < 0.75	844	47.9	1513	85.9	85	4.8
First-order (18)						
ICC ≥ 0.9	6	33.3	2	11.1	12	66.7
ICC ≥ 0.75–0.89	7	38.9	1	5.6	5	27.8
ICC < 0.75	5	27.8	15	83.3	1	5.6
Shape Metric (13)						
ICC ≥ 0.9	12	92.3	12	92.3	13	100.0
ICC ≥ 0.75–0.89	1	7.7	0	0.0	0	0.0
ICC < 0.75	0	0.0	1	7.7	0	0.0
Texture (74)						
ICC ≥ 0.9	19	25.7	4	5.4	47	63.5
ICC ≥ 0.75–0.89	24	32.4	2	2.7	25	33.8
ICC < 0.75	31	41.9	68	91.9	2	2.7
LoG (920)						
ICC ≥ 0.9	226	24.6	62	6.7	648	70.4
ICC ≥ 0.75–0.89	307	33.4	113	12.3	225	24.5
ICC < 0.75	387	42.1	745	81.0	47	5.1
Wavelet (736)						
ICC ≥ 0.9	135	18.3	29	3.9	590	80.2
ICC ≥ 0.75–0.89	180	24.5	23	3.1	111	15.1
ICC < 0.75	421	57.2	684	92.9	35	4.8

Figure 5.3: Number of features (n) and percentage of their groups (%) which fall into excellent ICC category (ICC≥0.9), good category (ICC≥0.75–0.89) and other (ICC<0.75) for all features and distinct feature types (first-order, shape, texture, LoG filtered and wavelet filtered).

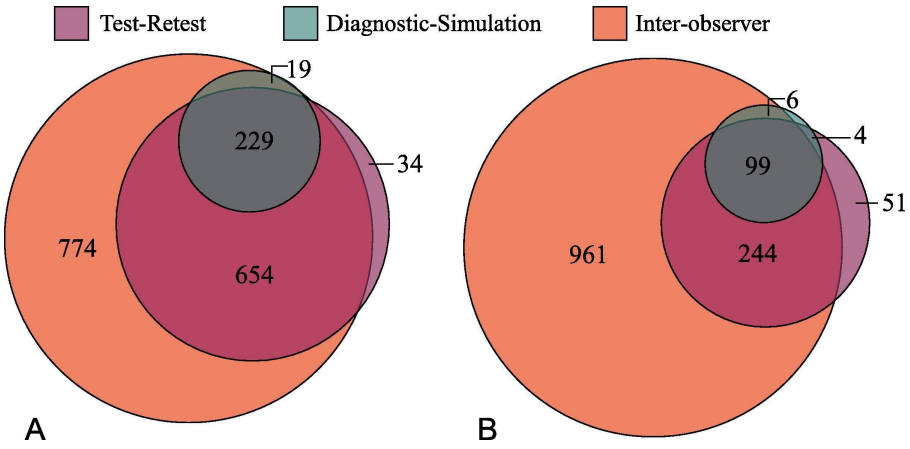


Figure 5.4: (A) Venn diagram illustrating the number of features which have ICC≥0.75 in the three cohorts (B) Venn diagram illustrating the number of features which have ICC≥0.9 in the three cohorts.

15.0% (45/300) having ICC≥0.75. The test–retest and inter-observers cohorts showed 51.7% and 95.6% of representative features have ICC≥0.75. Across all three cohorts, 229 common features out of the total 1761 features (including all image domains) had a *good* ICC value ≥0.75, and 99 features had an *excellent* ICC value ≥0.9 as illustrated in Fig. 5.4. Of the 229 features which had ICC≥0.75 in all three cohorts, 150 features also had a Pearson correlation coefficient of less than 0.9 with volume (i.e. not highly correlated with volume). Many of the features with both ICC≥0.75 and Pearson correlation coefficient <0.9 were repeated in multiple image domains. A list of the ICC values and 95% confidence interval for all radiomic features computed is provided in Table E3 (online). Table E4 provides their Krippendorff's alpha values and Pearson correlation coefficients. First-order and texture features were calculated in 19 image domains: the original image, 10 images with LoG kernel sizes ranging between 0 and 5mm, and 8 images from the wavelet decompositions. To explore any variation in feature stability (ICC) by image domains, ICCs for the 13 first-order and 74 texture features were combined by image domain

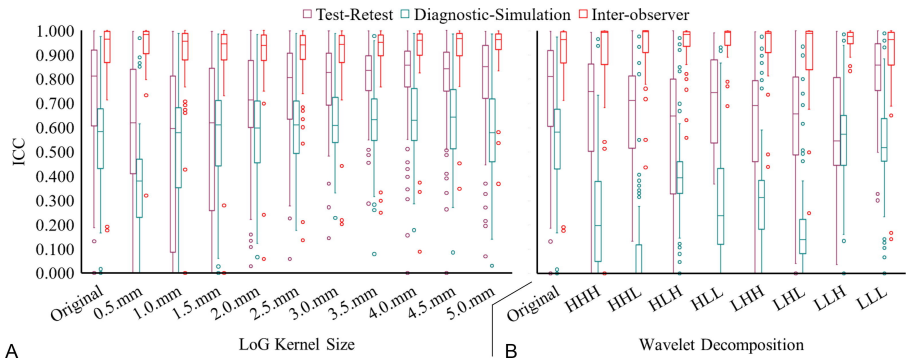


Figure 5.5: Box plot illustrating the distribution of intraclass correlation coefficients (ICC) for the first-order (n=13) and texture features (n=74) derived from the original image, Laplacian of Gaussian (LoG) filtered images with kernel sizes 0.5–5.0mm and each wavelet decomposition.

in Fig. 5.5. The ICCs from the features for the original image are included in each graph for comparison. The diagnostic–simulation cohort demonstrates significantly lower ICCs in all image domains. The range of ICC values varies between image domains with no clear image domain emerging as superior to the others. To compare feature stability in the original versus filtered images, the differences between the ICCs for each feature calculated on the original image and each filtered image were plotted for each of the three patient cohorts (Fig. 5.6). The LoG filtered images showed better ICCs than the original image for the diagnostic–simulation cohort, and worse ICCs for the test–retest and inter-observer cohorts. For the diagnostic–simulation cohorts, 31.9% of features demonstrated >10% higher ICCs with LoG filtered images when compared to the original image in contrast to 23.1% which showed ≥10% lower ICCs with LoG filtered images. The test–retest and inter-observer cohorts on the other hand demonstrated 24.4% and 3.4% of features with ICCs >10% higher in LoG filtered images, and 28.3% and 5.1% of features which demonstrated ICCs≥10% lower with LoG filtered images, respectively. The original image demonstrated better ICCs than the wavelet filtered images in

the diagnostic–simulation and test–retest cohorts. Respectively, 70.1% and 39.0% of features had $\geq$10% higher ICC in their original images when compared with wavelet filtered images. This is compared to 6.9% and 22.4% of features which had $>$10% lower ICC in their original image when compared with wavelet filtered images in the same cohorts. The inter-observer cohort demonstrated modestly higher ICCs from wavelet filtered images than from the original image domain (10.8% vs 5.6%). Further breakdowns of feature differences between original images and filtered images categorized by filter or texture feature type are supplied in Fig. E1 A-L.

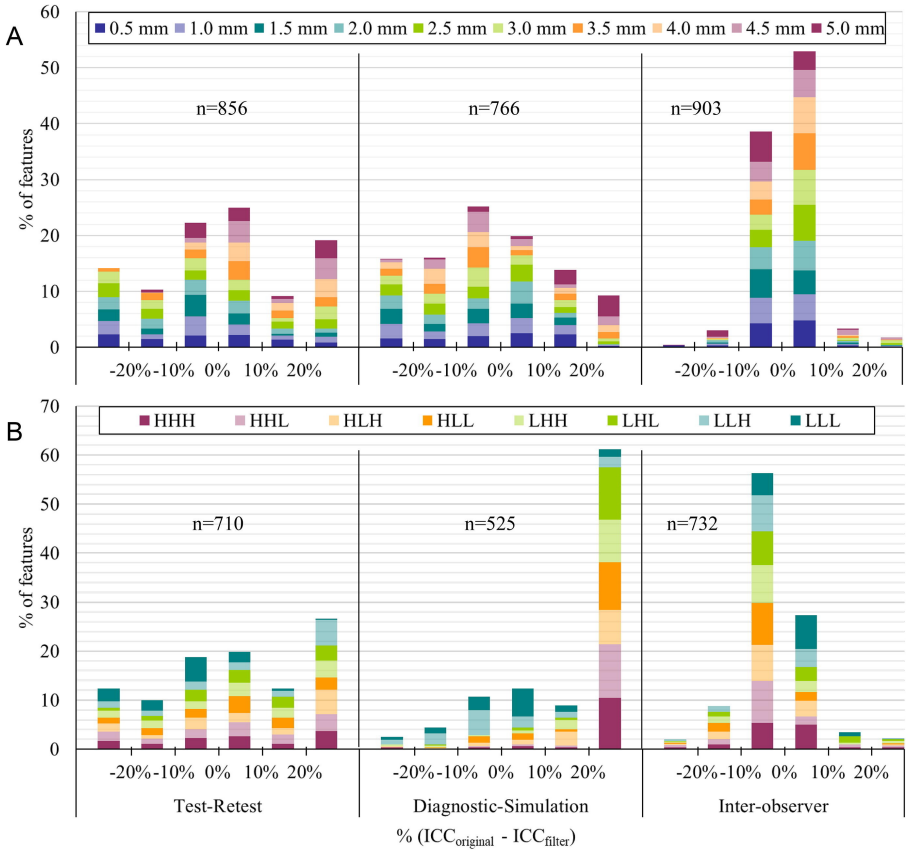


Figure 5.6: Histogram demonstrating the percentage difference between the ICC for the original vs. filtered feature for the three study cohorts. Only features which have ICC>0.5 in either the original image or the filtered image (*n* on the figure) are included. Each colour in the bars represents the ICC differences between the original image and individual filters. (A) The ICC differences between the original image and the various LoG kernel size filters. (B) The ICC differences between the original image and the wavelet decompositions.

5.4 DISCUSSION

Radiomics has emerged as a means of image-based prognostication. Ensuring radiomic feature stability is imperative to the external generalizability of such prognostic models. It is anticipated that this study will help guide the selection of stable radiomic features in future prognostic models by evaluating feature repeatability and reproducibility of radiomic features in three tests. Specifically, the test–retest cohort offers a controlled environment to identify radiomic features which most likely identify characteristics inherent to the tumour. The diagnostic–simulation cohort aims to identify features which are robust against differences in scanners and acquisition protocols, thus mimicking a clinical scenario. Both the test–retest and diagnostic–simulation evaluate errors originating from data acquisition. The inter-observer cohort, on the other hand, evaluates error originating from tumor delineation, another important clinical scenario. Combining the results from the three cohorts can represent a good strategy to perform feature dimensionality reduction. The importance of careful feature selection is first demonstrated in the test–retest cohort which resulted in 47.9% of the features with ICC <0.75 (below good reproducibility) despite the controlled setting. Likewise, even with the use of a phantom and identical imaging parameters, one study has shown that 4% of CT-radiomic features had a concordance correlation coefficient (a numerically similar but alternative popular agreement index to ICC which does not include ANOVA assumptions) of ≤ 0.85[3]. The inter-observer cohort demonstrated high ICCs, 74.4% of which were ≥ 0.9. Such a high ICC value in the inter-observer setting is expected given the high Dice coefficients (>0.9) between the observers' contours. In comparison, the literature reports Dice coefficients ranging from 0.86 for non-small cell lung cancer (NSCLC) CT to 0.26 for mesothelioma CT; 91% of features to have ICC>0.8 for NSCLC PET; and an average ICC of 0.77 for NSCLC CT-PET[33][22][21]. Inter-observer variability in MRI-radiomics has shown an average ICC of 0.85 for breast cancer and an ICC>0.95 for all entropy features (only features examined) from diffusion-weighted MRI for cervical cancers [7][25]. The diagnos-

tic–simulation test resulted in the fewest reproducible features, 14.1% of which have an ICC$\geq$0.75. From this study, we draw three conclusions. Firstly, shape features demonstrated the highest repeatability and reproducibility in all tests. Shape features are commonly reported as highly reproducible in the literature, and were shown to be less sensitive to CT slice thickness and reconstruction parameters in a phantom study[34]. Further, shape features were found to be repeatable in test–retest of rectal cancer and NSCLC[10]. A recent systematic review, mostly based on CT studies, concluded that shape features were more reproducible than texture features, but that first-order features are better than both[31]. Secondly, the diagnostic–simulation cohort was designed to test features in a clinically relevant setting across different MR scanners. Reasonably, the number of reproducible features was fewer than the other two cohorts, likely due to differing image acquisition parameters. Of the 248 reproducible features in the diagnostic–simulation cohort, 92.3% was also reproducible in the other two cohorts. Our findings are difficult to compare to the literature as specific features are uncommonly reported, especially given the sparse literature on MRI-based radiomics. Of the reproducible features identified in our study, coarseness has been reported as reproducible for breast cancer PET imaging[20]. Additionally, Fave et al. reported coarseness, gray length nonuniformity and run length nonuniformity as reproducible for NSCLC cone-beam CT [6]. Leijenaar et al. reported that GLCM and GLRLM were more reproducible than GLSZM, each of which encompasses at least one feature which appeared in our study as reproducible [17]. Thirdly, there is no substantial difference in feature stability between the original and filtered image domains. Wavelet and LoG-filtered images showed both better and worse reproducibility than the original images in the three cohorts tested in this study. Specifically with regard to the diagnostic–simulation cohort, this finding suggests that there is no filter or decomposition which overcame differences in acquisition parameters without losing the inherent tumour texture. Similarly, Schwier et al. demonstrated no significant improvement in reproducibility with a certain LoG-filter or wavelet decomposition [27]. Elsewhere, Timmeren et al. reported that wavelet

features were less reproducible than the unfiltered image features in a test–retest scenario [33]. We acknowledge limitations in our study. This was a single-institutional retrospective study with a modest number of patients that may not be representative of other institutions or patients. However, our cohort size is very similar to those reported in the literature [1][2][28][5] and provides important results which highlight the pressing need for radiomic studies with larger cohorts. Additionally, this study focused on cervical cancer and its applicability to other tumour sites is unconfirmed. Despite the validation of the PyRadiomics platform, results may differ from other radiomic feature extraction platforms. The fixed bin method employed in this study is limited due to the increased number of bins once wavelet filters are applied. Further investigation on the effect of fixed bin width versus dynamic bin width is required. There was no bias field correction applied to the images in this study. The impact of field variation across the bore on feature reproducibility requires further study. Finally, although we used commonly reported cut-offs from the literature for ICC categories (0.75 and 0.9), these may not represent the ideal threshold for feature inclusion in prognostic models. While this study presents limitations, it has systematically evaluated MR-based radiomic reproducibility in three clinically applicable settings which has scarcely been done previously. Future work will involve analyses of the dependencies between radiomic features and clinical variables to better understand which radiomic features are the most appropriate for inclusion in prognostic models. In conclusion, MRI-based radiomic features of cervical tumours were tested for their repeatability and reproducibility. Shape features emerged as the most reliable. The diagnostic–simulation resulted in the fewest reproducible features which highlights the importance of careful feature selection for radiomics generalizability. Further research in MRI-based radiomics is required to validate the use of reproducible features in prognostic models.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] Yoganand Balagurunathan, Yuhua Gu, Hua Wang, Virendra Kumar, Olya Grove, Sam Hawkins, Jongphil Kim, Dmitry B. Goldgof, Lawrence O. Hall, Robert A. Gatenby, and Robert J. Gillies. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Translational Oncology*, 7(1):72–87, February 2014.

[3] Roberto Berenguer, María del Rosario Pastor-Juan, Jesús Canales-Vázquez, Miguel Castro-García, María Victoria Villas, Francisco Mansilla Legorburo, and Sebastià Sabater. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*, 288(2):407–415, August 2018.

[4] Marie-Charlotte Desseroit, Florent Tixier, Wolfgang A. Weber, Barry A. Siegel, Catherine Cheze Le Rest, Dimitris Visvikis, and Mathieu Hatt. 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: Official Publication, Society of Nuclear Medicine*, 58(3):406–411, March 2017.

[5] Marie-Charlotte Desseroit, Dimitris Visvikis, Florent Tixier, Mohamed Majdoub, Rémy Perdrisot, Rémy Guillevin, Catherine Cheze Le Rest, and Mathieu Hatt. Development of a nomogram combining clinical staging with 18 f-fdg pet/ct image features in non-small-cell lung cancer stage i–iii. *European journal of nuclear medicine and molecular imaging*, 43(8):1477–1485, 2016.

[6] Xenia Fave, Dennis Mackin, Jinzhong Yang, Joy Zhang, David Fried, Peter Balter, David Followill, Daniel Gomez, A Kyle Jones, Francesco Stingo, et al. Can radiomics features be reproducibly measured from cbct images for patients with non-small cell lung cancer? *Medical physics*, 42(12):6784–6797, 2015.

[7] Yue Guan, Weifeng Li, Zhuoran Jiang, Ying Chen, Song Liu, Jian He, Zhengyang Zhou, and Yun Ge. Whole-Lesion Apparent Diffusion Coefficient-Based Entropy-Related Parameters for Characterizing Cervical Cancers: Initial Findings. *Academic Radiology*, 23(12):1559–1567, 2016.

[8] Akihiro Haga, Wataru Takahashi, Shuri Aoki, Kanabu Nawa, Hideomi Yamashita, Osamu Abe, and Keiichi Nakagawa. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. *Radiological Physics and Technology*, 11(1):27–35, March 2018.

[9] R Brian Haynes, K Ann McKibbon, Nancy L Wilczynski, Stephen D Walter, and Stephen R Werre. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ*, 330(7501):1179, May 2005.

[10] Panpan Hu, Jiazhou Wang, Haoyu Zhong, Zhen Zhou, Lijun Shen, Weigang Hu, and Zhen Zhang. Reproducibility with repeat CT in radiomics study for rectal cancer. *Oncotarget*, 7(44):71440–71446, November 2016.

[11] Qiao Huang, Lin Lu, Laurent Dercle, Philip Lichtenstein, Yajun Li, Qian Yin, Min Zong, Lawrence Schwartz, and Binsheng Zhao. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. *Journal of Medical Imaging*, 5(1):011005, January 2018.

[12] Terry K. Koo and Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, June 2016.

[13] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, November 2012.

[14] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, AndrĂ© Dekker, and Hugo J.W.L. Aerts. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, 48(4):441–446, March 2012.

[15] Ruben T. H. M. Larue, Janna E. van Timmeren, Evelyn E. C. de Jong, Giacomo Feliciani, Ralph T. H. Leijenaar, Wendy M. J. Schreurs, Meindert N. Sosef, Frank H. P. J. Raat, Frans H. R. van der Zande, Marco Das, Wouter van Elmpt, and Philippe Lambin. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice

thicknesses: a comprehensive phantom study. *Acta Oncologica*, 56(11):1544–1553, November 2017.

[16] T. M. Lehmann, C. Gönner, and K. Spitzer. Survey: interpolation methods in medical image processing. *IEEE transactions on medical imaging*, 18(11):1049–1075, November 1999.

[17] Ralph TH Leijenaar, Sara Carvalho, Emmanuel Rios Velazquez, Wouter JC Van Elmpt, Chintan Parmar, Otto S Hoekstra, Corneline J Hoekstra, Ronald Boellaard, André LAJ Dekker, Robert J Gillies, et al. Stability of fdg-pet radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica*, 52(7):1391–1397, 2013.

[18] Meghan G. Lubner, Andrew D. Smith, Kumar Sandrasegaran, Dushyant V. Sahani, and Perry J. Pickhardt. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, 37(5):1483–1503, October 2017.

[19] James P. B. O'Connor, Chris J. Rose, John C. Waterton, Richard A. D. Carano, Geoff J. M. Parker, and Alan Jackson. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 21(2):249–257, January 2015.

[20] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat. Tumor Texture Analysis in 18F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. *Journal of Nuclear Medicine*, 55(3):414–422, March 2014.

[21] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains,

Philippe Lambin, and Hugo J. W. L. Aerts. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. *PLoS ONE*, 9(7):e102107, July 2014.

[22] Matea Pavic, Marta Bogowicz, Xaver Würms, Stefan Glatz, Tobias Finazzi, Oliver Riesterer, Johannes Roesch, Leonie Rudofsky, Martina Friess, Patrick Veit-Haibach, Martin Huellner, Isabelle Opitz, Walter Weder, Thomas Frauenfelder, Matthias Guckenberger, and Stephanie Tanadini-Lang. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncologica (Stockholm, Sweden)*, 57(8):1070–1074, August 2018.

[23] Thomas Perrin, Abhishek Midya, Rikiya Yamashita, Jayasree Chakraborty, Tome Saidon, William R. Jarnagin, Mithat Gonen, Amber L. Simpson, and Richard K. G. Do. Short-term reproducibility of radiomic features in liver parenchyma and liver malignancies on contrast-enhanced CT imaging. *Abdominal Radiology (New York)*, 43(12):3271–3278, 2018.

[24] L. Presotto, V. Bettinardi, E. De Bernardi, M. L. Belli, G. M. Cattaneo, S. Broggi, and C. Fiorino. PET textural features stability and pattern discrimination power for radiomics analysis: An "ad-hoc" phantoms study. *Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB)*, 50:66–74, June 2018.

[25] Ashirbani Saha, Michael R. Harowicz, and Maciej A. Mazurowski. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. *Medical Physics*, 45(7):3076–3085, July 2018.

[26] Elisa Scalco and Giovanna Rizzo. Texture analysis of medical images for radiotherapy applications. *The British Journal of Radiology*, 90(1070):20160642, February 2017.

[27] Michael Schwier, Joost van Griethuysen, Mark G. Vangel, Steve Pieper, Sharon Peled, Clare M. Tempany, Hugo JWL Aerts, Ron Kikinis, Fiona M. Fennessy, and Andrey Fedorov. Repeatability of Multiparametric Prostate MRI Radiomics Features. *arXiv:1807.06089 [cs, eess]*, July 2018. arXiv: 1807.06089.

[28] Isaac Shiri, Hamid Abdollahi, Sajad Shaysteh, and Seied Rabi Mahdavi. Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features. *Iranian Journal of Radiology*, Special iss(5), April 2017.

[29] A. Talwar, J. M. Y. Willaime, L. C. Pickup, M. Enescu, D. Boukerroui, W. Hickes, N. M. Rahman, M. J. Gooding, T. Kadir, and F. V. Gleeson. Pulmonary nodules: Assessing the imaging biomarkers of malignancy in a "coffee-break". *European Journal of Radiology*, 101:82–86, April 2018.

[30] Florent Tixier, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 53(5):693–700, May 2012.

[31] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[32] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[33] Janna E. van Timmeren, Ralph T. H. Leijenaar, Wouter van Elmpt, Jiazhou Wang, Zhen Zhang, André Dekker, and Philippe Lambin. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*, 2(4):361–365, December 2016.

[34] Binsheng Zhao, Yongqiang Tan, Wei-Yann Tsai, Jing Qi, Chuanmiao Xie, Lin Lu, and Lawrence H. Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific Reports*, 6(1), September 2016.

[35] Rongrong Zhou, Zhongxing Liao, Tinsu Pan, Sarah A. Milgrom, Chelsea C. Pinnix, Anhui Shi, Linglong Tang, Ju Yang, Ying Liu, Daniel Gomez, Quynh-Nhu Nguyen, Bouthaina S. Dabaja, Laurence Court, and Jinzhong Yang. Cardiac atlas development and validation for automatic segmentation of cardiac substructures. *Radiotherapy and Oncology*, 122(1):66–71, January 2017.

[36] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

6

Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients

Adapted from: **"Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients"**. A Traverso, M Kazmierski, ML Welch, J Weiss, S Fiset, WD Foltz, A Gladwish, A Dekker, D Jaffray, L Wee, K Han. Radiotherapy and Oncology 143, 88-94. (2020).

Abstract

The aims of this study are to evaluate the stability of radiomic features from Apparent Diffusion Coefficient (ADC) maps of cervical cancer with respect to: (1) reproducibility in inter-observer delineation, and (2) image pre-processing (normalization/quantization) prior to feature extraction. Two observers manually delineated the tumour on ADC maps derived from pre-treatment diffusion-weighted Magnetic Resonance imaging of 81 patients with FIGO stage IB-IVA cervical cancer. First-order, shape, and texture features were extracted from the original and filtered images considering 5 different normalizations (four taken from the available literature, and one based on urine ADC) and two different quantization techniques (fixed-bin widths from 0.05 to 25, and fixed-bin count). Stability of radiomic features was assessed using intraclass correlation coefficient (ICC): poor (ICC$<$0.75); good (0.75$\leq$ICC$\leq$0.89), and excellent (ICC$\geq$0.90). Dependencies of the features with tumour volume were assessed using Spearman's correlation coefficient (ρ). The approach using urine-normalized values together with a smaller bin width (0.05) was the most reproducible (428/552, 78% features with ICC$\geq$0.75); the fixed-bin count approach was the least (215/552, 39% with ICC$\geq$0.75). Without normalization, using a fixed bin width of 25, 348/552 (63%) of features had an ICC$\geq$0.75. Overall, 26% (range 25–30%) of the features were volume-dependent ($\rho \geq$0.6). None of the volume-independent shape features were found to be reproducible. Applying normalization prior to features extraction increases the reproducibility of ADC-based radiomics features. When normalization is applied, a fixed-bin width approach with smaller widths is suggested.

6.1 INTRODUCTION

Cervical cancer is the fourth most frequent cancer in women, with an estimated 570,000 new cases in 2018, representing 6.6% of all female cancers worldwide. Cervical cancer still represents a significant burden for middle- and low-income countries [3]. Standard treatment for locally advanced (stage IB2-IVA) cervical cancer is concurrent chemoradiation. Computed Tomography (CT) and Magnetic Resonance (MR) are the standard imaging modalities for cervical cancer staging and evaluation of treatment response. Through appropriate choice of pulse sequences MR imaging provides greater soft tissue contrast than CT and enables assessment of physiological parameters and biochemical function. Diffusion-weighted imaging (DWI) in MR enables measurement of water diffusivity via generation of Apparent Diffusion Coefficient (ADC) maps, and ADC is an established biomarker of tumour cell density and related changes post-therapy [15]. DWI is increasing acquired in addition to T2-weighted MRI to detect cervical tumour [21], and pre-treatment tumour ADC has been shown to be predictive of recurrence in patients with cervical cancer treated with chemoradiation [11][8]. Radiomics refers to the extraction of additional information from the delineated GTV (Gross Tumour Volume) of patients' scans: including first order statistics of intensity values; morphological properties, and textural descriptors looking at local patterns (*textures*) [7]. Radiomics has been extensively applied to CT and Positron Emission Tomography (PET), with a growing number of studies addressing its role in MRI and specifically in DWI and ADC maps [13][26]. Two recent publications [17][18] have discovered and validated an ADC-derived radiomic feature (EntropyGLCM) as an independent predictor of disease-free survival and locoregional control in cervical cancer. These results highlight the promising role of MRI for outcome prediction in oncology and the urgent need to develop a robust methodology for feature extraction in DWI. Compared to CT, for example, radiomics analysis applied to ADC presents additional challenges intrinsic to the technology (e.g., more variable system and imaging parameters

limiting features reproducibilities; larger inter-observer variability in tumour delineation [12]). Feature reproducibility is a necessary, but insufficient, condition for high predictive power of a radiomic feature. However, if a radiomic feature has poor reproducibility, then its predictive power is likely low too. This has been deeply explained by Gudmundsson et al. [10] when analysing the stability of a physiological time series. While the Image Biomarker Standardization Initiative (IBSI) has set standards for feature extraction in CT and PET, there is still no established agreement on feature stability assessments and harmonization in MR [27]. Additional image pre-processing prior to feature extraction might be necessary, including: (a) image normalization which may reduce site-specific protocol differences; and (b) optimal configurations for the extraction of textural features, such as quantization of intensity values from which radiomic features are extracted. To extensively assess feature reproducibility, there is a need to investigate the sensitivity of radiomic features to inter-observer variability and image pre-processing in ADC. Finally, correlations between radiomic features (many of which seem to be a surrogate of tumour volume [24]) and tumour volume should also be investigated, to avoid the risk of introducing redundant information and over-fitting in radiomics-based prediction models. With the aim of speeding up the harmonization of radiomics in ADC and extending the work presented by the IBSI, in this manuscript a detailed methodology for evaluating the stability of radiomic features is proposed: (a) with respect to inter-observer variability; and (b) by comparing different normalizations and quantization approaches in ADC maps of cervical cancer patients. These analyses have zero overlap with prior considerations of this data, which tested factors affecting tumour ADC variability and association between pre-treatment ADC value with disease recurrence [12][10].

6.2 MATERIAL AND METHODS

6.2.1 Study population

This retrospective study was conducted using ADC maps derived from DWI from 81 women with stage IB–IVA cervical cancer treated with definitive chemoradiation in 2009–2013. This dataset has been described in [6]. The study was approved by the Institutional Review Board, with waiver of informed consent.

6.2.2 MRI methodology

Stacks of 2D T2-weighted images and ADC maps covering the cervical tumour were acquired on Siemens MRI systems [9]. The ADC maps were acquired according to the parameter sets in Table E1 (Supplementary Material).

6.2.3 Gross tumour volume (GTV) delineations

Two observers (a senior radiation oncology and a radiology research fellow) independently manually delineated the GTV in three dimensions directly on ADC maps co-registered to T2-weighted images (Pinnacle Treatment Planning System, Philips, The Netherlands). Their contours were reviewed by a gynaecologic radiation oncologist with 2.5 years of experience and a radiologist with 8 years of experience, respectively.

6.2.4 Image pre-processing

Digital pre-processing on ADC maps were applied prior to extracting radiomic features, including a combination of normalization and quantization. Features were also computed without normalization, termed 'baseline features' to test the sensitivity of first order and textural features. Table 6.1 summarizes the normalization and quantization approaches which were applied: no normalization +

fixed-bin width approach (Raw); no normalization + fixed-bin count approach (BinCount64); bladder-based normalization+fixed-bin width approach (BladderNorm), and 4 other normalization + fixed-bin width approaches (S1, S100BW15, S100BW5, S333).

6.2.5 Normalization

Normalization was applied to the whole image (not only within the ROI), by subtracting mean intensity centring it and dividing by the standard deviation. An additional biological-based normalization was considered by drawing a circular ROI of 10mm diameter in the middle of the bladder, where the urine signal was maximum for each patient [9]. Each intensity level was then normalized by dividing original values by the median intensity of the ROI.

6.2.6 Quantization

Besides normalization, different discretization (binning) of the image intensities was considered. In fact, textural feature computation requires the image intensities to be quantized into a discrete number of gray levels. We chose to compute textural feature using the fixed-bin width approach, as suggested by Leijenaar et al. [16], where intensity values are quantized in bins of fixed dimension. The bin width was chosen to produce between 30 and 128 bin counts as recommended by Tixier et al. [23]. The normalization S100 (Table 6.1) with bin width 5 (referred as S100BW5) produced a median of 25 bins (range 10–40) and thus was excluded from subsequent analysis as per [24]. While the aforementioned sources recommend a fixed-bin width approach, the IBSI [27] suggests a fixed bin count approach for raw MRI with arbitrary intensity units, as the fixed bin count approach introduces a normalizing effect. Since an optimal strategy for the extraction of texture features from ADC map (calibrated units) is not defined yet, we assessed both fixed bin width and fixed bin count approaches with respect to feature reproducibility. In addition,

Table 1
Patient demographics and image parameters for the three patient cohorts in this study (test–retest, diagnostic–simulation, inter-observer).

Characteristics	Patient cohorts			
	Test–retest	Diagnostic–simulation		Inter-observer
Number of patients	8	20		34
Age, mean ± SD (*yrs*)	54 (9.4)	51 (8.8)		49 (26 – 70)
Tumor characteristics				
FIGO Stage (*n*)				
1B	4	8		13
2A	2	2		5
2B	1	4		10
3A	0	1		0
3B	1	5		6
Max dimension on imaging, mean ± SD (*cm*)	4.1 ± 1.4	4.8 ± 1.6		5.0 ± 1.5
Volume, mean ± SD (cm^3)	29.2 ± 21.0	49.5 ± 44.1		46.0 ± 33.3
Time between images, median (range)	23 min (14–47 min)	8 days (0–14 days)		N/A
Image Parameters	Test–Retest*	Diagnostic	Simulation	Inter-observer
Number of images (*n*)	8	20	20	34
MR Scanner (*n*)				
GE Signa Excite	0	12	2	20
GE Signa HDx	7	1	14	0
Siemens Verio	1	3	2	9
Siemens Avanto	0	4	2	5
Magnetic Field (*n*)				
1.5 T	7	17	18	25
3.0 T	1	3	2	9
Sequence, median (range)				
Slice Thickness (*mm*)	4	4	4	4
Axial resolution (*mm*)	0.43 (0.43–0.69)	0.43 (0.43–0.78)	0.45 (0.39–0.78)	0.43 (0.43–0.69)
TE (*ms*)	102 (92–102)	98 (88–104)	104 (92–106)	98 (88–106)
TR (*ms*)	6100 (3850–6500)	3820 (3050–6340)	5842 (3500–6500)	3790 (3100–7667)

SD = standard deviation.
* All imaging parameters were the same between images from a single patient in the test–retest cohort.

Figure 6.1: Summary of the normalizations/quantizations considered in the experiment. The approach "S100BW5" produced a non-conformal (low) number of bins and thus was excluded from further analysis. For the Raw and BladderNorm scenarios the first bin starts at 0 intensity (as per ADC unit definition). The S333 scenario projects original values into values similar to CT range (-1000 is then the start value). For S1, S100BW5 and S100BW15 the start value is 0.

we explored if the "implicit" normalization embedded within the fixed-bin count approach was comparable to an approach using an "explicit" normalization combined with a fixed-bin width algorithm. A fixed bin number of 64 ("BINCOUNT 64") was chosen since it is commonly used in the literature, with good reproducibility in PET studies [23]. Outlier removal (values larger or smaller than 3 standard deviations with respect to the mean intensity value) was performed for this configuration, as it was shown to increase reproducibility in recent publications [4]. Finally, different filters were applied to the original images before feature extraction. The following options were considered: (a) Laplacian of Gaussian (σ= 3mm); (b) square; (c) square root; (d) exponential, and (f) gradient. The filters are explained in the PyRadiomics documentation (Link here). Filtering, as per PyRadiomics defaults, always happen before quantization. It is worth noting that filtering, normalization and quantization only affect first order and textural features, while shape features are left unchanged, since they are computed directly using the tumour segmentation mas. Therefore, shape features were extracted directly from the segmentation masks, one features for each of the 2 annotated GTVs for each subject.

6.2.7 Feature extraction

Radiomic feature extraction was performed with PyRadiomics version 2.0.1. The open-source PyRex extensions (Link here) were used to manage the conversions of DICOM and DICOM RT Structure files to binary segmentation masks. Prior to features extraction, images were resampled to 0.6×0.6×4mm, as standard procedure for radiomics studies [27]. A total of 565 radiomic features, of which 13 were shape features, were extracted from each subject. Mathematical definitions of these features are given on the PyRadiomics feature documentation page (Link here). Details used for computations are specified in the supplementary material (Section 1). The summary of this study workflow is shown in Fig. 6.2.

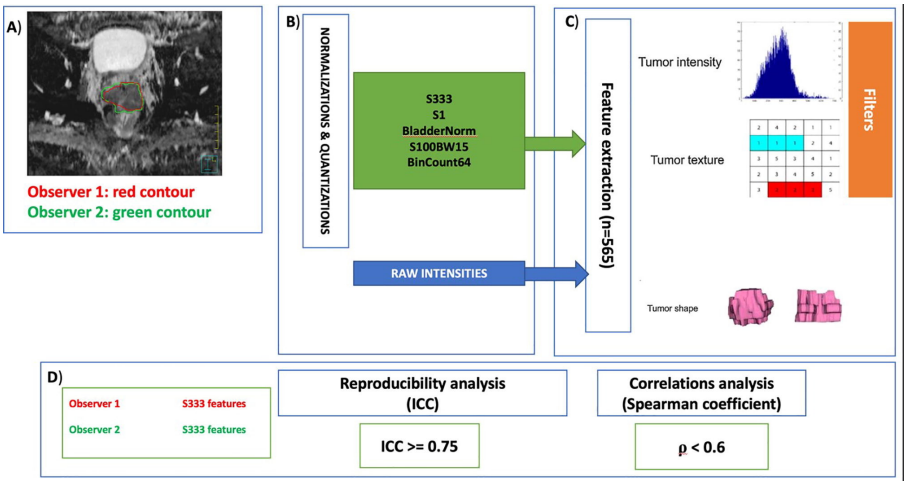


Figure 6.2: Summary of study workflow. (A) Two observers independently delineated the gross tumour volume on apparent diffusion coefficient maps. Contours were then exported as RTSTRUCT DICOM files. (B) Normalization and quantization procedure prior to feature extraction: 5 different approaches were applied prior to feature extractions. (C) Feature extraction: radiomic features were extracted from the two different contours and for all the different approaches. Also, features were extracted from raw intensities, without any prior normalization, using default PyRadiomics settings. Shape, statistics, and textural categories were considered. Image pre-processing filters were also applied to statistics and textural features. (D) Statistical analysis: to evaluate the stability of features with respect to inter-observer variability and the effect of normalization, the ICC metric was used. Feature values were compared between the two observers for all the configurations considered. Finally, dependencies between features and tumor volume were evaluated using the Spearman's correlation coefficient.

6.2.8 Statistical analysis

Statistical analysis was performed in R Studio (v1.1.383), R (v 3.5.1) and Python (v3.6.4). The Intraclass Correlation Coefficient (ICC) [2] was chosen as the reproducibility metric to evaluate the agreement of radiomic feature values with respect to interobserver variability for the different configurations considered. In particular, the definition of ICC (2,1), corresponding to two ways random effects with absolute agreement [14], was chosen for the analysis. An ICC of $\geq$0.75–0.89 was considered good reproducibility and an ICC$\geq$0.90 was considered excellent reproducibility as recommended by Koo et al. [14]. We considered a feature reproducible if ICC$\geq$0.75. The stability of the 13 shape features was evaluated only with respect to interobserver variability and not for all the quantization/normalization approaches, since they do not affect shape features. For clarity, the results for the shape features are presented in a dedicated subsection. The results of the analysis to test the impact of different approaches with respect to inter-observer variability refer to a total of 552 features, which corresponds to the total of computed features (565) minus the number of shape features (13). The DICE coefficient was used to measure spatial overlap between the two observers' contours. It ranges from a minimum of 0 (no spatial overlap) to a maximum of 1 (absolute agreement). Correlations between radiomic features and tumour volume were evaluated using the Spearman correlation coefficient (ρ). Correlations were computed for all the features and for all the considered approaches. The reported coefficient is the average coefficient between the two observers. ICCs were compared using repeated measures ANOVA, and pairwise p-values were adjusted using the false discovery rate (FDR) method [20]. Dichotomized ICCs were compared using a chi-squared test and pairwise p-values were adjusted using the FDR. Conditions of homogeneity of variance and normality of data were confirmed by performing the Levene's and Shapiro-Wilk's tests [5]. A sensitivity analysis was performed to investigate the possible effects magnetic field strength (1.5T and 3T) or b-values (0, 100, 800 vs 0, 400, 800 vs 0, 50, 400, 800 vs 0, 50, 400, 1000 s/mm2). The significance of the configura-

tions was tested using repeated-measures ANOVA controlling for the magnetic field strength and b-values. Statistical significance was set to $p<0.05$.

6.3 RESULTS

The median DICE similarity coefficient for all the patients between the two observers' GTV was 0.73 ± 0.12. ICCs values and the corresponding lower limits of the 95% confidence interval are provided in Table E2. Out of the 13 shape features, 6 features had "excellent" reproducibility (ICC$\geq$0.9), 3 features had "good" reproducibility (ICC$\geq$0.75–0.89), and the remaining 4 features had poor/moderate reproducibility (ICC$<$0.75). Overall, shape features had a mean ICC of 0.85 ± 0.13 (range 0.49–0.95). All the 9 reproducible (ICC$\geq$0.75) features besides tumour volumes had Spearman correlation with tumour volume $\rho \geq 0.6$ and this might explain their high reproducibility. The number of features within the "excellent" ($\geq$0.9), "good" ($\geq$0.75–0.89) or "poor/moderate" ICC category for each approach is summarized in Table 6.3. Overall, the urine-based approach ("BLADDERNORM") produced the largest number (193/552, 34%), while the fixed-bin count approach ("BINCOUNT64") resulted in the lowest number (65/552, 12%) of highly reproducible features (ICC$\geq$0.9). If only features obtained from the original, unfiltered images are considered (first order+textures, n=92), the most reproducible configuration remains the urine-based approach (70/92, 76% of features with ICC$\geq$0.75); the least remains the fixed bin-count approach (42/92, 46% of features with ICC$\geq$0.75). Fig. 6.4 summarizes the median ICC values and their interquartile ranges related to interobserver variability for each of the approaches considered and for all the 552 features. Repeated-measures ANOVA test showed significant differences in ICC among the various approaches ($F(5, 546)=7.69$, $p<0.001$). Pairwise comparisons revealed that the mean ICC from the fixed-bin count approach (0.68 ± 0.18) was statistically significantly lower than that from all the other

Table 2
Number of features (n) which fall into excellent ICC category (≥0.9), good category (≥0.75–0.89) and poor category (<0.75) for each approach applied to all features, distinct feature types (first order, textural) and different filters (logarithm, exponential, square, square root and gradient). This analysis refers to all the features (n = 552), excluding shape features.

Approach legend	RAW	S333	S1	S100BW15	BINCOUNT64	BLADDERNORM
Feature category (n)						
All features (552)						
ICC ≥ 0.9	111	134	119	129	65	193
0.75 ≤ ICC < 0.9	237	209	233	201	150	235
ICC < 0.75	204	209	200	222	337	124
First order (18)						
ICC ≥ 0.9	3	2	2	2	5	6
0.75 ≤ ICC < 0.9	9	6	6	6	6	7
ICC < 0.75	6	10	10	10	7	5
Textural (74)						
ICC ≥ 0.9	6	6	6	9	7	14
0.75 ≤ ICC < 0.9	39	36	32	29	24	43
ICC < 0.75	29	32	36	36	43	17
Laplacian of Gaussian (92)						
ICC ≥ 0.9	17	34	16	33	11	22
0.75 ≤ ICC < 0.9	31	24	41	27	7	43
ICC < 0.75	44	34	35	32	74	27
Exponential (92)						
ICC ≥ 0.9	12	4	11	4	3	66
0.75 ≤ ICC < 0.9	37	32	33	30	22	17
ICC < 0.75	43	56	48	58	67	12
Square (92)						
ICC ≥ 0.9	8	13	18	12	7	22
0.75 ≤ ICC < 0.9	59	49	47	42	35	48
ICC < 0.75	25	30	27	38	50	22
Square root (92)						
ICC ≥ 0.9	12	20	14	20	13	15
0.75 ≤ ICC < 0.9	36	37	41	37	24	43
ICC < 0.75	44	35	37	35	55	34
Gradient (92)						
ICC ≥ 0.9	53	53	52	49	19	48
0.75 ≤ ICC < 0.9	26	27	33	30	32	34
ICC < 0.75	13	12	7	13	41	10

Figure 6.3: Number of features (n) which fall into excellent ICC category (≥0.9), good category (≥0.75–0.89) and poor category (<0.75) for each approach applied to all features, distinct feature types (first order, textural) and different filters (logarithm, exponential, square, square root and gradient). This analysis refers to all the features (n=552), excluding shape features.

approaches: S333 (0.76±0.17, $p<0.001$); S1 (0.78±0.13, $p<0.001$); S100BW15 (0.76±0.15, $p<0.001$); RAW (0.76±0.15, $p<0.001$); and the urine-based approach (0.82±0.16, $p<0.001$). The mean ICC from the urine-based approach was statistically significantly higher than all the other approaches ($p<0.001$ for all comparisons). On sensitivity analysis controlling for magnetic field strength and b-values, the same trend was observed: the mean ICC from fixed-bin count approach was statistically significantly lower, and the mean ICC from urine-based normalization was statistically significantly higher than all the other configurations (Tables E3 and E4). Fig. 6.5 is the Venn

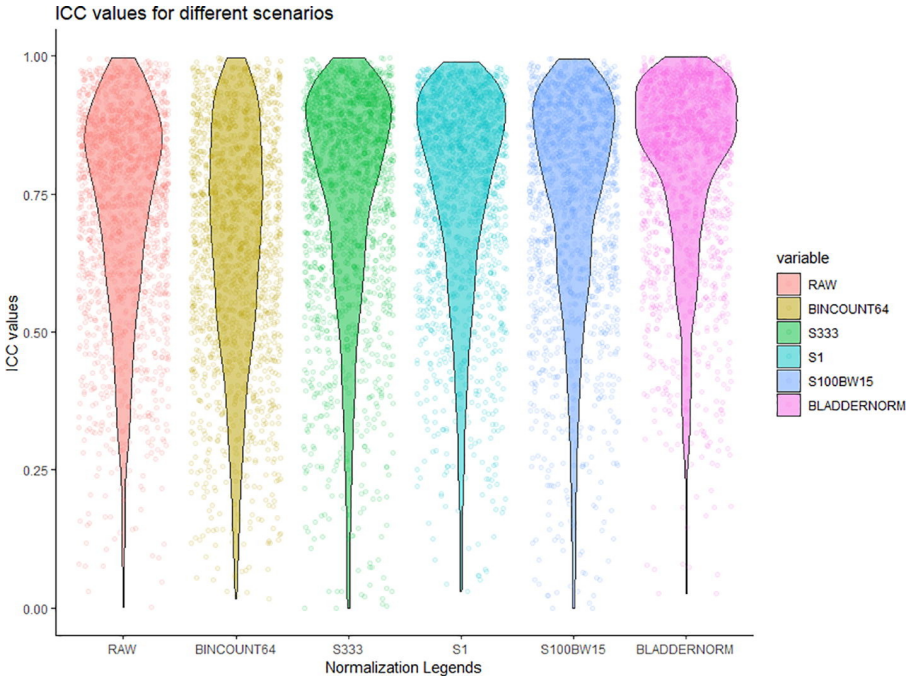


Figure 6.4: Box plots showing intraclass correlation coefficient (ICC) values for all the features with respect to inter-observer variability, for all the approaches and for features extracted directly from raw intensities. The box boundaries represent the 25th–75th percentile range, the middle horizontal line indicates the 50th percentile, open circles represent outliers, and error bars represent the maximum and minimum values excluding the outliers.

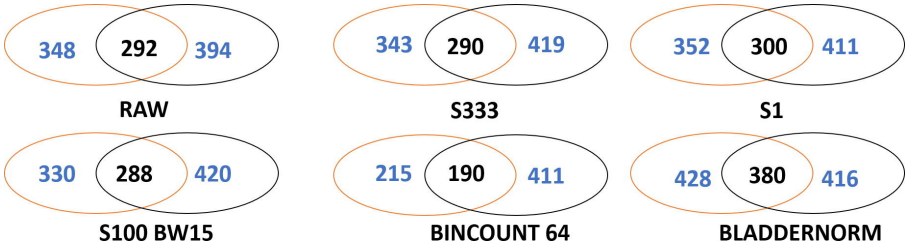


Figure 6.5: Venn diagram combining the reproducibility analysis (ICC≥0.75) in orange, with correlations between radiomic features and tumour volume ($\rho < 0.6$) in gray.

diagram summarizing the features with at least good reproducibility (ICC≥0.75) and a weak correlation with tumour volume ($\rho < 0.6$). Of note, 27% (143/552, range 25–30%) of the features on average showed strong correlations with tumour volume ($\rho > 0.6$). Urine-based approach resulted in the largest number of reproducible features with weak correlation with tumour volume (n=380), while the fixed-bin count approach minimized the reproducible feature count (n=190). Between these two sets, n=90 features overlapped. The pairwise chi-squared test showed statistically significant differences between the urine-based and the other approaches: S333 (n=290, $p<0.001$); S1 (n=300, $p<0.001$); S100BW15 (n=288, $p<0.001$); RAW (n=292, $p<0.001$).

6.4 DISCUSSION

6.4.1 Overall summary and comparison to prior studies

In this study we proposed a methodology for evaluating features' reproducibilities with respect to interobserver variability and image preprocessing (normalization and quantization) in ADC maps of cervix cancer patients. Despite the relatively small sample size, our work is the first study to propose a strong methodology to assess the robustness of radiomic features in ADC maps of cervical cancer patients, with

the aim of extending the effort of harmonization carried by the IBSI for radiomics in PET and CT. At the best of our knowledge, prior studies focused mainly on repeatability of radiomic features for prostate [22], brain [25] and/or dedicate phantoms [1] for T1- and T2-weighted MR images, except for the prostate study by Schwier et al. which also examined ADC maps. Our results are in line with the Schwier study, which stated that: (a) digital image pre-processing prior to features extraction sensitively changed features values; and (b) normalization prior to features extraction improved the features repeatabilities. However, it is important to note that Schwier reported opposite results for point (b) when looking at prostate T2 maps. Therefore, it is important to tune the optimal strategy according to the particular sequence considered. Dependence of shape features to inter-observer variability Since shape features measure morphological and topological properties of the GTV, features such as elongation, flatness or sphericity are potentially more prone to inter-observer variability. Conversely, a subset of shape features strongly correlated with tumour volume (ρ¿0.6) have excellent reproducibility, including tumour volume which is the most reproducible feature. Reducing the inter-observer variability between clinicians via introducing semi or fully automated algorithms for contouring might improve the reproducibility of the shape features that are poorly correlated with tumour volume. However, semi and automated contouring for MRI, and specifically for ADC maps, is challenging due to the poorer signal to noise ratio compared for example to CT.

6.4.2 Normalization

ADC acquisition parameters are particularly variable. Magnetic field strengths, image geometric features (e.g. field-of-view, spatial resolution, slice thickness), RF coils, b-values, diffusion-sensitizing gradient timings, post-processing filters and in-line ADC processing algorithms, and corrections for gradient non-linearity can present large variations between institutions [19]. To reduce

the above-mentioned effects two solutions can be adopted: (a) homogenization of acquisition protocols across different institutions; or (b) digital pre-processing of acquired images prior to features extraction to reduce as much as possible systematic biases. The first methodology has been proposed by several initiatives such as the QIBA (Quantitative Imaging Biomarker Alliance), including design of an ADC standardization phantom which demonstrated that inter-institutional ADC variability remains around 5% under ideal conditions. However, guidelines for radiomics calculation still are not clear, and any interventions which improve variability should impact radiomics reproducibility favorably. The second approach considers image operations like normalization, that may reduce the above-mentioned effects. This approach has been suggested by the IBSI with guidelines for CT [27]. In this view, our work is the natural extension for MRI. In principle, there is no strict recipe for choosing a normalization method. In our work, we applied several techniques previously presented in the literature, including use of urine as an internal ADC reference. We compared the impact of different normalization techniques with respect to inter-observer variability. As Fig. 6.4, Fig. 6.5 show, urine-based normalization provided a more stable strategy for feature reproducibility than approaches which directly extracted features from non-normalized ADC maps. This result is consistent with a prior study which showed that ADC normalized to urine was more reliable than non-normalized ADC for estimating the histological grade of bladder cancer [5].

6.4.3 Quantization

The computation of textural features requires discretization (binning) of the image intensities within a limited number of gray levels. We decided to use the fixed-bin width approach, since it is the most common strategy applied in published PET and CT studies. In our study, the quantization ranged from 0.01 to 25 bin widths. Considering the recommendations available in the literature [27], the approach S100BW5

was discarded because it produced a very small number of bin counts for all the patients, which might make textural features more sensitive to noise. We suggest using a smaller bin width when applying normalizations, in order to guarantee enough counts to preserve signal fidelity. For example, the urine-based normalization provided a bin count of 90 using a bin width of 0.01. An alternative approach is the fixed-bin count method, where relationship between image intensity and physiological meaning (if any) is broken [27]. It introduces a normalizing effect which may be beneficial when intensity units are arbitrary (e.g. for T1 or T2-weighted images), as suggested in [27], although ADC maps have defined units of mm2/s. In this study, features extracted without prior normalization and with a fixed-bin count of 64 were the least reproducible (Fig. 6.2, 6.4), and even less so than features extracted from non-normalized ADC maps. Nonetheless, these results may be modality- and/or disease-dependent and might not be transferable. Furthermore, the value chosen for the fixed-bin count was taken from the PET and CT literature, and might not be the optimal solution for ADC. Dependencies with tumour volume We used ρ¡0.6 (between each radiomic feature and the tumour volume) as the threshold to define a feature as 'volume-independent', despite not being able to find an optimal cut-off in the literature. The selected cut-off eliminates features with strong to moderate correlation with tumour volume, without being so restrictive. On average, 26% (range 25–30%) of the features were found to be volume-dependent. In general, applying normalization reduced the number of volume-dependent features. Filtered features did not show significant differences in terms of volume-dependence, with the only exception of gradient filtering. It is worth noting that some of the most stable features (ICC$\geq$0.9), were highly correlated ($\rho \geq$0.9) with tumour volume. Several limitations need to be highlighted. The relatively small sample size, compared to the largest number of computed features (and then the elevated number of degrees of freedom), might have influenced the statistical significance of the results. Additional data would have enhanced the sub-analyses, especially considering more configurations of b-values. The urine-based normalization should be tested in exter-

nal datasets. In fact, to test the robustness and the generalizability of the proposed methodology, the need of additional external datasets is mandatory. In particular, the presented methodology needs to be tested with respect to different acquisition protocols. Further experiments should include evaluating inter-observer variability for the delineation of this ROI. Additional studies are needed to verify the prognostic power of reproducible features.

6.5 CONCLUSION

We highlighted the importance of image normalization and quantization before feature extraction from ADC maps, and emphasize an urgent need for harmonization. Based on our results, normalizing using values of the urine in the bladder led to most reproducible features compared to no-normalization or extracting features using a fixed bin count approach.

Bibliography

[1] Bettina Baeßler, Kilian Weiss, and Daniel Pinto Dos Santos. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Investigative Radiology*, November 2018.

[2] John J. Bartko. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, 19(1):3–11, August 1966.

[3] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, November 2018.

[4] G. Collewet, M. Strzelecki, and F. Mariette. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magnetic Resonance Imaging*, 22(1):81–91, January 2004.

[5] Betty J. Feir-Walsh and Larry E. Toothaker. An empirical comparison of the anova f-test, normal scores test and kruskal-wallis test under violation of assumptions. *Educational and Psychological Measurement*, 34(4):789–799, 1974.

[6] Sandra Fiset, Mattea L. Welch, Jessica Weiss, Melania Pintilie, Jessica L. Conway, Michael Milosevic, Anthony Fyles, Alberto Traverso, David Jaffray, Ur Metser, Jason Xie, and Kathy Han. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and Oncology*, 135:107–114, June 2019.

[7] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[8] Adam Gladwish, Michael Milosevic, Anthony Fyles, Jason Xie, Jaydeep Halankar, Ur Metser, Haiyan Jiang, Nathan Becker, Wilfred Levin, Lee Manchul, Warren Foltz, and Kathy Han. Association of Apparent Diffusion Coefficient with Disease Recurrence in Patients with Locally Advanced Cervical Cancer Treated with Radical Chemotherapy and Radiation Therapy. *Radiology*, 279(1):158–166, April 2016.

[9] Adam P. Gladwish, Kathy Han, and Warren D. Foltz. Variation in apparent diffusion coefficient measurements among women with locally advanced cervical cancer. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 117(3):532–535, December 2015.

[10] Steinn Gudmundsson, Thomas Philip Runarsson, and Sven Sigurdsson. Test–retest reliability and feature selection in physiological time series classification. *Computer Methods and Programs in Biomedicine*, 105(1):50–60, January 2012.

[11] Jennifer C. Ho, Pamela K. Allen, Priya R. Bhosale, Gaiane M. Rauch, Clifton D. Fuller, Abdallah S. R. Mohamed, Michael Frumovitz, Anuja Jhingran, and Ann H. Klopp. Diffusion-Weighted Magnetic Resonance Imaging as a Predictor of Outcome in Cervical Cancer After Chemoradiation. *International Journal of Radiation Oncology, Biology, Physics*, 97(3):546–553, 2017.

[12] Amit Jethanandani, Timothy A. Lin, Stefania Volpe, Hesham Elhalawani, Abdallah S. R. Mohamed, Pei Yang, and Clifton D. Fuller. Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. *Frontiers in Oncology*, 8:131, 2018.

[13] Farzad Khalvati, Yucheng Zhang, Phuong H. U. Le, Isha Gujrathi, and Masoom A. Haider. PI-RADS guided discovery radiomics for characterization of prostate lesions with diffusion-weighted MRI. In Horst K. Hahn and Kensaku Mori, editors, *Medical Imaging 2019: Computer-Aided Diagnosis*, page 146, San Diego, United States, March 2019. SPIE.

[14] Terry K. Koo and Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, June 2016.

[15] Sangjune Laurence Lee, Jenny Lee, Tim Craig, Alejandro Berlin, Peter Chung, Cynthia Ménard, and Warren D. Foltz. Changes in apparent diffusion coefficient radiomics features during dose-painted radiotherapy and high dose rate brachytherapy for prostate cancer. *Physics and Imaging in Radiation Oncology*, 9:1–6, January 2019.

[16] Ralph T.H. Leijenaar, Georgi Nalbantov, Sara Carvalho, Wouter J.C. van Elmpt, Esther G.C. Troost, Ronald Boellaard, Hugo J.W.L Aerts, Robert J. Gillies, and Philippe Lambin. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Scientific Reports*, 5(1), September 2015.

[17] François Lucia, Dimitris Visvikis, Marie-Charlotte Desseroit, Omar Miranda, Jean-Pierre Malhaire, Philippe Robin, Olivier Pradier, Mathieu Hatt, and Ulrike Schick. Prediction of outcome using pretreatment 18F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiother-

apy. *European Journal of Nuclear Medicine and Molecular Imaging*, 45(5):768–786, 2018.

[18] François Lucia, Dimitris Visvikis, Martin Vallières, Marie-Charlotte Desseroit, Omar Miranda, Philippe Robin, Pietro Andrea Bonaffini, Joanne Alfieri, Ingrid Masson, Augustin Mervoyer, Caroline Reinhold, Olivier Pradier, Mathieu Hatt, and Ulrike Schick. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(4):864–877, April 2019.

[19] David C. Newitt, Ek T. Tan, Lisa J. Wilmes, Thomas L. Chenevert, John Kornak, Luca Marinelli, and Nola Hylton. Gradient nonlinearity correction to improve apparent diffusion coefficient accuracy and standardization in the american college of radiology imaging network 6698 breast cancer trial: GNC of ADC for Breast Clinical Trials. *Journal of Magnetic Resonance Imaging*, 42(4):908–919, October 2015.

[20] William S Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, December 2009.

[21] Evis Sala, Andrea Rockall, Deepa Rangarajan, and Rahel A. Kubik-Huch. The role of dynamic contrast-enhanced and diffusion weighted magnetic resonance imaging in the female pelvis. *European Journal of Radiology*, 76(3):367–385, December 2010.

[22] Michael Schwier, Joost van Griethuysen, Mark G. Vangel, Steve Pieper, Sharon Peled, Clare M. Tempany, Hugo JWL Aerts, Ron Kikinis, Fiona M. Fennessy, and Andrey Fedorov. Repeatability of Multiparametric Prostate MRI Radiomics Features. *arXiv:1807.06089 [cs, eess]*, July 2018. arXiv: 1807.06089.

[23] Florent Tixier, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textu-

ral feature analysis in 18F-FDG PET. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 53(5):693–700, May 2012.

[24] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[25] Fei Yang, Nesrin Dogan, Radka Stoyanova, and John Chetley Ford. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth. *Physica Medica*, 50:26–36, June 2018.

[26] He Zhang, Yunfei Mao, Xiaojun Chen, Guoqing Wu, Xuefen Liu, Peng Zhang, Yu Bai, Pengcong Lu, Weigen Yao, Yuanyuan Wang, Jinhua Yu, and Guofu Zhang. Magnetic resonance imaging radiomics in categorizing ovarian masses and predicting clinical outcome: a preliminary study. *European Radiology*, April 2019.

[27] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

7

Learning from scanners: Bias reduction and feature correction in radiomics

Adapted from: **"Learning from scanners: Bias reduction and feature correction in radiomics"**. I Zhovannik, J Bussink, A Traverso, Z Shi, P Kalendralis, L Wee, A Dekker, R Fijten, R Monshouwer. Clinical and translational radiation oncology 19, 33-38. (2019). Contribution: second autorship,radiomic analysis, statistical analysis.

Abstract

Radiomics are quantitative features extracted from medical images. Many radiomic features depend not only on tumour properties, but also on non-tumour 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. 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. Sixty-two of 92 radiomic features showed high variance due to the scanner SNR. Of these 62 features, 47 showed at least a factor 2 significant 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. Phantom measurements show that roughly two third of the radiomic features depend on the exposure setting of the scanner. The dependence can be modelled and corrected significantly reducing the variation in feature values with at least a factor of 2. More complex models will likely increase the correctability. Scanner SNR correction will result in more reliable radiomics predictions in NSCLC.

7.1 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 [4]. 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][5], treatment toxicity [1], tumor habitat characterization [9]. Although the radiomics approach shows promising results, different feature definitions, image pre-processing methods, and imaging instruments make cross-institutional learning difficult [6][7][10]. The Image Biomarker Standardization Initiative (IBSI) standardized radiomics mathematical definitions and image pre-processing [13]. 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. 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.

7.2 MATERIAL AND METHODS

7.2.1 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 7.1. For radiomics extraction, we used open-source pyradiomics 2.1.2 software with 25 HU binning and no re-sampling [11].

7.2.2 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

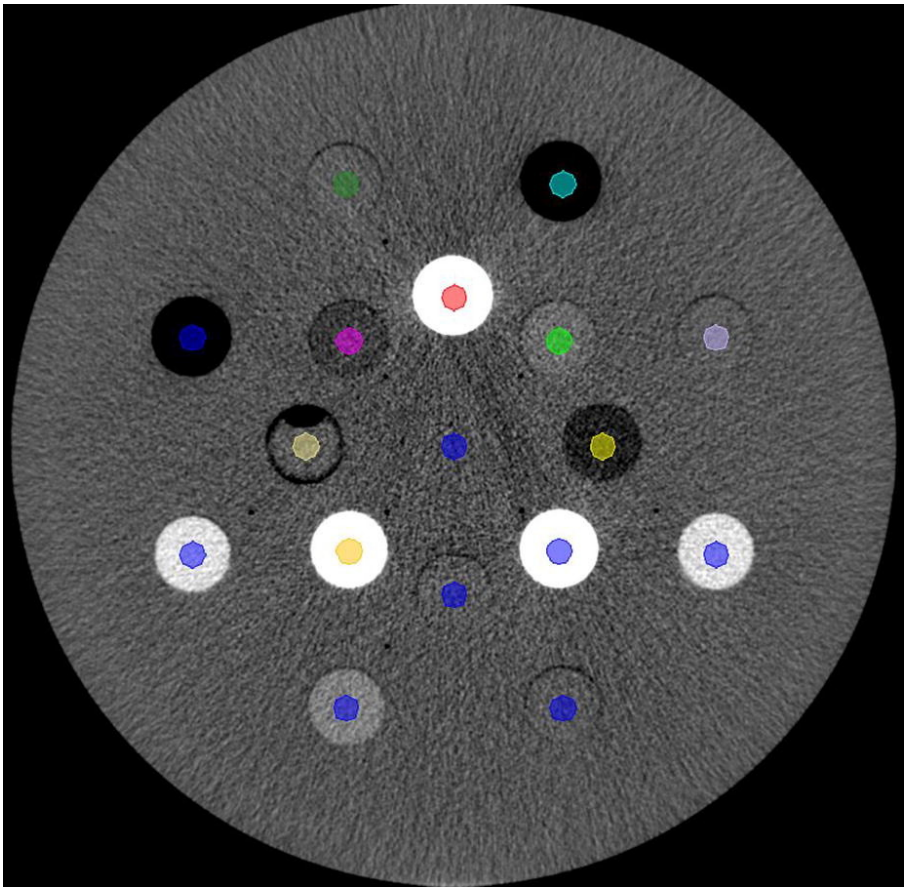


Figure 7.1: Gammex phantom configuration with cylindrical delineations. The 17 plug descriptions are in the supplementary A.

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

Figure 7.2: Formula 1

features were extracted from the gross tumor volume (GTV) of the primary tumor with the same pyradiomics extraction settings as in the phantom set.

7.2.3 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 7.4), 2) Subtracting TRV from radiomic feature's data, isolating the SNR trend in the data (figure 7.4), 3) fitting the correction function (figure 7.4), 4) Correcting the raw data (figure 7.4). 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 7.2, 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 [8].

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

Figure 7.3: Formula 2

7.2.4 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 7.3. To derive the score, we used TRV-shifted data (figure 7.4). The correctability score is a ratio: the numerator describes variability due to exposure (variance in means), the denominator describes intrinsic repeatability variance; ΔRF 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 ΔRF 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.

7.2.5 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 7.5):

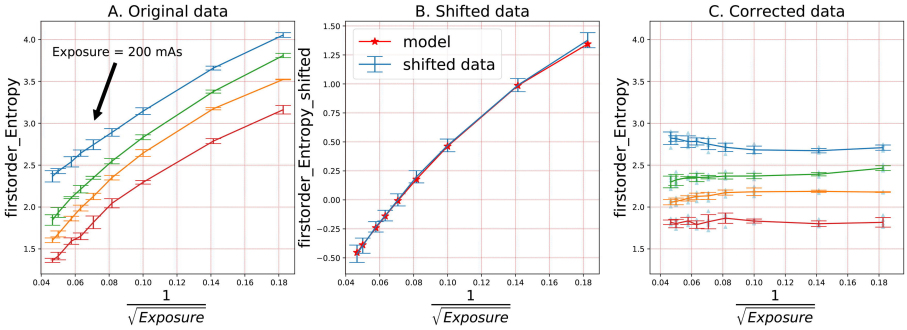


Figure 7.4: Radiomics correction model in three steps: 1) shift original data to 0 with TRV, 2) fit the model using the shifted data, 3) correct the original data using the model.

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

Figure 7.5: Formula 3

7.2.6 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. 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 7.6 and evaluates how large the difference between feature values in original ($RF_{original}$) and SNR-influenced ($RF_{-160mAs}$) images is if compared to the interquartile range in the feature distribution ($IQR_{0.75-0.25}(RF_{original})$):

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

Figure 7.6: Formula 4

7.3 RESULTS

7.3.1 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 7.7 shows CS for each selected radiomic feature as the step blue line.

7.3.2 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 7.7) 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 7.7. Interestingly, ROIs 9 and 15 (low density plugs, $<$ -600 HU) have low correctability, on the other hand, ROIs 2 and 8

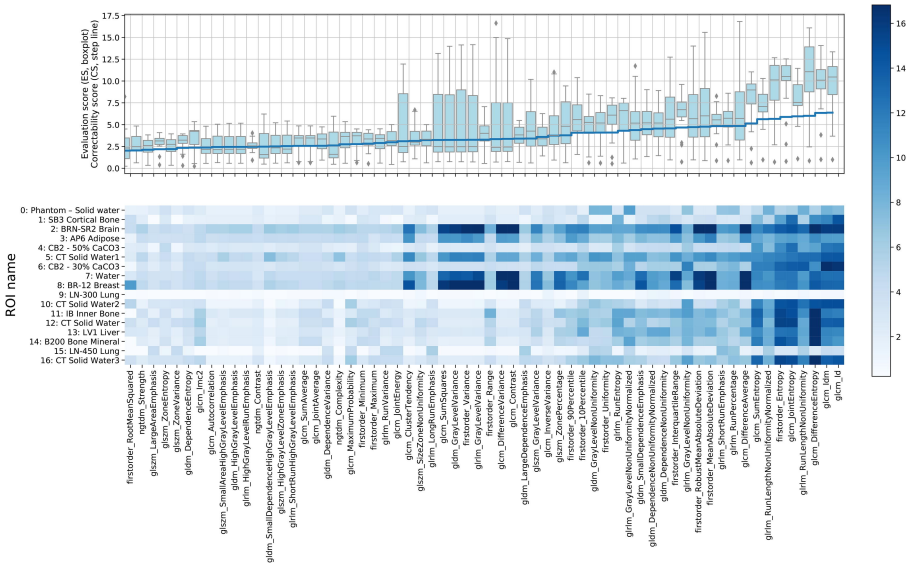


Figure 7.7: Correctability (2) and Evaluation (3) scores for the selected 62 radiomic features and 17 ROIs. The color bar represents the evaluation score (ES): the darker, the larger reduction in standard deviation was obtained.

(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.

7.3.3 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 7.8 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. 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. 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.

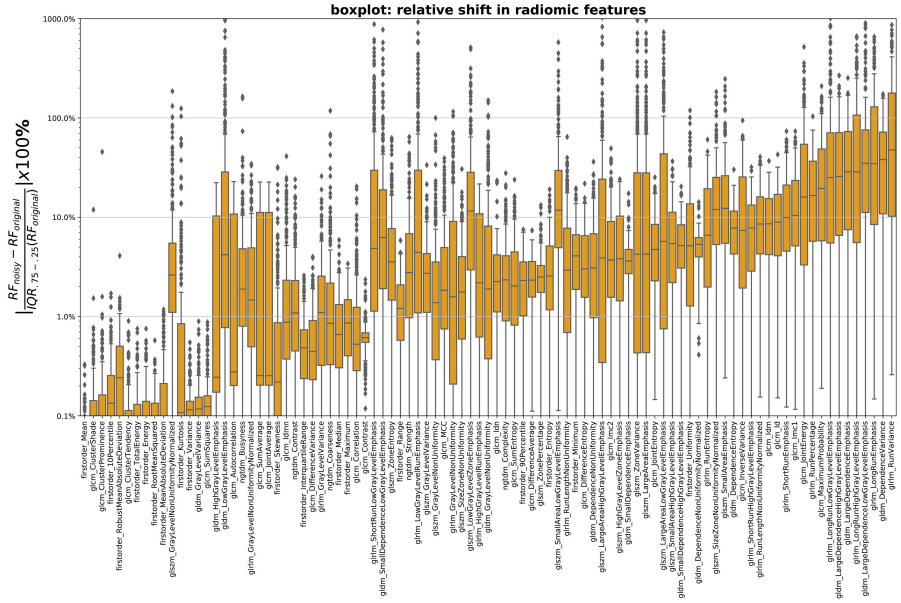


Figure 7.8: Relative shift (4) in radiomic features (in ascending order) versus feature names while applying Poisson noise (equivalent to decreasing scanner SNR, mAs) in the images of the NSCLC cohort.

7.4 DISCUSSION

7.4.1 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).

7.4.2 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.

7.4.3 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 [7]. 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 [12]. 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 [7]. 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 7.8). 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 [3]. Our finding is however in contrast with the conclusion of Mackin et al. [7]. 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.

7.5 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.

Bibliography

[1] Hamid Abdollahi, Seied Rabi Mahdavi, Bahram Mofid, Mohsen Bakhshandeh, Abolfazl Razzaghdoust, Afshin Saadipoor, and Kiarash Tanha. Rectal wall MRI radiomics in prostate cancer patients: prediction of and correlation with early rectal toxicity. *International Journal of Radiation Biology*, 94(9):829–837, 2018.

[2] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[3] Xenia Fave, Molly Cook, Amy Frederick, Lifei Zhang, Jinzhong Yang, David Fried, Francesco Stingo, and Laurence Court. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 44:54–61, September 2015.

[4] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Ra-

diomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[5] Ahmed Hosny, Chintan Parmar, Thibaud P. Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J. Gillies, Raymond H. Mak, and Hugo J. W. L. Aerts. Deep learning for lung cancer prognostication: A retrospective multicohort radiomics study. *PLOS Medicine*, 15(11):e1002711, November 2018.

[6] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, and Laurence Court. Measuring Computed Tomography Scanner Variability of Radiomics Features:. *Investigative Radiology*, 50(11):757–765, November 2015.

[7] Dennis Mackin, Rachel Ger, Cristina Dodge, Xenia Fave, Pai-Chun Chi, Lifei Zhang, Jinzhong Yang, Steve Bache, Charles Dodge, A. Kyle Jones, and Laurence Court. Effect of tube current on computed tomography radiomic features. *Scientific Reports*, 8(1):2354, dec 2018.

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[9] E. Sala, E. Mema, Y. Himoto, H. Veeraraghavan, J.D. Brenton, A. Snyder, B. Weigelt, and H.A. Vargas. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical Radiology*, 72(1):3–10, jan 2017.

[10] Muhammad Shafiq-ul Hassan, Geoffrey G. Zhang, Dylan C. Hunt, Kujtim Latifi, Ghanim Ullah, Robert J. Gillies, and Eduardo G. Moros. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *Journal of Medical Imaging*, 5(01):1, December 2017.

[11] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[12] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[13] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

8

Machine learning helps identifying volume-confounding effects in radiomics

Adapted from: **"Machine learning helps identifying volume-confounding effects in radiomics"**. A Traverso, M Kazmierski, I Zhovannik, M Welch, L Wee, D Jaffray, A Dekker, A Hope. Physica Medica 71, 24-30. (2020).

Abstract

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. 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. Over 80% of the features had large pairwise correlations. Nearly 30% of the features presented strong correlations with tumour 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. 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.

8.1 INTRODUCTION

Radiomics, the automated extraction of quantitative descriptors from medical images, has demonstrated promising prognostic and predictive results for overall survival [7], distant metastases [9] and cancer biology [16]. 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 overfitting, 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 [3]. 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 [8]. 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 [4], 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.

8.2 METHODS

8.2.1 Datasets

We used two retrospective public data sets for the analysis:

- 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. The dataset is the same used in Aerts et al. [1].

- 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 tumour 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. 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.

8.2.2 Radiomic features extraction

We used the open source software PyRadiomics v2.2.0 [12] to extract imaging features from each GTV. Pyrex (Link here), an extension of PyRadiomics was used to handle DICOM RTSTRUCT files as input, by generating a binary segmentation mask from the contour data [10].

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.

8.2.3 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.

8.2.4 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 [13]. 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)[13].

8.2.5 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 correlated with the first principal component, it is likely to be the true source of variation.

8.2.6 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 [6]. 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 8.1 (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 biased
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

Figure 8.1: Suggested workflow for radiomics signature developments that incorporates safeguards.

8.3 RESULTS

8.3.1 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 $|\rho| \geq 0.9$ (Fig. 8.2). Furthermore, nearly 30% showed correlation with tumour volume greater than 0.75 (Fig 8.2) . A similar correlation was observed in the HN1 dataset. Supplementary Table S1 lists all the radiomic features that presented a Spearman correlation $|\rho| \geq 0.8$ with GTV.

8.3.2 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. 8.3a). The difference in tumour volume distribution between clusters was highly significant (permutation P $<$ .001, Fig. 8.3a). Removing features moderately correlated with volume (with Spearman $|\rho| > 0.6$) still allowed for cluster separation by OS (P $<$ 10–6, Fig. 8.3b); however, the volume difference between clusters remained highly significant (P $<$ .001, Fig. 8.3b). Using only volume-independent features ($|\rho| < 0.1$) the groups could not be separated by survival (P = .8, Fig. 8.3c) or tumour volume (P

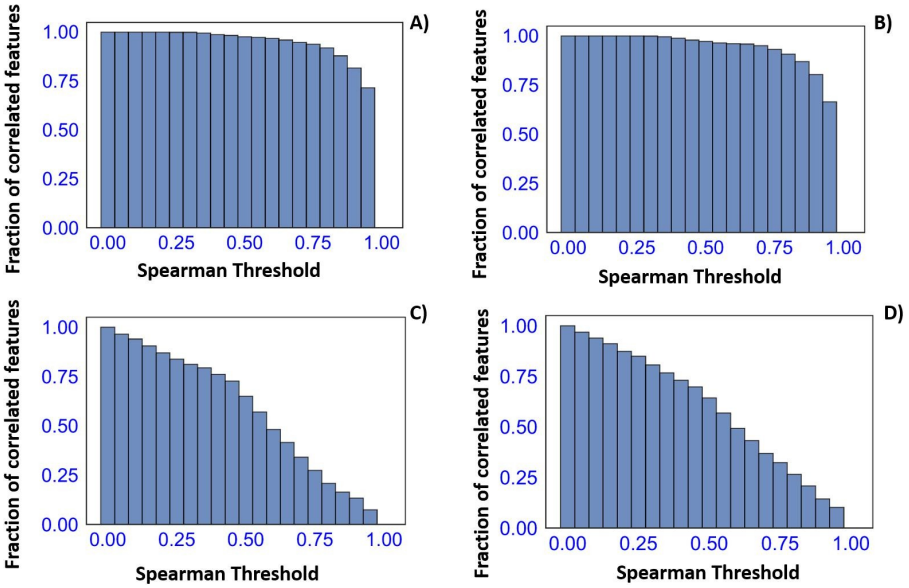


Figure 8.2: Proportion of correlated features as a function of Spearman rank correlation threshold. (a) shows the proportion of features with pairwise correlation greater than the threshold value in the lung dataset. Percentage of features correlated with volume at a given threshold is shown in (c). The results were similar in HN1 (b, d).

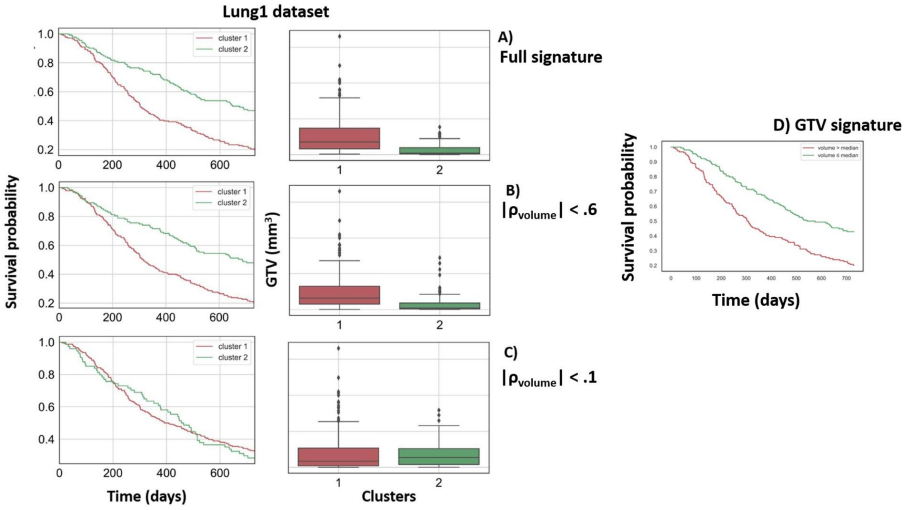


Figure 8.3: Kaplan-Meier OS estimates and volume distributions for each cluster identified in the NSCLC dataset. (a) all original and wavelet features, (b) features moderately correlated with volume (defined as Spearman $|\rho| < 0.6$ with volume), (c) features not correlated with volume ($|\rho| < 0.1$), d) GTV signature.

= .9, Fig. 8.3c). Groups could be separated by survival 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. 8.3d. 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 ($|\rho| < 0.1$) 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 $\rho = 0.78$, Fig. 8.4a). 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 ($|\rho| < 0.6$) correlated features were used (correlation with volume: $\rho = -0.37$ for PC 1 and $\rho = 0.79$ for PC 2, Fig. 8.4b), explaining the

significance between-cluster differences in tumour volume. Finally, there was no volume-dominant effect in features independent ($|\rho| < 0.1$) from volume (ρ = 0.01 for PC 1 and ρ = 0.05 for PC 2, Fig. 8.4c). 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. 8.4a 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 for PC 2, 8.4b 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. 8.4c right).

8.3.3 Feature selection and modelling

Fig. 8.5 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 and neck squamous carcinomas for OS. 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.

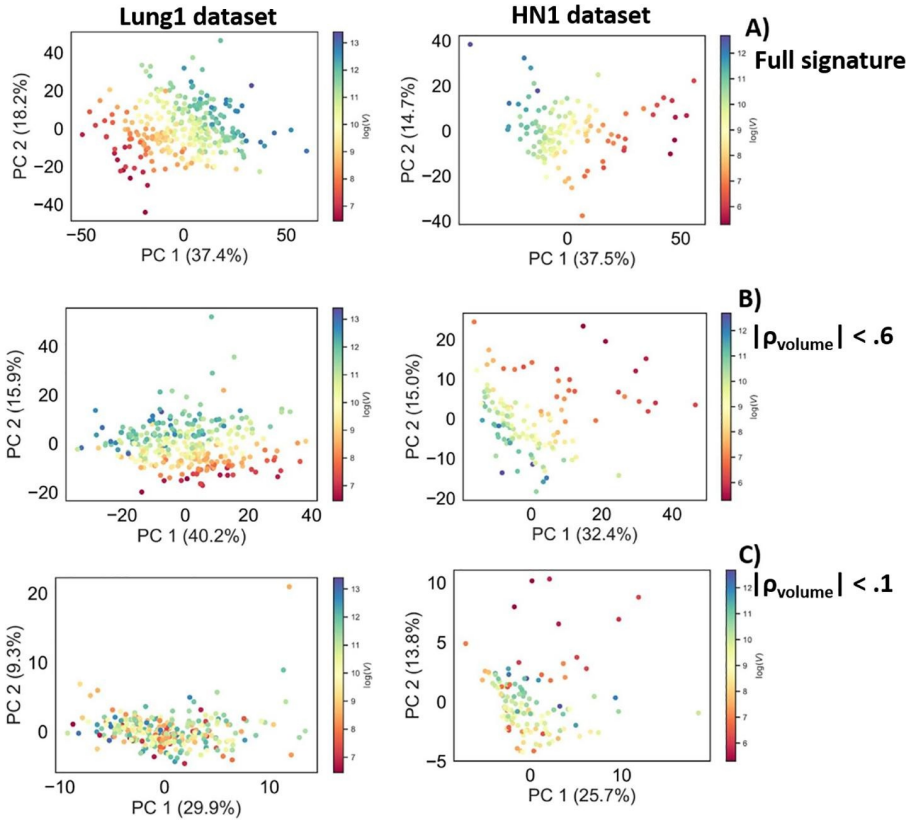


Figure 8.4: Principal component analysis of (a) full feature signatures, (b) features moderately correlated with volume ($|\rho| < 0.6$), (c) volume-independent features ($|\rho| < 0.1$) in lung and head and neck datasets. The data is shown projected on the first 2 principal components and the proportion of variance explained by each component is indicated. Colours correspond to tumour volume. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

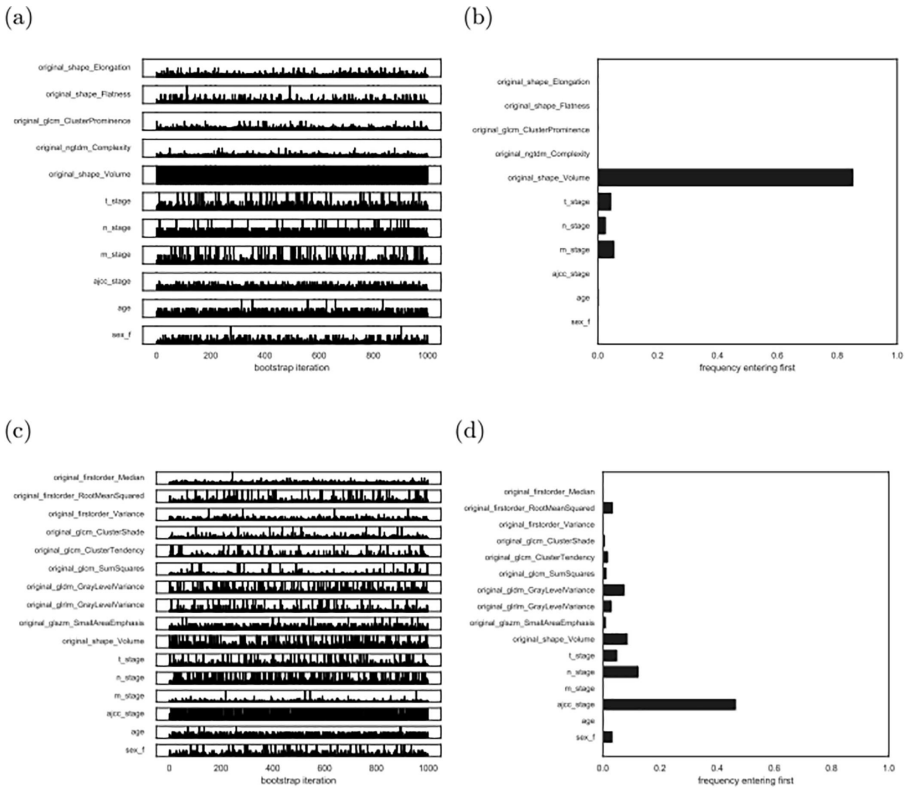


Figure 8.5: Bootstrap-based evaluation of predictive power and stability of imaging and clinical features. (a) and (c) show the order of each feature entering the model across 1000 resampling iterations in Lung1 and HN1, respectively. The height of the bar is inversely proportional to the order of selection (therefore, filled bar indicates higher importance). (b) and (d) show the frequency of each feature entering the model first in both datasets.

8.4 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. 8.2c 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 [15]. 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 explana-

tory 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. 8.3). 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. 8.3). 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) [14]. 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 fur-

ther 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 [15], 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. 8.5a 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. We provide safeguarding recommendations for signature developments in radiomics studies that build upon Welch et al. [15]: 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 8.1 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 8.1 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.

8.5 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.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] Mostafa Analoui, Joseph D Bronzino, and Donald R Peterson. *Medical imaging: principles and practices*. CRC Press, 2012.

[3] Anastasia Chalkidou, Michael J. O'Doherty, and Paul K. Marsden. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PLOS ONE*, 10(5):e0124165, May 2015.

[4] Michele De Palma and Douglas Hanahan. The biology of personalized cancer medicine: Facing individual complexities underlying hallmark capabilities. *Molecular Oncology*, 6(2):111–127, April 2012.

[5] Daniel Dejaco, Teresa Steinbichler, Volker H. Schartinger, Natalie Fischer, Maria Anegg, Jozsef Dudas, Andrea Posch, Gerlig Wid-

mann, and Herbert Riechelmann. Prognostic value of tumor volume in patients with head and neck squamous cell carcinoma treated with primary surgery. *Head & Neck*, 40(4):728–739, 2018.

[6] Issam El Naqa, Jeffrey Bradley, Angel I. Blanco, Patricia E. Lindsay, Milos Vicic, Andrew Hope, and Joseph O. Deasy. Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors. *International Journal of Radiation Oncology*Biology*Physics*, 64(4):1275–1286, March 2006.

[7] Elin Evans and John Staffurth. Principles of cancer treatment by radiotherapy. *Surgery (Oxford)*, 36(3):111–116, March 2018.

[8] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, May 2016.

[9] David Jaffray, Patrick Kupelian, Toufik Djemil, and Roger M Macklis. Review of image-guided radiation therapy. *Expert Review of Anticancer Therapy*, 7(1):89–103, January 2007.

[10] Zhenwei Shi, Alberto Traverso, Johan Soest, Andre Dekker, and Leonard Wee. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). *Medical Physics*, page mp.13844, October 2019.

[11] Tomoyoshi Takenaka, Koji Yamazaki, Naoko Miura, Ryo Mori, and Sadanori Takeo. The Prognostic Impact of Tumor Volume in Patients with Clinical Stage IA Non-Small Cell Lung Cancer. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 11(7):1074–1080, 2016.

[12] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode

the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[13] Sandro Vega-Pons and José Ruiz-Shulcloper. A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, May 2011.

[14] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to Use t-SNE Effectively. *Distill*, 1(10):10.23915/distill.00002, October 2016.

[15] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[16] Dennis Winkel, Gijsbert H. Bol, Petra S. Kroon, Bram van Asselen, Sara S. Hackett, Anita M. Werensteijn-Honingh, Martijn P.W. Intven, Wietse S.C. Eppinga, Rob H.N. Tijssen, Linda G.W. Kerkmeijer, Hans C.J. de Boer, Stella Mook, Gert J. Meijer, Jochem Hes, Mirjam Willemsen-Bosman, Eline N. de Groot-van Breugel, Ina M. Jürgenliemk-Schulz, and Bas W. Raaymakers. Adaptive radiotherapy: The Elekta Unity MR-linac concept. *Clinical and Translational Radiation Oncology*, 18:54–59, September 2019.

9

User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions

Adapted from: **"User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions"**. ML Welch, C McIntosh, A McNiven, S Hui Huang, BB Zhang, L Wee, A Traverso, B O'Sullivan, F Hoebers, A Dekker, DA Jaffray. Physica Medica 70, 145-152. (2020). Contribution: radiomic analysis, machine learning pipeline, data analysis, manuscript writing.

Abstract

Precision cancer medicine is dependent on accurate prediction of disease and treatment outcome, requiring integration of clinical, imaging and interventional knowledge. User controlled pipelines are capable of feature integration with varied levels of human interaction. In this work we present two pipelines designed to combine clinical, radiomic (quantified imaging), and RTx-omic (quantified radiation therapy (RT) plan) information for prediction of locoregional failure (LRF) in head and neck cancer (HN). Pipelines were designed to extract information and model patient outcomes based on clinical features, computed tomography (CT) imaging, and planned RT dose volumes. We predict HN LRF using: 1) a highly user-driven pipeline that leverages modular design and machine learning for feature extraction and model development; and 2) a pipeline with minimal user input that utilizes deep learning convolutional neural networks to extract and combine CT imaging, RT dose and clinical features for model development. Clinical features with logistic regression in our highly user-driven pipeline had the highest precision recall area under the curve (PR-AUC) of 0.66 (0.33–0.93), where a PR-AUC = 0.11 is considered random. Our work demonstrates the potential to aggregate features from multiple specialties for conditional-outcome predictions using pipelines with varied levels of human interaction. Most importantly, our results provide insights into the importance of data curation and quality, as well as user, data and methodology bias awareness as it pertains to result interpretation in user controlled pipelines.

9.1 INTRODUCTION

Prognostics are an important part of cancer care [15] [12] that estimates the risk of an individual's outcome based on multiple variables (e.g. tumour, patient and environmental). It aids in treatment decisions and differs from aetiological research where the goal is to explain whether an outcome can be attributed to a specific risk factor [33]. In addition to the traditional prognostic factors mentioned above, integration of treatment information to form a treatment specific-conditional prognosis is highly beneficial. We define treatment specific-conditional prognosis as the prediction of a treatment's effect, if administered as intended, on the patient's outcome [6]. The volume and variety of features available for inclusion in these types of predictions is expanding rapidly. This is in part the result of a hypothesis in cancer management that by analysing an extensive set of features that encompass the nuances of disease processes and treatments that we can achieve "Precision Medicine" [41][50][11]. Features can vary from highly cited and tested measurements designed to probe and describe the nuances of both a patient and corresponding disease [7][49][27][30], to more experimental imaging, tissue and treatment features that describe the activity of tumours before and during treatment [21]. These features can even include those generated through automation that explore disease outcome correlations with image signal values (i.e. radiomics) [47][14][48]. Quantified interventional features have the potential to be combined with these features for a truly comprehensive view of a patient's treatment-influenced course of disease. In head and neck (HN) radiation therapy (RT) it is known that the dose fractionation and quality of an RT plan can impact overall and locoregional failure (LRF) free survival [35][4][40]. Dose volume histograms (DVH) are calculated to evaluate a RT plan based on RT dose delivered to volumes of tissue [43][29]. Metrics calculated using the DVHs are known predictors of both toxicity and outcome [19], but lack spatial dose information that is indicative of a patient's disease and surrounding intrinsic anatomical

variations. Recent research quantifying spatial dose distributions for patients has found utility in toxicity prediction [22][32] and may similarly benefit treatment-specific conditional prognosis outcome prediction. However, as the number of prognostic factors that we consider increases, our methods for knowledge integration must change. The agglomeration of diverse features represents a movement towards precision medicine, but also the utilization of big data in cancer care [41][50][34]. Approaches and pipelines for big data feature integration would provide flexible solutions that could drive data exploration and clinical decision support systems forward; an idea that was demonstrated by Mobadersany et al. [31] who combined deep learning and traditional user defined features. Additionally, these 'big machines' can be thought of as user-controlled pipelines requiring a spectrum of user interaction and assurance while evaluating intermediate by products and tuning various operating parameters. In this work, we build two generalized feature integration pipelines for cancer treatment specific conditional prognosis; one of which is a highly user-driven process, while the other is substantially automated. Both pipelines leverage clinical, radiomic, and interventional features extracted from personalized RT plans (henceforth referred to as RTx-omic features). As a proof of concept, we applied our pipelines to a HN dataset to determine how the conditional-prognostic performance of clinical features may be impacted by RTx-omic and radiomic features during LRF prediction, and whether conclusions could be drawn regarding the influence of user bias on user-controlled pipelines.

9.2 METHODS

Our methods are designed to build and explore two pipelines for patient information integration and outcome prediction: 1) A machine learning pipeline that is inherently user-driven. Features are explicitly defined and informed by prior-knowledge, and classification models are finalized as a separate step in the pipeline; 2) a deep learning

pipeline that is more automated and allows spontaneous emergent features to be learned by the machine, while simultaneously developing a classifier. Both pipelines explored the impact of clinical, radiomic and RTx-omic features on outcome predictions. In this study we use LRF prediction at three years in HN cancer as our case study. Predictions are performed using different modelling methods, each of which has specific benefits to our research question. This section details the data curation, and pipelines used for our analysis.

9.2.1 Data curation and preparation

We used a single dataset from the Princess Margaret Cancer Centre with institutional research board approval. The dataset contained planning computed tomography (CT) DICOM images, DICOM RT Structures, DICOM RT Dose, and clinical variables for 190 patients. Gross tumour volumes (GTV) in the DICOM RT Structure file were contoured by radiation oncologists (experience levels ranging between 5 and 30 years) for intensity modulated radiation therapy (IMRT) treatment based on clinical-radiological evidence of disease extent. Often during contouring, simulation magnetic resonance imaging (MRI) was fused with the planning CT to aid in target delineation. Additionally, HN Radiation Oncology Quality Assurance Rounds occurred weekly for the opportunity to peer-review RT target volumes, including the GTV and clinical target volumes (CTV). Additional patient details can be found in Table 1 of the Supplementary Material. The inclusion criteria for this study was an oropharynx disease site, squamous cell carcinoma pathology, 70 Gy prescribed dose in 35 fractions to the primary GTV, and full delivery of prescribed dose. Application of inclusion criteria reduced our dataset from 190 patients to 160 patients with 18 LRF events at three years. This resulted in an imbalanced dataset with an event rate of 11%; a challenging problem for most modelling methods, but one we believe can be modelled using appropriate disease, image and treatment descriptors. Furthermore, for our highly user driven

pipeline (hence forth referred to as Machine Learning Pipeline and described below), only patients who did not have dental artifacts (DA) were included. This was to safeguard our radiomic features against spurious image signals [48] and reduced the dataset further to 64 patients with 7 LRF events. For our more automated pipeline (hence forth referred to as the Deep Learning Pipeline and described below) all 160 curated patients were used with the assumption that the convolutional neural network (CNN) would learn to distinguish between important and irrelevant machine generated features regardless of the DA status of the image.

9.2.2 Machine learning pipeline

Our Machine Learning Pipeline (MLP) (Fig. 9.1) was designed to allow a researcher to have control over all aspects of feature definition, feature space reduction, and model building, and validation. This is typical of the traditional clinical modelling or radiomics pipelines where features are defined based on prior knowledge of the disease, or hand-engineered and extracted from images.

9.2.3 Patient-specific features

The complexity of a patient's disease generates a variety of characteristics that may be of benefit when determining a treatment specific conditional-prognosis. In our MLP we aim to describe the disease using a variety of features that are known predictors of HN patient LRF (clinical features), as well as more exploratory features quantifying a patient's planning CT signal (radiomic features) and their personalized RT treatment plan designed based on their intrinsic anatomical variants (RTx-omics features). Following is a description of the features found in the different feature classes. It should be noted that due to the small number of events found in our dataset that our MLP suffers from the 'curse of dimensionality' – this is mitigated using feature space reduction which is described as a later step.

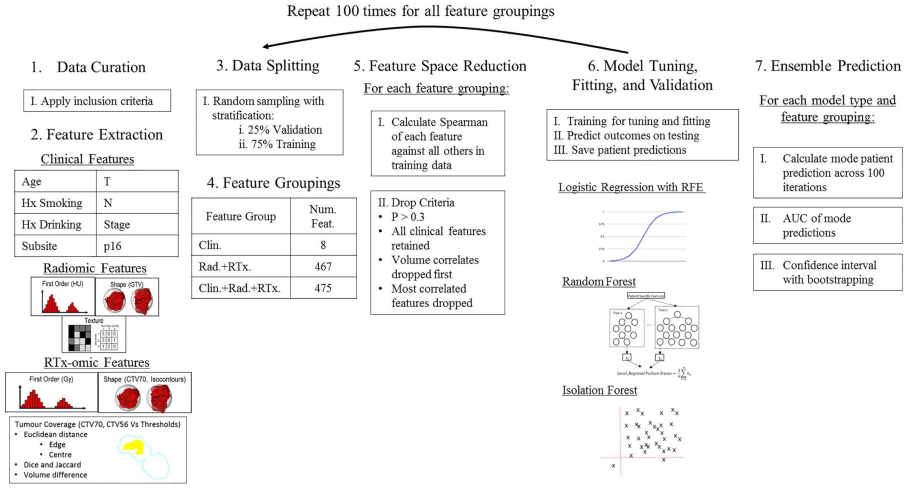


Figure 9.1: Machine Learning Pipeline for generalized feature analysis and outcome prediction in HN patients. Step 1: application of inclusion and exclusion criteria; Step 2: extraction of generalized features – clinical, radiomic and RTx-omic; Step 3: random sampling of patients into training and validation datasets; Step 4: feature grouping based on goal of determining added benefit of radiomic and RTx-omic features; Step 5: reduction of feature set based on spearman rank values calculated within a specific feature grouping; Step 6: tuning, fitting and validating of three different modelling techniques; Step 7: calculation of PR-AUC based on model prediction of patient outcome across 100 iterations of Steps 3–6.

9.2.4 Clinical features

Clinical features for our patients were collected from the Princess Margaret Cancer Center HN Anthology. Patient Age, smoking status, drinking status, disease subsite, T stage, N stage, overall stage, and p16 status were included as clinical features.

9.2.5 Radiomic features

PyRadiomics 2.0 [46] was used to extract radiomic features (n = 99) that quantified the planning CT (in Hounsfield Units, HU) within a patient's GTV. Images were resampled to an isotropic pixel size of 1 mm using BSpline interpolation, and a bin width of 25 was used for texture feature calculation [51]. All features from the first order statistics (n = 18), shape (n = 12) and texture (GLCM (n = 23), GLSZM (n = 16), GLDM (n = 14) and GLRLM (n = 16)) classes were extracted. For details on feature equations please see the extensive PyRadiomics documentation.

9.2.6 RTx-omic features

RTx-omic features were extracted using PyRadiomics 2.0 and a custom PyRadiomics module designed to quantify relationships between two ROIs. First order statistical features were extracted from the planned dose volume, where the voxels represent planned RT dose (Gy), instead of HU as was quantified with the radiomic features. These features were extracted from the GTV, CTV70 (clinical target volume at 70 Gy), CTV56 (elective clinical target volume at 56 Gy) and isocontours at 95 and 100% of 70, 63 and 56 Gy, which were generated by thresholding the planned dose volume. Shape features for the isocontours and CTV70 were also calculated. Tumour coverage was quantified using the custom PyRadiomics module. The module was developed to calculate the Euclidean distance between two ROI edges and centres, as well as Dice and Jaccard metrics, and volume differences. These

metrics were calculated to compare all isocontours against CTV70 and CTV56. RTx-omic features were defined in collaboration with a medical physicist, radiation therapist and radiation oncologist.

9.2.7 Ensemble LRF prediction and validation

We explored the impact of radiomic, RTx-omic and clinical feature groups on prediction of LRF at three years using a multi-step process that explored a variety of modelling methods.

9.2.8 Data splitting and feature grouping

Our dataset was split into 75% training and 25% testing sets. The data was randomly sampled and stratified to ensure equal distribution of LRF events in each set; this resulted in the training and testing sets containing 5-6 and 1-2 patients, respectively. Feature groups were combined to explore the added predictive value of radiomic and RTx-omic features on accepted clinical factors. This resulted in the following combinations of features 1) clinical, 2) radiomic and RTx-omic; and 3) clinical, radiomic and RTx-omic. (Step 4 in Fig. 9.1).

9.2.9 Feature space reduction

Each of the three training set feature groupings underwent feature space reduction to decrease the number of correlated features and the chances of overfitting to the training data. Feature space reduction involved calculating the Spearman rank value for each feature against all other features in the feature group of interest. If the Spearman rank value between two features was greater than or equal to 0.3 the features were considered correlated and one of them was dropped/removed. Clinical features were never dropped, since the goal was to determine added value above accepted clinical features, features correlated to volume were dropped first, and if two features

still remained, the feature that was correlated to the most number of other features was dropped (Step 5 in Fig. 9.1).

9.2.10 Model tuning, fitting and validation

After feature space reduction, model tuning, fitting and validation was performed using the training dataset. Three different modeling techniques available in Python's Scikit Learn package [38] were explored: a) logistic regression with recursive feature elimination (LOG) [39] – a highly interpretable method of modeling widely accepted in the clinical environment; b) random forest (RF) [20] – a more complex method aggregating multiple decision trees together to reduce bias and variance; c) isolation forest (IF) [28] – an ensemble of isolation trees designed to detect data anomalies, such as an LRF event in our dataset. Tuning parameters and methods can be found in our Supplementary Materials. After tuning based on the feature grouping of interest, a LOG, RF and IF model were fit to the training data for the same feature group of interest. The fit and tuned LOG, RF and IF models were used to predict the probability of a testing patient experiencing an LRF event, which was saved in an array (Step 6 in Fig. 9.1). Steps 3–6 of Fig. 9.1, Data Splitting and Feature Grouping, Feature Space Reduction, and Model Tuning, Fitting and Validation, were repeated 100 times for different splits of the data.

9.2.11 Ensemble prediction

After fitting 100 models for each of the feature groupings, and each of the modelling types, we performed an ensemble prediction of treatment specific conditional-prognosis for HN patient LRF at three years. Each combination of feature grouping and modelling method had an array where a row represented a patient and a column represented one of the 100 iterations. The average probability of a patient experiencing an LRF event across the 100 iterations was taken to be that patient's probability of experiencing an LRF event. The precision recall

area under the curve (PR-AUC) (described below) for a given feature grouping and modelling method was calculated on the average patient LRF event probability. Confidence intervals (CI)were calculated using bootstrapping.

9.2.12 Deep learning pipeline

In our Deep Learning Pipeline (DLP) a deep learning network was utilized to minimize user influence (Fig. 9.2). This gave the system control over what features to extract and how to combine them in the most beneficial way for LRF prediction. Three deep learning networks (DLNs) were trained: 1) Clinical, 2) Radiomic + RTx-omic, and 3) Clinical + Radiomic + RTx-omic. Patient image, RTDose and contour volumes were used in models 2 and 3.

9.2.13 Data encoding, generation and pre-processing

Clinical data encoding

The categorical clinical features (i.e. smoking status, drinking status, disease subsite, T stage, N stage, overall stage, and p16 status) were one-hot-encoded using the function "OneHotEncoder" from Python's scikit-learn 0.22 package [38] to obtain binary categorizations that are easier for the machine to interpret. Age is a continuous variable and remained unaltered.

Contour volume generation

The three-dimensional (3D) contour volumes were generated by combining each patient's GTV, CTV56 and CTV70 into a single volume, where the intersection of all three regions of interest was denoted by a 1, the intersection of CTV56 and CTV70 was denoted by a 2, and the remaining portion of CTV56 was denoted by a 3.

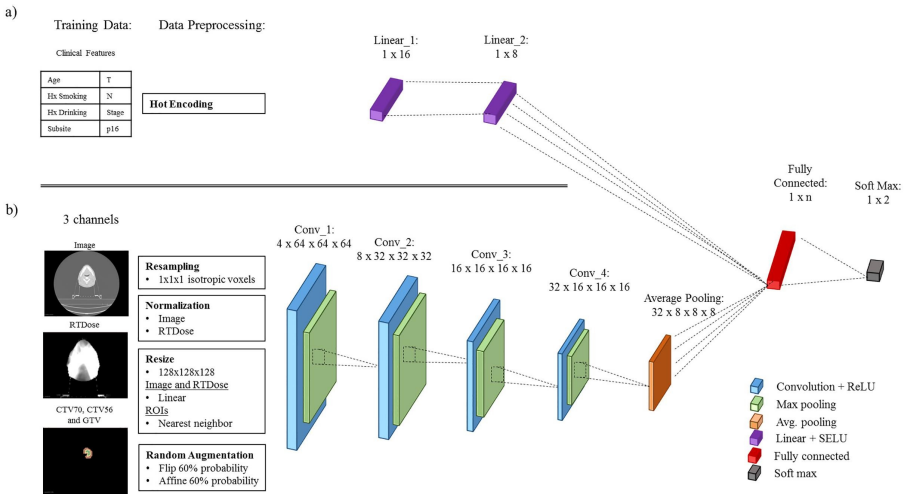


Figure 9.2: Deep Learning Pipeline for generalized feature analysis and outcome prediction in HN patients. a) the features, pre-processing, and linear layers used for our Clinical network. n in the fully connected layer is 8. b) the data, pre-processing steps and CNN layers for our Radiomic + RTxomic network. n in the fully connected layer is 16384. The final network described is the Clinical + Radiomic + RTx-omic network which combines both a) and b). In this network, n in the fully connected layer is 16392.

Image, dose and contour pre-processing

The 3D image, dose and contour volumes were processed prior to usage in CNN training or testing using a multistep procedure:

- Voxels in the CT image, RTDose volume and contour volume were interpolated to isotropic 1 mm3 sizes. SimpleITK's linear resampling image filter was used for the CT image and RTDose, and a nearest neighbour resampling filter was used for the contour volume. This reduced variability in the images and therefore improved processing by the CNN.

- The CT image and RTDose volume were normalized based on the mean and standard deviation of the population CT and RTDose volumes, respectively. Normalization ensures similar data distributions, allowing for fast convergence during network training.

- CT Images, RTDose volumes and contour volumes were resized to a grid size of 1283. Resizing was performed using the open-source scikit-image library [45], which preserves the image's HU distribution. The aspect ratio of the volumes were maintained by padding each of the volumes to a uniform size based on the largest dimension in the 3D volume.

- Two types of data augmentation were performed to introduce randomness to the training data and minimize the chances of overfitting. 1) Flipping of the volumes in the lateral direction. 2) affine transformations with rotations between -16 and + 16 degrees, translation in vertical and horizontal directions by 15% or the volumes width and height, and scaling by factors between 0.85 and 1.25. Each of the two different data augmentation types had a mutually exclusive chance of occurring of 60%.

9.2.14 CNN architecture and training

We used the open-source python library, PyTorch [23], to train our deep learning networks. A virtual machine from VMware, Inc. with 10 Intel Xeon CPU E5-2690 processors and a NVIDIA Tesla K40m GPU was used for training and testing. Ten-fold stratified cross validation was performed using the 160 curated patients and 18 LRF events.

- Clinical DLN: Utilized only the clinical features described above (Fig. 9.2a). The one hot encoded feature representations, along with the unaltered age feature were pushed through a two linear neural network layers with weighted optimization to account for class imbalance. The first layer underwent scaled exponential linear units (SELU) activation [25], the output of the second layers was used as input to a single fully connected layer. Outcomes were predicted using softmax classification.

- Radiomic + RTx-omic DLN: Used the patient image, RTDose and associated contour volume in a three-dimensional, three-channel, four-layer CNN (Fig. 9.2b). The outputs of all layers, except the final layer, underwent batch normalization, rectified linear unit functioning (ReLU) activation and max pooling [25]. The output of the final layer of the CNN underwent average pooling followed by a fully connected layer and softmax classification. The first CNN layer had convolutional kernel sizes of 5 with a padding of 2; the remaining layers used a size of 3 and padding of 1. Weighted optimization was used to account for imbalanced class distributions.

- Clinical + Radiomic + RTxomic DLN: A combination of the two previously described networks (Fig. 9.2 a and b). The output of the final linear layer from Clinical and the output from the final CNN layer from Radiomic + RTxomic are combined in the fully connected layer prior to softmax classification. Weighted optimization was used to account for imbalanced class distributions.

9.2.15 Scoring metric

In order to take into account the large class imbalance found in our dataset, the area under the PR-AUC was used for performance evaluation. PR-AUCs are more sensitive to class imbalances, and therefore provide a better metric of evaluation for our study compared to the more commonly used receiver operator characteristic curves [42]. Precision is the ratio of the number of true positives divided by the sum of true positives and false positives. Recall is the ratio of the number of true positives divided by the sum of true positives and false. When determining whether a PR-AUC is better than random the balance of classes must be considered. This is achieved by determining the probability of randomly guessing a positive event, given by the number of positive events divided by the sum of the positive and negative events, which is equivalent to the event rate of the dataset. For our dataset, a PR-AUC of 0.11 is considered random performance. PR-AUCs were calculated for our work using Python's Sci-kit learn library [38]. For additional comparison, the PR-AUC of univariate GTV volume was calculated, a known prognostic factor for HN cancer [8].

9.3 RESULTS

PR-AUC values above 0.11 are considered to have better than random performance. Our MLP with clinical features and LOG modelling had the overall highest PR-AUC for LRF prediction at three years of 0.66 (0.33–0.93). RF modelling performed best with clinical features only (PR-AUC = 0.61 (0.25–0.96)), and IF also performed best when utilizing only clinical features (PR-AUC = 0.42 (0.18–0.75)). Our DLP performed best with only clinical features as well (0.38(0.23–0.54)). PR-AUC values for all modelling methods and feature combinations can be found in Table 9.3. All of the above mentioned models and feature groups performed better than our univariate GTV volume predictor, which had a PR-AUC of 0.21. Table 9.4 presents the number of features that were retained after feature set reduction and model fitting in our MLP.

Table 1
PR-AUC and corresponding confidence intervals (CI) for both pipelines.

Random performance PR-AUC = 0.11	User-Driven Pipeline			Automated Pipeline
	Random Forest • Random stratified subsampling • 75/25% split DA patients excluded	Logistic Regression • Random stratified subsampling • 75/25% split DA patients excluded	Isolation Forest • Random stratified subsampling • 75/25% split DA patients excluded	Deep Learning • 10 fold cross validation • 20 epochs DA patients included
Clin.	0.61 (0.25–0.96)	0.66 (0.33–0.93)	0.42 (0.18–0.75)	0.38 (0.23–0.54)
Rad. + RTx.	0.12 (0.05–0.22)	0.19 (0.07–0.56)	0.26 (0.15–0.62)	0.36 (0.17–0.54)
Clin. + Rad. + RTx.	0.33 (0.12–0.73)	0.15 (0.08–0.48)	0.20 (0.12–0.50)	0.32 (0.20–0.45)

Figure 9.3: PR-AUC and corresponding confidence intervals (CI) for both pipelines.

Table 2
The number of features remaining after feature set reduction (FSR) and model fitting for the user-driven pipelines. The average number of features and standard deviations are presented.

	User-Driven Pipeline					
	Random Forest		Logistic Regression		Isolation Forest	
	FSR	Modeling	FSR	Modeling	FSR	Modeling
Clin.	8 ± 0	3 ± 1	8 ± 0	6 ± 2	8 ± 0	6 ± 0
Rad. + RTx.	409 ± 0	3 ± 1	409 ± 0	31 ± 56	409 ± 0	6 ± 0
Clin. + Rad. + RTx.	95 ± 13	4 ± 1	128 ± 21	27 ± 29	150 ± 15	6 ± 0

Figure 9.4: The number of features remaining after feature set reduction (FSR) and model fitting for the user-driven pipelines. The average number of features and standard deviations are presented.

The number of features was averaged across all 100 fittings for each of the feature groupings and modelling methods. It can be seen that all clinical features are retained after feature set reduction in the clinical feature grouping, as is expected based on the design of the feature set reduction method. Additionally, radiomic and RTx-omic features are known to correlate to clinical and volume features; therefore, more features were retained in the radiomic + RTxomic model than the clinical + radiomic + RTxomic model.

9.4 DISCUSSION

The ability to conditionally prognosticate a cancer patient's outcome based on their treatment is foundational to making personalized cancer medicine a reality. To accommodate existing and rapidly emerging

patient and treatment information, processes are required to integrate the variety of disease features available, including RTx-omic features that precisely quantify the treatment. Our work presents two user controlled pipelines where clinical features with LOG had the highest PR-AUC when predicting LRF at three years for HN cancer patients. More importantly, our results provide insight pertaining to the development of user-controlled pipelines for outcome prediction. In particular, the importance of curation, and user, data and methodology bias awareness as it pertains to result interpretation. The clinical features selected for this study provided the highest PR-AUC for HN LRF prediction at three years when combined with LOG modelling in our highly bespoke user driven pipeline. Although a promising result, large CIs indicate that subsampling was important and too few LRF events were present in our data. Additionally, the large CIs prevent us from definitively stating one model is better than another. Both of these observations suggest that a larger dataset may have resulted in a different final observation. These results are not to say that imaging and RT treatment features do not provide additional information important to the prediction of LRF, only that with the current data and our current features they do not draw immediate conclusions. When utilizing imaging and RT treatment information only, our DLP performed better than all three MLP modelling methods. This result may indicate that the machine was able to detect and extract features that were not seen and more informative than the hand-engineered/user-knowledge-informed features present in our MLP. Additionally, in our DLP, the Radiomic + RTx-omic DLN had comparable performance to the Clinical DLN (0.36 (0.17–0.54) vs. 0.38 (0.23–0.54), respectively, p-value = 0.97). This indicates that information could be extracted from images and RT treatment plans that is useful for conditional prognostics; we just have yet to obtain enough data to strengthen this signal. Future work may also be able to utilize larger resampling grids to retain more imaging and treatment details, providing more nuanced information to the machine. Despite this promise, LOG prediction with clinical features still performed better than both of these networks, and could be due to the breadth of knowledge included in the curation of

clinical feature definition [36][5][44] therefore requiring less complicated modelling techniques. Additionally, the DLP had more consistent PR-AUCs and smaller CIs across all feature combinations when compared to MLP modelling methods. This may be affected by differences in training/validation data, but it also seems to indicate that by using a less user-driven approach we are able to obtain more consistent information out of all data types when using our defined topology. These observations also lead the authors to suggest that various modelling methods, feature selection techniques, topology configurations, and levels of human interaction are tested during model development to determine the optimal performance for a given research question. This type of testing has been performed by other groups when utilizing radiomic features for outcome prediction [26][37] and would ensure that the best results for that given research question are achieved. Predictions utilizing quantitative image analysis and pattern recognition has been an area of study for close to two decades [16], [17]. Recent utilization of these methods in cancer prognostics with hand-engineered features has found promising results, particularly in HN cancer [24][10][1][3]. Deep learning is also being researched for its utility in this area [31]. In a recent study by Diamant et al. [9]., it was determined that deep learning methods were capable of identifying traditional radiomic features, in addition to newly generated features, that were beneficial in HN outcome prediction. Although the above mentioned work is promising, a recent study by Ger et al. [13] found that consistent associations between radiomic features and outcome in HN patients could not be found, even when utilizing large datasets (n ¿ 600) with standardized imaging practices. Obtaining large, high quality clinical datasets that are applicable to a given research question is challenging, as was seen in this study. However, if a strong biomarker or feature is embedded in the data and driving the outcome of interest it should be apparent, regardless of the dataset size, which has a stronger impact on the CIs than the overall performance [18]. When developing predictive models, it is understood that more data is often preferred. Larger datasets improve statistical analysis of the model and have a higher chance of containing heterogeneities that

models may encounter during clinical usage. More importantly, small datasets have increased potential for false positive and false negative errors [2] that are detrimental to health care resources and patient outcomes, respectively. The authors believe that the largest limitation for this study was the number of LRF events. The event rate for LRF was small, and in combination with our dataset size, this left very few examples to learn from during training. To account for the imbalance we used upsampling in our MLP and weighted optimization in our DLP; however the large CIs indicate the importance of subsampling in our study and the need for larger more diverse data. Additionally, utilization of uniform and high quality plans developed using the same planning criteria may have negatively impacted the final conclusions. Namely, it is possible that treatments were consistent enough that it was not possible to observe any LRF causing variations. Despite this, we were able to demonstrate the importance of benchmarking prognostic automated information generation pipelines against clinical variables which already achieve good predictions [48]. Another important limitation to the utilization of automated pipelines and data analysis is that imposed by the operator/human. Human knowledge is at the core of each step of an automated pipeline: data curation and collection, data pre-processing, feature definition – either through explicit definition or definition of a deep learning topology, feature selection, and model tuning, fitting and validation. Curation of the data in our study was guided by expert knowledge of clinical staff, as was definition of our RTx-omic features. Feature selection and modelling relied on prior author knowledge and experience. All of these steps will ultimately be biased by whomever is performing the experiments, which can be both a good and bad characteristic of the study. Until we are able to explore all permutations of potential features and machine learning methodologies within large datasets it is not possible to make definitive statements about the impact that automated pipelines will have on cancer care prognostics. By not fully understanding the risks associated with applied methods, we are likely to obtain unstable and misinformed results. From a user-driven pipeline perspective, some researchers [26][37] have done an excellent job of

publishing their results as a function of feature selection and modelling method performance. These types of publications are a good starting point when designing an experiment. However, researchers are urged to accurately publish all of their methods, not just the ones that had the best results. Additionally, it is important to understand the risk of data contamination that occurs in these studies. It is not common practice to have a true "Hold-Out" dataset [60], and therefore caution is warranted whenever interpreting the out of sample error rate, value and impact of a publications results. By exploring the rationale behind various steps of our processes we had important learnings regarding inherent biases present in current user-controlled pipelines; particularly when working with small datasets that contain only a few event of interest examples. There is a desire in this field to move towards the 'Big Machine' paradigm as a way to handle big data and provide a way to analyse and integrate the large and diverse data pools found within healthcare in a consistent and interoperable way. The processes that we have presented in this paper could be considered the 'little machine', a proprietary example of how the big machine would be operated. However, much larger and diverse datasets are needed to make true progress.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] David Jean Biau, Solen Kernéis, and Raphaël Porcher. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. *Clinical Orthopaedics and Related Research*, 466(9):2282–2288, September 2008.

[3] Marta Bogowicz, Stephanie Tanadini-Lang, Matthias Guckenberger, and Oliver Riesterer. Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Scientific Reports*, 9(1):15198, 2019.

[4] Jean Bourhis, Jens Overgaard, Hélène Audry, Kian K Ang, Michele Saunders, Jacques Bernier, Jean-Claude Horiot, Aurélie Le Maître, Thomas F Pajak, Michael G Poulsen, et al. Hyperfrac-

tionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. *The Lancet*, 368(9538):843–854, 2006.

[5] James Brierley, Brian O'Sullivan, Hisao Asamura, David Byrd, Shao Hui Huang, Anne Lee, Marion Piñeros, Malcolm Mason, Fabio Y. Moraes, Wiebke Rösler, Brian Rous, Julie Torode, J. Han van Krieken, and Mary Gospodarowicz. Global Consultation on Cancer Staging: promoting consistent understanding and use. *Nature Reviews Clinical Oncology*, 16(12):763–771, December 2019.

[6] Scott Buchanan. *The Doctrine of Signatures: A defense of theory in medicine*, volume 2. University of Illinois Press, 1991.

[7] Gabriella Cadoni, Luca Giraldi, Livia Petrelli, Manlio Pandolfini, Monica Giuliani, Gaetano Paludetti, Roberta Pastorino, Emanuele Leoncini, Dario Arzani, Giovanni Almadori, et al. Prognostic factors in head and neck cancer: a 10-year retrospective analysis in a single-institution in italy. *Acta Otorhinolaryngologica Italica*, 37(6):458, 2017.

[8] K.S.Clifford Chao, Gokhan Ozyigit, Angel I Blanco, Wade L Thorstad, Joseph O Deasy, Bruce H Haughey, Gershon J Spector, and Donald G Sessions. Intensity-modulated radiation therapy for oropharyngeal carcinoma: impact of tumor volume. *International Journal of Radiation Oncology*Biology*Physics*, 59(1):43–50, May 2004.

[9] André Diamant, Avishek Chatterjee, Martin Vallières, George Shenouda, and Jan Seuntjens. Deep learning in head & neck cancer outcome prediction. *Scientific Reports*, 9(1):2764, December 2019.

[10] Maximilian Diehn, Christine Nardini, David S. Wang, Susan McGovern, Mahesh Jayaraman, Yu Liang, Kenneth Aldape, Soonmee Cha, and Michael D. Kuo. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proceedings*

of the National Academy of Sciences of the United States of America, 105(13):5213–5218, April 2008.

[11] Georgia Doumou, Musib Siddique, Charalampos Tsoumpas, Vicky Goh, and Gary J. Cook. The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer. *European Radiology*, 25(9):2805–2812, September 2015.

[12] Ludwig Edelstein. *Ancient medicine; selected papers of Ludwig Edelstein*. Johns Hopkins Press, 1967.

[13] Rachel B. Ger, Shouhao Zhou, Baher Elgohari, Hesham Elhalawani, Dennis M. Mackin, Joseph G. Meier, Callistus M. Nguyen, Brian M. Anderson, Casey Gay, Jing Ning, Clifton D. Fuller, Heng Li, Rebecca M. Howell, Rick R. Layman, Osama Mawlawi, R. Jason Stafford, Hugo Aerts, and Laurence E. Court. Radiomics features of the primary tumor fail to improve prediction of overall survival in large cohorts of CT- and PET-imaged head and neck cancer patients. *PLOS ONE*, 14(9):e0222509, September 2019.

[14] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[15] Mary Gospodarowicz and Brian O'Sullivan. Prognostic factors in cancer. In *Seminars in surgical oncology*, volume 21, pages 13–18. Wiley Online Library, 2003.

[16] Ernest L Hall, Richard P Kruger, Samuel J Dwyer, David L Hall, Robert W Mclaren, and Gwilyu S Lodwick. A survey of preprocessing and feature extraction techniques for radiographic images. *IEEE Transactions on Computers*, 100(9):1032–1044, 1971.

[17] Charles A Harlow and Sharon A Eisenbeis. The analysis of radiographic images. *IEEE Transactions on Computers*, 100(7):678–689, 1973.

[18] Avijit Hazra. Using the confidence interval confidently. *Journal of Thoracic Disease*, 9(10):4124–4129, October 2017.

[19] Andrew J. Hope, Patricia E. Lindsay, Issam El Naqa, James R. Alaly, Milos Vicic, Jeffrey D. Bradley, and Joseph O. Deasy. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *International Journal of Radiation Oncology*Biology*Physics*, 65(1):112 – 124, 2006.

[20] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008.

[21] David A Jaffray. Image-guided radiotherapy: from current concept to future perspectives. *Nature Reviews Clinical Oncology*, 9(12):688, 2012.

[22] Wei Jiang, Pranav Lakshminarayanan, Xuan Hui, Peijin Han, Zhi Cheng, Michael Bowers, Ilya Shpitser, Sauleh Siddiqui, Russell H. Taylor, Harry Quon, and Todd McNutt. Machine Learning Methods Uncover Radiomorphologic Dose Patterns in Salivary Glands that Predict Xerostomia in Patients with Head and Neck Cancer. *Advances in Radiation Oncology*, 4(2):401–412, April 2019.

[23] Nikhil Ketkar. Introduction to PyTorch. In *Deep Learning with Python*, pages 195–208. Apress, Berkeley, CA, 2017.

[24] H. Kuno, M.M. Qureshi, M.N. Chapman, B. Li, V.C. Andreu-Arasa, K. Onoue, M.T. Truong, and O. Sakai. CT Texture Analysis Potentially Predicts Local Failure in Head and Neck Squamous Cell Carcinoma Treated with Chemoradiotherapy. *American Journal of Neuroradiology*, 38(12):2334–2340, December 2017.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[26] Stefan Leger, Alex Zwanenburg, Karoline Pilz, Fabian Lohaus, Annett Linge, Klaus Zöphel, Jörg Kotzerke, Andreas Schreiber,

Inge Tinhofer, Volker Budach, Ali Sak, Martin Stuschke, Panagiotis Balermpas, Claus Rödel, Ute Ganswindt, Claus Belka, Steffi Pigorsch, Stephanie E. Combs, David Mönnich, Daniel Zips, Mechthild Krause, Michael Baumann, Esther G. C. Troost, Steffen Löck, and Christian Richter. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Scientific Reports*, 7(1):13206, December 2017.

[27] Brian M Lin, Hao Wang, Gypsyamber D'Souza, Zhe Zhang, Carole Fakhry, Andrew W Joseph, Virginia E Drake, Giuseppe Sanguineti, William H Westra, and Sara I Pai. Long-term prognosis and risk factors among patients with hpv-associated oropharyngeal squamous cell carcinoma. *Cancer*, 119(19):3462–3471, 2013.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, Pisa, Italy, December 2008. IEEE.

[29] John T Lyman. Complication probability as assessed from dose-volume histograms. *Radiation Research*, 104(2s):S13–S19, 1985.

[30] Susan T Mayne, Brenda Cartmel, Victoria Kirsh, and W Jarrard Goodwin. Alcohol and tobacco use prediagnosis and postdiagnosis, and survival in a cohort of patients with early stage cancers of the oral cavity, pharynx, and larynx. *Cancer Epidemiology and Prevention Biomarkers*, 18(12):3368–3374, 2009.

[31] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A. D. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, March 2018.

[32] Serena Monti, Giuseppe Palma, Vittoria D'Avino, Marianna Gerardi, Giulia Marvaso, Delia Ciardo, Roberto Pacelli, Barbara A. Jereczek-Fossa, Daniela Alterio, and Laura Cella. Voxel-

based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. *Scientific Reports*, 7(1):7220, December 2017.

[33] Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Research methods & reporting-prognosis and prognostic research: what, why, and how?-doctors have little specific research to draw on when predicting outcome. this first article in a series explains why research into prognosis is important and how to design such research. *BMJ (CR)-print*, 338(7706):1317, 2009.

[34] Travis B. Murdoch and Allan S. Detsky. The Inevitable Application of Big Data to Health Care. *JAMA*, 309(13):1351, April 2013.

[35] Jens Overgaard, Hanne Sand Hansen, Lena Specht, Marie Overgaard, Cai Grau, Elo Andersen, Jens Bentzen, Lars Bastholt, Olfred Hansen, Jørgen Johansen, et al. Five compared with six fractions per week of conventional radiotherapy of squamous-cell carcinoma of head and neck: Dahanca 6&7 randomised controlled trial. *The Lancet*, 362(9388):933–940, 2003.

[36] Brian O'Sullivan, Shao Hui Huang, Bayardo Perez-Ordonez, Christine Massey, Lillian L. Siu, Ilan Weinreb, Andrew Hope, John Kim, Andrew J. Bayley, Bernard Cummings, Jolie Ringash, Laura A. Dawson, B.C. John Cho, Eric Chen, Jonathan Irish, Ralph W. Gilbert, Angela Hui, Fei-Fei Liu, Helen Zhao, John N. Waldron, and Wei Xu. Outcomes of HPV-related oropharyngeal cancer patients treated by radiotherapy alone using altered fractionation. *Radiotherapy and Oncology*, 103(1):49–56, April 2012.

[37] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo J. W. L. Aerts. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports*, 5(1):13087, October 2015.

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[39] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1):3–14, September 2002.

[40] Lester J. Peters, Brian O'Sullivan, Jordi Giralt, Thomas J. Fitzgerald, Andy Trotti, Jacques Bernier, Jean Bourhis, Kally Yuen, Richard Fisher, and Danny Rischin. Critical Impact of Radiotherapy Protocol Compliance and Quality in the Treatment of Advanced Head and Neck Cancer: Results From TROG 02.02. *Journal of Clinical Oncology*, 28(18):2996–3001, June 2010.

[41] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014.

[42] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015.

[43] William U Shipley, Joel E Tepper, George R Prout, Lynn J Verhey, Oscar A Mendiondo, Michael Goitein, Andreas M Koehler, and Herman D Suit. Proton radiation as boost therapy for localized prostatic carcinoma. *Jama*, 241(18):1912–1915, 1979.

[44] Temel Tirkes, Margaret A. Hollar, Mark Tann, Marc D. Kohli, Fatih Akisik, and Kumaresan Sandrasegaran. Response Criteria in Oncologic Imaging: Review of Traditional and New Criteria. *RadioGraphics*, 33(5):1323–1341, September 2013.

[45] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014.

[46] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[47] Mattea L Welch and David A Jaffray. radiomics: the new world or another road to el dorado?, 2017.

[48] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[49] Kim Wopken, Hendrik P Bijl, and Johannes A Langendijk. Prognostic factors for tube feeding dependence after curative (chemo-) radiation in head and neck cancer: a systematic review of literature. *Radiotherapy and Oncology*, 126(1):56–67, 2018.

[50] Po-Yen Wu, Chih-Wen Cheng, Chanchala D Kaddi, Janani Venugopalan, Ryan Hoffman, and May D Wang. –omic and electronic health record big data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering*, 64(2):263–273, 2016.

[51] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

10

The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques

Adapted from: **"The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques"**. A Traverso, J Van Soest, L Wee, A Dekker. Medical physics 45 (10), e854-e862.(2018).

Abstract

Personalized medicine is expected to yield improved health outcomes. Data mining over massive volumes of patients' clinical data is an appealing, low-cost and non-invasive approach toward personalization. Machine learning algorithms could be trained over clinical "big data" to build prediction models for personalized therapy. To reach this goal, a scalable "big data" architecture for the medical domain becomes essential, based on data standardization to transform clinical data into FAIR (Findable, Accessible, Interoperable and Reusable) data. Using Ontologies and Semantic Web technologies, we attempt to reach mentioned goal. We developed an ontology to be used in the field of radiation oncology to map clinical data from relational databases. We combined ontology with semantic Web techniques to publish mapped data and easily query them using SPARQL. The Radiation Oncology Ontology (ROO) contains 1,183 classes and 211 properties between classes to represent clinical data (and their relationships) in the radiation oncology domain following FAIR principles. We combined the ontology with Semantic Web technologies showing how to efficiently. When clinical FAIR data sources are combined (linked data) using mentioned technologies, new relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing.

10.1 INTRODUCTION

10.1.1 Motivation

Data-driven medicine has the potential to yield improved health outcomes [2] and is an integral component of value-based healthcare [5]. One of the biggest challenges for data driven medicine is to access and analyse clinical data with machine learning techniques to predict clinical outcomes combining all available information. Subsequently, these developed machine learning techniques can be used to build decision support systems for clinicians. The current obstacle to be addressed is the availability of outcome information (e.g., tumour control and treatment-related toxicity) that must be provided to "train" the machine learning models. Many models have been built based on data from clinical trials. However, clinical trials recruit only a small part of the presenting cases, therefore questions about applicability to underrepresented patient subgroups persist. In contrast, clinical data derived from routine care are known to have data quality issues (e.g., a high rate of missing values). To overcome the potential sensitivity to missing values as well as to provide sufficient training samples for machine learning, a scalable "big data" architecture for the medical domain becomes essential. For such a scalable architecture, data standardization is imperative. In particular, clinical data should be transformed following FAIR (Findable, Accessible, Interoperable, and Reusable) principles [19]. To make healthcare data FAIR, Ontologies and Semantic Web technologies play a key role, and hence will address the issue of semantic interoperability. In [1], the authors exploited the possibility to use ontological technologies to enable semantic interoperability with data coming from multicentre postgenomics clinical trials. In [16], the authors focused on applying Semantic Web technologies to the medical imaging domain, developing an ontology for medical image annotations. In [17], the authors investigated the possibility to use Semantic Web technology to store and represent metadata from DICOM image files. Both the studies showed the potential of ontologies technologies in allowing medical data interoperability. However,

the usage of ontologies and Semantic Web technologies applied to the field of radiation oncology are limited. In [14], the authors converted clinical data of prostate cancer patients from a local database using a dedicated ontology, but they did not exploit the possibility to merge different datasets from different diseases combining different ontologies. In [13], the authors stressed the concept of standardization of collected data (in rectal cancer) using ontological techniques to allow machine learning algorithms to build clinical prediction models. In addition, they strongly suggested using Semantic Web technologies in order to allow data sharing while respecting the privacy protection of individual patients. Finally, ontologies and Semantic Web techniques represent the required infrastructure for distributed learning [9] compared to traditional centralized learning approach, in distributed learning clinical data do not leave the hospital, but after being transformed into FAIR, they are queried during the training of the model, while the model is "learned" from different centres. Conversely, when looking at the radiation oncology domain, we could not find any study aiming at: (a) developing and validating a broader ontology to be used in the radiation oncology domain; (b) combining ontology and Semantic Web techniques to transform different clinical databases into FAIR and linked data. The role of the ROO is to provide a detailed and broad coverage of main concepts used in the radiation oncology domain such as: classification of neoplasms, patients' demographic characteristics; as well as clinical information like tumour's classification or treatment. The ontology is strongly focused on re-using published ontologies and/or terminologies. The added value of the ROO is to reused published ontologies/terminologies by defining new predicates, which establish relations between imported concepts. Combining different terminologies and expanding relationships between them is the path to guarantee the largest coverage. The ROO allows transforming unstructured clinical data from following FAIR principles. In particular, data will become:

- Findable (F): each data entity and their properties (F2), translated into universally concept via the ROO will have a globally unique

identifier (F1) and will be indexed on the Web (F3). Metadata will include specification of the data identifier (F4).

- Accessible (A): data will be retrievable by means of RDF triples and queryable using a universal language (A1). A permanent de-centralized storage point will be permanently available (A2), even when the original database could not be anymore.
- Interoperable (I): data are represented by universally adopted RDF language (I1). Queries rely on concept from imported ontologies/vocabularies that follow FAIR principles (I2).
- Reusable (R): several attributes specific data properties and the relations between different concepts via ROO predicates (R1). In this paper, we: (a) developed a broad ontology to cover the domain of radiation oncology; (b) combined ontology and semantic web techniques to transform clinical data from different disconnected databases into FAIR and linked data, allowing the discovery of new relationships.

10.1.2 Terminologies, vocabularies and ontologies

Before going into the details of ontologies' structure and properties, we provide the reader with some fundamentals regarding: terminologies, thesauri, vocabularies, and ontologies. Usually, a terminological system is an umbrella terms including the notions of: terminologies, thesauri, vocabularies, and ontologies [3]. Complexity increases from terminologies to ontologies:

- Terminology: a list of term referring to concept within a particular domain. For example, in the radiation oncology domain, concepts such as "patient" or "disease". The terminology can be seen as a list of concepts, but without providing any definition or introducing any structures/relations between the terms.
- Thesauri: a thesauri is a terminology, where concepts are indexed according to a certain rule (usually alphabetically). Example of

a thesauri is the International Classification of Diseases (ICD), which includes generic-related diagnostic terms (terminology), order alphabetically (thesauri).

- Vocabulary: in a vocabulary, indexed concepts are accompanied by a definition.

Conversely, an ontology is an explicit formal specification of the terms in the domain and relations among them11 expressed in machine-readable language; therefore, they can be processed automatically. An ontology adds more complexity than a dictionary, since it explicitly defines the relationship, i.e., predicates, between unique entities. Classes (i.e., concepts), subclasses, and predicates between concepts represent an ontology. Inference rules (also called automated reasoning) in ontologies supply further knowledge, since (new) relationships between concepts, which can be discovered, since not formally defined a priori. An ontology is commonly used to model consensus in understanding a domain between different partners (e.g., different medical centers). Major advantages of ontologies are: (a) sharing common understanding; (b) re-using of domain knowledge, analyzing domain knowledge, and (c) inferring new knowledge starting from relationship between defined concepts. The standard for developing ontologies is the Web Ontology Language (OWL) as recommended by the W3C (World Wide Web Consortium) to represent ontologies [7][12].

10.1.3 Semantic web technologies

Semantic Web is not a separate Web, but an extension of the current one, in which computers primarily interpret the data instead of humans. The current web provides rich-media content (e.g., written text, images, video's,) which is not easy to interpret for computers. In the Semantic Web extension, the information is represented in well-defined structures and semantics in order to enable automated processing of the contents by computers [11]. Hence, it can function

as a computer representation of already available web content, next to the human-readable web content. For the Semantic Web to function, computers must have access to structured collections of information. The basic building blocks are therefore provided by the Resource Description Framework (RDF) and the "SPARQL Protocol And RDF Query Language" (shorthand: SPARQL). Both RDF and SPARQL build on the existing web components of URIs and HTTP. URIs are the links to the actual resources, and can be represented as URLs (e.g., "http://mydomain.com/rdf/patient/12345"). These URIs are used to represent nodes (resources) and arcs (predicates) in the RDF graph. HTTP is used to publish RDF information on the web or to perform SPARQL queries on RDF stores. These RDF stores (also called SPARQL endpoints) are webpages which can be queried using the HTTP protocol. Most of these stores/endpoints also have human-readable web interfaces. By using RDF as a universal graph data structure, the Semantic Web relies on ontologies to give domain-specific structure and interpretation to the represented data. In these ontologies, hierarchies of concepts can be defined, as well as relationships between certain concepts; all written in RDF. It is a common practice to add human-readable attributes to the URIs, as it enables the creation of human-readable views on an RDF endpoint. By creating instances of concepts defined in the ontology, users can create graphs of data for representing real-life concepts (e.g., " http://mydomain.com/rdf/patient/12345 rdf:type http://mydomain.com/ontology/patient") where the resource 12345 is an instance of the class patient). In addition, ontologies can describe inferencing rules which are interpretable by inferencing-enabled RDF stores. In these stores, it is possible to query or show the inferenced information, which is not hard-coded (or materialized) in the RDF store. Hence, updating inferencing rules in the ontology would enable users to query or show additional information without updating the RDF store itself. This allows to uncover additional relationships in the actual data, and accommodates searches on different levels of data (e.g., patients are persons; therefore, searching for persons will include all patients in the database).

10.2 MATERIAL AND METHODS

10.2.1 Clinical database

We used a clinical database of oncological patients with a diagnosed rectal cancer from the THUNDER trial [18]. The goal of the trial was to develop a prediction model of rectal tumor response after chemo-radiotherapy that might be helpful in individualizing treatment strategies, i.e., selecting patients who need less invasive surgery or another radiotherapy strategy instead of resection. The database includes 80 patients and contains a diverse range of information, combining demographic and clinical outcomes. Due to its heterogeneous nature, it represents a good validation for the ROO. The ROO was applied to convert each values in the database, mapping them through the concepts available in the ontology. Relations between individuals were mapped using a graph structure. The graph output was then transformed into RDF triples, published on a dedicated endpoint and queryable, in line with FAIR principles.

10.2.2 Radiation oncology ontology (ROO) development

We developed a radiation oncology ontology (ROO). The ontology was designed using the editor tool Protégé [15] and publically published at the NCBO BioPortal (`https://bioportal.bioontology.org/ontologies/ROO`). The ROO adheres to the Ontology Web Language (OWL) 2 Query Language (QL) profile (`http://www.w3.org/TR/owl2-profiles/`). The ontology provides basic concepts, relationships, and properties/attributes for radiation oncology. The ontology was built following this procedure: (a) we identified variables of interest by collecting concepts and their definitions within the ontology using different datasets coming from several institutions belonging to the Euro CAT projects [4] . Due to its multicenter nature, we could allow a broad coverage for different diseases with the aim of making the ontology as much detailed as possible; (b) we published and

make publicly available on BioPortal several versions of the ontology during its development. This choice allowed users downloading, using and testing our ontology. In addition, a dedicated section on GitHub permitted users highlighting inconsistencies and/or requiring enhancements. In this way, our ontology became a dynamic body of knowledge with the aim of guaranteeing the broadest possible coverage for the radiation oncology domain. The high-level structure of the ROO is based on the Unified Medical Language System (UMLS) Semantic Network by the Semantic Types (classes) ontology (`http://bioportal.bioontology.org/ontologies/STY/?p=summary`) and the assertion of the Semantic Relations (properties) as specified by the UMLS (`https://uts.nlm.nih.gov/`). The ROO re-uses as much as possible entities from other ontologies such as the National Cancer Institute (NCIT) Thesaurus or the International Classification of Disease (ICD) ontologies. The ROO makes only use of ontologies published at NCBO's BioPortal and provided without any restrictions. Common re-used ontologies were: NCIT (National Cancer Institute Thesaurus); Units of Measurement Ontology (UO); Foundational Model of Anatomy (FMA); Semantic Types Ontology (STY); Semantic DICOM Ontology (SEDI), and International Classification of Diseases, Version 10 (ICD10). The ROO uses the original Unique Resource Identifiers (URIs) for these imported entities: an example is the concept of lung cancer that is inherited using the concept code C34 from the ICD ontology.

10.2.3 Ontology validation

Mapping between database schemas and ontology

One of the most important test to validate the ontology is guaranteeing that every element in a clinical relational database and its properties can be fully mapped with respectively the concepts and predicates in the ontology. The basic idea of the mapping process is linking each

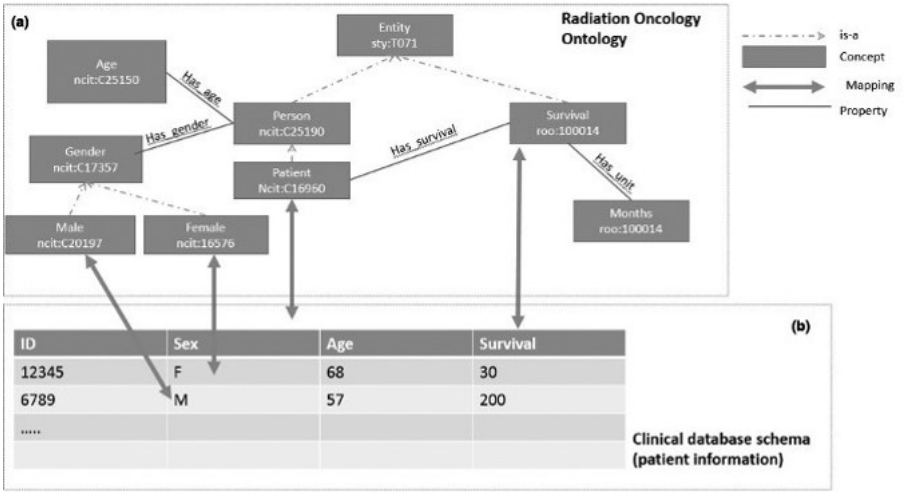


Figure 10.1: Overview of the ROO structure and the relational database. The hierarchical structure of the ROO is presented in the rectangle a. Hierarchical Relationships ("is subclass of") between classes, are expressed by dotted arrows. Mapping is performed to columns and values in a relational database (rectangle b).

component (row, columns, and values) of the database to its corresponding component (concept, property, relationship) of the ROO. The preliminary step is to identify a correspondence between the columns in the relational database and the ontology entities. A sketch representation of the mapping procedure is shown in Fig. 10.1. At the top, the hierarchical structure of the ROO is presented in the rectangle A. Hierarchical Relationships ("is subclass of") between classes, are expressed by dotted arrows. These relationships between more general classes (parents) and more specific classes (children) represent the ontology backbone since they allow properties inheritance. ROO concepts are expressed inside blue squares. Relationships between concepts (predicates) are expressed with arrows: they connect classes between each other. For example, patient and gender classes are connected by the property "has gender". A sample table of one of the datasets is shown

in the rectangle B. This table contains information about patient demographics (e.g., sex, age) as well as diagnosis (e.g., survival, tumor staging). The mappings are built between the table columns and the concepts in the ROO (shown as bold dotted double-headed arrows in the figure). For example, the column "Gen" is mapped to the concept gender (ncit:C17357) in the ROO. The link between a patient and the sex is made by the property "has gender". Several languages and software tools are available to perform the mapping procedure from relational databases to RDF triples [8]. We performed the mapping between the clinical data and the ontology using the D2RQ mapping language. D2RQ mapping language is a declarative language for mapping relational database schemas to RDF vocabularies and OWL ontologies. The language is read and interpreted by the D2RQ platform, which is written in Java and open-source available. We decided to use D2RQ because it represents one of the most common tools for database transformation from relational database to network structures [20]. In addition, the language is modular, easily allowing to link entities from the database to concepts and properties in ROO. The mapping defines a virtual RDF graph that contains instructions how to connect and map the information from the relational database. This is similar to the concept of views in SQL databases, except that the virtual data structure is an RDF graph instead of a virtual relational table. The mapping file, written in turtle (.ttl) syntax, contains the mapping between the database schema and the concepts defined in the ontology. The turtle syntax is the format for expressing data as RDF triples, then queryable using a dedicated language. An example of the mapping file is shown in Fig. 10.2. The mappings between table columns and their corresponding concepts are created using the command d2r:ClassMap. The mapping between the table columns to their corresponding properties is performed by using the command d2rq:PropertyBridge. In addition, in the mapping script each entity is associated with a Unique Resource Identifier (URI) to facilitate publishing on the Semantic Web and data linking. In the example, the entity patient is mapped to the concept C16960 from the NCIT. The bridge between a patient and his/her gender is mapped through the predicate 100018 ("has gender") from the

```
# PATIENT TABLE

map:patient a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:class ncit:C16960; #patient
        d2rq:uriPattern
"patient_@@derived_multidelineations.identifier@@";
        .

# PROPERTY BRIDGE FOR PATIENT

map:patient_label a d2rq:PropertyBridge;
                d2rq:BelongsToClassMap map:patient;
                d2rq:property rdfs:label;
                d2rq:column
"derived_multidelineations.collection_identifier";
                d2rq:datatype xsd:String;
                .
```

```
#CREATE NOW GENDER OBJECT

map:patient_gender_obj a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern
"gender_@@derived_multidelineations.identifier@";
        d2rq:condition "derived.gender IS NOT NULL";
        .

# Link to the gender object of the NCI thesaurus

map:patient_gender_uri_obj a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:patient;
        d2rq:property roo:100018; #has_gender
        d2rq:refersToClassMap map:patient_gender_obj;
        .
```

Figure 10.2: Example of the D2RQ mapping script. The first block (patient table) defines the mapping for each patient ID. The "ClassMap"property in a D2RQ script defines a mapping between a header in the relational database and the corresponding concept in the ontology. A "Property Bridge" is used in a D2RQ script to express relations between different concepts. In the example above, the "Property Bridge" has˙gender is used to link the patient concept to his/her gender.

ROO.

Publishing and querying data on the semantic web

The mapped data, transformed into URIs, are then stored in a RDF store, which is web-enabled (HTTP) and can be queried using SPARQL. Making these RDF stores web-enabled means that it is available internally or externally on a specific network, in the same way as webpages are. This does not per definition means that data are publically available, only that existing web techniques are used to represent semantically interoperable data. In our work, we used Blazegraph (`www.blazegraph.com`) as our RDF store (or SPARQL endpoint).

10.3 RESULTS

10.3.1 Radiation oncology ontology (ROO)

The ROO contains 1,183 classes, with an average number of four children per class; two classes have more than 25 children. The classes cover the most common concepts in radiation oncology, including cancer diseases, cancer-staging systems, and oncology treatments. Besides the classes, 211 predicates are introduced to express relationships between different classes. We divided the properties into five big categories: (a) conceptually related to; (b) functionally related to; (c) physically related to: (d) spatially related; and (e) temporally related to. Examples of mentioned categories are respectively: (a) diagnosed by; (b) has result; (c) connected to; (d) has location: (e) follows. All entities and predicates in the ROO have a URI, which can be resolvable as a link since hosted on `www.cancerdata.org`. A web RDF viewer allows the users inspecting a concept by typing on an internet browser the address `www.cancerdata.org/roo/[URI]`, where URI is the URI of the ontology entity. For example, the user will type `www.cancerdata.org/roo/100287` for the predicate "has pathological stage". In addition, the users are able to transverse the full tree of the ontology through the Web RDF viewer. The latest version of the ontology has been published on BioPortal, totally Open Source and available for the user to download. The ROO is available is the most common format, including OWL, which can be opened by the users using the software Protégé'.

10.3.2 Ontology validation

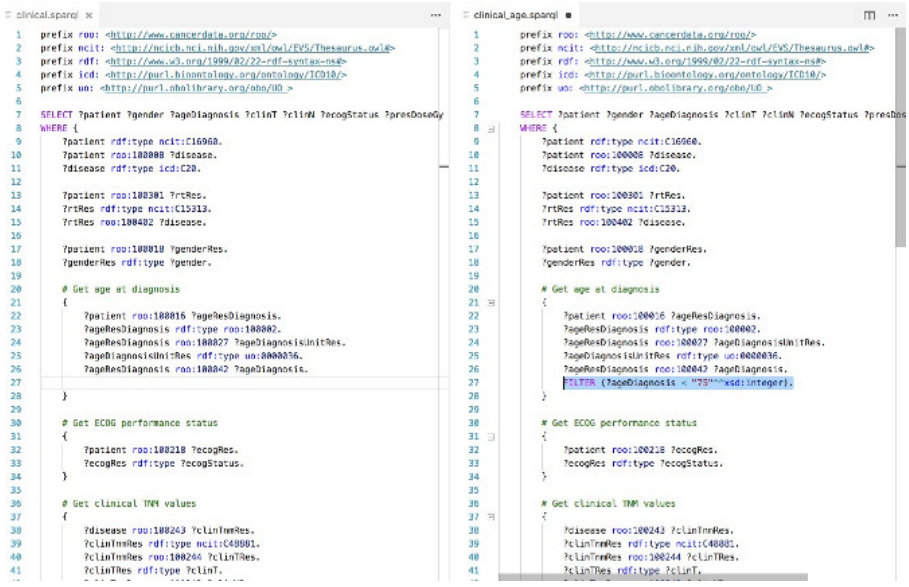
Mapping between database schemas and ontologies

A wiki page on how to perform the mapping between relational database schemas and the ontology is publically available on the GitHub (`https://github.com/jvsoest/Data-Integration-Tutorial/`

wiki/conversionClinicalData). The users can follow the guide to convert part of the Thunder dataset into RDF triples with the ROO using the example scripts provided.

Query formulation

After having mapped the data, it is possible to query them using SPARQL language. Users could query the data without having any prior knowledge of the relational database, since SPARQL queries are based on universal concepts defined by the ontology. Following the example in the Wiki (https://github.com/jvsoest/Data-Integration-Tutorial/wiki/queryClinicalData), let us suppose we want to search all the patients with rectal cancer and retrieve following information: age at diagnosis, ECOG (Eastern Cooperative Oncology Group) performance status score, clinical TNM stage, pathological TNM status, and prescribed dose in Gray. The example query is available at https://github.com/jvsoest/Data-Integration-Tutorial/blob/master/queries/queryClinicalData.sparql and it is shown in Fig. 10.3(a). The system returns all the patients and displays the results in the SPARQL result window on the web browser. Each object shown is associated with an URI, universally and unambiguously defining it when published on the Web. Furthermore, all triple patterns to find a certain variable are grouped in curly brackets. This creates the opportunity to make some variables optional or to specify some filters. For example, we could have asked for patients with an age at diagnosis below a certain value, by modifying the original query with a filter [highlighted in blue in Fig. 10.3(b)]. Finally, this example query can be used directly in programming languages/statistical languages to request valid data matrices. For example, in R using the SPARQL package, or in any other language using a representational state transfer (REST) interfacing package. Results from query shown in Fig. 10.3, where compared with data available in the original database to verify the



```
prefix roo: <http://www.cancerdata.org/roo/>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix icd: <http://purl.bioontology.org/ontology/ICD10/>
prefix uo: <http://purl.obolibrary.org/obo/UO_>

SELECT ?patient ?gender ?ageDiagnosis ?clinT ?clinN ?ecogStatus ?presDoseGy
WHERE {
    ?patient rdf:type ncit:C16960.
    ?patient roo:100008 ?disease.
    ?disease rdf:type icd:C20.

    ?patient roo:100301 ?rtRes.
    ?rtRes rdf:type ncit:C15313.
    ?rtRes roo:100402 ?disease.

    ?patient roo:100018 ?genderRes.
    ?genderRes rdf:type ?gender.

    # Get age at diagnosis
    {
        ?patient roo:100016 ?ageResDiagnosis.
        ?ageResDiagnosis rdf:type roo:100002.
        ?ageResDiagnosis roo:100027 ?ageDiagnosisUnitRes.
        ?ageDiagnosisUnitRes rdf:type uo:0000036.
        ?ageResDiagnosis roo:100042 ?ageDiagnosis.
    }

    # Get ECOG performance status
    {
        ?patient roo:100218 ?ecogRes.
        ?ecogRes rdf:type ?ecogStatus.
    }

    # Get clinical TNM values
    {
        ?disease roo:100243 ?clinTnmRes.
        ?clinTnmRes rdf:type ncit:C48881.
        ?clinTnmRes roo:100244 ?clinTRes.
        ?clinTRes rdf:type ?clinT.
```

```
prefix roo: <http://www.cancerdata.org/roo/>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix icd: <http://purl.bioontology.org/ontology/ICD10/>
prefix uo: <http://purl.obolibrary.org/obo/UO_>

SELECT ?patient ?gender ?ageDiagnosis ?clinT ?clinN ?ecogStatus ?presDos
WHERE {
    ?patient rdf:type ncit:C16960.
    ?patient roo:100008 ?disease.
    ?disease rdf:type icd:C20.

    ?patient roo:100301 ?rtRes.
    ?rtRes rdf:type ncit:C15313.
    ?rtRes roo:100402 ?disease.

    ?patient roo:100018 ?genderRes.
    ?genderRes rdf:type ?gender.

    # Get age at diagnosis
    {
        ?patient roo:100016 ?ageResDiagnosis.
        ?ageResDiagnosis rdf:type roo:100002.
        ?ageResDiagnosis roo:100027 ?ageDiagnosisUnitRes.
        ?ageDiagnosisUnitRes rdf:type uo:0000036.
        ?ageResDiagnosis roo:100042 ?ageDiagnosis.
        FILTER (?ageDiagnosis < "75"^^xsd:integer).
    }

    # Get ECOG performance status
    {
        ?patient roo:100218 ?ecogRes.
        ?ecogRes rdf:type ?ecogStatus.
    }

    # Get clinical TNM values
    {
        ?disease roo:100243 ?clinTnmRes.
        ?clinTnmRes rdf:type ncit:C48881.
        ?clinTnmRes roo:100244 ?clinTRes.
        ?clinTRes rdf:type ?clinT.
```

Figure 10.3: (a) On the left, example query without any filter; (b) on the right, example query introducing a filter on the diagnosis age. The query is written using the SPARQL language. Lines 1 to 5 are used to import the required ontologies. The query starts from line 7 (select query) asking for following information: patients gender, age at diagnosis, tumor T and N stages, overall health status, and prescribed dose. Then, each variable is queried in different blocks by making use of ROO concepts and predicates.

correctness of the mapping. Data comparison and visualization were performed and no differences were found when comparing the information available in the database with respect to the one available as SPARQL queries. The advantage with respect to a standard excel file, is that RDF data could be queried without any knowledge of the original data structures, by mean of SPARQL queries based on universal concepts defined by the ROO.

10.3.3 Combining different databases: linked data

One of the biggest benefits of Semantic Web and ontology technologies is the possibility to query different databases and make connections within them. For example, in radiation oncology it can be interesting for clinicians to investigate some properties (e.g., survival) of patients: (a) with a certain disease AND (b) treated according to a predetermined protocol AND (c) associating the publications of the clinical trial related to the protocol. Performing such a query using traditional relational databases is a real issue, since it not only requires combining different databases, but also a prior knowledge of their schemas. We solved the problem using ontologies and Semantic Web technologies.

Query formulation

In particular, we made use of Bio2RDF: an open-source project that uses Semantic Web technologies to build and provide the largest network of linked data for the life sciences. It contains among others the RDF versions of ClinicalTrials.gov and PubMed. In our query, the first part is equal to the query provided in the previous section. This query retrieves the patients available in the RDF store, and characteristics of these patients (e.g., age, gender, ECOG performance status), their disease (e.g., tumor classification), and the prescribed treatment. Based on this information, we linked

the patients to matching treatment protocols, as we defined the protocols and linked them to the correct ClinicalTrials.gov entry in Bio2RDF. Afterwards, the query contains a section to query the ClinicalTrials.gov linked data representation from Bio2RDF, and a URL generation for a PubMed query. To link the clinical information to public ClinicalTrials.gov (CTgov) information, we used the prescribed treatment variable (containing a unique URL) which was available on both internal (clinical) data, and the Bio2RDF CTgov linked data. From the CTgov linked data, we queried in which trials the same treatment protocol URLs were used. From this relation, we could retrieve information regarding the specific clinical trials, such as the CTgov identifier, the time period when the trial was conducted, which institutes were involved, and trial contact persons. Based on the CTgov identifier, we generated a link to the related manuscripts which have been indexed in PubMed. The full query to run this linked data example is available at `https://gist.github.com/jvsoest/eb015abfb0efd5c669fd36915ce2487d`. For example purposes, this query can be executed at `http://sparql.cancerdata.org/`.

10.4 DISCUSSION

10.4.1 Rationale for the ROO

Patients' demographics and clinical information are important for radiation oncology prediction/modeling studies. In particular, it is of interest of the radiation oncology community to explore the maximum amount of available clinical data to improve semantic interoperability during patient referral, and for models aiming at predicting outcomes such as overall survival or toxicities after a treatment. To reach this goal, data integration from different sources (internal/external relational databases) becomes a key factor, since most of the data are usually located in different relational databases. Since relational databases can present different structures, querying them to access information

without having a prior knowledge of the structures becomes a real issue. To tackle this issue, there is the need to transform clinical data following FAIR principles. Ontologies and Semantic Web technologies could represent the right choice to achieve this goal. We developed the Radiation Oncology Ontology (ROO) with the aim to provide an ontology of use within the radiation oncology field to be used to transform clinical data following FAIR principles.

10.4.2 Advantages of ontologies and semantic web data integration compared to relational databases

As presented in previous sections, the ROO has been used to transform clinical traditional database schemas into graph databases relying on ontologies. There are some differences between graph and database schemas. First, ontologies represent a domain on knowledge. Conversely, database schemas are conceived for (and linked to) particular applications, making their structures very diversified and difficult to be made interoperable. In fact, only users knowing the schemas structure (usually the owner of the data) can easily access them. On the contrary, ontologies transform data into universally concepts that can be queried by the users using SPARQL, without knowing the structures of the data themselves. In fact, data are transformed on universal concepts defined by the ontology itself, and available using URIs (and URLs). The usage of ontologies adds to transforming data from database schemas into FAIR data. An ontology, combined with Semantic Web technologies, is a stable conceptual interface on top of the relational database system. In fact, it can be scaled for data integration among multiple domains. Individual database schemas are mapped to the concepts of the ontology and it is relatively easy to integrate new database systems (when mapped/converted into Semantic Web data). The only modification required would be to update the mapping file. Overall, ontologies increase the semantic interoperability of already available data sources. This outcome has a direct impact on several clinical applications. In particular, it represents the underlying

infrastructure for developing multicenter prediction models for clinical outcomes in radiation oncology. In fact, if every medical center transformed their data into FAIR through the ontology, data analytics can be performed on a broader dataset reducing possibilities of overfitting. Ontologies and Semantic Web technologies will provide the infrastructure to query in an easy way the data needed by the model. Data will not need to leave the hospital, since being now FAIR, will be queried using SPARQL during the model training/validation. This application, known as distributed learning has been recently presented in literature as a promising application in radiation oncology [4][10][6]. In addition, Semantic Web and ontologies allow connecting different databases. In fact, data are transformed into universal concepts connected between each other: linked data. New relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing. In the examples we showed in the result section, we successfully integrated data coming from different sources: clinical databases, clinical trials bank, and scientific literature databases to answer questions of clinical interest. Finally, semantic databases (e.g., a collection of RDF records) have all the advantages from relational databases, but could provide the possibility of artificial intelligence to query and analyze the data, since these have been transformed into machine-readable records. Recently, we faced a transition from relational databases to semantic databases. The reason is that, semantic databases utilize an expanding semantic model that readily incorporates new varieties of data sources and more easily adjusts to changed requirements as they arise. Subsequently, linking disparate datasets is far easier in a semantic graph setting. In addition, semantic graphs allow to discover hidden relationships between underlying data. In fact, the granular nature of semantics allows to determine relationships between different elements.

10.4.3 Dynamic body of knowledge

We decided to put the ontology publically available on BioPortal, so that users could test and validate it with the aim of (a) developing a dynamic and growing body of knowledge; (b) guaranteeing the broadest coverage for the radiation oncology data domain. In addition, the latest version of the ROO is published on the GitHub: users are able to insert enhancements and open issues, making the ontology development a collaborative process.

10.4.4 Limitations

In this work, we explored the ontology-based data integration with data from rectal cancer databases. We were able to map all the entities present in the databases with concept and properties from the ROO ontology. However, the ROO should be tested also on larger databases, other diseases and routine clinical data to check if all the main information could be covered. In addition, this work lacks of the system evaluation. Further investigations on evaluating the system performance need to be considered such as comparing the query time between SPARQL and traditional databases.

10.4.5 Future developments

The first future development is to extend the ROO to guarantee a particular a broader coverage for an extensive use in the radiation oncology field. In particular, we would like to expand our ontology with:

- Detailed concepts for mapping radiation oncology annotations including organ at risks, nodals.
- Detailed concepts for mapping treatment-related concepts and properties such as Dose Volume histograms (DVH).

The second future development wants to expand the number of users. In this sense, we will continue proposing the ontology as underlying architecture for advance modeling applications such as distributed learning. In addition, we will try to use the ROO combined with other ontologies under development to combine and link: DICOM information, clinical data and quantitative features computed on patients' images and variables.

10.5 CONCLUSION

We successfully demonstrated that is possible to convert clinical data following FAIR principles using the combination of ontologies and Semantic Web technologies. We developed a broad Radiation Oncology Ontology that can be used in the domain of radiation oncology for data integration. In addition, we showed how Semantic Web technologies based on developed ontologies allows to efficiently and easily query data from different (relational database) sources without knowing a priori their structures. This outcome opens the possibility to use ontologies and Semantic Web technologies to further produce and analyze linked data in radiation oncology.

Bibliography

[1] Raul Alonso-Calvo, David Perez-Rey, Sergio Paraiso-Medina, Brecht Claerhout, Philippe Hennebert, and Anca Bucur. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. *Computer Methods and Programs in Biomedicine*, 118(3):322–329, March 2015.

[2] Mara G Aspinall and Richard G Hamermesh. Realizing the promise of personalized medicine. *Harvard business review*, 85(10):108, 2007.

[3] Nicolette F de Keizer, Ameen Abu-Hanna, and JHM Zwetsloot-Schonk. Understanding terminological systems i: terminology and typology. *Methods of information in medicine*, 39(01):16–21, 2000.

[4] Timo M. Deist, A. Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Andre Dekker, and Philippe Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*, 4:24–31, June 2017.

[5] M. J. Duffy and J. Crown. A Personalized Approach to Can-

cer Treatment: How Biomarkers Can Help. *Clinical Chemistry*, 54(11):1770–1779, November 2008.

[6] Issam El Naqa, Dan Ruan, Gilmer Valdes, Andre Dekker, Todd McNutt, Yaorong Ge, Q. Jackie Wu, Jung Hun Oh, Maria Thor, Wade Smith, Arvind Rao, Clifton Fuller, Ying Xiao, Frank Manion, Matthew Schipper, Charles Mayo, Jean M. Moran, and Randall Ten Haken. Machine learning and modeling: Data, validation, communication challenges. *Medical Physics*, 45(10):e834–e840, October 2018.

[7] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, June 1993.

[8] Matthias Hert, Gerald Reif, and Harald C. Gall. A comparison of RDB-to-RDF mapping languages. In *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*, pages 25–32, Graz, Austria, 2011. ACM Press.

[9] Arthur Jochems, Timo M. Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. 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(3):459–467, December 2016.

[10] Philippe Lambin, Erik Roelofs, Bart Reymen, Emmanuel Rios Velazquez, Jeroen Buijsen, Catharina M.L. Zegers, Sara Carvalho, Ralph T.H. Leijenaar, Georgi Nalbantov, Cary Oberije, M. Scott Marshall, Frank Hoebers, Esther G.C. Troost, Ruud G.P.M. van Stiphout, Wouter van Elmpt, Trudy van der Weijden, Liesbeth Boersma, Vincenzo Valentini, and Andre Dekker. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiotherapy and Oncology*, 109(1):159–164, October 2013.

[11] T Berners Lee and James Hendler. Publishing on the semantic web. *Nature*, 4(10):1023–1024, 2001.

[12] Ling Liu and M Tamer Özsu. *Encyclopedia of database systems*, volume 6. Springer New York, NY, USA:, 2009.

[13] Elisa Meldolesi, Johan van Soest, Anna Rita Alitto, Rosa Autorino, Nicola Dinapoli, Andre Dekker, Maria Antonietta Gambacorta, Roberto Gatta, Luca Tagliaferri, Andrea Damiani, et al. Vate: Validation of high technology based on large database analysis by learning machine. *Colorectal Cancer*, 3(5):435–450, 2014.

[14] Hua Min, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross, and J. Robert Beck. Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics*, 42(6):1035–1045, December 2009.

[15] Natalya Fridman Noy, Monica Crubézy, Ray W Fergerson, Holger Knublauch, Samson W Tu, Jennifer Vendetti, and Mark A Musen. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, pages 953–953, 2003.

[16] Daniel L Rubin, Cesar Rodriguez, Priyanka Shah, and Chris Beaulieu. ipad: Semantic annotation and markup of radiological images. In *AMIA annual symposium proceedings*, volume 2008, page 626. American Medical Informatics Association, 2008.

[17] Johan Van Soest, Tim Lustberg, Detlef Grittner, M. Scott Marshall, Lucas Persoon, Bas Nijsten, Peter Feltens, and Andre Dekker. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Studies in Health Technology and Informatics*, 205:166–170, 2014.

[18] Ruud G. P. M. van Stiphout, Vincenzo Valentini, Jeroen Buijsen, Guido Lammering, Elisa Meldolesi, Johan van Soest, Lucia Leccisotti, Alessandro Giordano, Maria A. Gambacorta, Andre

Dekker, and Philippe Lambin. Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: a multicentric prospective study with external validation. *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 113(2):215–222, November 2014.

[19] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[20] Arda Yunianta, Omar Mohammed Barukab, Norazah Yusof, Nataniel Dengen, Haviluddin Haviluddin, and Mohd Shahizan Othman. Semantic data mapping technology to solve semantic data problem on heterogeneity aspect. 2017.

11

The Radiomics Ontology (RO): standardizing radiomic studies following FAIR principles

Adapted from: **"The Radiomics Ontology (RO): standardizing radiomic studies following FAIR principles"**. A Traverso, M Vallières, A Zwanenburg, J van Soest, L Wee, J Seuntjens, M Hatt, O Morin, and A Dekker. Under review on Medical physics.

Abstract

The lack of adequate and standardized reporting strategy from radiomic computations strongly impact the transparency and reproducibility of radiomic studies. There is the need to transform radiomic data and metadata from individual experiments to adhere to FAIR (Findable Accessible Interoperable Reusable) principles in order to enable interoperability between different radiomic studies. Using ontologies and semantic web technologies, we attempt to reach the goal of allowing unambiguous description of radiomics features, extending beyond a purely lexical definition of individual features. The metadata allows us to capture the essential conditions under which a feature is extracted, that thereafter affects its reproducibility. We developed an ontology to be used in the field of radiomics to standardize and improve the reporting of data and metadata from radiomic computations. We combined the ontology with semantic web data standards and publish a subset of radiomic data with associated rich metadata from different radiomics packages, then we show that they are easily queried and compared using an appropriate query language (SPARQL). The Radiomics Ontology (RO) contains 458 classes, of which 187 are definitions of radiomic features. Besides classes, the RO has 76 object properties (predicates) that are used to fully document the whole spectrum of radiomic computations, from imaging pre/post processing to feature descriptions and computational details. We showed how to efficiently and easily integrate data and metadata from different radiomic packages, and then query these without a priori knowledge of the specific radiomics package outputs. The sharing of radiomic data containing rich metadata associated to their computational settings as web data objects now supports their use as FAIR data. This has the potential to accelerate the reproducibility and transparency of radiomic experiments and is a natural complement to the IBSI (Image Biomarker Standardization Initiative).

11.1 INTRODUCTION

11.1.1 Motivation

Radiomics, the automated extraction of quantitative information from medical images, has been deeply investigated for outcome predictions in, amongst others, oncology [1][2][4]. Unfortunately, only a small percentage of radiomic-based models are used in the clinic as decision support systems [18]. One of the reasons behind this discrepancy lies in the lack of reproducibility and generalizability of radiomic studies. The lack of reproducibility is strongly correlated with a lack of an adequate and satisfactory way of reporting radiomic studies. Most of the studies report only raw radiomic feature values but they omit the necessary computational steps that led to these values. The IBSI (Image Biomarker Standardization Initiative) is a multi-institutional effort to standardize radiomic feature definitions and their computations [19]. In their comprehensive document, they summarize the workflow of radiomic computations, therefore it includes not only the definitions of the most common radiomic features, but process steps that lead to these features such as image post-processing / pre-processing and algorithms used to tune the features' extraction. In a recent review, the authors pointed out how all the above computational steps potentially affect features' reproducibility [15]. By reporting only raw feature values, it becomes impossible for other users to fully reproduce and validate a radiomic experiment. However, a key to generalizability and usability of radiomic lies in being able to easily and extensively reproduce and validate these models in different institutions. Poor or insufficient quality of reporting limits this usability. Furthermore, an optimal configuration for radiomic features extraction might be modality or image pre-processing dependent, showing the need to provide data-driven evidence of computational settings to be preferred for a specific problem Again, the current inability to reach this stand in the way of adequate reporting strategy that enables interoperability. When developing a standardized radiomic reporting strategy, two points need to be taken into consideration: a) the number of computational radiomic

packages is growing; b) the features reported in studies should be interpretable by other researchers, without any prior knowledge of particular nomenclatures associated to features or computational details. To address the first point, a reporting solution must not limit the user to any one a particular computational package, but it should be interexchangeable for any package that might be used, even in-house and self-developed packages. For the second point, it is necessary to bind the solution with the concept of FAIR (Findable Accessible Interoperable Reusable) data management principles [17]. With this approach, data (radiomic feature values and their human readable labels) and associated metadata (describing in detail the computational settings used to derive the values) in a radiomics study becomes accessible and interoperable for everyone. The technologies to enable FAIR-ness for radiomic studies are open access ontologies and semantic web data objects. They have already been extensively used to standardize multisource clinical data in radiation oncology [14], but have not yet been applied for the standardization of radiomic computations. A first proof of concept study showing the power of ontologies in privacy preserving computational infrastructures was recently presented [12]. In this work: a) we present the RO (Radiomics Ontology) for reporting radiomic computations as FAIR-compliant data and metadata; b) we discuss use cases and applications of the RO.

11.1.2 Background: ontologies and semantic web techniques

In this section we offer the reader basic concepts related to ontologies and sharing data on the semantic web. An ontology is the formal specification of terms within a domain and their relations presented as machine-readable format [6], meant to be processed and mined automatically. Compared to vocabularies and terminologies [13], ontologies present more complexity via adding relationships, i.e. predicates, between unique entities. An ontology is formed by classes (i.e. concepts), subclasses, and predicates. Major advantages of ontologies are a) sharing common understanding of fundamental concepts related to

a field; b) inferring new knowledge starting from relations between the entities. The semantic web is an extension of the familiar web of http "pages" and web applications, where information is made available in specifically well-defined structures with metadata, so that it can be processed by machines [3]. The basic building blocks of the semantic web are the Resource Description Framework (RDF) and the "SPARQL Protocol And RDF Query Language' (shorthand: SPARQL) [10]. Databases transformed thorough the ontologies are published as RDF graph data structures on a dedicate SPARQL endpoint (i.e. RDF store) and can be queried using a web interface. By using RDF as a universal graph data structure, the semantic web relies on ontologies to give domain-specific structure and interpretation to the represented data. In addition, ontologies can describe inferencing rules which are interpretable by inferencing-enabled RDF stores. From any arbitrary number of remote and federated RDF stores, it is thus possible to query or show the inferenced relationships, which is not hard-coded (. materialized) in the RDF store. Hence, updating inferencing rules in the ontology would enable users to query or show additional information without updating the RDF store itself.

11.2 MATERIAL AND METHODS

11.2.1 Design of the radiomics ontology

We developed the radiomics ontology (RO). The ontology was designed using the editor tool Protégé (https://protege.stanford.edu/) and publicly published in at the NCBO BioPortal (`https://bioportal.bioontology.org/ontologies/RO`) repository. The RO adheres to the Ontology Web Language (OWL) 2 Query Language (QL) profile [5]. The ontology provides an extensive coverage of the radiomics computational workflow as described in detail within the IBSI reference manual [19]. Not only radiomics features are described, but also the main computational settings that apply to radiomics computations. The ontology was

built following this procedure a) variables of interests were identified looking at the IBSI reference document; b) main concepts of the ontologies were created for each of the sections in the IBSI manual. The radiomic features top class contains all the features described by the IBSI and found in all the most common available radiomics software. In addition, non-IBSI standardized features (available for example from the open source package Pyradiomics [16]) are mapped; c) the ontology was made available on BioPortal and GitHub (`https://github.com/albytrav/RadiomicsOntologyIBSI`) during its development phase so that radiomics users could verify and testing its level of coverage; c) the coverage of the ontology was considered satisfying when all the possible values / entities derived from the IBSI radiomic workflow were covered.

11.2.2 Template table for structured reporting and conversion to RDF

We developed a set of template tables to facilitate the standardization of the reporting of radiomic studies. The template tables are required not to limit a user to adopt a particular software for producing RDF triples. The template tables reflect the structure of the IBSI workflow, allowing a broad coverage of all the computational steps as described in the document. A total of 24 tables has been defined. In addition, detailed instruction for filling the table are provided. The template tables represent the input for the conversion to RDF triples via the radiomics ontology and are used to validate the ontology. The basic mapping procedure from template tables to RDF triples is to link each entity (row, column, values) of the table to its corresponding entity (class, property) in the RO. In Figure 11.1 we present a sketch of the conversion procedure between the template table "RadiomicsFeature.csv' and the ontology. The basic mapping procedure from template tables to RDF triples is to link each entity (row, column, values) of the table to its corresponding entity (class, property) in the RO. Several languages and software are available to perform the mapping procedure. Based

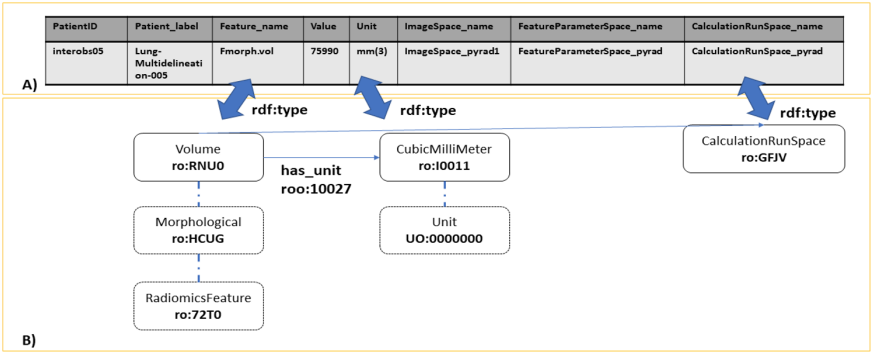


Figure 11.1: Overview of the template table structure and mapping with the RO. The hierarchical structure of the RO is presented in the rectangle B. Hierarchical Relationships ("is subclass of") between classes, are expressed by dotted arrows. Mapping is performed to columns and values in a relational database (rectangle a). In this case the mapping between the radiomic feature morphological volume, the corresponding unit and the calculation run space is shown as example. Similar procedure is adopted for all the tables and mapped concept are linked between each other using their URIs.

on our previous experience, we used the mapping language D2RQ (`http://d2rq.org/`). D2RQ is a declarative language allows mapping between relational databases to RDF vocabularies and OWL ontologies. We decided to opt for D2RQ since an open source software. The D2RQ mapping scripts are saved in turtle syntax (.ttl) and present a modular structure, easily allowing the map of different tables at the same time. One mapping script for each template database table was created.

11.2.3 Radiomic features as FAIR endpoint via the Semantic Web

At the end of the mapping procedure, the mapped data are uploaded and store into an RDF store, which is a public server (HTTP connection enabled) and can be queried using SPARQL language. In our experiment, we used both Blazegraph (`www.blazegraph.com`) and GraphDB (`http://graphdb.ontotext.com/`) as our RDF stores. They both work as public servers, but also as local storage.

11.2.4 Dataset and radiomic packages used

For testing the ontology, we used two different open source radiomic packages a) Pyradiomics, an open source python-based software developed at Harvard Medical School (Boston, USA) [16], and b) a Matlab-based software developed at McGill University, (Montreal, Canada) `https://github.com/mvallieres/radiomics`. The two software packages allow the customization of radiomic computations, but standard defaults are different. Two users independently extracted radiomic features from primary lung tumours with different delineations from 22 CT scans from the NSCLC-Radiomics-Interobserver1 as per the data available at the TCIA archive (`https://doi.org/10.7937/tcia.2019.cwvlpd26`). This dataset represents a good candidate for testing the ontology, since for example the presence of multiple delineations requires

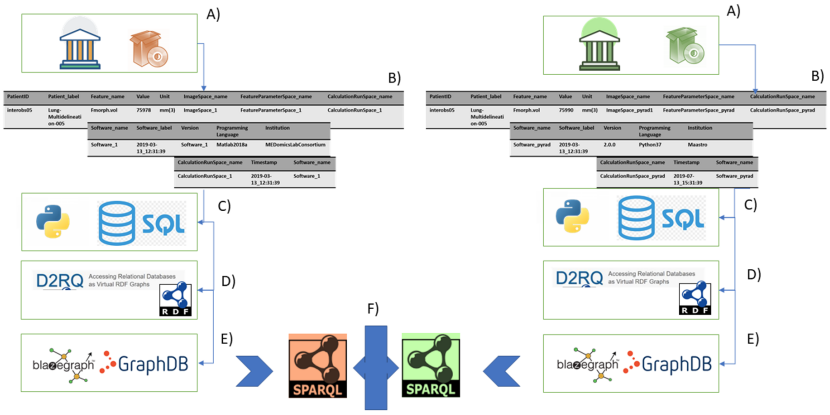


Figure 11.2: Sketch of the used workflow. A) Two users using two independent radiomic packages extract radiomic features from a common dataset. Details of the computation might different between the packages. B) The computation data and metadata are stored as .csv files using the IBSI-compliant template tables. C) Using the developed workflow, tables are stored in templated SQL databases. D) These tables are transformed into RDF triples using D2RQ scripts. The triples can be visualized locally on uploaded on a public server using for example Blazegraph or GraphDB. E) Each user can access the triples from the other one without prior knowledge of original labels as original output of the radiomic software.

the metadata related to segmentations to be associated to radiomic feature values. Their computations were then stored accordingly to the template tables and then the user converted them as RDF triples. Finally, they uploaded the data in a public SPARQL endpoint, for comparison. Figure 11.2 summarizes the adopted workflow.

11.3 RESULTS

To facilitate the replication of this experiment, as well as to disseminate and share with the community our developments, the workflow

is available at the following GitHub link (`https://github.com/albytrav/RadiomicsOntologyIBSI`).

11.3.1 Radiomics ontology

The RO contains 458 classes. The class with the largest number of children is RadiomicsFeature (`www.radiomics.org/RO/72TO`), which contains 187 definitions of radiomics features. Besides classes, the RO has 76 object properties (predicates) that are used to create the relationships between the different classes. All the entities and predicates have a URI, which is resolvable as a link hosted on `www.radiomics.org/RO/`. A web RDF viewer allows the inspection of a concept by typing its URI. Users can transverse the whole ontology tree through the Web RDF viewer. The RO is available in the most common formats, including .owl on `https://bioportal.bioontology.org/ontologies/RO` (bioportal) , which can be downloaded by the users with the software Protégé.

11.3.2 Template tables

Template tables were associated with instructions for the users on how to fill them. The central table is called "FeatureTable.csv" and it contains radiomic features with values and units. This table is then linked to: "ImageSpaceTable.csv"; "CalculationRunTable.csv"; "FeatureParameterSpaceTable.csv". The nested structure covers the whole spectrum of radiomic computations.

11.3.3 Mapping and comparison

Query example 1: Two users want to compare "all the radiomic features that have millimetre cubic as the unit", so that this query will return the radiomic features related to volume. The query example is available in the supplementary material, the results in Figure 11.3a. Here, without any prior knowledge of the original labels from the two

different software, it is possible to retrieve different categories of features with a particular property (unit in this case). From figure 11.3a it is possible to see that the user using Pyradiomics only computed one type of volume, while the user using the Matlab radiomic toolbox has different volumes (volume; approximate volume). Query example 2: Two users want to compare "how many morphological features are computed by the two software". For this query, we used one of the most useful properties of an ontology: inference (or automated reasoning). The idea is that properties embedded within the ontology can be used to query data, without that information being explicitly being uploaded as RDF triples. In this case it is not necessary to query for all the single classes corresponding to morphological features. We just queried for the top class of the RO "Morphological (HCUG)" and this automatically return all the subclasses (features) belonging to that category. The query example is shown in the supplementary material, the results in Figure 11.3b. Again, it is possible to see that Pyradiomics has less morphological features than the Matlab software. Query example 3: The two users want to compare "differences about their software". In this query we asked for the name of the radiomic package, the institution that developed it, and the programming language. The query example is shown in the supplementary material, the results in Figure 3c. As expected, the main difference between the two software is the programming language: python vs Matlab. Query example 4: We want to query "the volume of the tumour of the patients and some additional information such as gender and tumour T stage. This query is meant to show how the RO integrates with other ontologies such as the ROO (Radiation Oncology Ontology). It is an example of "linked data", where radiomic data (features) can be easily combined with clinical data. The query example is shown in the supplementary material, the results in Figure 11.3c.

	feature	unit	value
1	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"73005.66667"^^xsd:double
2	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"59897"^^xsd:double
3	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"71762"^^xsd:double
4	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"76970.66667"^^xsd:double
5	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"83454.66667"^^xsd:double
6	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"79926.33333"^^xsd:double
7	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"69583.66667"^^xsd:double
8	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"68983"^^xsd:double
9	http://localhost/rdf/feature_Fmorph.vol_interobs05	http://localhost/rdf/unit_mm{3}	"79098.33333"^^xsd:double

	feature	value	unit
1	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"67.46851117"^^xsd:double	http://localhost/rdf/unit_mm
2	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"66.12110102"^^xsd:double	http://localhost/rdf/unit_mm
3	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"64.56004957"^^xsd:double	http://localhost/rdf/unit_mm
4	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"66.48308055"^^xsd:double	http://localhost/rdf/unit_mm
5	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"79.32212806"^^xsd:double	http://localhost/rdf/unit_mm
6	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"64.49806199"^^xsd:double	http://localhost/rdf/unit_mm
7	http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_interobs05	"65.11528238"^^xsd:double	http://localhost/rdf/unit_mm

	software	institution	language
1	http://localhost/rdf/software_Software_1	http://localhost/rdf/institution_maastro	http://localhost/rdf/programminglanguage_python_3.7.3

	patient	gender	tstage	volume	value	unit
1	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"73005.66667"^^xsd:double	http://localhost/rdf/unit_mm{3}
2	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"59897"^^xsd:double	http://localhost/rdf/unit_mm{3}
3	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"71762"^^xsd:double	http://localhost/rdf/unit_mm{3}
4	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"76970.66667"^^xsd:double	http://localhost/rdf/unit_mm{3}
5	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"83454.66667"^^xsd:double	http://localhost/rdf/unit_mm{3}
6	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"79926.33333"^^xsd:double	http://localhost/rdf/unit_mm{3}
7	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"69583.66667"^^xsd:double	http://localhost/rdf/unit_mm{3}
8	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"68983"^^xsd:double	http://localhost/rdf/unit_mm{3}
9	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"79098.33333"^^xsd:double	http://localhost/rdf/unit_mm{3}
10	http://localhost/rdf/patient_interobs05	http://localhost/rdf/patient_interobs05/biologicalsex/male	http://localhost/rdf/patient_interobs05/neoplasm/ctstage/T2	http://localhost/rdf/feature_Fmorph.vol_interobs05	"103489.3333"^^xsd:double	http://localhost/rdf/unit_mm{3}
11	http://localhost/rdf/patient_interobs06	http://localhost/rdf/patient_interobs06/biologicalsex/male	http://localhost/rdf/patient_interobs06/neoplasm/ctstage/T1	http://localhost/rdf/feature_Fmorph.vol_interobs06	"12737"^^xsd:double	http://localhost/rdf/unit_mm{3}
12	http://localhost/rdf/patient_interobs06	http://localhost/rdf/patient_interobs06/biologicalsex/male	http://localhost/rdf/patient_interobs06/neoplasm/ctstage/T1	http://localhost/rdf/feature_Fmorph.vol_interobs06	"6136"^^xsd:double	http://localhost/rdf/unit_mm{3}

Figure 11.3: A) Results of query 1: all the radiomic features with unit millimetre cubic are retrieved. While Pyradiomics software only computes one type of morphological volume; the IBSI compliant software has an additional volume feature (approximate volume). Multiple values of volume are present for the patients, since as expected, radiomic features were extracted from multiple delineations B) Results of query 2: using the inference property of the ontology it is possible with few lines of code to retrieve all the radiomic features belonging to a particular class (morphological in this case). Results of query 3: properties about the two different radiomic packages are retrieved; Results of query 4: clinical data and radiomic data are combined.

11.4 DISCUSSION

We developed the radiomic ontology and we validated using a publicly available dataset with radiomic features computed using two different radiomic computational packages. The two users were independently able to compare data and metadata associated to their computations, without prior knowledge of the original labels of the software.

11.4.1 Rationale of this study

Radiomics is a rapidly developing field of study over recent years. Several open and closed source computational packages are now available such as LifeX [8], PyRadiomics [16], RaCaT [9]. Some of them allow full customization of feature extraction at different levels, such as for example the inclusion of different pre-processing techniques (filtering, resampling) to increase signal fidelity. The immense variety of handcrafted radiomics features makes it possible to mine for the optimal combination of features ie signature, to accurately predict an outcome class. However, such proliferation of possible features and feature processing settings has brought forward the need of standardization and harmonization especially for the problem of replication and validation [7]. Interoperability between radiomic computational packages remains a significant issue to this day: not only the feature nomenclature is different, but all the steps of the computations are still not always reported. When reported, they are expressed using non-standardized nomenclatures and format. All of the above makes it difficult for other users to reproduce and validate a radiomic experiment. In an earlier review, we pointed out the general unsatisfactory level of reporting in radiomic studies, which made impossible the conduction of any meta-analysis between different studies and was limiting consensus [15]. The IBSI initiative has committed itself to provide standards and recommendations for radiomic computations, together with extensively describing all the possible steps involved in a radiomic study. Guidelines and suggestions for reporting are identified in the last part of the document; but no final consensus has yet been proposed at time

of writing of this article. This work should be identified as natural complement of the IBSI effort, building upon their standardized global terminology and definitions, then by providing tools and methodologies for enhanced reporting of radiomic analyses with significantly enriched metadata. First, we provided a set of template tables, which can be used to cover the radiomic workflow. The template tables do not force a user to adopt radiomic software able to directly produce RDF triples [11]. Each radiomic software can be used to produce standardized output in the form of template tables and then the users can use our pipeline to produce RDF triples. The proposed workflow can be ported into other systems that can import structured reports and convert to RDF through an ontological schema. Only relying on tables would have limited the power of cross-correlating and comparing different radiomics computations. This is mainly due to the *static* nature of tables and relational databases. Therefore, we developed a dedicated radiomics ontology and mapped the above-mentioned schemas into RDF triples. Finally, this strategy allows the production of data and metadata from radiomic computations as FAIR-compliant, opening the door for more transparent and reproducible radiomics. For example, the proposed workflow can be used to investigate the predictive / prognostic power of texture features by comparing different computational settings or different software. In fact, texture features are prone to changes as soon as different settings to compute texture matrices are applied (e.g. different quantization algorithms). The optimal configuration for a specific problem can only be investigated by comparing multiple computation scenarios and / or software. This requires interoperability, which is guarantee by our ontology. It is important noticing that the proposed schema is not rigid, but each user can build a graph, based on the radiomics ontology, according to the needs.

11.4.2 Rationale for the radiomics ontology and semantic web

As presented in the previous sections, the RO was used to convert flat tables into graph data. The main difference between ontologies and relational databases is that ontologies represent a dynamic body of knowledge, by creating relations between entities through predicates. Conversely, relational databases are difficult to make interoperable. Only the owner of the database, as creator of the schema structures can access the information. Ontologies techniques transform data into relationships between universal concepts defined by unique identifiers (URIs), that can be queried without any prior knowledge of the original structure. The RO combined with semantic web technologies represents a powerful tool to allow integration and comparison of radiomics studies performed by multiple institutions. In fact, the results of the radiomics analysis, transformed using the RO and published on the Semantic Web, can be queried and compared between each other. As per ontology development, not only features values can be queried, but all the details related to the computation. In our results for example (query 3), we showed how it is possible to extract properties related to the radiomic software package. The combination of RO and semantic web can be used as reference framework for features' benchmarking by the IBSI. Each user, owner of a different radiomic software, can upload on the semantic web the results of its computation on a common, standardized set. These results can then be queried and differences in features values or software implementation can be investigated. Different open source libraries allow directly querying, storing and analysing RDF data such as "rdflib" for Python (`https://rdflib.readthedocs.io/en/stable/`) or the SPARQL wrapper for R (`https://cran.r-project.org/web/packages/SPARQL`). Our wiki is regularly updated with tutorials and examples to reach the broadest community as possible. Finally, the ontology has the flexibility to integrate with other ontologies. In our results (query4), we showed the integration between the RO and the ROO (Radiation Oncology Ontology). This example showed the integration of radiomic data and metadata (RO) with clinical data (ROO). A similar concept

was proposed in our recent publication, where an original radiomic study was reproduced using a privacy-preserving distributed learning infrastructure [12].

11.4.3 Limitations and future work

The radiomics ontology is a dynamic body of knowledge, which is meant to be updated with the same pace as new radiomics features are added or new computational packages are made available. We tested the ontology using two different software packages and one dataset. However, the ontology should be tested by as many different institutions as possible. Accordingly, the ontology is shared on GitHub making it a collaborative project. A second limitation regards the optimization of conversion from relational databases to RDF triples. We use D2RQ as mapping language, however different new languages have been proposed by the literature. It was out of topic of this preliminary proof-of concept to evaluate different mapping strategies, but it is in the plans to align our tools with the best and fastest mapping languages.

11.5 CONCLUSIONS

We developed a dedicated radiomics ontology to harmonize and improve quality of reporting in radiomic computations. The combination of our ontology and semantic web introduces FAIR principles to radiomic computations, for enabling transparency and reproducibility.

Bibliography

[1] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[2] Thibaud P Coroller, Vishesh Agrawal, Vivek Narayan, Ying Hou, Patrick Grossmann, Stephanie W Lee, Raymond H Mak, and Hugo JWL Aerts. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiotherapy and oncology*, 119(3):480–486, 2016.

[3] J. Davies, Dieter Fensel, and Frank Van Harmelen. *Towards the semantic web: ontology-driven knowledge management*. J. Wiley, Chichester, England ; Hoboken, NJ, 2003. OCLC: ocm50561857.

[4] Ming Fan, Hui Li, Shijian Wang, Bin Zheng, Juan Zhang, and Lihua Li. Radiomic analysis reveals dce-mri features for prediction of molecular subtypes of breast cancer. *PLoS One*, 12(2):e0171683, 2017.

[5] A Grigoris and Frank Van Harmelen. Web ontology language: Owl. handbook on ontologies in information systems, 2004.

[6] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, June 1993.

[7] Olivier Morin, Martin Vallières, Arthur Jochems, Henry C. Woodruff, Gilmer Valdes, Steve E. Braunstein, Joachim E. Wildberger, Javier E. Villanueva-Meyer, Vasant Kearney, Sue S. Yom, Timothy D. Solberg, and Philippe Lambin. A Deep Look Into the Future of Quantitative Imaging in Oncology: A Statement of Working Principles and Proposal for Change. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1074–1082, November 2018.

[8] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Frédérique Frouin, and Irène Buvat. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Research*, 78(16):4786–4789, August 2018.

[9] Elisabeth Pfaehler, Alex Zwanenburg, Johan R. de Jong, and Ronald Boellaard. RaCaT: An open source and easy to use radiomics calculator tool. *PLOS ONE*, 14(2):e0212223, February 2019.

[10] Bastian Quilitz and Ulf Leser. Querying distributed rdf data sources with sparql. In *European semantic web conference*, pages 524–538. Springer, 2008.

[11] Zhenwei Shi, Alberto Traverso, Johan Soest, Andre Dekker, and Leonard Wee. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). *Medical Physics*, page mp.13844, October 2019.

[12] Zhenwei Shi, Ivan Zhovannik, Alberto Traverso, Frank J. W. M. Dankers, Timo M. Deist, Petros Kalendralis, René Monshouwer,

Johan Bussink, Rianne Fijten, Hugo J. W. L. Aerts, Andre Dekker, and Leonard Wee. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Scientific Data*, 6(1):218, December 2019.

[13] Peter Spyns, Robert Meersman, and Mustafa Jarrar. Data modelling versus ontology engineering. *ACM SIGMOD Record*, 31(4):12, December 2002.

[14] Alberto Traverso, Johan van Soest, Leonard Wee, and Andre Dekker. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. *Medical Physics*, 45(10):e854–e862, October 2018.

[15] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[16] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[17] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene

van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[18] Alex Zwanenburg. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(13):2638–2655, December 2019.

[19] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

12

FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, Interobserver, Lung1 and Head-Neck1 TCIA collections

Adapted from: **"FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, Interobserver, Lung1 and Head-Neck1 TCIA collections"**. P Kalendralis, Z Shi, A Traverso, A Choudhury, M Sloep, I Zhovannik, MPA Starmans, D Grittner, P Feltens, R Monshouwer, S Klein, R Fijten, H Aerts, A Dekker, J van Soest, and L Wee. Medical Physics. doi:10.1002/mp.14322. (2020). Contribution: shared first authorship.

Abstract

One of the most frequently cited radiomics investigations showed that features automatically extracted from routine clinical images could be used in prognostic modelling. These images have been made publicly accessible via The Cancer Imaging Archive (TCIA). There have been numerous requests for additional explanatory metadata on the following datasets — RIDER, Interobserver, Lung1, and Head–Neck1. To support repeatability, reproducibility, generalizability, and transparency in radiomics research, we publish the subjects' clinical data, extracted radiomics features, and digital imaging and communications in medicine (DICOM) headers of these four datasets with descriptive metadata, in order to be more compliant with findable, accessible, interoperable, and reusable (FAIR) data management principles. Overall survival time intervals were updated using a national citizens registry after internal ethics board approval. Spatial offsets of the primary gross tumour volume (GTV) regions of interest (ROIs) associated with the Lung1 CT series were improved on the TCIA. GTV radiomics features were extracted using the open-source Ontology-Guided Radiomics Analysis Workflow (O-RAW). We reshaped the output of O-RAW to map features and extraction settings to the latest version of Radiomics Ontology, so as to be consistent with the Image Biomarker Standardization Initiative (IBSI). Digital imaging and communications in medicine metadata was extracted using a research version of Semantic DICOM (SOHARD, GmbH, Fuerth; Germany). Subjects' clinical data were described with metadata using the Radiation Oncology Ontology. All of the above were published in Resource Descriptor Format (RDF), that is, triples. Example SPARQL queries are shared with the reader to use on the online triples archive, which are intended to illustrate how to exploit this data submission. The accumulated RDF data are publicly accessible through a SPARQL endpoint where the triples are archived. The endpoint is remotely queried through a graph database web application at http://sparql.cancerdata.org. SPARQL queries are intrinsically federated, such that we can efficiently cross-reference

clinical, DICOM, and radiomics data within a single query, while being agnostic to the original data format and coding system. The federated queries work in the same way even if the RDF data were partitioned across multiple servers and dispersed physical locations. The public availability of these data resources is intended to support radiomics features replication, repeatability, and reproducibility studies by the academic community. The example SPARQL queries may be freely used and modified by readers depending on their research question. Data interoperability and reusability are supported by referencing existing public ontologies. The RDF data are readily findable and accessible through the aforementioned link. Scripts used to create the RDF are made available at a code repository linked to this submission: `https://gitlab.com/UM-CDS/FAIR-compliant_clinical_radiomics_and_DICOM_metadata`

12.1 INTRODUCTION

Clinical radiological imaging, such as computed tomography (CT), is a mainstay modality for diagnosis, screening, intervention planning, and follow-up for cancer patients worldwide [6]. Radiomics refers to high-throughput automated characterization of the tumour phenotype by analysing quantitative features derived from a radiological image [8]. Aerts et al. showed that CT radiomics features by themselves could contain information that is potentially prognostic of overall survival in non small cell lung (NSCLC) and head-and-neck (HN) cancer [3]. The radiomics hypothesis is that computationally derived features extract more information than can be processed by an unaided human eye, and therefore offers up new image biomarkers to speed up the research of personalized medicine. Radiomics has the potential to be a highly cost-effective option for retrospective observational clinical studies, since it can process routinely collected clinical radiological images residing in institutional archives. There remain significant challenges in regards to developing generalizable models that are based on reproducible and repeatable radiomics signatures [10][28][27][14]. Recent studies have suggested that harmonization of radiomics features across multiple institutions and different scanner parameters may be needed to realize its full potential [20] [23][11][29]. Computed tomography images for some frequently cited studies [3][30] in the digital imaging and communications in medicine (DICOM) format, have been made available via The Cancer Imaging Archive (TCIA)[30][4][24][2][1]. The DICOM standard incorporates metadata about image acquisition settings and it extends to regions of interest (ROIs) delineations (i.e., radiotherapy structure set, or RTSTRUCT), but many non radiology researchers remain unfamiliar with this conjoined data-metadata format. Pixel data only formats such as Neuroimaging Informatics Technology Initiative (NIfTI) and Nearly Raw Raster Data (NRRD) may be more intuitive for direct computation, but these have been stripped of imaging metadata. Imaging metadata is the essential context to understand why radiomics features from different scanners may or may not be

reproducible [31][13][15][7]. Software libraries are available that easily change from DICOM to NIfTI/NRRD [12] but in keeping with FAIR (Findable, Accessible, Interoperable, and Reusable) data stewardship principles [26], the imaging metadata needs to be preserved in such a way that links to the source images and post acquisition analyses will be retained. A similar argument holds for patients' clinical metadata and extracted radiomics features. Publishing tables of values as open access data does not by itself comply with FAIR principles, because there may be no metadata that richly describe what the data fields are, what its contents signify, and how it relates to other data. The point of FAIR principles is not only humans should grasp enough context about the data to use it meaningfully, but that the data must be made amenable for machine algorithms to automatically search and process, even on a massive global scale. Consider an example specific to radiomics. For a given feature, it is essential to describe how this feature is uniquely defined, which radiomics software (and version) was used to extract it, and what (if any) digital image pre-processing had been applied prior to extraction. Semantic ontologies [9] were developed in order to add descriptive metadata and hierarchical relationships on top of the data. Ontologies make explicit the formal meaning of concepts within its proscribed domain and the essential relationships between its set of concepts. The present work reuses the Radiation Oncology Ontology (ROO) [19], Semantic DICOM ontology (SeDI) [22], and the radiomics ontology (RO) (`https://bioportal.bioontology.org/ontologies/RO`). These ontologies themselves reuse existing terminologies and thesauri, such as the image biomarker standardization initiative (IBSI) [32], National Cancer Institute Thesaurus (NCIT) (`https://bioportal.bioontology.org/ontologies/NCIT`), the units of measurement ontology (UO) (`https://bioportal.bioontology.org/ontologies/UO`),and the DICOM data dictionary (`http://dicom.nema.org/medical/dicom/current/output/html/part06.html`), to identify its concepts.

Other advantages of ontologies include knowledge representation and the support for automated logical inferencing. A hierarchical structure is abstracted as directed acyclic graphs, wherein concepts and relationships are represented as vertices and edges of the graph, respectively. Any graph, regardless of complexity, can be written out in full as a series of machine-readable sentences consisting of strictly three pieces; subject (start vertex) — predicate (edge) — object (end vertex). Such "triples" are the basis of the resource descriptor format (RDF) that is a type of universal data storage format on the World Wide Web. Machine-based data mining and inferencing tasks are thus feasible in a highly efficient manner, being simplified to a "pattern matching" problem. The objective of this open data submission is to stimulate studies into repeatability, reproducibility, replication, and reusability of radiomics features from multiple datasets. The core collection being made publicly available here consists of (a) improvements to the four clinical imaging datasets described in the seminal radiomics publication by Aerts et al. [3] (b) extracted radiomics features described in line with IBSI recommendations [33] and (c) updates to the subject clinical data associated with the aforementioned image collections.

12.2 ACQUISITION AND VALIDATION METHODS

12.2.1 Description of the dataset

The metadata published in this submission links to four image collections, available under a Creative Commons license (Attribution-NonCommercial Unported; CC BY-NC 3.012), in DICOM format on TCIA and has been previously investigated by Aerts et al. [3]. These collections are described in detail elsewhere. In each of these collections, primary Gross Tumour Volumes (GTVs) had been delineated by experienced radiation oncologists; ROIs are included in the TCIA collections as RTSTRUCT and SEGMENTATION objects. In the original TCIA submission, some ROIs were vertically displaced due to the how treatment couch offsets were being reported by legacy

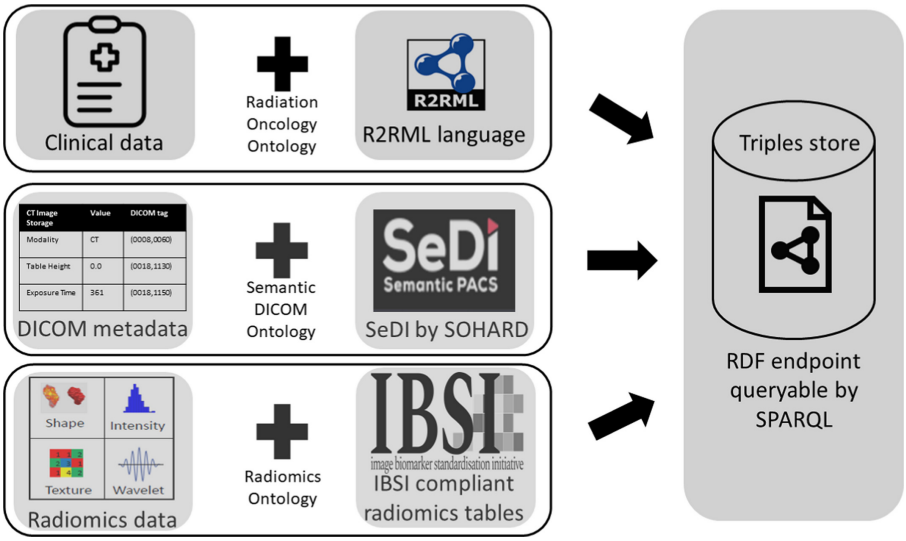


Figure 12.1: Representation of the conversion of the clinical data, digital imaging, and communications in medicine headers and radiomics features to resource descriptor format (RDF). The procedures are outlined in the text sections. The RDF triples can be queried from a publicly accessible endpoint using the SPARQL language.

radiotherapy treatment planning software – these have now been corrected. Clinical data have been extracted from patients' electronic medical records and, where applicable, survival intervals from commencement of radiotherapy treatment till date of death or loss to follow-up were updated using a national registry after internal review board approval. The clinical data have been made available with the imaging collections on TCIA.

12.2.2 Data format and usage notes

The workflow of the conversion of clinical data, DICOM metadata, and radiomics features to RDF triples is represented in Fig. 12.1.

12.2.3 Clinical metadata as RDF

Clinical tables (in CSV format) from TCIA were imported as standard relational databases (e.g., in PostGreSQL) (`https://www.postgresql.org/`) and then converted into RDF triples using a serializing scripting language such as R2RML (`https://www.w3.org/ns/r2rml`). R2RML allows the expression of an arbitrary relational database as an equivalent graph data object using a suitable target ontology (in this case, the ROO) which can be controlled by specifying a mapping file. The values of, and relationships between, the clinical data concepts were mapped onto a graph structure. A visual representation of an example ROO graph has been given by Traverso et al. [19]. The graph was exported as RDF triples and archived on a publicly query-able SPARQL endpoint. The mapping files used for the RDF triples acquisition in this particular data submission are made available for the reader on a public GitLab repository (`https://gitlab.com/UM-CDS/FAIR-compliant_clinical_radiomics_and_DICOM_metadata`).

12.2.4 DICOM metadata as RDF

The DICOM headers present in the abovementioned TCIA image collections were processed into graph objects using SeDI as the target ontology. A research-only version of the Semantic DICOM conversion service of SOHARD GmbH (Fuerth, Germany) was used to automatically extract the headers from DICOM files and subsequently export these as RDF triples to the same aforementioned SPARQL endpoint. This semantic representation of imaging metadata supports cross-referenced queries of DICOM tags against radiomics features for use in repeatability and reproducibility studies [22].

12.2.5 Radiomics metadata as RDF

The radiomics feature values of the primary GTV in the abovementioned image collections were extracted using the

Ontology-Guided Radiomics Analysis Workflow (O-RAW) [17] a PyRadiomics [21] — based FAIR-ification tool. Acquisition of the radiomics RDF triples required a two-stage process. The results of a radiomics extraction software application (in our case O-RAW, but the same holds for other software) must first be transferred into a set of inter-related tables needed for the IBSI. For this submission, we prepared a python script to fill these tables more efficiently; this is provided as an example for the reader on the repository (`https://gitlab.com/UM-CDS/FAIR-compliant_clinical_radiomics_and_DICOM_metadata`). Details of radiomics ontology development and its integration with the IBSI exceed the scope of this data article, but will be covered in detail in a separate publication. Radiomics RDF triples were saved to the same aforementioned SPARQL endpoint.

12.2.6 SPARQL public endpoint

The SPARQL query language is used to interrogate the clinical, DICOM, and radiomics triples that are archived in RDF as a publicly accessible internet resource referred to by the Universal Resource Locator (URL), (`http://sparql.cancerdata.org/`). The RDF triples are maintained in a persistent online graph database through a Blazegraph (`https://blazegraph.com/`) software application, which also supplies a user interface through which remote SPARQL queries may be entered. A public query may be executed as follows: after accessing the above URL, the Namespaces tab is selected and Nat_Com_Collections_final database is set to use. Queries may then be typed by hand or copy-pasted in the Query tab.

12.2.7 Example SPARQL queries

The first hypothetical example we consider is a researcher who wishes to get the data for a univariate model for overall survival in the Lung1 collection, such as Welch et al. [25], using a single radiomics feature

patientID	Fmorph_vol	Funits	deathStatus	time	Tunits
LUNG1-375	400106.666666666674	<http://localhost/rdf/unit_mm(3)>	1	120.0	<http://localhost/rdf/patient_LUNG1-375/days>
LUNG1-019	114154.666666666669	<http://localhost/rdf/unit_mm(3)>	1	336.0	<http://localhost/rdf/patient_LUNG1-019/days>
LUNG1-301	128059.0	<http://localhost/rdf/unit_mm(3)>	1	217.0	<http://localhost/rdf/patient_LUNG1-301/days>
LUNG1-374	38801.666666666666	<http://localhost/rdf/unit_mm(3)>	1	10.0	<http://localhost/rdf/patient_LUNG1-374/days>
LUNG1-317	13483.0	<http://localhost/rdf/unit_mm(3)>	0	3362.0	<http://localhost/rdf/patient_LUNG1-317/days>
LUNG1-320	145931.0	<http://localhost/rdf/unit_mm(3)>	1	544.0	<http://localhost/rdf/patient_LUNG1-320/days>
LUNG1-324	51210.333333333334	<http://localhost/rdf/unit_mm(3)>	1	1963.0	<http://localhost/rdf/patient_LUNG1-324/days>
LUNG1-079	41461.666666666666	<http://localhost/rdf/unit_mm(3)>	1	255.0	<http://localhost/rdf/patient_LUNG1-079/days>
LUNG1-389	20616.666666666668	<http://localhost/rdf/unit_mm(3)>	1	371.0	<http://localhost/rdf/patient_LUNG1-389/days>
LUNG1-315	11306.333333333336	<http://localhost/rdf/unit_mm(3)>	1	313.0	<http://localhost/rdf/patient_LUNG1-315/days>

Figure 12.2: Example of a SPARQL query for matching a radiomics feature called "Fmorph.vol" in the IBSI terminology to the overall survival status and survival time of the patients in the LUNG1 collection. Purely for illustrative purposes, we limited the rows of output to 10. The result of the query is shown in Fig. 12.3.

that is known by its IBSI text label "Fmorph.vol." We have setup the example query in Box 12.2. In brief, a SPARQL query consists of:

- Shorthand prefixes for namespaces referring to data, schema, syntax, and ontologies that are needed;
- SELECT and FILTER commands that allow us to shape the contents to be returned; and,
- a sequence of pattern matching rules that allow us to link patients to radiomics features and overall survival outcome.

The contents of Box 1 may be copied and pasted into the query window of Blazegraph (`http://sparql.cancerdata.org/#query`). Note that a patient study identifier links both the radiomics and clinical triples, such that we can query into both domains and cross-reference them within a single SPARQL query. The result of this example query that is limited (for display purposes) to ten subjects can be seen in Fig. 12.3. As another purely radiomics-based example, we may examine if distinct radiomics intensity discretization algorithms had been used during the extraction of a radiomics feature. If one were to execute the example query in Box 2, it would be seen that the specific radiomics

```
prefix rr: <http://www.w3.org/ns/r2rml#>
prefix ex: <http://example.com/ns#>
prefix sty: <http://purl.bioontology.org/ontology/STY/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix roo: <http://www.cancerdata.org/roo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix uo: <http://purl.obolibrary.org/obo/UO_>
prefix ro: <http://www.radiomics.org/RO/>

SELECT ?patientID ?Fmorph_vol ?Funits ?deathStatus ?time ?Tunits
WHERE {
 ?patient a ncit:C16960.                  #locate objects that are patients (unique ID is C16960 in the NCIT
 ?patient roo:P100042 ?patientID.         #match patients to a literal value which will be a research study ID
 ?patient ro:P00088  ?featureObj.         #match the patients to the corresponding objects in the radiomics
domain

 ?featureObj roo:100042 ?Fmorph_vol; roo:100027 ?Funits FILTER contains(str(?featureObj), "Fmorph.vol").
                                  #return only features called "Fmorph.vol" according to IBSI terminology
                                  #retrieve a metadata label indicating if the feature has any associated physical units

 ?patient roo:P100254 ?death.                    #locate patients that has a clinical "finding" for death by any
cause
 ?death roo:P100042 ?deathStatus.                #retrieve the literal value for the clinical finding as a death
status
 ?patient roo:has ?survivaldayssinceRT.          #retrieve the overall survival time object
 ?survivaldayssinceRT rdf:type ncit:C125201; roo:P100042 ?time; roo:P100027 ?Tunits.
                                  #obtain the value of the survival time interval
                                  #retrieve a metadata label indicating the time interval physical units

 FILTER regex(?patientID, "^LUNG1").
          #purely for the example, we only consider the patients in the LUNG1 collection
}
LIMIT 10  #purely for the example, we have limited the number of rows of output to 10
```

Figure 12.3: The result of ten patients' cases of the example query given in Box 1. We can see the research study IDs of patients from the public The Cancer Imaging Archive collections, the value of a radiomics feature, the value of the survival time, and the vital status of each patient. Additionally, we have displayed the units of the radiomics feature (if any, in this case it is cubic millimetres) and the survival time (days).

feature labelled as RO:Y1RO40 had been computed with 12 unique feature extraction settings, but only three discretization settings were used, all of which employed a fixed bin size (FBS) method. In our final example, we bring elements of the previous examples together into a single SPARQL query that cross-references DICOM, radiomics, and clinical follow-up. In the example provided in Box 3, we index the imaging modality (CT) with its Series Instance UID and Slice Thickness to the subset of morphological (ROI-dependent) radiomics features that were computed for the Lung1 dataset, along with the corresponding survival time and survival status.

```
prefix rr: <http://www.w3.org/ns/r2rml#>
prefix ex: <http://example.com/ns#>
prefix sty: <http://purl.bioontology.org/ontology/STY/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix roo: <http://www.cancerdata.org/roo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix uo: <http://purl.obolibrary.org/obo/UO_>
prefix ro: <http://www.radiomics.org/RO/>

SELECT ?patientID ?Fmorph_vol ?Funits ?deathStatus ?time ?Tunits
WHERE {
 ?patient a ncit:C16960.                  #locate objects that are patients (unique ID is C16960 in the NCIT
 ?patient roo:P100042 ?patientID.         #match patients to a literal value which will be a research study ID
 ?patient ro:P00088  ?featureObj.         #match the patients to the corresponding objects in the radiomics
domain

 ?featureObj roo:100042 ?Fmorph_vol; roo:100027 ?Funits FILTER contains(str(?featureObj), "Fmorph.vol").
                                  #return only features called "Fmorph.vol" according to IBSI terminology
                                  #retrieve a metadata label indicating if the feature has any associated physical units

 ?patient roo:P100254 ?death.                    #locate patients that has a clinical "finding" for death by any
cause
 ?death roo:P100042 ?deathStatus.                #retrieve the literal value for the clinical finding as a death
status
 ?patient roo:has ?survivaldayssinceRT.          #retrieve the overall survival time object
 ?survivaldayssinceRT rdf:type ncit:C125201; roo:P100042 ?time; roo:P100027 ?Tunits.
                                  #obtain the value of the survival time interval
                                  #retrieve a metadata label indicating the time interval physical units

 FILTER regex(?patientID, "^LUNG1").
          #purely for the example, we only consider the patients in the LUNG1 collection
}
LIMIT 10  #purely for the example, we have limited the number of rows of output to 10
```

Figure 12.4: Example of a SPARQL query for examining the different intensity discretization algorithm (i.e., histogram binning) for textural radiomics feature for a single arbitrarily selected subject in the Head–Neck1 collection.

```
prefix rr: <http://www.w3.org/ns/r2rml#>
prefix ex: <http://example.com/ns#>
prefix map: <http://mapping.local/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix roo: <http://www.cancerdata.org/roo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix ro: <http://www.radiomics.org/RO/>

SELECT DISTINCT ?paramspace ?discretisationparam ?discretisationAlgorithm
WHERE{
 ?patient a ncit:C16960.
 ?patient roo:P100042 ?patientID.
 ?patient ro:P00088 ?featureObj.

 ?featureObj rdf:type ro:Y1RO.
 #the Radiomics Ontology defines "ro:Y1RO" as a grey-level size zone matrix textural feature, specifically grey-
level nonuniformity normalized
 # i.e.
https://bioportal.bioontology.org/ontologies/RO/?p=classes&conceptid=http%3A%2F%2Fwww.radiomics.org%2
FRO%2FY1RO
 #the same feature is called Fszm.glnu.norm according to the IBSI terminology.

 ?featureObj ro:P00578 ?paramspace.                    #obtain the feature parameter space
 ?paramspace ro:P00009 ?discretisationparam.   #for each feature parameter space, what intensity discretization
algorithm was used
 ?discretisationparam ro:P0295212521 ?discretisationAlgorithm.

          #for a given discretization settings, what type of algorithm was used

 FILTER regex(?patientID, "^HN1067").                    #purely for this example, we arbitrarily selected one
subject to examine
}
```

Figure 12.5: Example of a SPARQL query for directly cross-referencing DICOM headers, radiomics features, and survival outcome into a single query. The result of the query is shown in Fig. 12.6.

```
prefix rr: <http://www.w3.org/ns/r2rml#>
prefix ex: <http://example.com/ns#>
prefix sty: <http://purl.bioontology.org/ontology/STY/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
prefix roo: <http://www.cancerdata.org/roo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix ro: <http://www.radiomics.org/RO/>
PREFIX sedi: <http://semantic-dicom.org/dcm#>
PREFIX seq: <http://semantic-dicom.org/seq#>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?patientID ?seriesUID ?modality ?sliceThickness ?featureObj ?Fvalue ?time  ?deathStatus
WHERE {
   ?patient rdf:type ncit:C16960.
   ?patient roo:P100042 ?patientID FILTER regex(?patientID, "^LUNG1-").
   ?patientSedi sedi:ATT00100020 ?patientID. #the patient research ID is used to link across to the DICOM
headers

   # Get DICOM study (linked to this patient)
   ?patientSedi sedi:hasStudy ?study.
   ?study sedi:ATT0020000D ?studyUID.
   OPTIONAL { ?study sedi:ATT00081030 ?studyDesc. }

   # Get the DICOM series (linked to this study)
   ?study sedi:containsSeries ?series.
   ?series sedi:ATT0020000E ?seriesUID;
         sedi:ATT00080060 ?modality FILTER regex(?modality, "^CT$").
   OPTIONAL { ?series sedi:ATT0008103E ?seriesDesc. }

  # Get the radiomics features defined as grey-level size zone matrix non-uniformity normalized
  #(linked to this patient)
   ?patient ro:P00088 ?featureObj.
            ?featureObj ro:P00578 ?paramspace; roo:100042 ?Fvalue FILTER regex(str(?paramspace),
"FeatureParameterSpace_1$").

   ?patient roo:P100254 ?death.
   ?death roo:P100042 ?deathStatus.
   ?patient roo:has ?survivaldaysinceRT.
   ?survivaldaysinceRT rdf:type ncit:C125201; roo:P100042 ?time.

   # Get image objects (image objects or RTStruct objects)
   ?series ?contains ?image.
   FILTER (?contains IN (sedi:containsImage, sedi:containsStructureSet)).
   ?image sedi:ATT00080018 ?sopInstanceUID.
            ?image sedi:ATT00180050 ?sliceThickness.

            # Additional series info (not always available in every combination)
   ?equipmentObj sedi:isEquipmentOf ?series.
   OPTIONAL { ?equipmentObj sedi:ATT00080070 ?manufacturer }
   OPTIONAL { ?equipmentObj sedi:ATT00081090 ?model }
} LIMIT 100
```

Figure 12.6: A partial snapshot of the example query given in Box 3. Given as a result of the query are: the subject research ID, the computed tomography series instance unique identifier (UID), the imaging modality and the slice thickness. Each of these are associated with 13 distinct morphological feature concepts (in column featureObj) and the numerical value of each radiomics feature (in column Fvalue). The digital imaging and communications in medicine and radiomics data are cross-referenced to the vital status and survival time interval as per the example in Box 1.

12.3 DISCUSSION

12.3.1 Advantage of using ontologies and storing data on the World Wide Web

Patients' data and specifically demographics or clinical details play a crucial role in prediction modelling studies. Transparent and reproducible radiomics research requires availability of data and metadata associated with a particular study. In the case of prediction modelling, these tend to be source images and the clinical outcomes, for example, survival status and survival time interval. One of the ways to render data FAIR and easily available to be queried remotely over well-established World Wide Web technology is to archive them as RDF data on a persistent online SPARQL endpoint. This requires existing domain ontologies in order to unambiguously define concepts, and relationships between concepts, by mapping them to standardized terminology. The use of publicly defined ontologies and machine-readable lexicons overcome the potential barriers of human language understanding and unknown data encodings. The ontologies further apply some level of knowledge representation that follows in the tracks of human logic and inferencing, such that we can use machine-based queries to discover and process data, without having to first develop extensive knowledge of the relational database structure of the original data. Lastly, we were able to exploit the intrinsically federated pattern matching nature of SPARQL queries to show how to efficiently cross-reference data from across the clinical, DICOM header, and radiomics domains.

12.3.2 Potential applications

By making this data available on the SPARQL endpoint, we offer a version of the combined DICOM data, clinical information, and radiomics features in a manner that is in closer alignment with FAIR data principles. In this way, we hope to facilitate the investigation of radiomics reproducibility research across different institutions,

each of which may speak different human languages, use different imaging protocols, and extract radiomics features in subtly different ways. The queries demonstrated here work in the same way even if these RDF data had been partitioned over multiple databases, irrespective of its geographical location. As has been shown in other publications, the proposed methodology here can be used prospectively for exchanging radiomics prediction models for training or validation, in accordance with a paradigm known as distributed (or equivalently, federated) machine learning [18][5][16]. We have provided examples of SPARQL queries, primarily as a form of guidance notes on how to use this data submission. We would encourage the academic community to adjust them according to their own questions and potentially utilize this methodology for multicentre studies. Radiomics researchers that derive immediate benefit from this open resource could be data scientists and medical physicists with some database query experience. Publishing this as a semantic web resource allows real-time queries and answers about the data. This follows an overall trend toward a growing amount of linked open data with on-demand access. Online SPARQL tutorials are available: (`https://www.w3.org/2009/Talks/0615-qbe/`), (`https://jena.apache.org/tutorials/sparql.html`), (`http://www.ontobee.org/tutorial/sparql`). We anticipate that the aforementioned audience could build user-friendly search interfaces on top of this resource, so as to make it more easily used by others with less programming experience. The reusability of the datasets is strongly supported by the usage of publicly available ontologies, such that the reader is able to look up the ontologies online to search for concepts of interest to them. We have also shared mapping files and RDF conversion scripts on a public code repository, that can also be reused in future.

12.3.3 Limitations of the present submission

One of the major and potentially time-consuming tasks on the way to publishing the RDF data is the mapping of data fields and data values. We have tried to streamline the process in the current submission by preparing mapping files as templates and, wherever possible, using scripting to control serialization applications such as R2RML. However, it is acknowledged that there is no single universally "correct" mapping to a given target ontology. It is likely that persons working independently could apply the same ontologies but produce quite different (and potentially incompatible) knowledge representations. In the analogy of graphs, there is no single unique graph to represent a given dataset; it is possible to derive many different such graphs that are still logically plausible. In semantic data circles, this is well-known as the "open-world" paradigm that is commonly expressed as "anyone can say anything about anything." The solution of such a problem is not up to any one piece of investigation nor any one data scientist. As with all conventions and normative standards in healthcare, convergence gradually emerges over time through numerous cycles of usage, refinement, and dissemination. Our methodology and RDF database are therefore not static, so it is intended to be improved and refined together with developing methodology over time.

12.3.4 Possibilities for future development

The question of comparing and then reconciling different data graphs is an ongoing and active line of research in data science. These so-called shape expressions do not fall within the present scope of submission, but could lead to promising opportunities for improvement. This potentially makes it possible to query data graphs independently of the norms assumed by its publisher. There is also strong research activity toward stricter standardization of data collection and top-down imposition of knowledge representation. Unlike the approach used in this work, where we the first had the data and then cast it toward a target ontology, the top-down approach requires data elements and a

data structure to be rigidly defined first of all before the data are collected. This would be very useful for mapping prospective data, but it is less clear how such rigid standards should be applied to legacy data and retrospective studies. Research is currently in progress toward a modular mapping process, where mappings for generic information that is common for many disease types (e.g., patient demographics) can be rigidly defined and reused often. At the opposite end, highly study-specific mappings may need to be more dynamic or performed on an ad hoc basis. Modular and piece-wise reusable mappings for closely related disease types may significantly reduce the overall RDF preparation time, however, at time of writing such a modular process was not yet ready.

12.4 CONCLUSIONS

We have updated and improved four imaging datasets on TCIA. We converted and published clinical data, radiomics features and DICOM headers as online RDF from these four datasets using ontologies and standard web technology. These RDF triples are stored in a public endpoint giving an opportunity to the radiomics community to query these datasets using the SPARQL language. We have demonstrated the realizability of this approach of making the combined data available as FAIR data, in order to incentivize multicentre research into reproducibility of radiomics features across multiple datasets.

Bibliography

[1] HJ Aerts, ER Velazquez, RT Leijenaar, C Parmar, P Grossmann, S Cavalho, J Bussink, R Monshouwer, B Haibe-Kains, D Rietveld, et al. Data from nsclc-radiomics [data set]. *The Cancer Imaging Archive. Available online: https://doi. org/10.7937 K*, 9, 2019.

[2] HJWL Aerts, E Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, S Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Data from nsclc-radiomics. *The cancer imaging archive*, 2015.

[3] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[4] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Pub-

lic Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013.

[5] Timo M. Deist, A. Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Andre Dekker, and Philippe Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*, 4:24–31, June 2017.

[6] Leonard Fass. Imaging and cancer: a review. *Molecular oncology*, 2(2):115–152, 2008.

[7] Xenia Fave, Molly Cook, Amy Frederick, Lifei Zhang, Jinzhong Yang, David Fried, Francesco Stingo, and Laurence Court. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 44:54–61, September 2015.

[8] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[9] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, June 1993.

[10] Yanqi Huang, Zaiyi Liu, Lan He, Xin Chen, Dan Pan, Zelan Ma, Cuishan Liang, Jie Tian, and Changhong Liang. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (i or ii) non—small cell lung cancer. *Radiology*, 281(3):947–957, 2016.

[11] Ruben T H M Larue, Gilles Defraene, Dirk De Ruysscher, Philippe Lambin, and Wouter van Elmpt. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *The British Journal of Radiology*, 90(1070):20160665, February 2017.

[12] Xiangrui Li, Paul S Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of neuroscience methods*, 264:47–56, 2016.

[13] Dennis Mackin, Xenia Fave, Lifei Zhang, Jinzhong Yang, A. Kyle Jones, Chaan S. Ng, and Laurence Court. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PloS One*, 12(9):e0178524, 2017.

[14] Ji Eun Park, Seo Young Park, Hwa Jung Kim, and Ho Sung Kim. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean journal of radiology*, 20(7):1124–1137, 2019.

[15] Muhammad Shafiq-ul Hassan, Geoffrey G. Zhang, Dylan C. Hunt, Kujtim Latifi, Ghanim Ullah, Robert J. Gillies, and Eduardo G. Moros. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *Journal of Medical Imaging*, 5(01):1, December 2017.

[16] Zhenwei Shi, Kieran Foley, Juan Pablo De Mey, Emiliano Spezi, Philip Whybra, Tom Crosby, Johan van Soest, Andre Dekker, and Leonard Wee. External validation of radiation-induced dyspnea models on esophageal cancer radiotherapy patients. *Frontiers in oncology*, 9:1411, 2019.

[17] Zhenwei Shi, Alberto Traverso, Johan Soest, Andre Dekker, and Leonard Wee. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). *Medical Physics*, page mp.13844, October 2019.

[18] Zhenwei Shi, Ivan Zhovannik, Alberto Traverso, Frank J. W. M. Dankers, Timo M. Deist, Petros Kalendralis, René Monshouwer, Johan Bussink, Rianne Fijten, Hugo J. W. L. Aerts, Andre Dekker, and Leonard Wee. Distributed radiomics as a signature validation

study using the Personal Health Train infrastructure. *Scientific Data*, 6(1):218, December 2019.

[19] Alberto Traverso, Johan van Soest, Leonard Wee, and Andre Dekker. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. *Medical Physics*, 45(10):e854–e862, October 2018.

[20] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[21] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[22] Johan Van Soest, Tim Lustberg, Detlef Grittner, M. Scott Marshall, Lucas Persoon, Bas Nijsten, Peter Feltens, and Andre Dekker. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Studies in Health Technology and Informatics*, 205:166–170, 2014.

[23] Janna E. van Timmeren, Ralph T. H. Leijenaar, Wouter van Elmpt, Jiazhou Wang, Zhen Zhang, André Dekker, and Philippe Lambin. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*, 2(4):361–365, December 2016.

[24] L Wee and A Dekker. Data from head-neck-radiomics-hn1 [data set]. *The Cancer Imaging Archive.*, 2019.

[25] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui

Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[26] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[27] Jia Wu, Khin Khin Tha, Lei Xing, and Ruijiang Li. Radiomics and radiogenomics for precision radiotherapy. *Journal of Radiation Research*, 59(suppl_1):i25–i31, March 2018.

[28] Bin Yang, Lili Guo, Guangming Lu, Wenli Shan, Lizhen Duan, and Shaofeng Duan. Radiomic signature: a non-invasive biomarker for discriminating invasive and non-invasive cases of lung adenocarcinoma. *Cancer management and research*, 11:7825, 2019.

[29] Stephen S F Yip and Hugo J W L Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, July 2016.

[30] Binsheng Zhao, M Kris, and L Schwartz. Data from rider lung ct. the cancer imaging archive, 2015.

[31] Ivan Zhovannik, 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*, 19:33–38, November 2019.

[32] Alex Zwanenburg, Stefan Leger, Martin Vallières, Steffen Löck, and for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative. *arXiv:1612.07003 [cs]*, December 2016. arXiv: 1612.07003.

[33] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J. R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*, 295(2):328–338, May 2020.

13

Distributed radiomics as a signature validation study using the Personal Health Train infrastructure

Adapted from: **"Distributed radiomics as a signature validation study using the Personal Health Train infrastructure"**. Z Shi, I Zhovannik, A Traverso, FJWM Dankers, TM Deist, P Kalendralis, R Monshouwer, J Bussink, R Fijten, HJWL Aerts, A Dekker, L Wee. Scientific data 6 (1), 1-8. (2019). Contribution: second authorship, radiomic analysis, infrastructure development, manuscript writing.

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. In a previous study, these were referred to as Lung1 (n=421) and Lung2 (n=221). 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. 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). 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.

13.1 INTRODUCTION

Images from radiological examinations are presently one of the largest underutilized resources in healthcare "big data" [18]. 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 [8][13][14]. Radiomic data have now been investigated in CT [5][21][10], magnetic resonance imaging (MRI) [19][27] and positron emission tomography (PET) [15][7]. 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 [3][23][29][20]. 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. 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`)[4], 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 [28], known as Personal Health Train (PHT; `https://vimeo.com/143245835`)[24] has been successfully used to develop and validate models on non-image clinical data [12][11]. 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)[25]. 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) [30]. 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). The first of these datasets (hereafter referred to as "Lung1") [1] 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.

13.2 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 13.1. 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 Fig. 13.2). Within each cohort, the subjects were stratified into two risk groups, based on the median of the risk score distribution in Lung1. 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). 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`). 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.

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

Figure 13.1: 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.

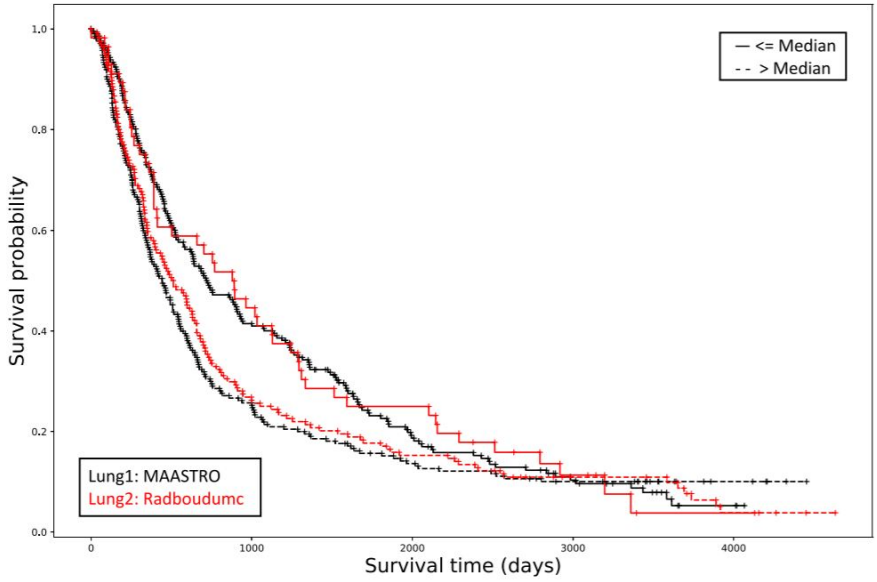


Figure 13.2: The performance of radiomic signature ZS2019 according to Kaplan-Meier survival analysis. The signature was developed in Lung1 (MAASTRO; black line) and then distributedly validated in Lung2 (Radboudumc; red line). The upper and lower survival curves were split according to the median of 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.

13.3 DISCUSSION

In this paper, a model (ZS2019) derived from radiomic features and overall survival locally within one institution was able to be exchanged inter-operably 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. 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 [2]. 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 (DataSHIELD28 and WebDISCO29) 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. 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. 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`)[26], WebDisco (`https://omictools.com/webdisco-tool`)[17] 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 van Griethuysen et al. [23] 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.

13.4 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.

13.5 METHODS

13.5.1 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.

13.5.2 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.

13.5.3 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.

13.5.4 Data processing

The study steps are shown schematically in Fig. 13.3 for MAASTRO and Radboudumc. The core of the radiomic feature extraction process utilizes free and open-source PyRadiomics [2] (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

Figure 13.3: 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.

PyRadiomics; these were based on the SimpleITK (v1.0.1) toolkit [16]. 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 publication2; 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 [2] that accompanied the original publication:

- Energy from the intensity histogram feature class, which estimates the overall density of the region of interest,

- Compactness from the morphological feature class, which describes the volume of the object relative to that of a perfect sphere,
- 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
- 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).

13.5.5 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 [22], 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.

13.5.6 Distributed approach

The VLP distributed learning architecture has been described in deep detail elsewhere [12][11][6]. 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.

13.5.7 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.

13.5.8 Cox model evaluation

Anonymous timepoints for Kaplan-Meier survival curves 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) [9] 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.

13.5.9 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.

Bibliography

[1] HJWL Aerts, E Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, S Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Data from nsclc-radiomics. *The cancer imaging archive*, 2015.

[2] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), December 2014.

[3] Aditya P Apte, Aditi Iyer, Mireia Crispin-Ortuzar, Rutu Pandya, Lisanne V Van Dijk, Emiliano Spezi, Maria Thor, Hyemin Um, Harini Veeraraghavan, Jung Hun Oh, et al. Extension of cerr for computational radiomics: a comprehensive matlab platform for reproducible radiomics research. *Medical physics*, 45(8):3713–3720, 2018.

[4] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Can-

cer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013.

[5] Thibaud P. Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H. Mak, and Hugo J.W.L. Aerts. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3):345–350, March 2015.

[6] Timo M. Deist, A. Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Andre Dekker, and Philippe Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*, 4:24–31, June 2017.

[7] Kieran G Foley, Robert K Hills, Beatrice Berthon, Christopher Marshall, Craig Parkinson, Wyn G Lewis, Tom DL Crosby, Emiliano Spezi, and Stuart Ashley Roberts. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of pet in patients with oesophageal cancer. *European radiology*, 28(1):428–436, 2018.

[8] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, February 2016.

[9] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[10] Yanqi Huang, Zaiyi Liu, Lan He, Xin Chen, Dan Pan, Zelan Ma, Cuishan Liang, Jie Tian, and Changhong Liang. Radiomics signature: a potential biomarker for the prediction of disease-free sur-

vival in early-stage (i or ii) non—small cell lung cancer. *Radiology*, 281(3):947–957, 2016.

[11] Arthur Jochems, Timo M Deist, Issam El Naqa, Marc Kessler, Chuck Mayo, Jackson Reeves, Shruti Jolly, Martha Matuszak, Randall Ten Haken, Johan van Soest, 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(2):344–352, 2017.

[12] Arthur Jochems, Timo M. Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. 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(3):459–467, December 2016.

[13] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Schabath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, November 2012.

[14] Philippe Lambin, Ralph T. H. Leijenaar, Timo M. Deist, Jurgen Peerlings, Evelyn E. C. de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T. H. M. Larue, Aniek J. G. Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest, Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M. Mottaghy, Joachim E. Wildberger, and Sean Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews. Clinical Oncology*, 14(12):749–762, December 2017.

[15] Ralph TH Leijenaar, Sara Carvalho, Emmanuel Rios Velazquez, Wouter JC Van Elmpt, Chintan Parmar, Otto S Hoekstra, Corne-

line J Hoekstra, Ronald Boellaard, André LAJ Dekker, Robert J Gillies, et al. Stability of fdg-pet radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica*, 52(7):1391–1397, 2013.

[16] Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. The Design of SimpleITK. *Frontiers in Neuroinformatics*, 7, 2013.

[17] Chia-Lun Lu, Shuang Wang, Zhanglong Ji, Yuan Wu, Li Xiong, Xiaoqian Jiang, and Lucila Ohno-Machado. Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219, 2015.

[18] John McKnight, Brian Babineau, and J Gahm. North american health care provider information market size & forecast. *ESG-Enterprise Strategy Group*, 2011.

[19] Ke Nie, Liming Shi, Qin Chen, Xi Hu, Salma K Jabbour, Ning Yue, Tianye Niu, and Xiaonan Sun. Rectal cancer: assessment of neoadjuvant chemoradiation outcome based on radiomics of multiparametric mri. *Clinical cancer research*, 22(21):5256–5264, 2016.

[20] Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, Claire Pellot-Barakat, Michael Soussan, Frédérique Frouin, and Irène Buvat. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Research*, 78(16):4786–4789, August 2018.

[21] Chintan Parmar, Ralph TH Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo JWL Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports*, 5:11044, 2015.

[22] Alberto Traverso, Johan van Soest, Leonard Wee, and Andre Dekker. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. *Medical Physics*, 45(10):e854–e862, October 2018.

[23] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[24] Johan Van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. In *MIE*, pages 581–585, 2018.

[25] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[26] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, 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(5):1372–1382, 2010.

[27] Bin Zhang, Jie Tian, Di Dong, Dongsheng Gu, Yuhao Dong, Lu Zhang, Zhouyang Lian, Jing Liu, Xiaoning Luo, Shufang Pei, et al. Radiomics features of multiparametric mri as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clinical Cancer Research*, 23(15):4259–4269, 2017.

[28] Jane Zhang. Ontology and the semantic web. 2007.

[29] Lifei Zhang, David V Fried, Xenia J Fave, Luke A Hunter, Jinzhong Yang, and Laurence E Court. Ibex: an open infrastructure software platform to facilitate collaborative work in radiomics. *Medical physics*, 42(3):1341–1353, 2015.

[30] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary J. R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela

Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*, 295(2):328–338, May 2020.

14

From multisource data to clinical decision aids in radiation oncology: the need for a clinical data science community

Adapted from: **"From multisource data to clinical decision aids in radiation oncology: the need for a clinical data science community"**. J Kazmierska, A Hope, E Spezi, S Beddar, WH Nailon, B Osong, A Ankolekar, A Choudhury, A Dekker, K Røe Redalen, A Traverso. Radiotherapy and Oncology. https://doi.org/10.1016/j.radonc.2020.09.054. (2020).

Abstract

Big data are no longer an obstacle; now, by using artificial intelligence (AI), previously undiscovered knowledge can be found in massive data collections. The radiation oncology clinic daily produces a large amount of multisource data and metadata during its routine clinical and research activities. These data involve multiple stakeholders and users. Because of a lack of interoperability, most of these data remain unused, and powerful insights that could improve patient care are lost. Changing the paradigm by introducing powerful AI analytics and a common vision for empowering big data in radiation oncology is imperative. However, this can only be achieved by creating a clinical data science community in radiation oncology. In this work, we present why such a community is needed to translate multisource data into clinical decision aids.

14.1 INTRODUCTION

14.1.1 The clinic as a learning health care system

Many large commercial enterprises are redirecting their business approaches to exploit the new knowledge they can gain from the data they collect daily. This strategy arose from the need to mine a large amount of diverse data, often unstructured and coming from multiple sources: so-called "big data." Whereas initially big data seemed to present an obstacle, now it is becoming more evident that leveraging massive data collections using novel techniques can reveal previously undiscovered knowledge [25][17]. These techniques include analytic methods spanning from traditional statistics and hypothesis testing to more advanced algorithms inspired by machine learning (ML), a branch of artificial intelligence (AI), in which powerful computational systems augment our brain's learning capacity by employing complex mathematical algorithms to reveal patterns in data, mainly for the purpose of generating new knowledge [6][31]. AI is not a new concept in oncology. Recent reviews described two major applications of AI in the medical field: automation and decisions' augmentation[2][27]. The former includes applications such as auto contouring of both organs at risk (OAR)s and target volumes; the latter covers the whole spectrum of decision support systems. However, by comparing applications in the enterprise domain, AI is often referred to as "data analytics". In this view, two powerful applications of AI in medicine are often forgotten, namely the ability of AI to retrieve data belonging to multiple sources and spread across different locations. Most of the big-companies' data analytics include powerful tool that can collect data from multiple sources, such as for example social media or health wearables. This part of AI is very useful, because of the intrinsic nature of multisource data: they are sparse and unstructured. The second powerful aspect of AI is that, after having retrieved multisource data, automated QA can be performed. This aspect of AI is often forgotten, but in radiation oncology data quality is fundamental for applying AI in the clinic. After automated QA, which starts from unstructured

data (often referred to as "veracity" of big data), data are transformed into a network of knowledge, the so-called "linked-data", resulting in new knowledge. We believe this broader view of AI is key to a new era in our healthcare systems. Translating this new learning paradigm into radiation oncology will improve the classification of disease and reveal new ways to improve cancer treatment and predict patients' clinical events. Unfortunately, the radiation oncology community lags far behind in the adoption of big data approaches for providing the patient-centered, individualized care often referred to as personalized medicine [48]. With the term radiation oncology community, we do not only refer to radiation oncologists, but the extensive community concerned with treating cancer patients with radiotherapy. The treatment is rarely only consisting of radiotherapy but is often multimodal and consisting of radiotherapy in combination with chemotherapy, immunotherapy and/or surgery involving multiple professionals such as radiologists, pathologists, surgeons, medical physicists, and medical- and radiation oncologists. While the community agrees that the future of medicine as a whole and radiation oncology in particular is in learning health care systems, where data are transformed into new knowledge as part of clinical routine, there remain gaps in our ability to rapidly learn from data generated in the clinic during the course of patient care [2]. By definition a learning health care system is a system that has been designed to generate and apply the best evidence generated from a collaborative effort among patients and care providers." The central point of a learning healthcare system is that knowledge discovery becomes a natural outgrowth of patient care. A learning healthcare system is meant to push forward evidence-based medicine by: a) fast translation from knowledge produced in clinical research to clinical practice; (b) empowerment of a shared responsibility culture between the different stakeholders involved in the clinic; and (c) facilitating engagement of patients and doctors for evidence production and dissemination[14]. In radiation oncology, we still mainly learn by narrowing and simplifying our research questions, in the process often moving them far from the complexity of real-world clinical practice. For example, most support for clinical decisions comes from clinical

trial data. On one hand, clinical trials can provide high-quality data, but on the other hand, they have several major drawbacks: a) their exclusion of patients with complex cases that do not fit their strict inclusion criteria; b) their narrow focus on just one research question or a limited number of questions that often determine the choice of collecting specific variables; c) their high cost; d) the long time required to recruit sufficient patients to reach statistically significant results; and e) their infrequent exploration of how combinations of several factors might influence patients' outcomes. Conversely, patients produce a vast amount of data, from diagnosis to treatment and follow-ups. Only a small percentage of this data is actively used to produce new insights that can push our clinical practice and lessons learnt from clinical trials towards personalized medicine. In this view, big data empowered with AI is not a strategy to substitute randomized clinical trials, but rather a strategy to augment the knowledge from clinical trials. For example, AI can be used to explore multisource big data to better stratify patients and optimize clinical trials enrollment by defining group of patients for which the introduction of a new treatment is more likely to be found beneficial. Finally, by leveraging multisource big data a large spectrum of prognostic as well as confounding factors can be examined. This data integrates and considerably expand the original collections from randomized clinical trials. A recent communitarian effort is being carrying on boosting the efficacy of RCTs. This effort foresees the possibility to increase the "pragmaticism" of RCTs. A detailed review [23] pointed out the prominent role of AI in supporting this translation. By combining the expertise brought by clinical data scientists and medical doctors it is possible to use robust, validated and well understood AI tools to improve trial success rates starting from trial design and preparation (e.g. a better recruitment strategy) to execution." A recent investigation providing updating guidelines for more "pragmatic" RCTs, the SPIRIT-AI [53], is supporting the abovementioned transition and it is currently adopted in recent clinical trial protocols that included AI-driven intervention. These guidelines were needed considering the increasing number of RCTs making use of AI tools. It is important to highlight how specific attention and dedicated

methodologies need to be adopted when performing casual inferencing from both randomized clinical trials and multisource data [55]. The same issues that exist for casual inferencing from observational studies, such as the presence of confounding factors, sampling selection and cross-population biases also exist for inferencing from aggregated data (e.g. multisource data). A recent study [3] recommended the extension of parametric causal inferencing mathematical models specifically developed for clinical trials to non-parametric models specifically developed for aggregated data. The authors claimed that this methodology cannot be separated from the AI analytical tools used to process that data. This is indeed a key point, which need to be combined with the need of improving data quality (see coming statements) for robust casual inferencing. The fuel of a learning health care system in radiation oncology is the data that are generated every day in the clinic. This requires us to reimagine the clinic as a source of big data. Currently, most of the big data generated in the clinic are wasted as a source of research because we have been unable to equip the clinic with big analytics. Nevertheless, we cannot support a learning paradigm in radiation oncology solely by borrowing technologies from other fields, such as business enterprises. The path towards a learning health care system in radiation oncology needs to pass several milestones, which are summarized in Figure 14.1. All these milestones represent shifts in our concept of traditional clinical medicine. These milestones are:

- M1: Understanding the clinic as a source of big data: Where do the data come from and why are the data "big?" Data are not only produced directly by daily routine clinical activities, but also indirectly, for example when researchers process clinical data. This produces a combination of data and metadata, which are logically connected but might be sparse, even within the same institution. Combining and reunifying these data is the largest challenge to be tackled.
- M2: Identifying data types and involved stakeholders. The generation of multisource big data involves different professions and users. It is fundamental to identify not only the users and stake-

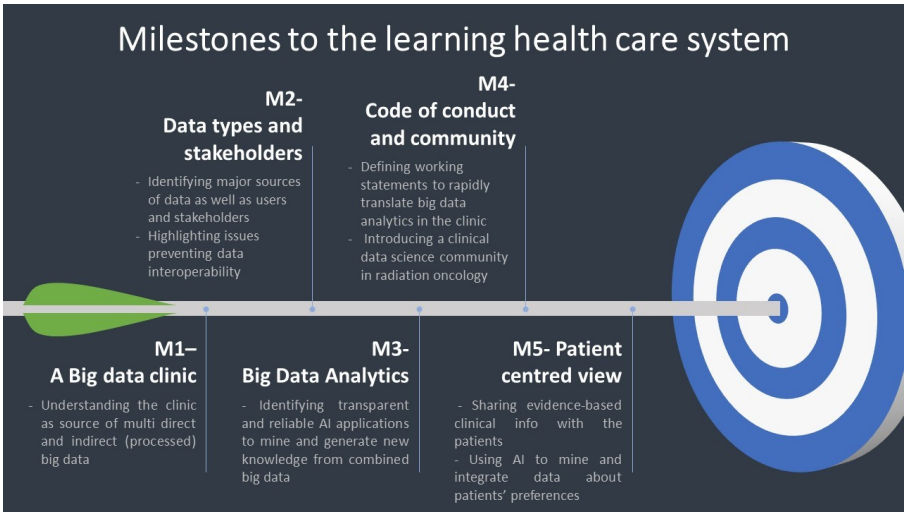


Figure 14.1: Overview of the milestones for moving the clinic towards a learning health care system. Five milestones have been identified. Milestones 1 and 2 involve developing a deep understanding of clinical data as a source of big data and metadata and the need to involve stakeholders and users and to address issues that limit data interoperability. M3 introduces robust and collaborative AI-driven analytics for the development of clinical decision aids. M4 introduces a new clinical data science community for radiation oncology to harmonize existing initiatives and define a code of conduct. M5 introduces a patient-centered view and decision-making processes for the learning health care system.

holders of these data but also the major constraints that limit the interoperability of these data. Interoperability and data housekeeping are the keys for boosting data quality. High-quality data will have a strong impact on the robustness and integrity of our data-driven clinical decisions.

- M3: Defining which data analytics can be used to extract unique insights from multisource data. After we learn how to correctly retrieve, curate, mine, and combine multisource big data, it is possible to use AI as the engine to burn the data fuel. However, use of AI per se does not guarantee success. Strong transparency and robust methodology will enable meaningful applications of AI to discover new knowledge in the data. This methodology should comprise both analytics for verifying data quality, as well as methods for tackling the issues related to causal inferencing from aggregated data.

- M4: Defining working statements and a code of conduct to rapidly translate data analytics into the clinic as decision aids. To fill the gap between AI developments and their translation into the clinic as decision aids, a global effort to involve all the professional figures, stakeholders, and users identified at M2 is needed. This effort requires the creation of a clinical data science community in radiation oncology. Such a community would not replace previous efforts or already-existing focused work and task groups, but instead act as a harmonizer by defining a code of conduct and a shared vision.

- M5: A patient-centered learning health care system. The brand-new learning health care system has to be made patient-centered by a) developing AI analytics to include patient perspective data; b) improving the expandability of AI analytics, and c) using decision aids in combination with shared decision making.

14.1.2 Multisource data, data types, and stakeholders (M1 and M2)

The clinic is a source of big data. Common data types include medical images, electronic health records (EHRs), and patient-reported outcomes [59]. However, the clinic also indirectly produces metadata associated with traditional data types from the algorithms that process data. Examples are quantitative imaging biomarkers and radiomic data (large amounts of features extracted from medical images and analyzed using data characterization algorithms), which generate predictive or prognostic factors from source data. In Table 14.1, we summarize the main types of these highly variable multisource data and provide descriptions of the commonly available formats, the data owners, stakeholders, and users, and issues with or barriers to interoperability of the data.

14.1.3 AI to empower multisource data (M3)

One of the largest issues faced when dealing with multisource big data is that the ability to process these data is beyond our human brain capacity. However, recent developments in AI algorithms have emerged as attractive and much-needed tools to empower multisource data analysis. AI and ML have created opportunities to build powerful computational facilities and a surge in data sharing, data collection, and advanced data mining algorithms. The use of ML algorithms in radiation oncology is rapidly growing; their main applications are quality assurance, organ segmentation, treatment planning, image guidance, motion tracking, and treatment response modeling. However, radiation oncology has not yet fully exploited the enormous potential of AI for analyzing multisource data that integrate variables from time-dependent sources, such as sequential quantitative imaging or genetic biomarkers. These developments could change the classical paradigms for radiotherapy by automating and optimizing clinical processes and quality control to provide decision support for personalized patient care, for instance by altering radiotherapy prescriptions

Data type	Description and Common format	Stakeholders	Data users	Major issues for Interoperability
EHRs	EHRs are computerized medical information systems that collect, store, and display patient information.	Patients Hospitals	Clinicians Nurses Medical physicists Researchers Radiation therapists (RT/RTT) and dosimetrists Administrative staff	Free-text entries
Outcome measures	Data on survival, recurrence, and toxicities are commonly found in the EHR, but when part of clinical trials, data can also be found in spreadsheets or electronic/paper-based case report forms.	Patients Hospitals Clinical trial units Study funders Regulatory agencies	Clinicians Researchers Research nurses Clinical data coordinators	Lack of standardization Free-text entries Consistency and completeness of collected data
Laboratory data	Software and databases used to manage and store results from laboratory tests and pathology data	Patients Hospitals	Clinicians Nurses Engineers Researchers	Lack of standardization in acquisition and analysis Non digital format Image storage, management, transmission and sharing
Genomics	Separate databases for large-scale genomic data	Patients Hospitals	Clinicians Researchers	
Medical images	Medical images are acquired for diagnosis, staging, and treatment planning. The most common modalities include PET, CT, CBCT, MRI, and ultrasonography. Medical images are regulated by a commonly accepted standard (DICOM).	Patients Hospitals	Clinicians Medical physicists Radiation therapists Researchers	Lack of standardization in acquisition and analysis Duplication of data within same Institution
Radiotherapy TPS & Verification Systems	Dose-volume histograms, metrics for radiation dose delivered to the tumors and organs at risk at single treatments and over the whole treatment course, are saved in the TPS. Plan information Dose distribution Treatment delivery data	Patients Hospitals	Clinicians Medical physicists Radiation therapists	Institutional and clinician bias in treatment Data accessibility and full use of data for data analytics
Patient-reported outcomes	Patient-reported outcomes, such as treatment-induced side effects, can be found in the EHR if part of standard treatment. For clinical trials, there may be various types of electronic or paper forms.	Patients Hospitals	Clinicians Nurses Researchers	Lack of standardization Free-text entries Consistency and completeness of collected data
National cancer registries	Population-based registries of cancer incidence, treatment, and outcomes are often recorded in national databases.	Health authorities	Clinicians Researchers Government	Institutional bias diagnosis, treatment, and follow-up. Free-text entries.
Nonmedical information	Environment, income, socioeconomic status, race, ethnicity, education, housing	Local government	Government Researchers	Information bias

Figure 14.2: Summary of the main data types available for multisource data analysis. Abbreviations: EHR, electronic health record; PET, positron emission tomography; CT, computed tomography; CBCT, cone beam computed tomography; MRI, magnetic resonance imaging; TPS, treatment planning system.

and fractionation schedules. Hence, AI-based analysis of multisource data could dramatically change the way radiotherapy is approached and will likely play a central role in the future development of personalized, precision medicine. Despite the great potential of AI, the current situation in radiation oncology is that only a small percentage of the data collected is used for decision-making in the clinic owing to several obstacles that hinder the sharing, processing, and deployment of data in the clinic. By throwing these data in the "trash," we risk losing unique insights that could radically change our clinical practice. We need to realize that human capabilities are not sufficient to process big clinical data and that clinicians need the help of AI to fully translate the large amounts of data collected in the clinic into decision-making about routine clinical practice. Artificial intelligence in clinical care is recently being recognized as a medical device by the FDA, with applications spanning from medical imaging analysis, clinical decision aids and tools to optimize patient care[28][24]. These applications not only apply to the USA, but similar evidence is seen in Europe and Asia. The FDA has developed a complete product lifecycle for AI applications, like what was conceived for medical devices[8]. A nice example of combining big data with AI is presented in the study by Mayo et al, where a decision support system is used to improve dose delivery to spare health tissues [38].

14.1.4 A patient-centered clinical data science community in radiation oncology (M4 and M5)

The key to success at achieving the above-mentioned milestones is the creation of a new community: one focused on clinical data science in radiation oncology. Because multiple stakeholders are involved, the problems of big data cannot be solved by only one professional discipline; instead, they will require a joint effort bringing together broad, multidisciplinary expertise and including clinicians, medical physicists, data scientists, biologists, patients, and other stakeholders. However, instead of proposing an independent community, we recommend

building upon already-existing working groups and task groups that touch these professional roles. To coordinate these communities and working groups and to speed the realization of a learning health care system, we propose the development of a collection of statements and a code of conduct. Finally, we underline the importance of introducing tools that enable not only the collection and elaboration of data reporting patients' perspectives, but also a synergy between clinicians, decision aids, and patients.

14.1.5 Vision and statements

With this position paper, we offer a basis for shifting the current paradigm in radiation oncology towards the clinic as a learning health care system. In the subsequent sections of the paper, we will elaborate on five supporting statements that are fundamental to reach the above milestones. For each statement, we have identified already-existing activities, efforts, or smaller communities that will be our main interlocutors in coordinating these efforts within the community. An overview of the statements is provided in Figure 14.3.

Statement 1: FAIR principles for data management plans

Over the past ten decades, numerous patient registries and databases have been established worldwide. However, few people know of their existence, let alone how to access the information in them. This lack of exposure and accessibility limits the power and, hence, the potential benefits of ML/AI tools, since a model's performance directly correlated with the amount of information it learns (trains) on, as seen in Statement 4. The need to improve the infrastructures that support the use and reuse of all these pieces of information (multisource data) in their respective silos is therefore paramount. To lay the groundwork for accomplishing this, a diverse group of stakeholders—both private and academic—jointly designed a set of principles referred to as the

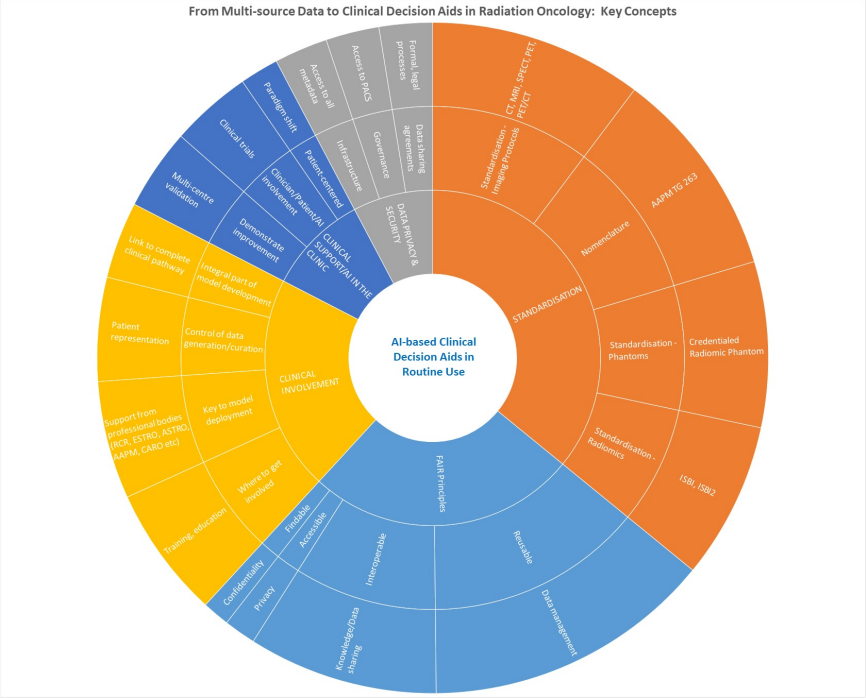


Figure 14.3: Overview of the vision of the community, the milestones and the code of conduct statements.

FAIR data principles [65]. The FAIR principles urge that all data sets should be FAIR: findable, accessible, interoperable, and reusable, while respecting data privacy and patient confidentiality principles. The goal is to improve data (re)use by providing detailed metadata descriptions that are readable to both humans and machines, thereby making data findable, accessible, and interoperable. Therefore, for multisource data to work as intended, all data sources must adhere to the FAIR Guiding Principles and respect data privacy and confidentiality. With the dawn of this new, data-centric era, data can be regarded as the new oil. Like crude oil, which differs in its physical and chemical characteristics from region to region, data also differs from one source or format to another. Because multisource data stem from various sources and are collected using various methods, information can be incomplete, inconsistent, biased, or imprecise. Therefore, in using multisource data, answers to the following five W-questions will facilitate the transparency of the data-generation procedure:

- Who generated the data?
- Why were the data generated?
- When were the data generated?
- Where were the data generated?
- What generated the data?

The phrase ”you are what you eat” applies not only to humans but to models as well. The data science implication of this axiom means that when a model feeds (trains) on ”bad” data, the resulting model is inevitably bad—in other words, less accurate—and the converse is almost always true. Therefore, no matter how complicated a model might be, it will never catch extra information in the data more effectively than a simple, explainable model would. Ensuring the integrity and quality of multisource data is even more essential if inferencing is envisaged. In a multisource AI-driven radiotherapy system, clear guidelines need to be laid out for data stewardship. The health care organization should incorporate metadata containing data provenance at

the source, and proper data lineages should be maintained at all stages of data management. Adopting health care standards that address data stewardship at the source and enable an audit trail from data acquisition to data curation will ensure better traceability for AI models. Finally, data must be interoperable. Data from one source should be semantically as well as syntactically interpretable across different systems. Health care data structure and exchange standards like HL7 FHIR, OHDSI OMOP, and XDS provide the means to structure data in globally acknowledged and accepted formats. Use of clinical coding terminology systems and vocabulary (WHO ICD, ICF, SNOMED CT, LOINC, etc.) should be encouraged. The focus of adopting these standards should be to make the implementations of shared terminologies and vocabularies as generic as possible while permitting specificity as needed. In all cases, including those in which adoption of a health care data exchange standard is not possible, data should be sufficiently supported by metadata.

Data quality and effects on AI applications

Data quality assurance is an essential exercise at all stages of data curation, although the definition of "quality" is context dependent and adopting a single measure to gauge quality is challenging. The elements of data quality are accuracy, completeness, consistency, credibility, and timeliness. By accuracy, we mean that the intended value of the data is both correct and unambiguous. A very preliminary way of ensuring accuracy at the source is by using validation rules at the time of data acquisition. However, as data are shared across domains, a validation rule can itself become inconsistent, thereby increasing the chance of the data's being inaccurate. Jack Olson, in his book Data Quality: The Accuracy Dimension, argues that data can never be 100% accurate [49]. This is because the content of data can be validated against permissible values but not against the actual occurrence. He gives the example of how the value "brown" for eye color can be a valid entry but not an accurate one, simply because the person's eye

color may actually be blue. Thus, inaccuracy can create bias in AI systems that may affect clinical outcomes. However, the very AI systems that demand quality data for better decision support may in fact contribute to improving the quality of data at the source. While AI is largely seen as a tool to extract value from data, it can also act as an instrument to add value back to the data.

Statement 2: Standardization of methodologies

Standardization of radiomic algorithms

The lack of standardization in image processing methods and quantitative radiomic feature extraction, as well as the lack of validated and verifiable reference values, is hampering the clinical implementation of quantitative radiomic imaging biomarkers [68][15]. A lack of standardized, consistent, clear, and sufficient detail in reporting radiomic features, in addition to intrinsic issues with repeatability and reproducibility [60], make radiomic findings difficult to reproduce [69] and trust. Standardization of radiomic algorithms and image processing pipelines is essential for the development of the field and should be strongly encouraged. The approach proposed by the Image Biomarker Standardization Initiative [70], which includes the standardization of a set of 174 radiomic features, the definition of a general radiomics image processing scheme, and the publication of imaging data sets and reference feature values (`https://theibsi.github.io`), is the most advanced effort to date. It is expected to continue and to be widely accepted and used in the future [20].

Standardization of image acquisition and phantoms

To be useful, quantitative imaging biomarkers must be both repeatable and reproducible [46]. Radiomic features are affected by acquisition, reconstruction, and image preprocessing settings [18][37][36][9][63].

This applies to all imaging modalities. Standardization and harmonization of imaging procedures are essential requirements for the development of robust, repeatable, and valid imaging biomarkers. The standardization of imaging protocols for all imaging modalities (computed tomography [CT], magnetic resonance imaging, magnetic resonance spectroscopy, single-photon emission CT, and positron emission tomography) should be strongly encouraged within the same institution and across different institutions, as it would facilitate the interoperability of quantitative imaging biomarkers [22][13]. This would be particularly important when assessing treatment or tumor response on a large scale and as part of clinical trials [45]. The standardization of imaging protocols should be accompanied by the development of new phantoms specifically designed to address the challenge of providing reproducible reference values for textural features [37].

Standardization of nomenclature within the radiation oncology community

The adoption of a standard radiation oncology nomenclature would enable and facilitate extraction and sharing of all types of data from EHRs across different institutions, states, provinces, countries, and continents. Such a standardized nomenclature would support large international clinical trials, ease collaborations across borders, and contribute to improvements in clinical practice and patient care [41]. Moving forward, it is essential that new clinical trial protocols use standardized nomenclatures for capturing their data. The question remains which standard nomenclature should be used. At present, the standardized nomenclature for radiotherapy proposed by the American Association of Physicists in Medicine (AAPM) Task Group (TG) 263 [39] seems to be the most likely standard to become accepted and widely used. The Global Quality Assurance of Radiation Therapy Clinical Trials Harmonization Group (`https://rtqaharmonization.org`) has recently unified the contouring of organs at risk by compiling, in line with AAPM TG 263

and the American Society for Radiation Oncology (ASTRO), guidance for delineation and a standard nomenclature for integration into clinical trial protocol [41].

Statement 3: Privacy-preserving collaborative big data infrastructures

FAIR data principles ensure that data are syntactically and semantically interoperable, thereby promoting seamless data sharing among health care providers. While sharing patient-level data for better decision making is important, protecting patient privacy is essential. Ethical, legal, and societal issues regarding data sharing bar hospitals and clinics from sharing data. When there is too little data shared, ML and AI technologies starve themselves with little or no data. However, if we cannot bring data to the algorithms, it is possible to send algorithms to the data. Infrastructures built around these data silos can connect and provide a way to send algorithms to the data sources. While the data stays well protected within the jurisdiction of the healthcare provider or the patient themselves, the algorithms via the infrastructure can fetch results. This way privacy of patient data is protected while at the same time, research is promoted. This section explores big data infrastructures that enables privacy preserving collaborative research. We talk about two types of infrastructure: centralized collaborative big data infrastructures and federated big data infrastructures. In centralized infrastructures, different hospitals and healthcare providers enter a collaboration and upload patient data to a secured centralized repository. Researchers can use data from the repository either train their algorithms and perform analysis. Additionally, the centralized repository can also provide a compute environment where algorithms can be sent and computation performed. A researcher initiating a data analysis process will have no direct access to the data and can only retrieve the result of the analysis. However, it is important to mention that the data will still be located outside individual hospitals. As such, patient's consent in sharing data to a central repository and use for secondary purposes

needs to be addressed properly. Another initiative, Informatics for Integrating Biology and the Bedside (I2B2), aims to integrate data from different biomedical disciplines and to deliver these data to researchers. I2B2 provides tools and frameworks for merging and linking genomic and biological data to clinical data in a health data warehouse (`https://i2b2.cchmc.org/`) [12]. Similarly, the HMO Cancer Research Network connects more than 11 million patient records from 14 health care providers in a virtual data warehouse (`https://healthcaredelivery.cancer.gov/crn/`). Federated big data infrastructure emphasizes keeping the data at the source while pushing the analytics to the source. Each hospital maintains a local data repository to which researchers can send their algorithms and from which they can fetch results. In a collaborative environment, each hospital would act as an individual data provider, generating sets of results that can then be aggregated to obtain global results. An example is the Personal Health Train (PHT) [10]. PHT shifts the focus from sharing data to sharing algorithms to the source of the data, essentially within the jurisdictional environment of the hospital. A researcher using the infrastructure is agnostic about the data schema and distribution at the source and as such relies heavily upon the FAIR data principles. Each hospital hosts a data station containing FAIR data, and provides a computation environment for the train (metaphor for package containing algorithms and data retrieval query). PHT is platform independent and the researcher can autonomously choose the technology for implementing the algorithm (e.g., Python, R, Matlab, Java). The communication between the data stations and researchers occurs through a secured and centralized message broker. Study showed that PHT is scalable so that federated, privacy-preserving analyses involving many thousands of patients can be conducted [19][19][54][16]. Another example is DataSHIELD, a collaborative and privacy-preserving data analysis environment connecting multiple hospitals. This infrastructure enables researchers to send algorithms to the data without having to retrieve data locally. Unlike PHT, DataSHIELD sends algorithms packaged in the R statistical programming environment to an Opal database hosted

at each hospital. PCORnet is a network of several clinical research institutes that supports pragmatic trials and comparative effectiveness research across one or several of the participating institutes [42]. More recently, MedCO provided a privacy preserving federated data analysis platform (`https://medco.epfl.ch/`). MedCO focusses on keeping data at the source and provides multi-party homomorphic encryption to all data sources, providing an additional layer of security and privacy. OpenSAFELY initiative in the UK enables trusted analysts to run large scale computation on pseudonymised patient records inside environments managed by electronics health records software company (`https://opensafely.org`). It is important to mention that creating a collaborative environment connecting many different hospitals and clinics while preserving patient data privacy is a multifaceted challenge. Keeping data at the source may not be sufficient when the amount of data is small. The infrastructure needs to be adaptive, flexible, scalable, and secure. It should be transparent to the patient and to society in general to maintain trust. A balance between respecting privacy and creating maximal societal benefits needs to be ensured. While the data need to be FAIR, the analysis should be fair, accurate, confidential, and transparent (FACT). Finally, it is important to acknowledge the important role that legal and professional bodies have in ensuring that there are appropriate legislative frameworks in place that public and commercial stakeholders can adopt and follow. At the heart of any centralized or federated multi-source data science initiative in radiation oncology must therefore be full engagement with regional data protection regulations such as those set out in the European Union General Data Protection Regulations (GDPR), which are now the cornerstone of data sharing initiatives in Europe.

Statement 4: Involvement of clinicians in the data science community

The clinician is an essential member of the data science community

For decades, decisions in medicine have been based on clinical guidelines that are carefully developed and based on the highest-level evidence from large randomized controlled trials. Recently, individualized approaches to treatment have become an increasingly compelling research area. This trend is particularly prominent in oncology, where the discovery of new prognostic and predictive factors including viral infection, hypoxia signatures, driver mutations, and many others has enabled more precise treatment selection to match the characteristics of each patient and tumor [7]. However, greater personalization makes generating level-1 evidence difficult, if not impossible, as the number of matched patients in each subgroup decreases, ultimately coming down to a single individual. Predictive modelling using AI and multisource data offers a way to address this conundrum. The multidisciplinary field of clinical ML attracts researchers from diverse disciplines, including clinicians, computer scientists, medical physicists, and biostatisticians. Unfortunately, these different research communities often work in isolation, with separate jargon, specialized publications, and hermetic knowledge. Often, groups of scientists access partial data but lack the full clinical context or a complete understanding of the limitations of the data (e.g., embedded treatment effects) because their expertise may not lie in the clinical domain. Thus, to overcome these obstacles to the clinical implementation of AI tools, close cooperation among specialties is mandatory. The role of clinician is—and will remain—crucial to the clear definition of a relevant clinical problem and the identification of appropriate prediction targets, e.g. biologically relevant mechanisms (hypoxia, gene expression) or cancer- and treatment-specific outcomes like survival, relapse, and treatment toxicity. Clinicians must be involved in both baseline data review and model generation to detect garbage-in garbage-out situations arising from malformed or poorly designed models. Data sci-

ence approaches often highlight previously known clinical factors that are already used clinically to select patients for treatment, which can confound the interpretation of outcomes data. AI models that detect and latch onto novel details of individual patient cases are required so that these approaches can supplement, rather than reiterate, current clinical practice. Finally, and perhaps most importantly, clinicians will be the end users of any deployed multisource-based ML tool; they will be the ones to interpret the output of such tools to make responsible decisions about patient care and provide feedback to improve the database. Most clinicians trust their own experience and intuitions developed over years of practice and might find it difficult to rely on a model's prediction, especially if they do not understand the reasoning behind it. Thus, close collaboration between algorithm developers and clinicians is necessary to create models that clinicians can trust. For example, many studies focus on the interpretability of data science tools, the lack of which is one of the key obstacles to the wide clinical adoption of predictive models. Ultimately, the clinician is a critical bridge between the patient and the treatment team. This bridge is even more important as it allows patient preferences to be integrated into the planning process and may in turn change the way the ML/AI model is deployed. One example would be the development of multiple pareto-optimal treatment plans that integrate patient preferences into the final decision-making to select the outcomes most valuable to the patient. Furthermore, this integration stands to bring about a shift in the training of medical professionals in radiotherapy; for example, knowledge about how AI tools work and how to use them in personalized medicine will replace skills in, for instance, delineation of organs at risk [35][56].

The clinician should be involved in data generation and data curation

The constantly increasing power of computers has made collecting and analyzing large amounts of data relatively easy and allows the building of searchable and expandable databases for research, modeling,

and generation of new hypotheses. Often these data sets are used only once and then discarded or kept internally by the originating institution, which limits the power and capabilities of ML/AI for using these data sets. Therefore, the clinician must play a crucial role in designing and maintaining dedicated databases for ML model training. Choosing the relevant features for a clinical problem, considering the defined outcomes, and identifying possible biases are all still in the domain of clinical expertise [21]. For example, clinical decisions about radical versus palliative approaches in localized advanced head and neck tumors are to some extent affected by the clinician's personal experience of successful treatments. The data must be understood by the clinician before any modeling can take place, and clinicians are more willing to use models if the input features are aligned with evidence-based practice [58]. An example of such a model based on data routinely collected and updated every hour from electronic records of intensive care unit patients has been described by Thorsen-Meyer et al. [57]. However, models should not only use patient characteristics known to be important, but also uncover previously unknown associations. Here again, the clinician can help distinguish a truly novel predictive variable from biases, data set artefacts, or confounding factors. Additionally, clinicians can easily provide feedback in case of a false prediction and follow up with a misidentified patient, especially in cases with an unusual trajectory or medical history. To enable searchable and expandable databases for modeling, research, and generation of new hypotheses, all data sets must be shared (adhering to the FAIR principles described above) and accessible for other institutions to use for training, validation, or additional analysis. In this respect, several approaches may help clinicians become more engaged in data collection and curation. The most important is integration of this process into standard clinical workflows and standard operating procedures. This will allow clinical data, treatment planning information, diagnostic imaging, and outcomes to be seamlessly collected as part of clinical practice. Such integrated data collection will incentivize physicians to contribute high-quality data on all patients. Even simple synoptic endpoint collection can provide a powerful backbone for large data set

generation. When leveraged properly, rapid learning and automatic data collection will be crucial for clinicians in this era of fast progress in new therapeutics as well as technology and information overload. Crucially, simple methods to share data safely are necessary. If such methods are implemented properly, clinicians can derive visible benefits from sharing their efforts, which will convince them to contribute willingly. Examples of this could include simple quality assurance and second opinions, rapid outcome estimates, speed up evaluation of new technologies or automated workflow acceleration. Moreover, there is evidence that publication where associated data are shared in accordance with FAIR principles are cited more frequently [51], creating an additional incentive for the clinicians. Publication with open FAIR data available for readers should also be promoted as such by journal editors, meaning safe repositories have to be provided for authors. Building a culture of data sharing, not only within research institutions, but also hospitals and biotechnology companies is the most important challenge for the future. Policy makers will have an important role to play in creating a global structured policy of data sharing. A good example is the Final Report and Action Plan from the European Commission's Expert Group on FAIR data "Turning FAIR into reality", which paves a way to build infrastructure, recognition of obstacles and benefits and creates incentives for European research institutions to participate in data sharing. General concept of transparency and data sharing concerns also pharmaceutical companies, where process of sharing data obtained in clinical trials is still in it's infancy, mostly due to lack of policies of data sharing. A step forward to change this situation has been done by Miller et al who developed dedicated score -The Good Pharma Scorecard - to monitor of transparency in process of sponsored research and data sharing process. To have such policy in place is very important for industry itself due to increasing pressure of external stakeholders including patients and clinicians to speed up gathering knowledge and evidence by transparent collection of data [44]. Another initiative with potential to facilitate routine clinical data collection and sharing is the Real World Data (RWD) framework developed and proposed by the US Food and Drug Administration. The aim

is to collect post-approval data from electronic health records (EHR) and other clinical data repositories to generate Real World Evidence (RWE) of risk and benefits of currently approved products. Such approach promotes shared learning and encourages stakeholders to use RWE in their research, as well as to use common data models, unified terminology and data encoding for different sources. Clinicians are crucial not only for defining clinical problems and relevant outcomes, but also for the dynamic expansion and adaptation of databases, accounting for biases and unusual patient trajectories. We cannot forget that all this work should be focused on individual patient benefits but may also provide population-level benefits if resources are constrained. The clinician should be involved in all steps of model development and deployment An increasing number of clinicians is interested in cooperating with AI/ML scientists [17]. However, others remain reluctant and do not yet trust AI models, preventing deployment of these tools in the clinic [26]. One commonly stated reason for this distrust is unsatisfactory predictive performance, especially of prognostic models. However, what level of predictive accuracy is clinically acceptable is unclear. Moreover, accurate prediction of complex endpoints like overall survival is very difficult, even for an experienced clinician. A related question is how much better than a human a model must perform to be considered useful, especially if the human baseline is low. Many published models perform well for well-defined, simple outcomes such as prediction of local control or extracapsular extension in involved head and neck lymph nodes [11][30]. The most-used metric to measure predictive performance on a binary classification task is the area under the receiver operating characteristic curve. To generate predictions for new data, a single operating point needs to be selected. The standard approach is to give equal weights to specificity and sensitivity, but in many real clinical situations the cost of error may vary and may differently affect patients' outcomes. For example, a model that incorrectly suggests a patient will have a very high risk of toxicity may deprive the patient of the possibility to receive curative therapy. Only by knowing the holistic clinical picture can one decide how to define the expected parameters of models, including the desired specificity

and sensitivity of predictions. The role of the clinician in defining these optimal operating points is crucial. Clinicians should also be involved in the model review and development to prevent 'blind alleys' and other problems. The accuracy of a model is said to be inversely proportional to its explainability [34]. The trade-off between explainability and accuracy is still an unsolved problem, as the best-performing models based on deep learning are "black boxes"—synonymous with a lack of transparency and understanding—and are the least explainable. The results of successful attempts to improve the explainability of neural networks in health care were published by Yang [67]. However, even an explainable model may not be clinically applicable/actionable if the output is not additive or meshed into existing clinical approaches. It is vital that all models grow from and additively expand to fit existing clinical knowledge. Rediscovering, for instance, that a tumor's size predicts patient outcomes can be avoided by incorporating clinical knowledge early in the problem domain and establishing target areas to enhance [62]. Hence, ideally, models should be organically integrated with practice to provide continuous feedback, allowing clinicians to monitor and understand their effects and limitations. Any model, no matter how accurate and interpretable in the development stage, needs to be thoroughly validated in a controlled trial before clinicians can trust it. Validating AI models will require new trial design, especially in terms of endpoints and evaluation criteria. Nagendran et al found only two completed and published randomized controlled trials (RCTs), of AI algorithms in gastroenterology and ophthalmology [47][33][61], while the FDA has approved more than 16 deep learning algorithms in ophthalmology, radiology and cardiology. The most often used endpoint in such studies is the performance of AI/ML tools on some metric (e.g. receiver operating characteristic) versus human experts. However, ever if the AI outperforms human experts, it is not clear whether replacing the clinician's experience by an automated algorithm translates into benefit for patients in real-world use. Additionally, many clinical tasks have no well-defined ground truth, making an objective, direct comparison difficult. Endpoints such as performance of clinician supported by an algorithm, im-

proved workflow efficiency or time and financial savings could better reflect the real-world impact of clinical AI. Another challenge in model validation is how to evaluate and update models under data distribution shift, for example when the treatment guidelines or patient population change. Policy changes, such as the regulatory framework for AI/ML Software as a Medical Device recently proposed by the FDA, can help build clinician trust in AI models by enforce transparency and continuous performance monitoring as part of the approval process. Good performance in a retrospective test set is not sufficient; for example, a model can perform well in data from the institution where it was developed in but fail when tested on data from a different hospital, despite seemingly identical input data and targets. This can happen because of covariate shift—a change in the distribution of input data between different institutions, such as different CT scanner models and protocols—and because of unobserved confounders that have real impact on outcomes, like the quality of the health care system, the provision of supportive care, and even the approach of individual clinician treating the patient. Continuously reporting model accuracy at each deployed institution with individual physicians' feedback will be critical to maintaining and monitoring models and will help to ensure that physicians maintain confidence in the approach. To fully take advantage of this opportunity, clinician involvement is necessary at every step—from formulating the problem, through the selection of appropriate input data and prediction targets, to model validation in a prospective clinical trial. Data collection, curation and model development will require both financial and human resources. Recent rapid progress in AI has led to increased public interest and expectations of many stakeholders, including patients, regulators and governments. Resource allocation to AI research and implementation — both on central level, like EC and local institutional boards — is therefore expected to increase in the near future. Moreover, systemic solutions and infrastructure like automatic data collection, rapid learning systems, standardized format of collection, easy retrieval and seamless integration with the clinical workflow will reduce the load on clinicians and make the shift smooth and effective. To be trustworthy and actionable, the

predictions need to be interpretable, although the balance between interpretability and accuracy is still a subject of research. Interpretation and prospective validation, with identification of biases and confounding factors, preferably in a clinical trial, will increase clinician trust and allow for deployment of ML tools in the clinic.

Statement 5: From AI to a patient-oriented view

Realizing AI's potential in clinical practice calls for a patient-centered perspective in the development, design, and implementation of AI tools. First, AI tools must be oriented towards addressing clinical questions that matter to patients. Second, the output of AI tools must be integrated into decision aids that present relevant clinical information in a format that is clear, understandable, and actionable [42]. Finally, AI-based model outputs must be explainable and be combinable with patient preferences in a shared decision-making process. AI to enable retrieving patient data Orienting AI towards the patient perspective involves determining what is relevant to patients in clinical terms as well as how they experience their condition. Certain aspects of these data are routinely collected as patient-reported outcome measures (PROMs) and stored in patients' EHRs. It is unclear to what extent PROMs are analyzed and used [4], particularly due to doctors' lack of time, resources, and expertise [64]. This presents an opportunity for AI/ML techniques to harness and analyze EHR data and PROMs to identify relationships between treatments and patient-relevant outcomes [32]. Aside from patient data stored in hospital records, increasing amounts of data are also generated externally, as more patients, through use of the internet, are taking an active role in managing their health decisions. Consequently, the role of patient organizations is also shifting from providing information to building platforms and online communities in which patients can share experiences and knowledge that go beyond the data captured in PROMs. For instance, the patient organization PatientsLikeMe, an online community that connects over 650,000 patients across nearly 3000 health conditions, is based on the

principle of seeing the patient as a person rather than as a disease, and accordingly collects data on patients' definitions of health and outcomes. PatientsLikeMe is actively involved in AI initiatives to generate insights from this vast and rich data source (`https://www.patientslikeme.com`). Initiatives such as these can help target AI tools to clinically relevant questions.

Clinical decision aids for doctors and patients

AI tools must be built into decision aids that support shared deliberation and decision-making processes between patients and doctors. Decision aids provide a means to inform patients about their conditions, reflect on their own values, and weigh their treatment options in the context of their preferences. Poor design is one of the main factors that hinders decision aid uptake [1]. Implementation of AI-enhanced decision aids is more likely to succeed when development follows a user-centered design process that takes into account end-users' contexts, needs, goals, and decision-making [66]. We have previously emphasized the importance of including doctors in the development process, as well as the patient focus in determining relevant clinical questions. Once developed, it is critical that decision aids be embedded into the clinical workflow, for instance through integration into the hospital's EHR system, to minimize the amount of time and manual work required in entering a patient's data [70]. In addition, integrating the decision aid into the clinical consultation itself can pave the way for data-driven shared decision-making in which AI-based recommendations are discussed in the context of the doctor's clinical knowledge/experience and the patient's preferences. Bringing AI into the consultation Traditional shared decision-making consists of a two-way information exchange between doctors and patients; doctors share their clinical expertise on the treatment options and their benefits and risks, and patients share their values and preferences [5]. When both sides understand each others' perspectives, they can deliberate on the available options from a common ground and make a choice that

is rooted in the best clinical evidence as well as the patient's individual circumstances. The introduction of AI-based decision aids represents a third angle from which treatment information can be personalized according to the patient's individual characteristics. The "black box" nature of certain AI tools may make it challenging for the doctor to articulate the reasoning behind a given diagnosis or treatment recommendation [52], so the explainability of the AI tool is a crucial factor in ensuring that decision-making does not shift back towards paternalism [40]. This includes interpreting AI model outputs, such as risk estimates or prognoses, and communicating them in a way that is understandable to patients [43]. Moreover, little is known about the patient perspective on receiving AI-supported care. Preliminary findings from skin cancer screening suggest that patients are open to the potential of AI in improving care quality as long as it functions as decision support rather than decision replacement [50] and the doctor-patient dyad is maintained [67]. More research is needed to understand the shift in roles and responsibilities that accompanies AI implementation and how to use AI models to empower patients. In particular, the perspectives of social scientists and anthropologists are needed to bring AI into alignment with human decision-making [29].

14.2 DISCUSSION

In this paper, we identified the barriers that are currently limiting the adoption of big data analytics in the clinic toward the development of a learning health care system. The main barrier is the ability to handle the large amount of data and metadata produced in the clinic as result of daily clinical and research activities. As we discussed, this big data involves different stakeholders and users and presents significant interoperability issues. We therefore identified the need to analyze the major sources of multisource data and metadata and the limits on their interoperability (milestones M1 and M2). We next discussed how, when it becomes capable of fully connecting these sparse multisource data, AI will provide powerful analytics to develop data-driven

clinical decision aids (milestone M3). However, because of the variety of data types and stakeholders involved, multiple professionals need to be involved and coordinated. For this reason, we presented the need to define a clinical data science community in radiation oncology, to act as harmonizer of the different professional figures with a common vision sustained by a code of conduct and working statements and with a strong orientation toward patient-centered care (milestone M5). This community will not be an independent actor, but will build upon already existing communities, efforts, and working groups. Clinicians will have a prominent leading role both in determining the requirements of the technical developments and in continuously interfacing with the more technical professionals. This is meant to guarantee that technical developments are in line with unmet clinical needs. We envision this community to be fully embedded within the major global radiation oncology societies, such as ESTRO, ASTRO, CARO, AAPM, EFOMP, RANZCR and FARO and to include patient societies such as CRC and PatientsLikeMe. Our future activities will be to engage with the above-mentioned societies to define working groups, as briefly depicted in Figure 14.4.

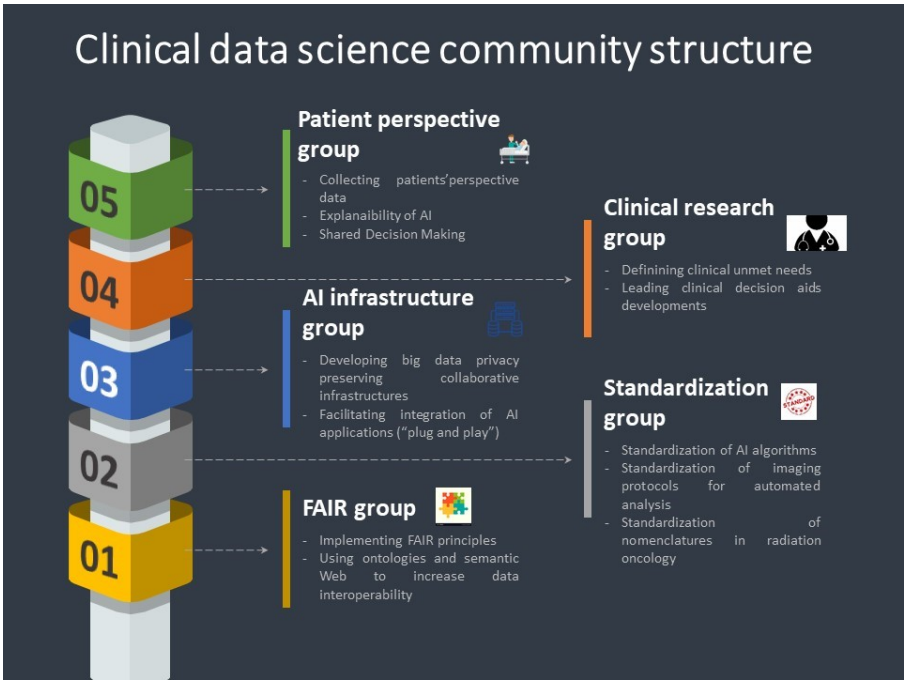


Figure 14.4: Overview of the structure of the new data science community in radiation oncology. In line with the presented milestones, five major working groups are identified: FAIR principles group (M1); standardization group (M 2); AI applications and big data collaborative infrastructures group (M3); clinical research and definition of unmet clinical needs group (M4); and patient-centered decision aids and shared decision-making group (M5). The role of each group is to coordinate with similar existing task forces and working groups from European and American societies active in radiation oncology.

Bibliography

[1] Thomas Agoritsas, Anja Fog Heen, Linn Brandt, Pablo Alonso-Coello, Annette Kristiansen, Elie A Akl, Ignacio Neumann, Kari AO Tikkinen, Trudy van der Weijden, Glyn Elwyn, Victor M Montori, Gordon H Guyatt, and Per Olav Vandvik. Decision aids that really promote shared decision making: the pace quickens. *BMJ*, page g7624, February 2015.

[2] null Amisha, Paras Malik, Monika Pathania, and Vyas Kumar Rathaur. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, 8(7):2328–2331, July 2019.

[3] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7345–7352, 2016.

[4] Paul J. Barr, Scott A. Berry, Wendolyn S. Gozansky, Deanna B. McQuillan, Colleen Ross, Don Carmichael, Andrea M. Austin, Travis D. Satterlund, Karen E. Schifferdecker, Lora Council, Michelle D. Dannenberg, Ariel T. Wampler, Eugene C. Nelson, and Jonathan Skinner. No date for the PROM: the association between patient-reported health events and clinical coding in primary care. *Journal of Patient-Reported Outcomes*, 4(1):17, March 2020.

[5] Michael J. Barry and Susan Edgman-Levitan. Shared Decision Making — The Pinnacle of Patient-Centered Care. *New England Journal of Medicine*, 366(9):780–781, March 2012.

[6] David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs (Project Hope)*, 33(7):1123–1131, July 2014.

[7] Michael Baumann, Mechthild Krause, Jens Overgaard, Jürgen Debus, Søren M. Bentzen, Juliane Daartz, Christian Richter, Daniel Zips, and Thomas Bortfeld. Radiation oncology in the era of precision medicine. *Nature Reviews Cancer*, 16(4):234–249, April 2016.

[8] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(1):118, December 2020.

[9] Roberto Berenguer, María del Rosario Pastor-Juan, Jesús Canales-Vázquez, Miguel Castro-García, María Victoria Villas, Francisco Mansilla Legorburo, and Sebastià Sabater. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology*, 288(2):407–415, August 2018.

[10] Oya Beyan, Ananya Choudhury, Johan van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, et al. Distributed analytics on sensitive medical data: The personal health train. *Data Intelligence*, 2(1-2):96–107, 2020.

[11] Luca Boldrini, Davide Cusumano, Giuditta Chiloiro, Calogero Casà, Carlotta Masciocchi, Jacopo Lenkowicz, Francesco Cellini, Nicola Dinapoli, Luigi Azario, Stefania Teodoli, Maria Antonietta Gambacorta, Marco De Spirito, and Vincenzo Valentini. Delta radiomics for rectal cancer response prediction with hybrid 0.35 T

magnetic resonance-guided radiotherapy (MRgRT): a hypothesis-generating study for an innovative personalized medicine approach. *La radiologia medica*, 124(2):145–153, February 2019.

[12] Abdelali Boussadi and Eric Zapletal. A fast healthcare interoperability resources (fhir) layer implemented over i2b2. *BMC medical informatics and decision making*, 17(1):120, 2017.

[13] Luciano R. F. Branco, Rachel B. Ger, Dennis S. Mackin, Shouhao Zhou, Laurence E. Court, and Rick R. Layman. Technical Note: Proof of concept for radiomics-based quality assurance for computed tomography. *Journal of Applied Clinical Medical Physics*, 20(11):199–205, November 2019.

[14] Andrius Budrionis and Johan Gustav Bellika. The Learning Healthcare System: Where are we now? A systematic review. *Journal of Biomedical Informatics*, 64:87–92, 2016.

[15] Irène Buvat and Fanny Orlhac. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. *Journal of Nuclear Medicine*, 60(11):1543–1544, November 2019.

[16] Ananya Choudhury, Johan van Soest, Stuti Nayak, and Andre Dekker. Personal health train on fhir: A privacy preserving federated approach for analyzing fair data in healthcare. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 85–95. Springer, 2020.

[17] Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58:161–167, August 2019.

[18] G. Collewet, M. Strzelecki, and F. Mariette. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magnetic Resonance Imaging*, 22(1):81–91, January 2004.

[19] Timo M. Deist, A. Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Andre Dekker, and Philippe Lambin. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*, 4:24–31, June 2017.

[20] Adrien Depeursinge, Vincent Andrearczyk, Philip Whybra, Joost van Griethuysen, Henning Müller, Roger Schaer, Martin Vallières, and Alex Zwanenburg. Standardised convolutional filtering for radiomics. *arXiv:2006.05470 [cs, eess]*, June 2020. arXiv: 2006.05470.

[21] Issam El Naqa, Dan Ruan, Gilmer Valdes, Andre Dekker, Todd McNutt, Yaorong Ge, Q. Jackie Wu, Jung Hun Oh, Maria Thor, Wade Smith, Arvind Rao, Clifton Fuller, Ying Xiao, Frank Manion, Matthew Schipper, Charles Mayo, Jean M. Moran, and Randall Ten Haken. Machine learning and modeling: Data, validation, communication challenges. *Medical Physics*, 45(10):e834–e840, October 2018.

[22] Rachel B. Ger, Shouhao Zhou, Pai-Chun Melinda Chi, Hannah J. Lee, Rick R. Layman, A. Kyle Jones, David L. Goff, Clifton D. Fuller, Rebecca M. Howell, Heng Li, R. Jason Stafford, Laurence E. Court, and Dennis S. Mackin. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Scientific Reports*, 8(1):13047, December 2018.

[23] Stefan Harrer, Pratik Shah, Bhavna Antony, and Jianying Hu. Artificial Intelligence for Clinical Trial Design. *Trends in Pharmacological Sciences*, 40(8):577–591, 2019.

[24] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36, January 2019.

[25] Ira S. Hofer, Eran Halperin, and Maxime Cannesson. Opening the Black Box: Understanding the Science Behind Big Data and Predictive Analytics. *Anesthesia and Analgesia*, 127(5):1139–1143, 2018.

[26] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), July 2019.

[27] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. Artificial intelligence in radiology. *Nature Reviews. Cancer*, 18(8):500–510, 2018.

[28] Nicolas Houy and François Le Grand. Personalized oncology with artificial intelligence: The case of temozolomide. *Artificial Intelligence in Medicine*, 99:101693, August 2019.

[29] Geoffrey Irving and Amanda Askell. AI Safety Needs Social Scientists. *Distill*, 4(2):10.23915/distill.00014, February 2019.

[30] Benjamin H. Kann, Daniel F. Hicks, Sam Payabvash, Amit Mahajan, Justin Du, Vishal Gupta, Henry S. Park, James B. Yu, Wendell G. Yarbrough, Barbara A. Burtness, Zain A. Husain, and Sanjay Aneja. Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma. *Journal of Clinical Oncology*, 38(12):1304–1311, April 2020.

[31] Harlan M. Krumholz. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Affairs*, 33(7):1163–1170, July 2014.

[32] Camillo Lamanna and Lauren Byrne. Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm. *AMA Journal of Ethics*, 20(9):E902–910, September 2018.

[33] Haotian Lin, Ruiyang Li, Zhenzhen Liu, Jingjing Chen, Yahan Yang, Hui Chen, Zhuoling Lin, Weiyi Lai, Erping Long, Xiaohang Wu, Duoru Lin, Yi Zhu, Chuan Chen, Dongxuan Wu, Tongyong Yu, Qianzhong Cao, Xiaoyan Li, Jing Li, Wangting Li, Jinghui Wang, Mingmin Yang, Huiling Hu, Li Zhang, Yang Yu, Xuelan Chen, Jianmin Hu, Ke Zhu, Shuhong Jiang, Yalin Huang, Gang Tan, Jialing Huang, Xiaoming Lin, Xinyu Zhang, Lixia Luo, Yuhua Liu, Xialin Liu, Bing Cheng, Danying Zheng, Mingxing Wu, Weirong Chen, and Yizhi Liu. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine*, 9:52–59, March 2019.

[34] Yi Luo, Huan-Hsin Tseng, Sunan Cui, Lise Wei, Randall K. Ten Haken, and Issam El Naqa. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR|Open*, 1(1):20190021, July 2019.

[35] Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, and Andre Dekker. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*, 126(2):312–317, February 2018.

[36] Dennis Mackin, Xenia Fave, Lifei Zhang, David Fried, Jinzhong Yang, Brian Taylor, Edgardo Rodriguez-Rivera, Cristina Dodge, Aaron Kyle Jones, and Laurence Court. Measuring Computed Tomography Scanner Variability of Radiomics Features:. *Investigative Radiology*, 50(11):757–765, November 2015.

[37] Dennis Mackin, Rachel Ger, Cristina Dodge, Xenia Fave, Pai-Chun Chi, Lifei Zhang, Jinzhong Yang, Steve Bache, Charles Dodge, A. Kyle Jones, and Laurence Court. Effect of tube current on computed tomography radiomic features. *Scientific Reports*, 8(1):2354, dec 2018.

[38] Charles S. Mayo, Michelle Mierzwa, Jean M. Moran, Martha M. Matuszak, Joel Wilkie, Grace Sun, John Yao, Grant Weyburn, Carlos J. Anderson, Dawn Owen, and Arvind Rao. Combination of a Big Data Analytics Resource System With an Artificial Intelligence Algorithm to Identify Clinically Actionable Radiation Dose Thresholds for Dysphagia in Head and Neck Patients. *Advances in Radiation Oncology*, page S2452109420300142, January 2020.

[39] Charles S. Mayo, Jean M. Moran, Walter Bosch, Ying Xiao, Todd McNutt, Richard Popple, Jeff Michalski, Mary Feng, Lawrence B. Marks, Clifton D. Fuller, Ellen Yorke, Jatinder Palta, Peter E. Gabriel, Andrea Molineu, Martha M. Matuszak, Elizabeth Covington, Kathryn Masi, Susan L. Richardson, Timothy Ritter, Tomasz Morgas, Stella Flampouri, Lakshmi Santanam, Joseph A. Moore, Thomas G. Purdie, Robert C. Miller, Coen Hurkmans, Judy Adams, Qing-Rong Jackie Wu, Colleen J. Fox, Ramon Alfredo Siochi, Norman L. Brown, Wilko Verbakel, Yves Archambault, Steven J. Chmura, Andre L. Dekker, Don G. Eagle, Thomas J. Fitzgerald, Theodore Hong, Rishabh Kapoor, Beth Lansing, Shruti Jolly, Mary E. Napolitano, James Percy, Mark S. Rose, Salim Siddiqui, Christof Schadt, William E. Simon, William L. Straube, Sara T. St. James, Kenneth Ulin, Sue S. Yom, and Torunn I. Yock. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *International Journal of Radiation Oncology*Biology*Physics*, 100(4):1057–1066, March 2018.

[40] Rosalind J McDougall. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, 45(3):156–160, March 2019.

[41] Christos Melidis, Walther R. Bosch, Joanna Izewska, Elena Fidarova, Eduardo Zubizarreta, Kenneth Ulin, Satoshi Ishikura, David Followill, James Galvin, Annette Haworth, Deidre Besuijen, Clark H. Clark, Elizabeth Miles, Edwin Aird, Damien C. Weber, Coen W. Hurkmans, and Dirk Verellen. Global Harmo-

nization of Quality Assurance Naming Conventions in Radiation Therapy Clinical Trials. *International Journal of Radiation Oncology*Biology*Physics*, 90(5):1242–1249, December 2014.

[42] B. Middleton, D. F. Sittig, and A. Wright. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. *Yearbook of Medical Informatics*, Suppl 1:S103–116, August 2016.

[43] D. Douglas Miller. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. *npj Digital Medicine*, 2(1):62, December 2019.

[44] Jennifer Miller, Joseph S. Ross, Marc Wilenzick, and Michelle M. Mello. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. *The British Medical Journal*, 366:l4217, 2019.

[45] Jean Moran, Andrea Molineu, Jon Kruse, Mark Oldham, Robert Jeraj, James Galvin, Jatinder Palta, and Arthur Olch. Guidance for the Physics Aspects of Clinical Trials. Technical report, AAPM, January 2018.

[46] Olivier Morin, Martin Vallières, Arthur Jochems, Henry C. Woodruff, Gilmer Valdes, Steve E. Braunstein, Joachim E. Wildberger, Javier E. Villanueva-Meyer, Vasant Kearney, Sue S. Yom, Timothy D. Solberg, and Philippe Lambin. A Deep Look Into the Future of Quantitative Imaging in Oncology: A Statement of Working Principles and Proposal for Change. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1074–1082, November 2018.

[47] Myura Nagendran, Yang Chen, Christopher A. Lovejoy, Anthony C. Gordon, Matthieu Komorowski, Hugh Harvey, Eric J. Topol, John P. A. Ioannidis, Gary S. Collins, and Mahiben Maruthappu. Artificial intelligence versus clinicians: systematic

review of design, reporting standards, and claims of deep learning studies. *The British Medical Journal*, 368:m689, 2020.

[48] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2011.

[49] Jack E. Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, San Francisco, 2003.

[50] Christopher Pearce, Adam McLeod, Natalie Rinehart, Robin Whyte, Elizabeth Deveny, and Marianne Shearer. Artificial intelligence and the clinical world: a view from the front line. *Medical Journal of Australia*, 210(S6), April 2019.

[51] Heather A. Piwowar, Roger S. Day, and Douglas B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS One*, 2(3):e308, March 2007.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA, August 2016. ACM.

[53] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, and SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *The British Medical Journal*, 370:m3210, 2020.

[54] Zhenwei Shi, Ivan Zhovannik, Alberto Traverso, Frank J. W. M. Dankers, Timo M. Deist, Petros Kalendralis, René Monshouwer, Johan Bussink, Rianne Fijten, Hugo J. W. L. Aerts, Andre Dekker,

and Leonard Wee. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Scientific Data*, 6(1):218, December 2019.

[55] Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1):1–21, February 2010.

[56] Reid F. Thompson, Gilmer Valdes, Clifton D. Fuller, Colin M. Carpenter, Olivier Morin, Sanjay Aneja, William D. Lindsay, Hugo J.W.L. Aerts, Barbara Agrimson, Curtiland Deville, Seth A. Rosenthal, James B. Yu, and Charles R. Thomas. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiotherapy and Oncology*, 129(3):421–426, December 2018.

[57] Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Kirstine Belling, Søren Brunak, and Anders Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179–e191, April 2020.

[58] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv:1905.05134 [cs, stat]*, August 2019. arXiv: 1905.05134.

[59] Alberto Traverso, Frank JWM Dankers, Leonard Wee, and Sander MJ van Kuijk. Data at scale. In *Fundamentals of Clinical Data Science*, pages 11–17. Springer, Cham, 2019.

[60] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and Reproducibility of Radiomic Features:

A Systematic Review. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1143–1158, nov 2018.

[61] Pu Wang, Tyler M. Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

[62] Mattea L. Welch, Chris McIntosh, Benjamin Haibe-Kains, Michael F. Milosevic, Leonard Wee, Andre Dekker, Shao Hui Huang, Thomas G. Purdie, Brian O'Sullivan, Hugo J.W.L. Aerts, and David A. Jaffray. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology*, November 2018.

[63] Philip Whybra, Craig Parkinson, Kieran Foley, John Staffurth, and Emiliano Spezi. Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. *Scientific Reports*, 9(1):9649, December 2019.

[64] Paul Wicks, Michael Massagli, Jeana Frost, Catherine Brownstein, Sally Okun, Timothy Vaughan, Richard Bradley, and James Heywood. Sharing Health Data for Better Outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2):e19, June 2010.

[65] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene

van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.

[66] Holly O Witteman, Selma Chipenda Dansokho, Heather Colquhoun, Angela Coulter, Michèle Dugas, Angela Fagerlin, Anik MC Giguere, Sholom Glouberman, Lynne Haslett, Aubri Hoffman, Noah Ivers, France Légaré, Jean Légaré, Carrie Levin, Karli Lopez, Victor M Montori, Thierry Provencher, Jean-Sébastien Renaud, Kerri Sparling, Dawn Stacey, Gratianne Vaisson, Robert J Volk, and William Witteman. User-centered design and the development of patient decision aids: protocol for a systematic review. *Systematic Reviews*, 4(1):11, December 2015.

[67] Yinchong Yang, Volker Tresp, Marius Wunderle, and Peter A. Fasching. Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 152–162, New York, NY, June 2018. IEEE.

[68] Stephen S F Yip and Hugo J W L Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, July 2016.

[69] Alex Zwanenburg. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(13):2638–2655, December 2019.

[70] Alex Zwanenburg, Martin Vallières, Mahmoud A. Abdalah, Hugo J. W. L. Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J. Beukinga, Ronald Boellaard, Marta Bogowicz, Luca Boldrini, Irène Buvat, Gary

J. R. Cook, Christos Davatzikos, Adrien Depeursinge, Marie-Charlotte Desseroit, Nicola Dinapoli, Cuong Viet Dinh, Sebastian Echegaray, Issam El Naqa, Andriy Y. Fedorov, Roberto Gatta, Robert J. Gillies, Vicky Goh, Michael Götz, Matthias Guckenberger, Sung Min Ha, Mathieu Hatt, Fabian Isensee, Philippe Lambin, Stefan Leger, Ralph T.H. Leijenaar, Jacopo Lenkowicz, Fiona Lippert, Are Losnegård, Klaus H. Maier-Hein, Olivier Morin, Henning Müller, Sandy Napel, Christophe Nioche, Fanny Orlhac, Sarthak Pati, Elisabeth A.G. Pfaehler, Arman Rahmim, Arvind U.K. Rao, Jonas Scherer, Muhammad Musib Siddique, Nanna M. Sijtsema, Jairo Socarras Fernandez, Emiliano Spezi, Roel J.H.M. Steenbakkers, Stephanie Tanadini-Lang, Daniela Thorwarth, Esther G.C. Troost, Taman Upadhaya, Vincenzo Valentini, Lisanne V. van Dijk, Joost van Griethuysen, Floris H.P. van Velden, Philip Whybra, Christian Richter, and Steffen Löck. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*, 295(2):328–338, May 2020.

15
Discussion

15.1 Executive summary

In this thesis we have investigated the role of radiomics in RT, by focusing on presenting and tackling the issues that are limiting the translation of radiomics-derived models to the clinic as decision support systems. Three main issues have been identified: lack of robustness of radiomics with respect to image acquisition settings, pitfalls and methodological issues in the use of machine learning in radiomic studies, and lack of standardization and a common infrastructure to enable multi-centre radiomic studies. The first issue is deeply connected with the concepts of radiomic reproducibility and repeatability. In chapter 3, a systematic review of the major sources impacting radiomic reproducibility and repeatability has been presented. Three major outcomes arose from this study: A) radiomic feature classes present different grade of sensitivity to image acquisition settings, digital image pre-processing, and contouring variability; B) the level of the investigations is more mature in CT (Computed Tomography) than in MRI (Magnetic Resonance Imaging), despite the fact that the latter imaging modality offers better soft tissue contrast compared to the first one and therefore might be more suitable to fully catch tumour biology and heterogeneity, and C) poor quality of reporting of radiomic studies mainly due to lack of standardization. Points A) and B) have motivated the work presented in Chapters 3 and 4 where we investigated radiomic reproducibility in different MR sequences including diffusion weighted imaging for rectal and cervix cancer patients. As expected, similar results to CT studies were found in MRI. Scanners' variability seems to have the largest impact on features' reproducibility; different manufacturers' models or differences in uniform magnetic field strength (1.5 vs 3T) being the major drivers of feature instability. This present us with a major problem, considering the desire of having radiomic biomarkers that can be validated across multiple hospitals without any degradation of their prognostic power. Also, inter observer variability in contouring had a major impact in features' stability especially for ADC (Apparent Diffusion Coefficient) maps of rectal cancer patients. This evidence opens the debate whether auto-

contouring might play a role in reducing disagreements on contours and whether it might provide better radiomic stability. In chapter 3, we also showed how digital image pre-processing, while being a common practice in the radiomic workflow, should be used with caution and only after it has been proved that such techniques can improve the signal-to-noise ratio of radiomic features. In fact, according to our investigation, tuning radiomic computations had a strong impact on radiomic reproducibility, more marked for texture features, which are meant to measure tumour heterogeneity. After these investigations, we provided a possible method to improve the stability of radiomic features by normalizing them according to biological ROIs present in the image. This was shown in Chapter 5, for ADC maps of cervix cancer patients using a ROI drawn inside the bladder. We also developed a method which showed how, by isolating known acquisition parameters that impact radiomic stability, it is possible to find the functional dependence of radiomic features to these parameters and correct for them. This was applied in Chapter 6 for the imaging parameter "radiation exposure", thus closely related to signal-to-noise ratio, in CT images of imaging phantoms. Unfortunately, we did not have the availability of dedicated imaging phantoms for MR, but the presented methodology can be extended to MR. Future work in the acquisition correction domain includes providing corrections for radiomic features with respect to major sources of uncertainties in MR, starting from the strength of the magnetic field. One additional point that arose from the previously mentioned studies is that many radiomic features embed strong mathematic dependencies to tumour size (GTV) and their apparent large reproducibility and repeatability might be driven by this fact. GTV has been found to be one of the most reproducible features in all the literature for CT, PET and MRI. Furthermore, tumour size is a valid prognostic factor in almost all cancer types at least for overall or disease-free survival as demonstrated in Chapter 8. These hidden correlations are dangerous if not investigated. They might lead us to think that a feature is measuring a texture of the tumour (possibly related to a certain biological property), while it could just be a "replication" of a traditional clinically

accepted prognostic factor. The main take home message from the first part of this thesis is that even before focusing on the predictive or prognostic role of radiomic features, it is relevant to deeply investigate concepts like reproducibility and repeatability, which can have a strong impact on the generalizability and validity of radiomic models. In addition to correlations with simple features such as volume, radiomic features might present strong correlations with clinically accepted prognostic factors (e.g. HPV status for head and neck cancers). Therefore, there is the need to re-think machine learning as tool to investigate these issues, rather than a "black-box" where some numbers (i.e. radiomic features) are thrown in it and predictions are returned. This was presented in the second part of this thesis, namely in chapters 7 and 8. In chapter 7 we posed strong focus on the "volume effect" in radiomic studies. This problem was already raised in a previous publication, which called for the urgency of "safeguards" for radiomics. With our study we showed how machine learning, and more specifically unsupervised methods, can be used to reveal confounding effects in radiomics. We focused on tumour volume, but the presented framework is not just for volume but other variables as well. While we focused only on CT imaging and head and neck and lung cancers, the presented framework is strongly suggested for all the radiomic studies. In chapter 8 we expanded our previous work by highlighting the importance of benchmarking radiomic models against accepted clinical prognostic factors (above all TNM staging) and by comparing not only traditional machine learning classifiers, but also fully automated pipelines based on deep learning architectures. It is worth offering guidance to interpret these results, to avoid frustration, especially from users of radiomic models (i.e. clinicians). These investigations did not aim at showing the non-utility of radiomics, but rather they offer a change of paradigm to raise the bar of radiomic studies. Often, AI pipelines are applied without cautions, finding correlations between input data and outcomes of interest that might not be supported by causality. This problem increases the risk of spurious associations and false discoveries. While we recognize the growing trend of fully optimizing data and more specifically image analysis, we warrant cau-

tions in a "black-box" approach. Therefore, in these chapters we rediscovered the role of human intervention in computational pipelines both for data QA as well as for methodological QA. One of the main messages from the first two parts of this thesis is also that only by performing multi-institutional studies we can provide more evidence that a specific methodology works for a radiomic study. However, this methodology might not be universal and unique and can be dependent on the specific image modality or cancer types considered. Also, generalizability of radiomic models calls for performing a large number of experiments among institutions as suggested by the TRIPOD statement. Unfortunately, the lack of standardization and harmonization of radiomics is a strong limiting factor. First, many radiomic computational packages are available and the number will increase. Recently, several companies have started developing tools for automated quantitative image analysis. Most of these tools are protected by copyright or patents and it is not possible to access all the details of the radiomic computations and mining of the features. Each software has its own standards both for naming conventions of features and settings used to extract features. Stand-alone standards also exist for clinical data, which are often used to correlate radiomic features with the outcomes of interest. Finally, the well-known problems related to data sharing do not allow to easily perform multi-centre radiomic studies. In chapters 9-11 we presented the building blocks of a framework based on distributed learning. These building blocks are based on ontologies and semantic web techniques, which permit to introduce FAIR (Findable Accessible Interoperable Reusable) principles to radiomic studies boosting transparency and generalizability. We presented the results of these methodologies in chapter 12, where we were the first extending distributed learning to radiomic studies. We successfully showed that distributed results cannot be distinguish from centralized ones. We strongly believe this framework will allow faster and large centre radiomic studies. Finally, chapter 13 remarks again how radiomic should not be thought as living a vacuum, but it should be part of a larger effort lead by a clinical data science community, with a central role of clinicians in leading and giving the directions of research.

15.2 Limitations of this work

Several limitations of this work need to be highlighted. First, we specifically focused on radiomics in CT and MR images. The reasons behind this are two fold: A) CT was the first imaging modality investigated in radiomic studies, and B) as we pointed out in chapter 3, the lack of evidence of radiomic studies in MR called for an urgency of extending previously published literature for this imaging modality, which might provide a better soft-tissue contrast compared to CT and therefore more suitable for quantitative measurements of textures. We did not investigate the role of radiomics in PET imaging. Therefore, it remains open the debate whether the results available in this work can be extended to this imaging modality. Nevertheless, it seems that the results found for CT and MR are in line with literature on PET studies. This evidence shows again the importance of investigating a robust and agreed methodology to improve the quality of radiomic studies. Finally, the frameworks proposed in chapters 9-10-13 can be considered as independent from a specific imaging modality, as they have to be perceived as safeguards or data infrastructures to accelerate the deployment of radiomic models in the clinic. Second, in this work we briefly touched upon deep learning applications in radiomics. Conversely, deep learning has somewhat stolen the scene in medical image analysis, with several applications spanning from automation to modelling. In chapter 9 we proposed an application of a CNN for prognostication of head and neck cancer patients. While we agree that the network of this architecture is far from being the latest available in the literature, it was out of topic of that publication to provide the readers with a comprehensive analysis of top performing CNN architectures for radiomics. Nevertheless, we pointed out how using fully automated pipelines does not free the radiomic community from adopting safeguards in their study. A detailed comparison of hand crafted image analysis approaches with fully automated pipelines as well as meticulous analysis of hidden relations among radiomic and ”deep radiomic” features is needed.

15.3 Future outlook

This thesis has contributed to raise awareness on some of the *pitfalls* encountered in radiomics. It is worth to remark again that this thesis does not give any judgement on the quality and role of radiomics in radiation oncology. Conversely, the author of this thesis strongly believes that radiomics can be a powerful tool to augment the power of decision making in patient care. Furthermore, radiomics has the power to *revitalize* the vast amount of standard of care medical images, mainly addressed for visual inspection or human tasks, but sadly put aside at completion of these tasks. Therefore, radiomics has to be conceived as a *cost effective* approach to introduce meaningful AI-driven applications in medical imaging to improve the RT workflow. Unfortunately, the term *Artificial* in AI seems to have caused a general sense of moving away from the responsibilities and role of humans in the development of AI solutions. The advent of powerful fully automated data analysis pipelines can have the risk to augment this gap. Conversely, as this thesis has shown, the quality of a radiomic model is mainly determined by a meticulous design of the whole radiomic workflow, from data acquisition, to computations and modelling. A full control of all the steps of this workflow as well as a consciousness of the possible hidden factors leading to the risk of false discoveries is mandatory. Finally, from my work performed in these last 4 years it emerged a gap among the radiomic community and the clinical world. While a large effort has been put in designing radiomic studies, we can debate whether the same effort has been put to engage the clinical community. It is still an open question whether a successful radiomic model will be implemented in the clinic. There might be the risk that this model, despite having high performances, might not be addressing a desired change in clinical practice. What it is found to be statistically significant not necessarily it will be clinically significant. There is the need to re-think radiomics and more in general AI as not living in a vacuum of being an academic exercise, but rather part of a common vision shared among the different stakeholders that intervene in patient care in radiation oncology: doctors, researchers, technicians, funding agencies,

insurance companies, patients and the government. All these concepts have been well described in chapter 14 of this thesis. To conclude, the future of radiomics will be brighter only if we re-think radiomics, and more in general AI and medical image analysis, as a community effort from a worldwide clinical data science community with the above-mentioned stakeholders. The future of radiomics has to pass through this paradigm shift, a necessarily *radiomics renewal* phase.

Research impact and utilisation summary

1 Societal impact

In this thesis, we analysed and proposed solutions to some major issues currently limiting the application of AI to medical imaging in radiation oncology. These issues are strongly impacting the possibility to translate research prototypes as decision support systems in the clinic. Decision support systems bring a benefit not only to clinicians, but also to patients. The ongoing challenge for radiation oncology is to provide patients with the best possible treatment. The treatment strategy must be as good as possible, boosting curative intent, while decreasing the risk of radiation-induced side effects and disease relapse. When achieved, cancer patients will live longer and better. Many of our treatment decisions are based on the evidence arising from randomized clinical trials. While we recognize the importance of clinical trials, we also highlight how the evidence from clinical trials should be expanded with a "real life" evidence. It is a well-known problem that strict requirements for trials' accrual can lead to results that only apply to a small population, which might not be representative of the variety of patients walking into the clinic every day. In this thesis, I present how AI applied to medical imaging is key to improve patient care in the near future. I discuss how medical images are a source of unique patient-centred data, which can reveal insights about our patients. I show and propose AI-driven techniques for automated image analysis and biomarker discovery that can lead to the development of robust decision support systems (DSSs). These DSSs redefine and augment our clinicians' prior knowledge because they consider unique information based on the patient-level. Biomarkers are derived from medical images, which contain fingerprints of patients' anatomy and tumours. By proposing a robust methodology for automated medical

image analysis (i.e., *radiomics*), this thesis is contributing in the near future to speed up the translation of image-derived biomarkers in the clinic with the societal impact of supporting better decisions for the treatment of cancer patients. While the work has been mainly devoted to lung and head neck cancers, it is scalable to many other anatomical sites and even diseases not necessarily related to cancer, since medical images are the most frequent type of data acquired in a clinic.

2 Economic Impact

In the last years, radiation oncology has faced an "explosion" of treatment options. Besides conventional radiation therapy and surgery, the recent advances in biology have opened the path to therapies that interact with our immune systems (immunotherapy) to suppress cancer cells or by targeting specific molecular profiles of tumours (molecular or targeted therapies). Even for more traditional therapies like surgery, hardware advances such as robotics is improving our surgeons' abilities to be more precise and therefore, reduce post-surgery complications. For treatment with radiation, alternatives to conformal radiotherapy are for example represented by proton/ion therapy, which can improve radiation delivered to the targets, while reducing damage to surrounding healthy tissues. Doctors are struggling to pick the best option since sometimes there is not enough evidence that on a patient-level one treatment should be preferred over another. An additional problem also comes into the game: many of these techniques are still very expensive. This problem connects to the optimization of the health care system. In the Netherlands, healthcare is regulated by health insurance providers. These providers will be willing to reimburse the above-mentioned treatment options, but only after showing them the efficacy of these treatments on a large scale. Re-defining the evaluation of the best treatment options for a patient is therefore the key to improve cost-effectiveness. Using image derived biomarkers as an additional tool to support clinicians in finding the best treatment option, as stated in this thesis, will, in the long term, boost cost-

effectiveness. Using medical imaging to objectively quantify treatment response, will reduce the risk of providing a treatment which could have been stated to be beneficial at the time of planning, but could become harmful with a strong impact not only of patients' health, but also on unnecessary costs. Even if it was out of scope for this thesis to perform a cost-effective analysis, the developments presented in this work will have an economic impact on the long term. Finally, the approaches presented in this thesis are directly applied to standard of care medical imaging modalities (e.g., PET/CT, MRI). Scans that are acquired anyway in the radiotherapy workflow, therefore no additional investment of money is required to gain more information useful for care decisions.

3 Cultural Impact

This thesis tackles some of the problems related to the introduction of AI to improve our ability to make better-informed decisions. When it comes to the application of AI for the automation of time-consuming tasks, we are more willing to accept that an autonomous artificially intelligent "entity" can replace us. However, when it comes to be supported by AI in our decisions, we are more reluctant to accept this shift in paradigm. I believe that two of the major causes behind these discrepancies are: A) the poor performance of AI-based prognostic and predictive models in radiotherapy, and B) the false myth that AI applications in radiotherapy are meant to replace the users, requiring only negligible human interaction. With regard to the former, this thesis has been devoted to investigating the issues that are causing the degrade of radiomics-based models' performances when validated on multiple datasets. I have proposed methods that can support more transparent and robust developments of image-derived biomarkers and shown how cautions are required when using AI, or more specifically ML and DL, to draw conclusions. With regard to the latter, I have shown that even in the presence of fully or semi-automated image analysis and

modelling pipelines, a human interaction is required to verify the correctness of assumptions, as well as the benchmarking of newly discovered biomarkers with traditionally accepted prognostic of predictive factors in radiation oncology. Finally, in the last chapter of the thesis, I have discussed solutions to improve the acceptance of AI in the clinic, with a dedicated focus on the role of multiple stakeholders. This thesis is proposing the paradigm to re-shift human-centricity when using AI. This will have a strong cultural impact, and, in my opinion, it will boost the acceptance of AI.

4 Technological impact

This thesis did not per se developed a new hardware technology. Nevertheless, it contains two promising potential technological products. The first product is an image-analysis framework which consists of multiple processing pipelines, image harmonization, extraction of image derived biomarkers, quality assurance of these biomarkers and modelling. This pipeline can easily be inserted in the clinical workflow or within the clinical workstations. Potential users are scanner manufactures, as well as companies developing clinical workstations or AI solutions. Second, in the third part of the thesis, a framework based on a distributed learning solution for radiomics was presented. This data infrastructure has the impact to alleviate the effort required to collect and process data on a centralized repository. The technology is not limited to traditional machine learning algorithms, but it can be extended to distributed deep learning. Overall, this technology has the impact that even smaller centres, with availability of fewer data compared for example to large institutions, will be able to perform large scale experiments, with the benefit of training and validating algorithms on data with an order of magnitude larger than the sample size simply available in the single clinic. Finally, this thesis introduced the concept of FAIR withing medical imaging studies. The re-think of radiomic studies as FAIR-compliant experiments enriches the reproducibility of

such experiments and enables inter-operability of multiple radiomic computational packages.

Summary

Medical imaging plays a key role in radiation oncology. Patients' scans are used for diagnosis and tumour staging, treatment planning, delivery and monitoring; and disease follow up. They offer a non-invasive tool to extrapolate not only biological properties of the tumour, but also the relations between cancer cells and surrounding tissues, which are important for example to evaluate the risk of treatment-induced toxicities. Unfortunately, we are still facing a sub-optimal use of medical imaging. Radiological findings from medical images are mainly analysed in a (semi)qualitative fashion using visual inspections. Medical images are then discarded when a specific task is completed. In recent years, the research community has started to re-think the role of medical imaging, considering patients' scans as a source of big data. The hypothesis is that medical images contain quantitative information that is invisible to the human eye, referred to as "radiomics". The availability of automated imaging processing pipelines, based on the AI (Artificial Intelligence) branches of ML (Machine Learning) and DL (Deep Learning), will allow to retrieve this information and use it to develop non-invasive image-derived biomarkers. These biomarkers represent a fingerprint of our patients, and when translated into DSSs (Decision Support System) can move patient care towards personalized treatment. After the rapid hype following the introduction of this technology in radiation oncology, a bottleneck has been reached since several issues limit the rapid translation of radiomics-derived models in the clinic as DSSs. This thesis identifies and proposes solutions to three of these major challenges: A) the lack of robustness of developed biomarkers; B) the absence of a robust methodology for ML in radiomics, and C) privacy-related barriers that impede the validation of developed biomarkers. The work poses a new paradigm to re-think the role of AI in medical imaging, to open a new era for a *radiomics-renewal*.

Published work on international journals

1. AK Jha, S Mithun, V Jaiswar, UB Sherkhane, NC Purandare, K Prabhash, V Rangarajan, A Dekker, L Wee, **A Traverso**. "Repeatability and reproducibility study of radiomic features on a phantom and human cohort". (2021). Scientific reports 11 (1), 1-12

2. P Kalendralis, Z Shi, **A Traverso**, Choudhury, M Sloep, I Zhovannik, MPA Starmans, D Grittner, P Feltens, R Monshouwer, S Klein, R Fijten, H Aerts, A Dekker, J van Soest, L Wee. "FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections". (2020). Radiotherapy and Oncology 152, S834-S835

3. M Welch, **A Traverso**, C Chung, D Jaffray. "Quantiative Radiomics in radiation oncology". (2020). The modern technology of radiation oncology", Volume 4. Edited by Jacob van Dyk. Medical Physics Publishing

4. J Kazmierska, A Hope, E Spezi, S Beddar, WH Nailon, B Osong, A Ankolekar, A Choudhury, A Dekker, K Røe Redalen, **A Traverso**. "From multisource data to clinical decision aids in radiation oncology: The need for a clinical data science community". (2020). Radiotherapy and Oncology 153, 43-54

5. C Rao, S Pai, I Hadzic, I Zhovannik, D Bontempi, A Dekker, Jonas Teuwen, **A Traverso**. "Oropharyngeal Tumour Segmentation using Ensemble 3D PET-CT Fusion Networks for the HECKTOR Challenge". (2020). 3D Head and Neck Tumor Segmentation in PET/CT Challenge, 65-77

6. **A Traverso**, M Kazmierski, I Zhovannik, M Welch, L Wee, D Jaffray, A Dekker, A Hope. "Machine learning helps identifying volume-confounding effects in radiomics". (2020). Physica Medica 71, 24-30

7. ML Welch, C McIntosh, **A Traverso**, L Wee, TG Purdie, A Dekker, B Haibe-Kains, DA Jaffray. "External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification". (2020). Physics in Medicine and Biology 65 (3), 035017

8. ML Welch, C McIntosh, A McNiven, S Hui Huang, BB Zhang, L Wee, A Traverso, B O'Sullivan, F Hoebers, A Dekker, DA Jaffray. "User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions". (2020). Physica Medica 70, 145-152

9. **A Traverso**, M Kazmierski, ML Welch, J Weiss, S Fiset, WD Foltz, A Gladwish, A Dekker, D Jaffray, L Wee, K Han. "Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients". (2020). Radiotherapy and Oncology 143, 88-94

10. ML Welch, C McIntosh, TG Purdie, L Wee, **A Traverso**, A Dekker, B Haibe-Kains, D A Jaffray. "Automatic classification of dental artifact status for efficient image veracity checks: effects of image resolution and convolutional neural network depth". (2020). Physics in Medicine and Biology 65 (1), 015005

11. Z Shi, **A Traverso**, J van Soest, A Dekker, L Wee. "Ontology-guided radiomics analysis workflow (O-RAW)". (2019). Medical physics 46 (12), 5677-5684

12. I Zhovannik, J Bussink, **A Traverso**, Z Shi, P Kalendralis, L Wee, A Dekker, R Fijten, R Monshouwer. "Learning from scanners: bias reduction and feature correction in radiomics". (2019). Clinical and translational radiation oncology 19, 33-38

13. Z Shi, I Zhovannik, **A Traverso**, FJWM Dankers, TM Deist, P Kalendralis, R Monshouwer, J Bussink, R Fijten, HJWL Aerts, A Dekker, L Wee. "Distributed radiomics as a signature validation study using the Personal Health Train infrastructure". (2019). Scientific data 6 (1), 1-8

14. S Fiset, ML Welch, J Weiss, M Pintilie, JL Conway, M Milosevic, A Fyles, **A Traverso**, D Jaffray, U Metser, J Xie, K Han. "Repeatability and reproducibility of MRI-based radiomic features in cervical cancer". (2019). Radiotherapy and Oncology 135, 107-114

15. **A Traverso**, M Kazmierski, Z Shi, P Kalendralis, M Welch, H Dahl Nissen, D Jaffray, A Dekker, L Wee. "Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing". (2019). Physica Medica 61, 44-51

16. P Kalendralis, **A Traverso**, Z Shi, I Zhovannik, R Monshouwer, MPA Starmans, S Klein, E Pfaehler, R Boellaard, A Dekker, L Wee. "Multicenter CT phantoms public dataset for radiomics reproducibility tests". (2019). Medical physics 46 (3), 1512-1518

17. L Wee, SMJ van Kuijk, FJWM Dankers, **A Traverso**, M Welch, A Dekker. "Reporting Standards and Critical Appraisal of Prediction Models". (2019). Fundamentals of Clinical Data Science, 135-150

18. **A Traverso**, FJWM Dankers, L Wee, SMJ van Kuijk. "Data at Scale". (2019). Fundamentals of Clinical Data Science, 11-17

19. **A Traverso**, FJWM Dankers, B Osong, L Wee, SMJ van Kuijk. "Diving deeper into models". (2019). Fundamentals of clinical data science, 121-133

20. FJWM Dankers, **A Traverso**, L Wee, SMJ van Kuijk. "Prediction modeling methodology". (2019). Fundamentals of clinical data science, 101-120

21. SMJ van Kuijk, FJWM Dankers, **A Traverso**, L Wee. "Preparing data for predictive modelling". (2019). Fundamentals of clinical data science, 75-84

22. L Vassallo, **A Traverso**, M Agnello, C Bracco, D Campanella, G Chiara, ME Fantacci, E Lopez Torres, A Manca, M Saletta, V Giannini, S Mazzetti, M Stasi, P Cerello, D Regge. "A cloud-based computer-aided detection system improves identification of lung nodules on computed tomography scans of patients with extrathoracic malignancies". (2019). European radiology 29 (1), 144-152

23. **A Traverso**, L Wee, A Dekker, R Gillies. "Repeatability and reproducibility of radiomic features: a systematic review". (2018). International Journal of Radiation Oncology* Biology* Physics 102 (4), 1143-1158

24. **A Traverso**, J Van Soest, L Wee, A Dekker. "The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques". (2018). Medical physics 45 (10), e854-e862

25. AAA Setio, **A Traverso**, et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge". (2017). Medical image analysis 42, 1-13

26. **A Traverso**, EL Torres, ME Fantacci, P Cerello. "Computer-aided detection systems to improve lung cancer early diagnosis: state-of-the-art and challenges". (2017). Journal of Physics: Conference Series 841 (1), 012013

Curriculum Vitae

Alberto Traverso was born in Italy on March 2^{nd} 1988. He holds a bachelor's in theoretical physics and a master's in medical physics from University of Turin. He joined the department of Radiology in the Oncological Institute (IRCCS) Candiolo in Italy, working as a clinical physicist on lung cancer screening. In 2017, he joined Maastro Clinic and started as a PhD student at Maastricht University at the Faculty of Health, Medicine and Life Sciences, under the supervision of Prof. Andre Dekker. He has been a visiting researcher at several European and American hospitals including: Veijle Hospital (Denmark), Princess Margaret Cancer Centre (Toronto, Canada), Moffitt Cancer Centre (Tampa, Florida), MD Anderson Cancer Centre (Houston, Texas), University of Michigan Hospital (Ann Arbor, Michigan). He has authored several publications in the field of medical image analysis in radiation oncology and clinical data science. He has received several personal fellowships including: ESTRO travel grant, Rene Vogel travel grant, Early career scholarship from AAPM, multiple ZonMw travel grants. He is currently employed as a senior researcher at Maastro Clinic. Alberto's vision is that medical images empowered with AI will provide patients' unique biomarkers to personalize cancer treatment.