Maastricht University

Artificial intelligence applications in radiotherapy

Citation for published version (APA):

Kalendralis, P. (2022). *Artificial intelligence applications in radiotherapy: The role of the FAIR data principles*. [Doctoral Thesis, Maastricht University]. Maastricht University. https://doi.org/10.26481/dis.20221010pk

Document status and date:
Published: 01/01/2022

DOI:
10.26481/dis.20221010pk

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 05 Oct. 2023

Artificial intelligence applications in radiotherapy-The role of the FAIR data principles

Petros Kalendralis

Cover

The cover was designed by the Italian artist Davide Magliacano.

Printing

This thesis was printed by Ridderprint, www.ridderprint.nl

ISBN: 978-94-6458-519-3

Copyrights

@ Copyright Petros Kalendralis, Maastricht 2022

Artificial intelligence applications in radiotherapy-The role of the FAIR data principles

DISSERTATION

to obtain the degree of Doctor at the Maastricht University, on the authority of the Rector Magnificus, Prof. dr. Pamela Habibović in accordance with the decision of the Board of Deans, to be defended in public on

Monday, 10th of October 2022, at 10:00 hours

by Petros Kalendralis

Supervisor:

Prof. dr. ir. Andre Dekker

Co-Supervisors:

Dr. Rianne Fijten

Dr. Johan van Soest

Assessment Committee:

Prof. dr. ir. Frank Verhaegen (Chair)

Prof. dr. ir. Michel Dumontier

Dr. Aiara Lobo Gomes

Dr. Joanna Kazmierska, Greater Poland Cancer Centre, Poznan, Poland

Prof. dr. Vincenzo Valentini, Università Cattolica del Sacro Cuore, Rome,Italy

The work of this thesis has been funded through the Dutch Cancer Society KWF(grant number 30943405N)

Table of contents

Chapter 1

General introduction and thesis structure

1.1 Radiotherapy and medical imaging

Radiotherapy (RT) plays a crucial role in the treatment of numerous cancer types[1]. In fact in Europe, a significant percentage of cancer patients (47%-53% depending on the diagnostic tumour stage and the anatomical tumour site of treatment) undergo RT as a main or concurrent treatment modality[2]. The main goal of RT is to deliver maximum radiation dose to the malignant cells of the tumour while minimising the radiation dose to the surrounding healthy tissues which are referred to as organs at risk (OARs). This approach gives the highest chance of destroying or eliminating the tumour cells, while allowing normal tissues to recover and suffer from as little toxicities as possible[3].

Different modalities and doses of RT treatment are chosen depending on the tumour characteristics, such as the TNM stage and anatomic tumour site. For instance, for the head and neck RT treatment, studies proposed and compared different RT treatment modalities such as the intensity modulated radiation therapy (IMRT) and Volumetric Modulated Arc therapy (VMAT)[4–6]. Curative RT is intended to cure and eliminate a significant large amount of the malignant cancer cells, while palliative RT aims to alleviate pain and delay the tumour growth[7].

Linear accelerators (LINACs) are the devices most commonly used for the delivery of external beam radiotherapy (EBRT) where the radiation source is located externally to the patient's body during treatment. LINACs have been used in RT since the 1950s[8]. The LINAC produces particle beams consisting of photons and electrons with a range from 4 to 25 MegaVolts (MV). Some superficial tumours are being treated using X-ray tubes but tumours that are located deeper in the organs require more energetic beams produced by LINACs.

Relatively novel techniques such as proton RT are also introduced for clinical purposes. Proton therapy's greatest advantage over conventional RT is its accuracy and its ability to spare healthy tissues. Conventional X-rays are made of photons that pass through the body and deposit a considerable amount of energy to the tumour target and the surrounding healthy tissue. Proton particles have a particular physical property called the Bragg peak[9]. They are able to deposit less amount of energy on their way to the tumour target and the surrounding tissues beyond it. This property allows to confine the radiation dose to the tumour and radically decreases the risk of radiation induced toxicities to the healthy tissues.

The proton particles originate from the ion source, where hydrogen atoms are separated into electrons and protons. The protons are injected into a machine called cyclotron where they are accelerated. Acquiring high velocity, the protons are sent through an energy selection system and a degrader that adjust their energy. The proton beam "transport" system conducts the proton particles with the correct trajectory and energy. Finally, the proton beam arrives in the treatment room using a gantry that revolves 360 degrees around the patient for the delivery of the beam.

The RT workflow pipeline can be divided in four parts: (i) diagnosis, (ii) treatment planning, (iii) treatment delivery and (iv) follow-up (figure 1.1). During diagnosis, imaging is mainly used to determine the stage (TNM classification), location and the size of the tumour. Furthermore, there is an important amount of demographic baseline information that has to be taken into account for the selection of the appropriate treatment that will most benefit the patient. Demographic characteristics such as the age or the past medical history of patients influence the treatment selection decision. For instance, minority and aged (cancer) patients have an increased risk of undertreatment and underestimation of pain[10,11]. Usually, depending on the treatment protocol and the fractionation schema (the daily RT dose delivery), the time period between the first diagnostic imaging scan and the end of RT treatment of a patient differs depending on the treatment, the tumour location and the fractionation schedule selected according to the hospitals' treatment protocols[12].

The RT treatment planning procedure is a crucial component of the therapeutic process of patients. Each patient's case is being analysed in a multidisciplinary team meeting where the tumour pathology and staging are determined. Moreover, decisions are taken regarding the RT treatment intent (palliative or radical) and RT treatment modality (e.g. photons, protons, electrons or brachytherapy), the dose prescription and fractionation regime from the radiation oncologists. In most cases Computed Tomography (CT) scans are preferred for the treatment planning procedure as they provide an electron density map which is necessary for the RT dose calculation in the treatment planning system (TPS) software. In the TPS, the anatomical tumour target(s) is (are) delineated as well as with the different OARs in the treatment planning scans for the dose calculation and plan simulation.

In the past years, the introduction of AI algorithms enabled the automatic delineation of several anatomical structures as well as the implementation of automated treatment planning[13]. Moreover, the quality assurance (QA) tests for the efficient execution and delivery of a RT treatment planning include the audit of several technical and physical parameters. For instance, the RT dose prescription labelling, physical parameters, treatment scheduling, patients' set-up instructions and dose volume histogram (DVH) parameters[12,14] are being checked. Specifically for the DVH parameters, due to the high complexity of the RT treatment planning procedure and the goal of achieving as much as possible dose sparing to the OARs, significant efforts have been made for automated-individualised QA using knowledge-based DVH predictions[15].

Furthermore, during the treatment planning verification phase, the RT treatment plan calculated in the TPS undergoes QA checks before its delivery by the treatment machines (LINACs). These QA checks usually include phantoms, which are devices used for the calibration of the delivery machines verifying and ensuring that there are no dosimetric differences between the TPS calculated plan and the delivered one to the patients[16]. Moreover statistical metrics such as the gamma index (GI)/ gamma (γ) pass rate in the pretreatment patient plan verification are used for the quantitative evaluation of dose distribution to the patients before their treatment. Finally, during the follow up stage, medical images are acquired after the end of the treatment for tumour monitoring purposes. Specifically, the follow up stage is important for identification of a recurrent tumour or a tumour spread in other anatomical sites.

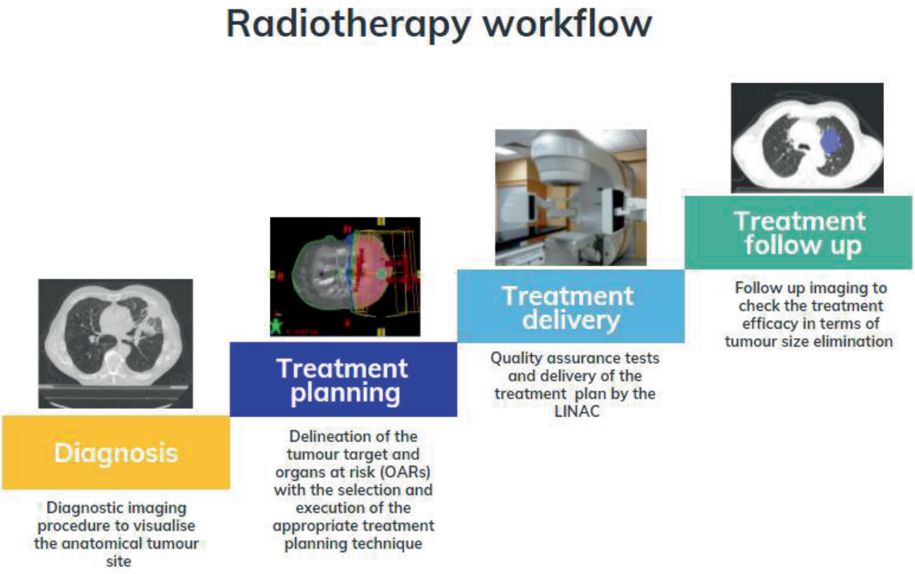
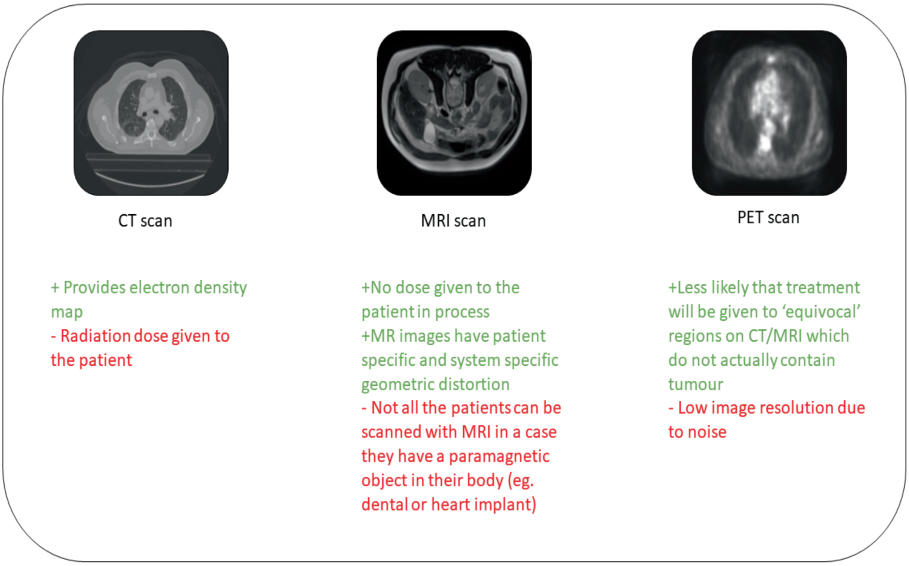

Radiotherapy workflow
Diagnosis
Diagnostic imaging procedure to visualise the anatomical tumour site
Treatment planning
Delineation of the tumour target and organs at risk (OARs) with the selection and execution of the appropriate treatment planning technique
Treatment delivery
Quality assurance tests and delivery of the treatment plan by the LINAC
Treatment follow up
Follow up imaging to check the treatment efficacy in terms of tumour size elimination

Figure 1.1: Synopsis of the four components of the RT workflow. Starting with the diagnosis of the disease on the left and the treatment planning procedure using various imaging modalities depending on the anatomical tumour site that has to be treated. Subsequently, the treatment delivery follows where the treatment plan is “transferred” from the TPS to the LINAC after passing the quality assurance (QA) tests. Finally, in the follow up phase, the treatment efficacy is verified in terms of tumour elimination after the end of the RT course using different imaging modalities.

Medical imaging is tightly connected with each of the four components of the RT workflow. Different imaging modalities are being used for the diagnostic, treatment planning and follow-up purposes. The most common imaging modalities used for diagnostic and treatment planning purposes are CT, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) (figure 1.2). Mainly CT scans are used for treatment planning due to the electron density maps which are necessary for the dose calculation, while a small amount of radiation dose is given to the patient. MRI is mainly used for the better representation of the soft tissues having the advantage of not delivering radiation dose to the patients. Patients with paramagnetic objects in their body cannot be scanned with MRI according to the safety protocols of MRI due to the strength of the magnetic field. PET or PET/CT imaging is used mainly for the better visualisation/inspection of the targets that are not well visualised by CT or MRI. The low resolution due to the imaging noise of PET scans constitutes a disadvantage.



+ Provides electron density map
- Radiation dose given to the patient

+No dose given to the patient in process
+MR images have patient specific and system specific geometric distortion
- Not all the patients can be scanned with MRI in a case they have a paramagnetic object in their body (eg. dental or heart implant)

+Less likely that treatment will be given to 'equivocal' regions on CT/MRI which do not actually contain tumour
- Low image resolution due to noise

Figure 1.2: Overview of the three main imaging modalities used for diagnostic and treatment planning purposes in RT specifying the different advantages (+ green colour) and disadvantages (- red colour) of each imaging modality. The scans included in the figure are part of the following publicly available imaging collections stored in xnat.bmia.nl. Specifically, the CT scan is part of the STW_STRATEGY_MAASTRO_LUNG1 collection[17], the MRI scan is part of the WORC imaging collection[18] and the PET scan is available through the STW_STRATEGY_MULTIDELINEATION imaging collection[19].

1.2 Artificial intelligence applications in Radiotherapy

During the last years, significant technological advances in the field of computer science introduced the implementation of Artificial Intelligence (AI) in different disciplines. RT is one of the disciplines in medicine where AI techniques have started to manifest their potential. AI can be defined as the ability of machines to demonstrate and display actions that usually are performed by humans[20]. These actions include problem-solving, decision making, learning and classification tasks. The emerging need to improve the quality of each part of the RT workflow as well as the RT-based patient outcomes have stimulated the introduction of AI techniques in RT. In addition, the fact that RT is an information technology (IT) driven discipline in medicine, makes AI somewhat easier to introduce in RT.

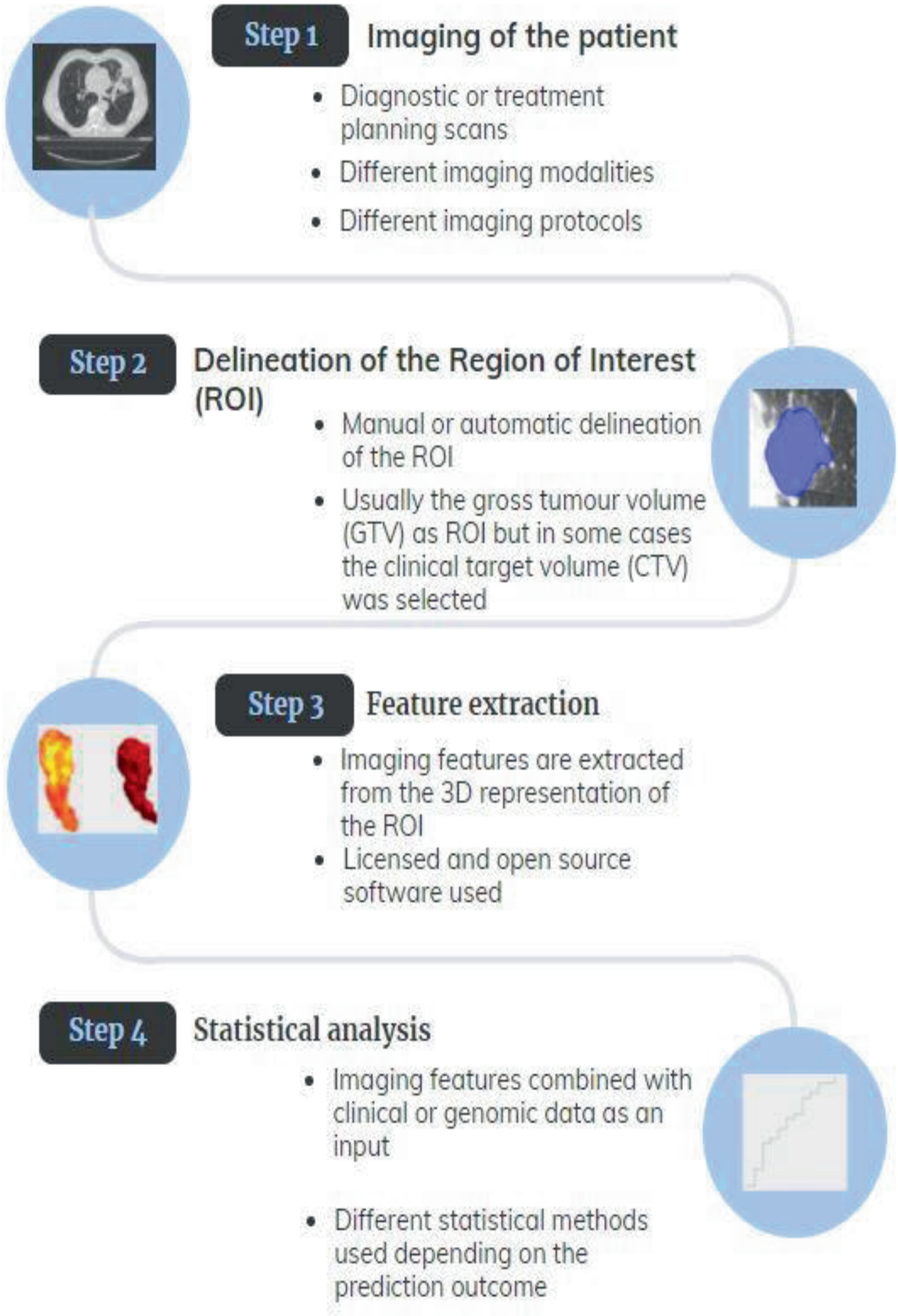
Taking into account that a plethora of RT routine clinical tasks include time-consuming and labour-intensive tasks such as the delineation of the different OARs during the RT treatment planning phase or the LINAC QA checks before the execution of treatment plans, AI applications can potentially play a significant role in the field. Machine learning (ML) and deep learning (DL) applications as branches of AI can substantially improve the quality and accuracy of treatment delivery as a supporting tool to RT professionals such as radiation oncologists, medical physicists and radiation technologists (RTTs)[21].

ML can be defined as the branch of AI that has the ability to perform tasks by learning from data using computational power without being explicitly programmed[22]. DL is a branch of ML and can be defined as a group of algorithms or methods that is used for supervised and unsupervised learning based on artificial neural networks (ANNs)[23]. For supervised learning tasks, DL translates data into representations similar to principal components deriving layered structures, while for unsupervised learning, generative adversarial networks (GANs) and autoencoders are the main components for the generation of synthetic data based on different datasets[24]. Applications such as the automatic delineation of the different anatomical targets [25,26] or DL-based automated treatment planning[27] have been proven to improve the RT process and in some cases to have better performance than humans[28]. Prognostic modelling and classification tasks are some of the applications of ML in RT implemented using mathematical models and algorithms.

1.3 Radiomics

The integration of AI and ML techniques in RT, in combination with the vast amount of multimodality imaging data resulted in the start of a new era in RT. Specifically, the information provided visually from medical images can be enriched with quantitative information of the tumour, extracted from the pixel information in the medical images. This information is known as imaging features and can potentially reveal significant information regarding the tumour phenotype. In combination with ML techniques, it can compose an ML-based image analysis framework towards patients' personalised treatment approach. This framework that transforms diagnostic or treatment planning patients’ scans into a mineable knowledge is known as “radiomics”[29].

The concept of radiomics has as its main components the identification and quantification of tumour characteristics and statistical modelling, aiming to adapt and personalise patients’ treatment in terms of treatment planning, outcome prediction, treatment response and decision-making. Radiomics was first introduced in 2012[30] as a new concept and currently constitutes an active research discipline in the field of radiation oncology and RT. The radiomics pipeline includes four main stages. (i) Imaging of the patient, (ii) region of interest (ROI) delineation, (iii) imaging features extraction and (iv) statistical analysis/modelling (figure 1.3). Each of these stages has different technical characteristics. Radiomics have been applied to different imaging modalities such as CT[31], PET[32] or MRI[33], with manual and semi-automatic ROI delineation[34], with the latter having promising results in terms of the robustness of radiomics features using AI-based semi-automatic algorithms[35], while the statistical analysis can be adjusted and differs depending on the feature selection method or the predicted outcome. It should be highlighted that a significant amount of radiomics studies have been applied on CT images due to their high availability in RT departments for treatment planning purposes. Furthermore, limitations such as the plethora of different MRI scanning protocols and the technical limitations of PET imaging (ie. low resolution and noise) influence the reproducibility of radiomics models based on these two imaging modalities[36,37].



Step 1 Imaging of the patient

- Diagnostic or treatment planning scans
- Different imaging modalities
- Different imaging protocols

Step 2 Delineation of the Region of Interest (ROI)

- Manual or automatic delineation of the ROI
- Usually the gross tumour volume (GTV) as ROI but in some cases the clinical target volume (CTV) was selected

Step 3 Feature extraction

- Imaging features are extracted from the 3D representation of the ROI
- Licensed and open source software used

Step 4 Statistical analysis

- Imaging features combined with clinical or genomic data as an input
- Different statistical methods used depending on the prediction outcome

Figure 1.3: The four different steps of the radiomics pipeline. The first step includes the imaging of the patients. Different imaging modalities during the different phases of the RT workflow can be used in radiomics depending on the scope of the radiomics study. Secondly, the ROI delineation takes place. The delineation can be done manually or automatically using AI algorithms. The third radiomics step includes the imaging features extraction using the 3D ROI representation. The extraction can be done using various open source or licensed software. Last, the imaging features extracted are often combined with other types of data and used as an input for statistical analysis and prediction modelling.

One of the most important and promising aspects of the radiomics concept is the possibility to acquire information regarding the tumour phenotype using quantitative imaging features. This information cannot be easily observed and acquired by the human eye of radiologists or radiation oncologists[38]. It is worth highlighting that the imaging features included in a radiomics study can be enriched with other feature types such as clinical data from the electronic health record (EHR) systems containing demographic, treatment response and follow-up information.The combination and inclusion of imaging and clinical (or other feature types) in the statistical analysis of the radiomics workflow may contribute to the robustness of the radiomics statistical output[29]**.** Similarly, AI plays a significant role in the radiomics framework via the ML algorithms used in statistical model building[39,40]. Hence, AI can be considered one of the key components of the transformation of the patients scans into mineable knowledge towards personalised patient care without the need for human observation.

Currently, the oxymoron fact with the radiomics-based prediction models is that they are not yet introduced in a clinical environment, despite the high amount of radiomics publications in the literature. In the last decade, several studies investigated the potential of radiomics in different imaging modalities, having as a starting point the publication of Aerts et al.[31] in 2014 that introduced the potential value of radiomics in survival prognosis of lung and head and neck patients.

In a later stage, radiomics literature focused on several applications such as the classification of malignant tissues for diagnostic purposes[41], the “virtual” biopsy approach combining radiomics and genomics data (ie.“radiogenomics”) [42] and the investigation of the scanners variability influence on radiomics features values using phantoms[43]. Although there were significant efforts of the radiomics community, the incorporation of radiomics models in the clinical routine of RT professionals is still an unmet milestone due to significant barriers. Specifically, some of the characteristic barriers are the inconsistencies in the delineations of the imaging input data, absence of standardisation of the imaging features extraction and computation pipeline and the difficulties regarding radiomics-based models exchange between different centres, hampering the reproducibility of radiomics research.

In this thesis, in chapter 2, the main pitfalls of the radiomics pipeline will be identified as well as the key points for the standardisation and reproducibility of the radiomics framework. A main component in this chapter will focus on data sharing and data interoperability among radiomics researchers.

1.4 Artificial Intelligence-based quality assurance in radiotherapy

Quality assurance (QA) checks in RT are part of the daily routine undertaken by medical physicists and RTTs of RT departments. These QA checks are performed during all the stages of the RT workflow and constitute a time-consuming and labour intensive procedure. They occupy a significant portion of machines downtime and part of the clinical routine day of RT professionals[44]. The three main categories of QA checks are (i) QA of delivery machines and medical equipment, (ii) QA of patients treatment plans and (iii) QA checks for errors detection before the execution of the treatment plan. All these QA checks categories require special equipment or analytics such as QA phantoms, radiochromic dosimetry films and gamma (γ) pass rate statistical analysis.

To reach the ultimate goal to reduce the number of and time needed for QA checks in RT, such as replacing a patient’s treatment plan QA with an automated test predicted by a ML algorithm, the RT professionals community should have validated and robust tools in their hands[45]. During the past years, several studies introduced various ML algorithms for QA checks having as a main goal to reduce the QA workload of the RT professionals. Specifically, they focused on the investigation of the prediction of gamma (γ) index pass rate for the patient's specific treatment plan verification using different ML algorithms including convolutional neural networks (CNNs)[46,47], Poisson regression[48], Random forest[49] and support vector classifiers[50(p)]. Moreover, ML algorithms were used for the prediction of errors regarding the position of multileaf collimators into TPS computations using random forest and regression algorithms [51,52(p)].

Another promising ML concept regarding the treatment plan verification was introduced by Luk et al.[53] using Bayesian networks. This Bayesian networks approach proposed a probabilistic model for the early detection of RT treatment planning errors using simulated errors concerning LINAC mechanical, patient positioning and general treatment planning errors. The transparency of Bayesian networks based on the network query possibility, that gives the flexibility to explain the different decisions made via the different connections of the network, makes Bayesian networks a strong candidate of a ML algorithm that has the potential to support clinical aids in RT regarding the early detection of RT treatment planning errors.

However, despite the fact that the introduction of ML in QA checks has the potential to eliminate the human effort and time needed, there are still a lot of steps to be done for the introduction of this novel technology into the daily clinical routine of RT professionals. The reasons behind this hesitant introduction of the above mentioned ML techniques are mainly connected to the high and advanced data or computer science skills required for the development of these technologies. Specifically, it is a common phenomenon that a significant percentage of RT professionals in Europe such as the medical physicists lack AI knowledge while a high percentage of them recognise its significance and added value in the clinical routine procedures[54]. Moreover, the introduction and implementation of these novel techniques by the RT professionals is accompanied by the “black-box” approach as the understanding and functionalities of AI algorithms is quite problematic.

In chapter 7 of this thesis, we externally validated a Bayesian network approach developed in the University of Washington Medical Centre for the early detection of RT treatment plan errors using an external validation dataset of Maastro Clinic in the Netherlands.

1.5 "Big-data" and prognostic modelling in radiotherapy-the need for validation

As discussed in the previous section, the new era of AI in the RT QA procedures requires the acquisition and curation of large scale datasets[55]. Likewise, the implementation of AI techniques in RT requires the acquisition and use of a high amount of data for training and validation purposes. As mentioned, diagnostic and planning images in combination with demographics or baseline clinical characteristics and patients' records are used for the determination of the treatment strategy from the multidisciplinary group (MDT) of radiation, medical and surgical oncologists.This amount of data is stored digitally for every patient across the different departments involved in the cancer treatment of each hospital such as the RT, medical oncology, radiology, nuclear medicine or surgery department. Moreover, these data are stored in different EHR systems and imaging archives such as the picture archiving and communication system (PACS) in different formats (.txt, .doc , .pdf, JPEG, DICOM)[56]. An overview of the different data types of each different stage of the RT workflow is shown in figure 1.4.

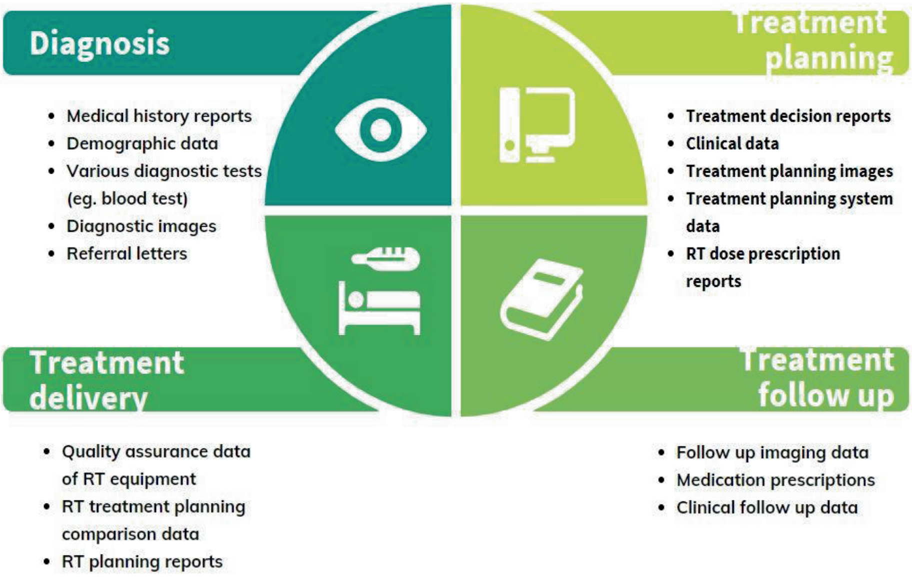

Diagnosis
• Medical history reports
• Demographic data
• Various diagnostic tests (eg. blood test)
• Diagnostic images
• Referral letters
Treatment planning
• Treatment decision reports
• Clinical data
• Treatment planning images
• Treatment planning system data
• RT dose prescription reports
Treatment delivery
• Quality assurance data of RT equipment
• RT treatment planning comparison data
• RT planning reports
Treatment follow up
• Follow up imaging data
• Medication prescriptions
• Clinical follow up data

Figure 1.4: An overview of the multisource data of each stage of the RT workflow. Different imaging or data from different department's reports are "produced" in every stage. These multisource data are stored in different sources such as EHR systems, PACS and TPS databases.

A recent study estimated that the number of cancer patients worldwide was 19.3 million (GLOBOCAN 2020)[57]. In terms of data volume a minimum of 0,1 GBs of data is generated per patient[56]. This vast amount of data, which is also referred as “Big-data”, has the potential to transform the radiation oncology landscape with their responsible and appropriate exploitation adding a significant contribution towards personalised data-driven RT. Yet, this enormous amount of data cannot be processed fully manually by humans.

Therefore, there is a need to develop ML algorithms that can translate this data into meaningful knowledge that could potentially support the RT professionals in their routine operations. RT outcomes prediction modelling, automated delineation of OARs, automated treatment planning, imaging guidance RT and the quality assurance (QA) checks of RT delivery devices constitute some of the applications that can potentially benefit from this new era of data-driven RT[58].

Nonetheless, the meaningful translation of these data into knowledge that can improve or support the decision making process in RT is more than a time consuming and labour intensive procedure. RT data come from different sources such as TPS, radiology/RT reports and imaging data and are stored in different archive systems. Moreover, the different data are labelled and registered with hospitals’ specific languages and terminologies which leads to an interoperability issue. Furthermore, multicentric data exchange constitutes a prerequisite for responsible and reproducible RT research. Especially for RT data-based prediction modelling studies, external validation based on the TRIPOD statement, is a crucial step before the clinical implementation of a model[59]. Nevertheless, there are some barriers that make RT data exchange between different centres problematic. These barriers are usually related to the privacy and data security regulations of each hospital or data “owner” as RT data holds patients' sensitive personal information.

Prognostic modelling has as its main objective the improvement and the introduction of personalised treatment approaches in RT via the development and implementation of mathematical

models and algorithms. These algorithms can potentially calculate the risk or the probability of a specific outcome or toxicities rates during the RT treatment or the prediction of survival of a specific patient's cohort with a specific disease. For instance, one of the applications of these prognostic algorithms is used in the Netherlands is the model based approach (MBA).With this approach, ML algorithms predict and evaluate the patients who might benefit from proton or photon RT comparing normal tissue complication probabilities (NTCP) rates[60].

The oxymoron fact with the AI prognostic algorithms is that their construction and application is not a complicated procedure but on the contrary, their (external) validation can be problematic due to various factors. For example, different patient characteristics, different scanning protocols and different treatment fractionation schedules across different centres are some of the factors that make the generalisation, reproducibility and therefore clinical implementation of these AI-based prognostic models challenging. During the past years initiatives for standardised prediction models reporting guidelines have been demonstrated such as the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)[61] having as a goal the full evaluation of prediction models. Nonetheless, the fully transparent and adequate report and description of prediction models is not achievable due to barriers within RT. Despite the high and emerging need of validation of the novel field of AI-based prediction models in RT, data sharing complications such as legal or ethical administration protocols, absence of an adequate amount of data validation cohorts and technical issues such as code and software not being publicly available are some of the barriers that make it difficult.

On the other hand, taking into account the multiple imaging modalities and EHR systems or data archives, we can conclude that there is a multi-source data production in the RT workflow. This amount of data is usually stored in different formats with missing data items as it is challenging to have them structured in a standardised format and make them available for research or clinical purposes. This has a negative impact in the reproducibility of RT findings as the RT data are not usually interoperable and reusable from internal and external users. In addition, the privacy and ethical regulations due to the privacy-sensitive RT data contribute to the above-mentioned problem of data interoperability and reusability. Summing up, there is an emerging need for a standardised data acquisition-management framework within RT that enables responsible and interoperable data use and re-use.

1.6 The introduction and implementation of the FAIR data principles in radiotherapy

Taking into account the rapid developments of the AI and ML techniques accompanied with the “big-data” era, the RT landscape changes. The RT community should re-orientate its principles regarding the use of this valuable information from the patients data and the implementation of these novel technologies. One of the approaches that should be taken into account by the RT research and clinical community is the usage of data according to the Findable, Accessible, Interoperable and Reusable (FAIR) data principles (figure 1.5)[62]. These principles have the potential to transform the data into a machine-readable format assisting humans on data searching and processing, applying AI or ML algorithms on them and exchanging them via interoperable standards and terminologies.

In the RT field, the FAIR principles constitute a prerequisite standard format for federated learning studies. In chapters 4 and 5 of this thesis the FAIR principles are applied to radiomics publicly available datasets and distributed radiomics-based survival models studies between two different hospitals without the exchange of patients data towards achieving the goal of interoperable and reusable prognostic models. Moreover, in this thesis the creation of a FAIR data model approach is presented for the standardisation of the Dutch proton therapy data registry.

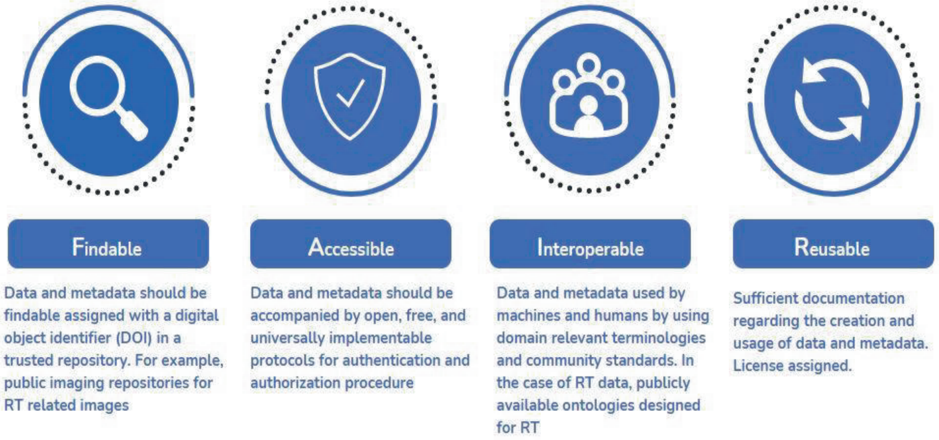

Findable
Data and metadata should be findable assigned with a digital object identifier (DOI) in a trusted repository. For example, public imaging repositories for RT related images
Accessible
Data and metadata should be accompanied by open, free, and universally implementable protocols for authentication and authorization procedure
Interoperable
Data and metadata used by machines and humans by using domain relevant terminologies and community standards. In the case of RT data, publicly available ontologies designed for RT
Reusable
Sufficient documentation regarding the creation and usage of data and metadata. License assigned.

Figure 1.5: Brief description of each FAIR data principle. In this thesis, in chapter 8 we will analyse further the implementation of the FAIR principles in the RT domain, proposing a collaboration framework of all the RT professionals that can accelerate the FAIR data concept in RT.

As mentioned, one of the main and challenging goals of FAIR data infrastructures is the interoperability of research findings so that several users can reuse and develop further knowledge. Technically, this can be achieved by the use of universal and publicly available vocabularies or terminologies that describe concisely and efficiently the several data items used in a RT database or a prognostic model. Specifically, the introduction of ontologies in the field of RT enables the interoperability required for a FAIR data model providing significant semantic domain knowledge. Ontologies can be defined as computer-compatible knowledge terminologies that “demand” relationships between the different data elements. Recently,the domain of ontologies has been introduced the field of radiation oncology by the ontologies focused working group of the American Association of Physicists in Medicine (AAPM) having as a main goal the introduction of ontologies in the RT community providing a sufficient usage and construction guide and underlining the main concepts of them[63].

The introduction of the previously mentioned AI techniques, including the introduction of the FAIR data principles in the modern data management plan introduces a new era in the RT professionals. An era where different professional disciplines such as computer scientists, data scientists and medical physicists need to collaborate with each other to supplement each other's lack of knowledge. Currently, there are initiatives in the RT community such as the working group in AI from the European Federation of Organisations For Medical Physics (EFOMP) where there is a discussion regarding the update of educational curricula of medical physicists and RTTs with integration of basic AI concepts of AI. Moreover, collaborations between the different RT stakeholders are emerging for the formation of a common vision regarding the future of the data-driven RT landscape. Figure 1.6 represents the common RT data workflow from the data storage systems to the final research or clinical data-driven output indicating the importance of the FAIR data principles.

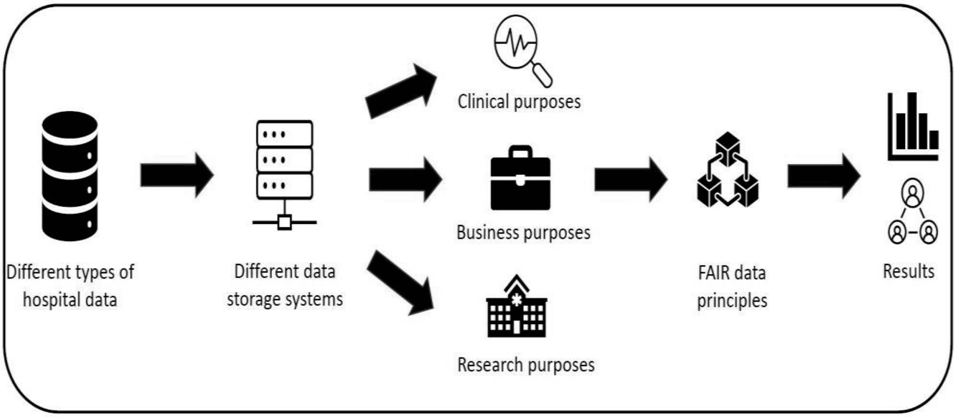


Figure 1.6: Overview of the typical RT data workflow. Multisource data are stored in different archive systems. Different extraction mechanisms in combination with the necessary legal actions make the data available to the different RT stakeholders for different purposes. The FAIR data principles stress mainly the interoperability and reusability aspect of the results output making it findable and accessible to the different stakeholders of RT as they are identified in chapter 8 of the thesis.

1.7 Thesis structure

The four main components of the thesis can be separated in (i) the introduction and implementation of the radiomics concept in RT (ii) the prediction modelling using "Big-data" in RT (iii) the application of AI techniques for QA in RT and the (iv) introduction of the FAIR data principles in RT. The structure of the thesis is as follows: Chapter 2 presents the radiomics concept in the radiation oncology landscape and its pitfalls and uncertainties, providing a roadmap for the standardisation of the radiomics workflow according to the Image Biomarker Standardisation Initiative (IBSI)[64]. Furthermore, the steps for the clinical integration of the radiomics concept are described. Chapter 3 provides a publicly available dataset of phantoms scanned in three different RT centres, having as a goal the reproducibility of radiomics studies. Chapter 4 introduces the FAIR data principles with publicly available radiomics datasets. The main goal of this chapter is to underline the importance of transparent radiomics research using publicly available datasets while using the FAIR principles to integrate multisource data. In chapter 5, the FAIR data principles constitute a prerequisite format for distributed learning radiomics-based prediction modelling using a federated infrastructure. Chapter 6 describes the external validation of ML-based RT prediction models evaluating NTCP models related to dysphagia using an external validation cohort of patients candidates for proton therapy in the Netherlands. Chapter 7 introduces the QA of RT treatment planning using AI algorithms and specifically bayesian networks underlining also the emerging need for external validation. Furthermore, chapter 8 provides a vision on how the RT community can integrate and implement the FAIR data principles in clinical and research studies providing an overview of action items for the RT stakeholders. Finally, in chapter 9 and 10 technical implementations of the FAIR principles are presented with semantic models created for the Dutch proton patients cancer registry purposes.

Bibliography

1. Elwood JM, Sutcliffe SB, eds. Cancer Control. Vol 1. Oxford University Press; 2013. doi:10.1093/med/9780199550173.001.0001
2. Borras JM, Lievens Y, Grau C. The need for radiotherapy in Europe in 2020: Not only data but also a cancer plan. Acta Oncologica. 2015;54(9):1268-1274. doi:10.3109/0284186X.2015.1062139
3. Washington CM, Leaver DT, eds. Principles and Practice of Radiation Therapy. Fourth edition. Elsevier Mosby; 2016.
4. Lee TF, Chao PJ, Ting HM, et al. Comparative analysis of SmartArc-based dual arc volumetric-modulated arc radiotherapy (VMAT) versus intensity-modulated radiotherapy (IMRT) for nasopharyngeal carcinoma. Journal of Applied Clinical Medical Physics. 2011;12(4):158-174. doi:10.1120/jacmp.v12i4.3587
5. Mashhour K, Kamaleldin M, Hashem W. RapidArc vs Conventional IMRT for Head and Neck Cancer Irradiation: Is Faster Necessary Better? Asian Pac J Cancer Prev. 2018;19(1). doi:10.22034/APJCP.2018.19.1.207
6. Stieler F, Wolff D, Schmid H, Welzel G, Wenz F, Lohr F. A comparison of several modulated radiotherapy techniques for head and neck cancer and dosimetric validation of VMAT. Radiotherapy and Oncology. 2011;101(3):388-393. doi:10.1016/j.radonc.2011.08.023
7. Walter J, Miller H, Bomford CK. A Short Textbook of Radiotherapy: Radiation Physics, Therapy, Oncology. 4th ed. Churchill Livingstone ; distributed by Longman; 1979.
8. Thwaites DI, Tuohy JB. Back to the future: the history and development of the clinical linear accelerator. Phys Med Biol. 2006;51(13):R343-R362. doi:10.1088/0031-9155/51/13/R20
9. Wilson RR. Radiological Use of Fast Protons. Radiology. 1946;47(5):487-491. doi:10.1148/47.5.487
10. Anderson KO, Mendoza TR, Valero V, et al. Minority cancer patients and their providers: pain management attitudes and practice. Cancer. 2000;88(8):1929-1938.
11. Tamayo-Sarver JH, Dawson NV, Hinze SW, et al. The effect of race/ethnicity and desirable social characteristics on physicians' decisions to prescribe opioid analgesics. Acad Emerg Med. 2003;10(11):1239-1248. doi:10.1111/j.1553-2712.2003.tb00608.x
12. Pitchford G. Radiotherapy Physics: in Practice (Second Edition). Phys Med Biol. 2001;46(3):899-899. doi:10.1088/0031-9155/46/3/701
13. Moore KL. Automated Radiotherapy Treatment Planning. Seminars in Radiation Oncology. 2019;29(3):209-218. doi:10.1016/j.semradonc.2019.02.003
14. Goyal S, Kataria T. Image Guidance in Radiation Therapy: Techniques and Applications. Radiology Research and Practice. 2014;2014:1-10. doi:10.1155/2014/705604
15. Tol JP, Dahele M, Delaney AR, Slotman BJ, Verbakel WFAR. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? Radiat Oncol. 2015;10(1):234. doi:10.1186/s13014-015-0542-1

16. Craig T, Brochu D, Van Dyk J. A quality assurance phantom for three-dimensional radiation treatment planning. International Journal of Radiation Oncology*Biology*Physics. 1999;44(4):955-966. doi:10.1016/S0360-3016(99)00070-X
17. STW_STRATEGY_MAASTRO_LUNG1. https://xnat.bmia.nl/app/template/XDATScreen_report_xnat_projectData.vm/search_element/xnat:projectData/search_field/xnat:projectData.ID/search_value/stwstrategyln1
18. WORC. https://xnat.bmia.nl/app/action/DisplayItemAction/search_value/worc/search_element/xnat:projectData/search_field/xnat:projectData.ID
19. STW_STRATEGY_MULTIDELINEATION. https://xnat.bmia.nl/app/template/XDATScreen_report_xnat_projectData.vm/search_element/xnat:projectData/search_field/xnat:projectData.ID/search_value/stwstrategymmd
20. Nilsson NJ. Artificial Intelligence: A New Synthesis. 5th print. Kaufmann; 2003.
21. Francolini G, Desideri I, Stocchi G, et al. Artificial Intelligence in radiotherapy: state of the art and future directions. Med Oncol. 2020;37(6):50. doi:10.1007/s12032-020-01374-w
22. Mitchell TM. Machine Learning. McGraw-Hill; 1997.
23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. doi:10.1038/nature14539
24. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative Adversarial Networks: An Overview. IEEE Signal Process Mag. 2018;35(1):53-65. doi:10.1109/MSP.2017.2765202
25. Pan K, Zhao L, Gu S, et al. Deep learning-based automatic delineation of the hippocampus by MRI: geometric and dosimetric evaluation. Radiat Oncol. 2021;16(1):12. doi:10.1186/s13014-020-01724-y
26. van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. Radiotherapy and Oncology. 2019;138:68-74. doi:10.1016/j.radonc.2019.05.010
27. Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. Front Oncol. 2020;10:580919. doi:10.3389/fonc.2020.580919
28. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol. 2018;126(2):312-317. doi:10.1016/j.radonc.2017.11.012
29. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-577. doi:10.1148/radiol.2015151169
30. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036
31. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):4006. doi:10.1038/ncomms5006
32. Foley KG, Shi Z, Whybra P, et al. External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiotherapy and Oncology. 2019;133:205-212. doi:10.1016/j.radonc.2018.10.033

33. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Physica Medica. 2019;61:44-51. doi:10.1016/j.ejmp.2019.04.009
34. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. Woloschak GE, ed. PLoS ONE. 2014;9(7):e102107. doi:10.1371/journal.pone.0102107
35. Owens CA, Peterson CB, Tang C, et al. Lung tumor segmentation methods: Impact on the uncertainty of radiomics features for non-small cell lung cancer. PLoS One. 2018;13(10):e0205003. doi:10.1371/journal.pone.0205003
36. Cook GJR, Azad G, Owczarczyk K, Siddique M, Goh V. Challenges and Promises of PET Radiomics. International Journal of Radiation Oncology*Biology*Physics. 2018;102(4):1083-1089. doi:10.1016/j.ijrobp.2017.12.268
37. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: A systematic review. International Journal of Radiation Oncology*Biology*Physics. Published online June 2018. doi:10.1016/j.ijrobp.2018.05.053
38. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol. 2016;2(12):1636. doi:10.1001/jamaoncol.2016.2631
39. Giraud P, Giraud P, Gasnier A, et al. Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers. Front Oncol. 2019;9:174. doi:10.3389/fonc.2019.00174
40. Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics. Med Phys. 2020;47(5). doi:10.1002/mp.13678
41. Starmans M, Klein S, van der Voort SR, Thomeer MG, Miclea RL, Niessen WJ. Classification of malignant and benign liver tumors using a radiomics approach. In: Angelini ED, Landman BA, eds. Medical Imaging 2018: Image Processing. SPIE; 2018:48. doi:10.1117/12.2293609
42. Tselikas L, Sun R, Ammari S, et al. Role of image-guided biopsy and radiomics in the age of precision medicine. Chin Clin Oncol. 2019;8(6):57-57. doi:10.21037/cco.2019.12.02
43. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and Translational Radiation Oncology. 2019;19:33-38. doi:10.1016/j.ctro.2019.07.003
44. Simon L, Robert C, Meyer P. Artificial intelligence for quality assurance in radiotherapy. Cancer/Radiothérapie. Published online June 2021:S1278321821001104. doi:10.1016/j.canrad.2021.06.012
45. Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy and Oncology. 2020;153:55-66. doi:10.1016/j.radonc.2020.09.008
46. Kimura Y, Kadoya N, Tomori S, Oku Y, Jingu K. Error detection using a convolutional neural network with dose difference maps in patient-specific quality assurance for volumetric modulated arc therapy. Physica Medica. 2020;73:57-64. doi:10.1016/j.ejmp.2020.03.022

47. Interian Y, Rideout V, Kearney VP, et al. Deep nets vs expert designed features in medical physics: An IMRT QA case study. Med Phys. 2018;45(6):2672-2680. doi:10.1002/mp.12890
48. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning: Virtual IMRT QA. Med Phys. 2016;43(7):4323-4334. doi:10.1118/1.4953835
49. Lam D, Zhang X, Li H, et al. Predicting gamma passing rates for portal dosimetry-based IMRT QA using machine learning. Med Phys. 2019;46(10):4666-4675. doi:10.1002/mp.13752
50. Granville DA, Sutherland JG, Belec JG, La Russa DJ. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. Phys Med Biol. 2019;64(9):095017. doi:10.1088/1361-6560/ab142e
51. Carlson JNK, Park JM, Park SY, Park JI, Choi Y, Ye SJ. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. Phys Med Biol. 2016;61(6):2514-2531. doi:10.1088/0031-9155/61/6/2514
52. Chuang K, Giles W, Adamson J. A tool for patient-specific prediction of delivery discrepancies in machine parameters using trajectory log files. Med Phys. 2021;48(3):978-990. doi:10.1002/mp.14670
53. Luk SMH, Meyer J, Young LA, et al. Characterization of a Bayesian network-based radiotherapy plan verification model. Med Phys. 2019;46(5):2006-2014. doi:10.1002/mp.13515
54. Diaz O, Guidi G, Ivashchenko O, Colgan N, Zanca F. Artificial intelligence in the medical physics community: An international survey. Physica Medica. 2021;81:141-146. doi:10.1016/j.ejmp.2020.11.037
55. McNutt TR, Moore KL, Wu B, Wright JL. Use of Big Data for Quality Assurance in Radiation Therapy. Seminars in Radiation Oncology. 2019;29(4):326-332. doi:10.1016/j.semradonc.2019.05.006
56. Lustberg T, van Soest J, Jochems A, et al. Big Data in radiation therapy: challenges and opportunities. Br J Radiol. 2017;90(1069):20160689. doi:10.1259/bjr.20160689
57. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer J Clin. 2021;71(3):209-249. doi:10.3322/caac.21660
58. F.I. Osman A. Radiation Oncology in the Era of Big Data and Machine Learning for Precision Medicine. In: Antonio Aceves-Fernandez M, ed. Artificial Intelligence - Applications in Medicine and Biology. IntechOpen; 2019. doi:10.5772/intechopen.84629
59. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiotherapy and Oncology. 2018;127(3):370-373. doi:10.1016/j.radonc.2018.03.004
60. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. Radiotherapy and Oncology. 2013;107(3):267-273. doi:10.1016/j.radonc.2013.05.007

61. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350(jan07 4):g7594-g7594. doi:10.1136/bmj.g7594
62. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18
63. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017
64. Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS. Image biomarker standardisation initiative. arXiv:161207003 [cs]. Published online December 21, 2016. Accessed June 8, 2018. http://arxiv.org/abs/1612.07003

Chapter 2

Radiomics: “Unlocking the potential of medical images for precision radiation oncology”

Adapted from: Radiomics: “Unlocking the potential of medical images for precision radiation oncology”

Petros Kalendralis, Martin Vallières, Benjamin H. Kann, Aneja Sanjay, Arif S. Rashid, Andre Dekker, Rianne Fijten

Book chapter in “Radiomics for clinicians” (submitted to Elsevier-final edits from the Editor in chief),

Contribution: First authorship

Abstract

During the last decade radiation oncology became one of the most data-driven medical specialties due to the rapid development of computational methods and artificial intelligence (AI) in medical imaging domain. The radiomics concept has converted medical images into minable data associated with clinical events used for personalized medicine. In this chapter we will present an overview of the fundamental principles of the radiomics pipeline as well as with a roadmap for responsible and reliable radiomics research studies. Furthermore, the major uncertainties and pitfalls of the radiomics pipeline are outlined with the most up-to-date solutions and recommendations of the Imaging Biomarker Standardization Initiative (IBSI) for responsible radiomics. Finally, we discuss the potential translation of radiomics into the clinic via the commissioning of radiomics models and the comparison between the operational excellence and the prediction outcome of the models.

2.1 Introduction

Medical imaging is routinely used for screening, monitoring, diagnostic and treatment purposes in the management of cancer[1]. Multimodality images contain a variety of significant information about the tumor characteristics. Traditionally, medical images have been analyzed visually. However, visual observation is time consuming and makes reliable elucidation of all the potential information embedded into an image difficult. Computers can address this challenge and have the potential to automate the task of extracting all information from medical images[2]. In fact, progress in data mining and machine learning (ML) have changed the way in which many radiologists observe medical images, where the increase in computational power has enabled the observation and extraction of high-dimensional quantitative features from medical images[3].These features can be used by researchers and clinicians to answer crucial clinical questions regarding tumor phenotypes, as well as the diagnosis and treatment of patients. Overall, the concept of the automated extraction of quantitative imaging features is known as "radiomics".

The term radiomics can be more specifically defined as the high-throughput computerized extraction of quantitative features from medical images such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). These features can be used for outcome prediction modelling and mainly are extracted from the Region of Interest (ROI) contour, such as the Gross Tumor Volume (GTV) represented in the medical images using various imaging modalities, The radiomics workflow was first described in 2012 by Lambin *et al.*[2] and Kumar *et al.*[4]The first comprehensive proof-of-concept study in radiomics is considered to be the one by Aerts et al.[5] in 2014, in which the authors described a radiomics approach to model overall survival of lung and head and neck patients using CT images. Since then, the number of publications of radiomics studies has grown exponentially as it is shown in figure 2.1.

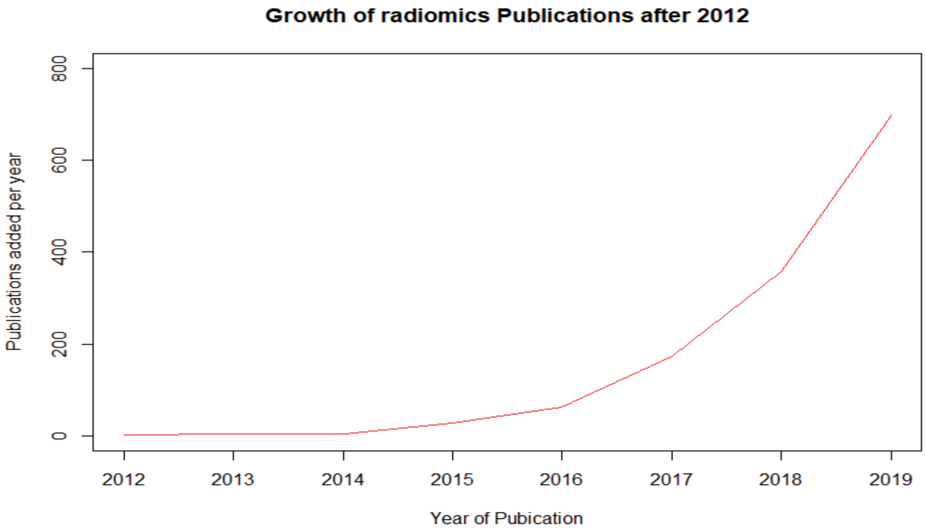


Figure 2.1: The growth of radiomics publications since 2012 (PubMed).

The purpose of this chapter is to provide an overview of radiomics in radiation oncology, by describing the ML applications and how they can be implemented in the hospital. Furthermore, this chapter will present the fundamental steps that should be followed by researchers working in the field of radiomics, in line with the recommendations of the Image Biomarker Standardization Initiative (IBSI)[6,7]. Following the integration of ML techniques in the radiomics field, the crucial challenge has now become the introduction and implementation of radiomics applications into the clinical environment. With this in mind, the potential barriers that need to be overcome for the clinical translation or radiomics will be discussed as well as with future recommendations for the general acceptance and commissioning of radiomics.

2.2 Implementation of radiomics in radiation oncology

In this section we will provide an overview of the fundamental radiomics principles in research as well as with the uncertainties and pitfalls of each part of the radiomics pipeline. Furthermore, following the IBSI recommendations, we present guidelines for the standardization of radiomics studies. In the last part of this section, the potential of distributed learning combined with radiomics studies is analysed, with the useful application for the exchange and validation of radiomics models.

2.2.1 Fundamentals of radiomics in research

- **Methodology**

Radiomics studies involve several steps that require the inclusion of different domain experts, such as radiation oncologists/clinicians, medical physicists and computer/data scientists. The steps of the radiomics workflow include (I) data acquisition, (II) ROI segmentation, (III) feature extraction, (IV) model development. A representation of the radiomics workflow is shown in the figure 2.2.1. Each of these steps poses different challenges and uncertainties that will be described further in this chapter.

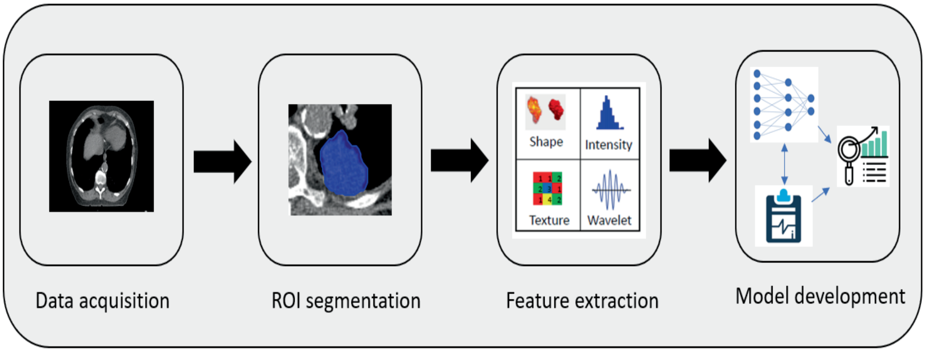


Figure 2.2.1:Representation of the typical radiomics work containing the data acquisition, the ROI segmentation, the feature extraction and the statistical analysis for model development.

- **Data acquisition and preparation**

There is a large amount of multi-source data across the different hospitals generated in the daily clinical routine that can be used for research studies. A large volume of images isstored in the Pictures Archive Communication System (PACS), which can potentially be connected with the corresponding clinical data from the patient's local Electronic Health Record (EHR) database. Radiomics have the potential to be a highly-cost effective line of inquiry leading to better treatment selection and improved stratification using this amount of data.

The applications and potential of radiomics have been demonstrated by several studies so far[8]. A wide variety of acquisition settings is used for the different imaging modalities.The combination of different vendors, reconstruction algorithms and scanner models across the different institutions is more than a usual phenomenon. These differences play a significant role in the computation of radiomics features. This impact can be translated into texture as the imaging noise level can be affected, resulting in a difference in the classification of the ROI and inconsistent results in the statistical analysis of radiomics features[9,10]. Besides the realistic fact that it is impossible to unify the image acquisition settings and techniques, efforts should be applied for the standardization of the appropriate pipeline for the development of accurate radiomics models derived from images obtained with different protocols.

Radiomics features are highly dependent on the imaging parameters. The most dependent parameters that could affect the radiomics features values and that need to be taken into account for every imaging modality are the number of grey levels, pixel or voxel size[11] and the range of grey level values[12]. Additionally, in the past years, MRI studies[13,14] suggested the removal of the signal intensity nonuniformity from the MR images. Several approaches and solutions have been proposed for managing these dependencies. Regarding the normalization of the grey level values, a study from Collewet et al.[15] suggested the $\pm 3\sigma$ $(\pm 3\ sigma)$method which proved to be promising. To deal with the size of pixels, interpolation methods such as linear cubic B-spline interpolation can be used for pixel resampling according to Parker et al.[16]. Besides the fact that several image preprocessing methods are integrated in some radiomics platforms which are not publicly available for commercial purposes, it should be acknowledged that many open-source softwares such as ImageJ or 3DSlicer can handle image data preprocessing methods.

Data preparation is one of the most important initial steps for radiomics studies as the quality of the data has an impact in the radiomics feature extraction and the development or validation of the radiomics model. Several conditions such as radiomics features scaling, over-sampling, randomization and discretization of the dataset, should be taken into acount before data analysis using ML or AI algorithms. As the radiomics features are extracted at various scales, feature scaling should be involved in the radiomics pipeline, as there is a potential interference with the ML model parameters. In other words, feature scaling is the change of the numeric feature values to a common scale without important distortions, that is categorized to normalization and standardization. The distribution of the data determined by the ML algorithm is the factor that will designate the feature scaling technique.The performance of the ML algorithms is highly correlated with the class balance of the datasets. Imbalance can lead to important misleading classification results. One of the potential solutions is the resampling of the training dataset used for the model development.

Radiomics researchers need high quality datasets[17]. Big and standardized radiomics datasets enriched with clinical metadata can accelerate the acceptability of radiomics by the clinicians. However, the data collection for the radiomics researchers constitutes a time consuming procedure as they need access to the medical images and patients' clinical data warehouse. Furthermore, data exchange between different institutions presents ethical or legal issues due to the sensitivity of medical data privacy. Some imaging public repositories such as the Cancer Imaging Archive (TCIA)[18] and the instance of the Extensible Neuroimaging Archive Toolkit (XNAT-https://xnat.bmia.nl) hosted within the within Dutch national research infrastructure (TraIT, www.ctmm-trait.nl) constitutes a valuable resource for radiomics researchers. These abovementioned public imaging repositories provide a variety of imaging datasets combined with clinical metadata tables available for radiomics studies.

- **Tumour segmentation (manual-semiautomatic-automatic)**

One of the crucial parts of the radiomics pipeline included in the clinical routine practice of radiologists and radiation oncologists is the segmentation of the different ROIs and the surrounding Organs at Risk (OAR). The radiomics feature values are extracted from the volume of the ROI we want to analyze computationally. The majority of the studies use the primary GTV as a ROI for feature extraction, as their main goal is the development of prediction models based on the primary tumor response to the treatment. Other studies presented a different approach[19–22], stating that the peripheral tissues of the primary GTV could potentially include predictive information regarding distant metastasis or tumor recurrences.

The manual segmentation of the different ROIs used to be the ”gold standard” during the past years in the clinical routine of radiologists and radiation oncologists. This method is time consuming for the clinicians, as a detailed observation of each different image slice is required. Furthermore, the morphological variations of the tumor region is a factor for inter-observer variability in segmentation, as the different observers may have different approaches regarding the tumor morphology. These differences can have an impact on the radiomics features reproducibility and repeatability[23–26]. Advances in the field of AI and radiation oncology introduced the semiautomatic[27,28] and automatic[29] Specifically, Shi et al.[27] used the ATLAAS based[30] semi-automatic segmentation method in PET images to externally validate a prognostic model for oesophageal cancer patients, while Chen et al.[29] used a multiscale 1-layer deep 3D Convolutional Neural Network (CNN), technique introduced by the study of Kamnitsas et al.[31]. Although the automated segmentation methods have raised the interest and the concern of the majority of the clinicians and researchers, there are still a lot of steps to achieve a fully automated pipeline for the detection of the tumor volumes in medical images[32–34].

- **Feature extraction/selection (different features and software)**

Following the data acquisition and image segmentation, a set of imaging features is computed from the delineated volume of the ROI, which is used as an input for the radiomics model. In other words, the radiomics feature extraction is the next step of the radiomics workflow as a connection between the medical images and the clinical parameters is established. There are two main categories of radiomics features depending on the way they are extracted: the “manually” extracted features from a radiomics feature extraction software, known as “traditional” features and the deep learning (DL) features.

Usually, the delineated ROIs of the medical images are analysed by simple statistics such as mean or standard deviation, entropy and kurtosis[35]. Although these statistical measurements are easily implemented by any computational method, they fail to provide full spatial information. For this reason, the radiomics approach is based on more sophisticated and mathematically formulae expressed imaging features. The “traditional” manually extracted radiomics features contain several feature groups. The first category contains the shape features that provide information regarding, for example, the shape, surface-volume-ratio and sphericity of the delineated ROI.The second group consists of the first, second and high-order features composed from the intensity of each voxel of the ROIs[36]. The distribution of the intensity (minimum, median, maximum and entropy) of the ROI segmentation is described by the first order features while the second order features present the statistical correlation between the voxels or pixels of the ROI such as the textural features. The high-order feature group represents features that consider relationships between three or more voxels or pixels. Several radiomics studies involved wavelet and model based features such as fractals in their features extraction pipeline[37–39]. Furthermore, texture features such as the Local Binary Patterns (LBP)[40,41] can be included on the third group. An overview of examples of the different “traditional” radiomics feature groups is presented on the table 2.2.1

Table 2.2.1: Example of radiomics features according to different classes

Feature categories	Feature names examples
Shape	Elongation
	Flatness
	Volume
	Sphericity
	Surface Area
	Surface to Volume Ratio
First-order	Energy
	Entropy
	Interquartile Range
	Kurtosis

	10th Percentile
	90th Percentile
	Skewness
Second order	Grey Level Non Uniformity
	Grey Level Variance
	High Grey Level Emphasis
	Large Area Low Grey Level Emphasis
High order	Wavelet
	Autoregressive model

As the popularity of DL has increased during the past years in the field of radiation oncology, DL algorithms have been developed to select and generate the features for a given task within the different layers, without any manual intervention[42]. A significant example of an unsupervised deep neural network is the auto-encoder. The input image is transformed in a feature vector using a stacked convolutional architecture of pooling and activation layers. As a next step, the feature vector is mapped by the decoder part to the input image space. The DL features are defined by the result of the last convolutional layer[43–45].

Recent studies stated the preeminence of DL features to “handcrafted” features[46,47].The advantageous point of DL features compared to the “traditional” features is based on the fact that they are instantly learned from data. For this reason, DL features can be adapted to specific properties of a dataset and correlated with clinical more easily.

- **Feature selection-Modeling -The significance of ML**

The reliability, performance and reproducibility of a model is strongly dependent on the features included on the model construction process. A big amount of features might lead to over-fitted models especially when the number of samples (patients, in the radiomics case) is higher than the free parameters (features). The selection of the most suitable-predictive features is a crucial part that will result in the reduction of the free parameters and elimination of the unnecessary features based on the clinical endpoint. There are several approaches of feature selection in radiomics studies, but mainly they are categorized in two categories: Filtering and embedding[48].

Filtering methods include the evaluation of the features without the involvement of the model. The filtering method is separated in two different categories: univariable and multivariable methods. With the univariable filter method, the most reproducible features are taken into account based on the Chi-squared test for instance. Multivariable methods use rankers and selectors of subsets such as the correlation based feature selection. With the embedded method, during the modelling procedure a feature subgroup is suggested and evaluated. Wrapper methods suggest the generation of the features as a first step and then the feature evaluation follows including the model.

The majority of radiomics studies based on filtering methods use feature reproducibility analysis for the evaluation of the features. This analysis has a goal to reduce the dimensionality and exclude the features with relatively low reproducibility. The most common statistical approach for reproducibility analysis that is used from important radiomics studies[5,25,43,49,50] is the intra-class correlation coefficient (ICC)[51] which is based on the investigation of the most reproducible features from the comparison of different annotations-segmentations of ROIs in radiomics case-from multiple observers. However this method requires the inclusion of radiomics datasets that contain multiple delineations from different observers which increases the workload as some of the ROIs need to be segmented multiple times. An example of an embedded method of feature selection is the least absolute shrinkage and selection operator (LASSO) which is for the generation of the selected features and the prediction model as an output[52–56].

After the elimination of the extracted features with the feature selection procedure, the development of the model will follow by including the mostly ranked reproducible features. ML plays a key role in model development. Depending on the availability of the clinical metadata that can be combined with the radiomics model and the outcome of the prediction model (e.g. overall survival, toxicity, lymph nodes metastasis etc.), supervised and unsupervised algorithms can be used. Supervised algorithms require labelled datasets for the training of the model including two steps: the training and testing of the model. During the training process the selected radiomics features are paired with the clinical metadata, and via a pre-defined loss function the correlation between the radiomics

features and clinical metadata is learned by the model. Supervised algorithms have been used for radiomics models development such as the support vector machine (SVM)[57], logistic regression[58] and random forest (RF)[43]. Unsupervised algorithms are used as an alternative in the case of unavailability of clinical labels (ie. patients that are not categorized). This algorithms category clusters the patients samples into several groups according to the similarity level of the different samples which is calculated by a distance measurement. K-means clustering[59], fuzzy clustering[60] and consensus clustering[61] are examples of some of the algorithms used for unsupervised learning. Although the selection of the algorithm for the model development is not included in a specific protocol according to the literature, the best approach includes multiple experiments taking into account the clinical endpoint that a radiomics study investigates.

The value and reliability of a radiomics model relies on its potential validation. The independent external validation of a radiomics model constitutes the basic factor that can transform it in a valuable clinically introduced model. Additionally, radiomics models that are prospectively validated increase the robustness of a radiomics study. The receiver operating characteristic curve (ROC) is a statistical tool that measures the performance of the models. Furthermore, the clinical and potential prognostic power evaluation of the radiomics model can be assessed by the area under the curve (AUC), specificity and sensitivity of the model. For the complex task of the overall survival prediction, concordance index (C-index) and ROC curve can be used for the validation process.Moreover, calibration for survival analysis is necessary for the investigation of a potential agreement between a prediction model and the clinical outcome[62].

- **Challenges of ML**

Each of the steps described on the previous paragraphs is important for the radiomics model development and evaluation. Medical imaging analysis with ML techniques requires specific expertise.The biggest challenge lies in the fact that ML extracts the correct information from the input data under the precondition of the correct preparation of the data. The data structure should be understood by the radiomics researcher before any ML technique should be applied. A potential fault in learning and understanding the data will produce wrong models with unreliable results. One of the major problems in radiomics models is the unbalanced datasets in which some of the clinical outcomes are represented with missing values. This imbalance on the dataset could induce a bias on the class that is represented in the training sample of the radiomics model. As there is an absence of a specific protocol-guide for all the steps of the radiomics pipeline, there are numerous uncertainties and pitfalls in the implementation of radiomics in research. In the next section we will try to map them, and for each step, to propose potential solutions.

2.2.2 Uncertainties and pitfalls in radiomics

- **Data acquisition during imaging**

As we have discussed in the previous section, the starting point for a reliable and potentially externally validated model is the data quality. In the ideal scenario, all the imaging data with the clinical metadata used for radiomics studies would be structured, without missing values and bias and understandable from the researchers and clinicians.The medical scans used for radiomics studies are usually acquired using different imaging protocols, imaging settings and reconstruction parameters. These differences in the imaging acquisition process may result in inconsistencies in the output of the radiomics model[63]. The influence of the various image acquisition parameters on the repeatability (same subject, imaging system and imaging acquisition parameters) and reproducibility (same subject, different scanners, reconstruction kernels, slice thickness, etc.) of the extracted radiomics features was discussed by the systematic reviews of Traverso et al[64]. and Larue et al[65].

CT is commonly used for treatment planning purposes in radiotherapy. As the main scope of many radiomics studies is the prediction of the overall survival after the treatment course, radiotherapy planning CT scans were widely used by several studies. Some of the main factors of influence of the radiomics features are the voxel size, slice thickness, exposure and reconstruction kernels according to phantom and patient cohort studies[11,12,66–72]. Several approaches have been suggested for the minimization of the influence of these parameters to the radiomics features. Pixel size resampling, normalization on the grey level and voxel size are techniques that could enhance the robustness of the radiomics features[11,12].Regarding the multicenter radiomics studies, as it is rarely common to have imaging datasets acquired with the same imaging protocol, scanner and reconstruction algorithms, it is suggested to include phantom dataset studies for the investigation of the parameters that influence the radiomics features reproducibility[73,74].

PET constitutes an additional source of information for radiomics studies, as several parameters have the potential to enrich the biological or volumetric morphology of the tumor.

Several studies investigated the differences of textural features due to the variations in the reconstruction algorithms and the number of iterations in PET imaging[26,75,76].

Furthermore, the respiratory motion of the patients suggested to be taken into consideration as a factor that could potentially affect the radiomics features in PET according to the studies of Oliver et al.[77] and Grootjans et al.[78]. Moreover, according to the same studies, the standardized uptake value (SUV) was found to be an influence factor for the textural features. In recent years, several initiatives have tried to standardize the imaging protocols for PET[79,80] accompanied with recommendations for the development of new biomarkers[10,82]. Despite the fact that these initiatives were a starting point for the standardization discussion of PET imaging protocols, the standardization of PET imaging acquisition is likely to be lacking from a multicenter (even an intracenter) perspective. This is in part due to issues from the PET scanner variabilities, the injected activity of ^{18}F-fluorodeoxyglucose (^{18}F-FDG), the time difference between the injection time and the image acquisition and the CT scanning parameters that are used from the attenuation correction of PET images[81–83]. Although PET imaging is promising for the interrogation of the specific mutations in tumour biology, there are limitations and lack of standardization.

MRI is the most appropriate imaging modality for screening soft anatomical tissues such as the brain due to the flexibility of the selection of several pulse sequences for image acquisition. Several phantom and patient studies investigated the impact of the different magnetic field strength scanners,different imaging protocols used and scanner manufacturers[84–89]. Specifically, the phantom study by Mayerhoefer et al.[86] found that the textural features were sensitive to variations in the MRI acquisition parameters such as the number of acquisitions (NA), the repetition time (TR), the echo time (TE) and the sampling bandwidth (SBW). Furthermore, according to the above-mentioned study, the reduction of the imaging resolution reduced the sensitivity of the imaging features to the acquisition parameters. Additionally, researche studies[90,91] suggested the approach of intensity normalization for the reduction of the differentiations caused by the acquisition parameters for brain scans. The influence of the normalization on texture was studied by Collewet et al.[15] using co-occurrence matrix, run-length matrix, gradient matrix and Harr wavelet energy features.

- **Segmentation**

The component of the radiomics workflow that is crucial for the determination of the ROI that will be analyzed is the segmentation method used for the delineation. Recent advances in the field of AI resulted in the introduction of the semi-automatic or automatic segmentation for some OARs. Although this new approach suggested for the clinicians eliminates the time consuming procedure of delineating manually the different anatomical structures, there is no golden standard for the delineation method used across the different departments. A study from Velazquez et al.[92] investigated the potential of semi-automatic segmentation of NSCLC patients. This study used a 3D slicer algorithm to compare three semi-automatic contours by three different observers to five manual contours in PET/CT scans. The semi-automatic delineations showed less uncertainty than manual delineations. The study of Kalpathy-Cramer et al.[93] compared three different algorithms for the semi-automatic segmentation of lung nodules. The intra-algorithm results presented less variability than the inter-algorithm results. This comparison indicated the need of using the same segmentation algorithm for all the time points of a multicenter or intracenter study due to the large differences between the segmentation algorithms.

- **Feature extraction - Feature selection algorithms**

In radiomics studies, a large number of features is extracted. Each of these features is extracted with different image preprocessing methods (eg. filtering/denoising), different ROI determination(eg. segmentation methods, 2D/3D) and image interpolation methods (eg. nearest-neighbors, b-splines). Specifically for the textural features, there are several parameters that are varied such as the design of the texture matrices (eg. number of directions, distances, normalization) and the grey-levels discretization methods (eg. relative, absolute, equalization). All of the above parameters result in the significant and difficult task of handling thousands of variables.

There are various open-source and commercial software applications for radiomics feature extraction developed in different programming languages. Each of these applications provides the extraction of different amounts of features classified in different classes. Some examples of open source software tools that can be easily used from the radiomics research community are the RaCaT[94], ontology-guided radiomics analysis workflow (O-RAW)[95], Pyradiomics[96], LIFEx[97] and IBEX[98].

As we described in the previous sections of the chapter, the feature selection procedure is an important step for the creation of the final radiomics model. Although there is no protocol or standardized guidelines for feature selection in radiomics studies, there is a rule-of-thumb to use at least ten times more patients than features for the avoidance of over-fitting.

- **Model development**

Radiomics models should be validated. The predictive ability of a model should be evaluated with a test model in a different independent external dataset that is preferably available after the development of the algorithm for modelling. The importance of model validation was illustrated by Zwanenburg et al.[99] according to the recommendations of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis(TRIPOD) statement[100]. Notwithstanding the significance of the algorithm selection for model building, most of the early PET and CT radiomics studies reported from the systematic review by Chalkidou et al.[101] performed inappropriate statistical analysis.

After the description of the above uncertainties in the radiomics workflow, we would like to give an overview of potential recommendations-solutions for each step of the workflow accompanied with some reference reporting points for radiomics studies. According to the study of Deist et al.[102] there was not a significant difference between the performance score of the different ML classifiers across twelve chemoradiotherapy datasets. This result concludes that the model quality is highly dependent on the quality of the dataset while the manipulation of the mathematical learning process has minor importance. Control of the image acquisition settings for all the imaging modalities as well as with harmonization techniques, especially for multicentric studies, could be beneficial for the data quality improvement of radiomics studies. Taking into account that the radiomics feature values are sensitive to the variations in the voxel size, interpolation methods should be applied to the images. Moreover, the detailed description of the feature definitions is suggested for the reduction of features redundancy and the increase of features robustness. Furthermore, the appropriate statistical analysis regarding the training, validation and test of the model should be taken into account. Finally, the contribution of prospective multicenter studies is significant as well as with the cautious exploration of DL.

- **IBSI**

A significant progress for the standardization and the harmonization of the radiomics workflow steps has been made from the IBSI consortium. The main goals of the IBSI are: (i) provide a standardized workflow for radiomic computations, from image processing to feature computation; (ii) provide benchmarking tests and associated reference values for radiomic computations on medical images; and (iii) provide reporting guidelines for radiomic studies. More details about this initiative are provided in the next section.

2.2.3 Guidelines for the standardization of radiomics computations

Radiomics research has already shown great promise for supporting clinical decision-making. However, the fact that radiomics-based strategies have not yet been translated to routine practice can be partly attributed to the low reproducibility of most current studies. The workflow for computing features is complex and involves many steps, often leading to incomplete reporting of methodological information (e.g., texture matrix design choices and grey-level discretization methods). As a consequence, few radiomics studies in the current literature can be reproduced from start to end.

To accelerate the translation of radiomics methods to the clinical environment, about 70 scientists from 25 institutions in 8 countries have participated since September 2016 to the Image Biomarker Standardisation Initiative (IBSI)[6,7] . Figure 2.3.1 presents the standardized radiomics workflow defined by the IBSI. The IBSI aims at standardizing both the computation of features and the image processing steps required before feature extraction. For this purpose, a simple digital phantom was designed and used in Phase 1 of the IBSI to standardize the computation of 174 features from 11 categories: 29 morphological, 2 local intensity, 18 statistical, 23 intensity histogram, 7 intensity-volume histogram, 25 grey level co-occurrence matrix, 16 grey level run length matrix, 16 grey level size zone matrix, 16 grey level distance zone matrix, 5 neighbourhood grey tone difference matrix and 17 neighbourhood grey level dependence matrix features. In Phase 2 of the IBSI, a set of CT images from a lung cancer patient was used to standardize radiomics image processing steps using 5 different combinations of parameters including volumetric approaches (2D vs 3D), image interpolation, re-segmentation and discretization methods. The initiative has now reached completion and benchmark values for Phase 1 and Phase 2 have been defined, along with a compliance check spreadsheet created for this purpose. The standardized workflow and benchmark values defined by the IBSI could now thus serve as a calibration tool for any radiomics software, and compliance with IBSI standards is strongly encouraged for any future radiomics study.

Overall, we want to reiterate that the use of standardized computation methods would greatly enhance the reproducibility potential of radiomics studies. Furthermore, it is essential to rely on supplementary material (usually allowed in most journals) to provide exhaustive methodological details, including the comprehensive description of image acquisition protocols, sequence of operations, image post-acquisition processing, tumor segmentation, image interpolation, image re-segmentation and discretization, formulas for the calculation of features, and benchmark calibrations. Table 2.3.1 provides guidelines on feature computation details to be reported in radiomics studies as defined by the IBSI[6,7] and Vallières et al[103]. Ultimately, we envision the use of dedicated ontologies to improve the interoperability of radiomics analyses via consistent tagging of features, image processing parameters and filters. For example, the Radiomics Ontology (www.bioportal.bioontology.org/ontologies/RO) could provide a standardized means of reporting radiomics data and methods, and would more concisely summarize the implementation details of a given radiomics workflow.

Finally, some guiding principles already exist to help radiomics scientists to further implement the responsible research paradigm into their current practice. A concise set of principles for better scientific data management and stewardship — the "FAIR guiding principles" — has been defined, stating that all research objects should be findable, accessible, interoperable, and reusable. Implementation of the FAIR principles within the radiomics field can undoubtedly facilitate clinical translation. In terms of the construction of radiomics-based prediction models via multivariable analysis, there are two basic requirements. First, all methodological details and clinical information must be clearly reported or described to facilitate reproducibility and comparison with other studies and meta-analyses. Second, models must be tested in sufficiently large patient datasets distinct from development sets to statistically demonstrate their efficacy over conventional models (e.g., existing biomarkers, tumor volume, cancer stage, etc.). To allow for optimal reproducibility potential and further independent testing, all data, final models and programming code related to a given study needs to be made available to the community. Table 2.3.2 provides guidelines that can help to evaluate the quality of radiomics studies[6,7].

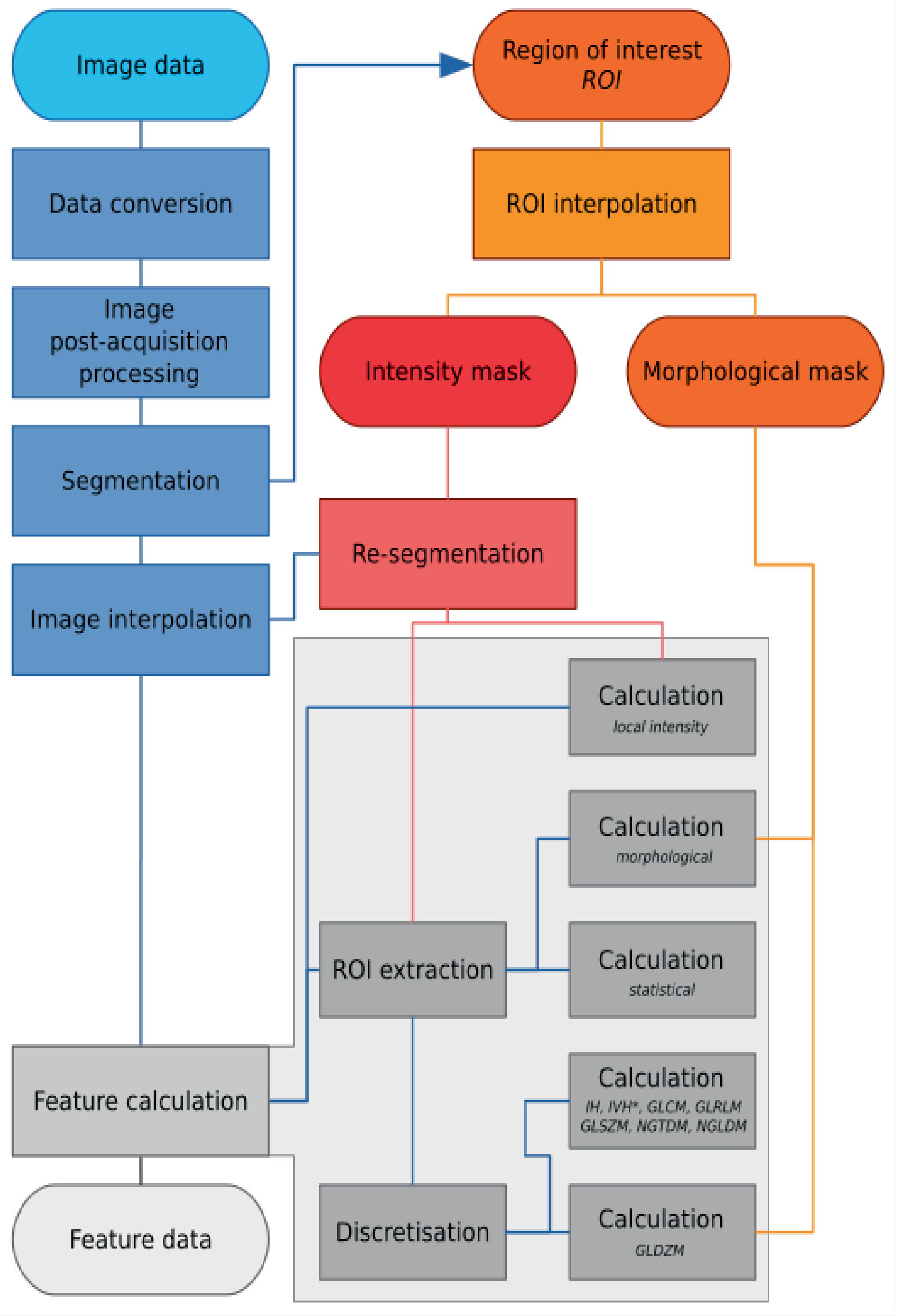


Figure 2.3.1: Radiomics computation workflow as defined by the IBSI[6,7].

Table 2.3.1: Reporting guidelines on the computation of radiomics features (adapted from Zwanenburg *et al.*[6,7] and Vallières *et al.*[103], with permissions from Ibrahim *et al.*[104] and Avanzo et *al.*[105]).

GENERAL	
Image acquisition	Acquisition protocols and scanner parameters: equipment vendor, reconstruction algorithms and filters, field of view and acquisition matrix dimensions, MRI sequence parameters, PET acquisition time and injected dose, CT x-ray energy (kVp) and exposure (mAs), etc.
Volumetric analysis	Imaging volumes are analyzed as separate images (2D) or as fully-connected volumes (3D).
Workflow structure	Sequence of processing steps leading to the extraction of features.
Software	Software type and version of code used for the computation of features.
IMAGE PRE-PROCESSING	
Conversion	How data were converted from input images: e.g, conversion of PET activity counts to SUV, calculation of ADC maps from raw DW-MRI signal, etc.

Processing	Image processing steps taken after image acquisition: e.g., noise filtering, intensity non-uniformity correction in MRI, partial-volume effect corrections, etc.
ROI SEGMENTATION[a,b]	How regions of interests (ROIs) were delineated in the images: software and/or algorithms used, how many different persons and what expertise (specialty, experience), how a consensus was obtained if several persons carried out the segmentation, in automatic or semi-automatic mode, etc.
INTERPOLATION	
Voxel dimensions	Original and interpolated voxel dimensions.
Image interpolation method	Method used to interpolate voxels values (e.g, linear, cubic, spline, etc.) as well as how original and interpolated grids were aligned.
Intensity rounding	Rounding procedures for non-integer interpolated grey levels (if applicable), e.g., rounding of Hounsfield units in CT imaging following interpolation.
ROI interpolation method	Method used to interpolate ROI masks. Definition of how original and interpolated grids were aligned.

ROI partial volume	Minimum partial volume fraction required to include an interpolated ROI mask voxel in the interpolated ROI (if applicable): e.g., a minimum partial volume fraction of 0.5 when using linear interpolation.
ROI RE-SEGMENTATION	
Inclusion/exclusion criteria	Criteria for inclusion and/or exclusion of voxels from the ROI intensity mask (if applicable), e.g., the exclusion of voxels with Hounsfield units values outside a pre-defined range inside the ROI intensity mask in CT imaging.
IMAGE DISCRETIZATION	
Discretization method	Method used for discretizing image intensities prior to feature extraction: e.g., fixed bin number, fixed bin width, histogram equalization, etc.
Discretization parameters	Parameters used for image discretization: the number of bins, the bin width and minimal value of discretization range, etc.
FEATURE CALCULATION	

Features set	Description and formulas of all calculated features.
Features parameters	Settings used for the calculation of features: voxel connectivity, with or without merging by slice, with or without merging directional texture matrices, etc.
CALIBRATION	
Image processing steps	Specifying which image processing steps match the benchmarks of the IBSI.
Features calculation	Specifying which feature calculations match the benchmarks of the IBSI.

[a] In order to reduce inter-observer variability, automatic and semi-automatic methods are favored.

[b] In multimodal applications (e.g., PET/CT, PET/MRI, etc.) ROI definition may involve the propagation of contours between modalities via co-registration. In that case, the technical details of the registration should also be provided.

Table 2.3.2: Quality factors in radiomics studies (adapted from Lambin *et al.*[106] and Vallières *et al.*[103], with permissions from Ibrahim *et al.*[104] and Avanzo et *al.*[105]).

IMAGING	
Standardized imaging protocols	Imaging acquisition protocols are well described and ideally similar across patients. Alternatively, methodological steps are taken towards standardizing them.

Imaging quality assurance	Methodological steps are taken to only incorporate acquired images of sufficient quality.
Calibration	Computation of radiomics features and image processing steps match the benchmarks of the IBSI.
EXPERIMENTAL SETUP	
Multi-institutional/external datasets	Model construction and/or performance evaluation is carried out using cohorts from different institutions, ideally from different parts of the world.
Registration of prospective study	Prospective studies provide the highest level of evidence supporting the clinical validity and usefulness of radiomics models.
FEATURE SELECTION	
Feature robustness	The robustness of features against segmentation variations and varying imaging settings (e.g., noise fluctuations, inter-scanner differences, etc.) is evaluated. Unreliable features are discarded.
Feature complementarity	The inter-correlation of features is evaluated. Redundant features are discarded.

MODEL ASSESSMENT

False discovery corrections	Correction for multiple testing comparisons (e.g., Bonferroni or Benjamini- Hochberg) is applied in univariate analysis.
Estimation of model performance	The teaching dataset is separated into training and validation set(s) to estimate optimal model parameters Example methods include bootstrapping, cross-validation, random subsampling, etc.
Independent testing	A testing set distinct from the teaching set is used to evaluate the performance of complete models (i.e., without retraining and without adaptation of cut- off values). The evaluation of the performance is unbiased and not used to optimize model parameters.
Performance results consistency	Model performance obtained in the training, validation and testing sets is reported. Consistency checks of performance measures across the different sets are performed.
Comparison to conventional metrics	Performance of radiomics-based models is compared against conventional metrics such as tumor volume and clinical variables (e.g., staging) in order to evaluate the added value of radiomics (e.g., by assessing the significance of AUC increase calculated with the DeLong test).

Multivariable analysis with non-radiomics variables	Multivariable analysis integrates variables other than radiomics features (e.g., clinical information, demographic data, panomics, etc.).
CLINICAL IMPLICATIONS	
Biological correlate	Assessment of the relationship between macroscopic tumor phenotype(s) described with radiomics and the underlying microscopic tumor biology.
Potential clinical application	The study discusses the current and potential application(s) of proposed radiomics-based models in the clinical setting.
MATERIAL AVAILABILITY	
Open data	Imaging data, tumor ROI and clinical information are made available.
Open code	All software code related to computation of features, statistical analysis and machine learning, and allowing to exactly reproduce results, is open source. This code package is ideally shared in the form of easy-to-run organized scripts pointing to other relevant pieces of code, along with useful sets of instructions.

Open models	Complete models are available, including model parameters and cut-off values.

2.2.4 Radiomics and distributed learning infrastructures

One of the potential solutions for the uncertainties related to radiomics studies can be based on the distribution or exchange of the knowledge among the radiomics researchers of different centers. Open code, data and models exchange could potentially accelerate the introduction of radiomics into the clinical routine for decision making or diagnostic purposes. However, privacy, legal and ethical issues make the multi-center model development and validation problematic due to the exchange of patients' medical scans that contain private information. Several public repositories such as the XNAT and TCIA host open access datasets for radiomics reproducibility and repeatability studies, nonetheless, the majority of data owners stay circumspect about exchanging or sharing datasets that contain patient scans in a public repository. Due to the barriers of the above-mentioned issues there is a need for the development of a privacy preserving infrastructure that permits data and models exchange for researchers.

The Personal Health Train (PHT)[107] concept constitutes a novel approach that has as a goal to maximise the availability of data for research in a distributed and rapid learning environment. This aim can be achieved by not transferring the data but transferring the research question to the data. This initiative has been successfully applied for the development and validation of clinical data models by recent studies[108–111]. The PHT concept can be applied to the radiomics models with the prerequisite of transforming the radiomics features data to a Findable, Accessible, Interoperable and Accessible (FAIR) format which was first introduced by Wilkinson et al.[112]

The infrastructure that enables the inclusion of radiomics models in the PHT is based in three parts. The first step is the harmonization of the definition and classes of the radiomics features recommended by the IBSI standards[7]. This specification relies on the fact that the radiomics features can be extracted with a radiomics software application that supports the extraction of different feature classes or different feature names. Secondly, the development of the Radiomics Ontology (RO) which is available to the BioPortal[113] enabled the semantic description of the radiomics features objects following the Semantic Web Standards, as every feature class object is labelled with unique identifiers according to the recommendations of the IBSI. The third part consists of the storage of the radiomics features in Resource Description Framework (RDF) format in FAIR data stations across the different institutes for the exchange of statistical models with "trains".The overall representation of the PHT concept is shown on the figure 2.4.1

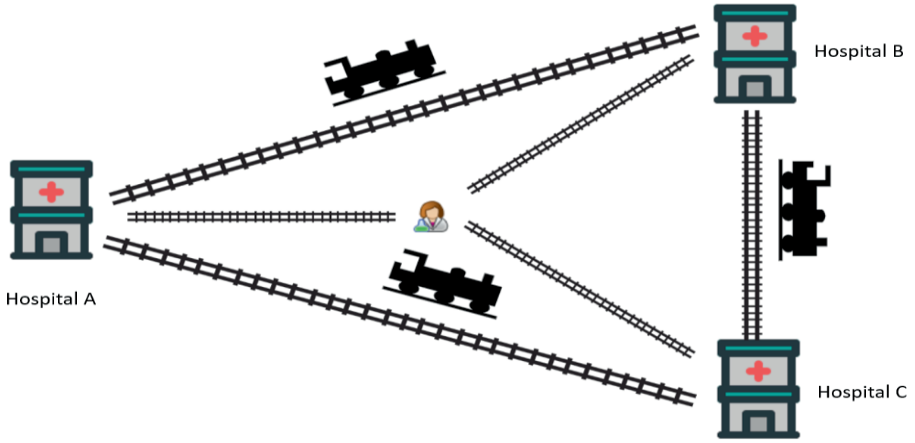


Figure 2.4.1: Representation of the Personal Health Train (PHT) concept across three different hospital where FAIR data stations are connected with statistical "trains" for the models exchange.

The recent study of Shi et al.[114] used the distributed learning approach with the PHT for the two years overall survival radiomics signature model of NSCLC patients using two different datasets from the innovative radiomics study of Aerts et al.[5]. The study succeeded to learn and validate the radiomics signature in two different institutes (Maastro clinic and Radboud University Medical Center) in a remote distributed way across the two institutes using the Lung1[114] as a training dataset and the Radboud's internally Lung2 set as a validation dataset using the Varian Learning Portal (VLP) application. Besides the advantage of the privacy preserving data models exchange by using the PHT infrastructure, the reproducibility and repeatability of the radiomics studies is supported by using the publicly available ontologies (ie. RO). Moreover, the flexibility of the RDF format enhances the distributed learning even in the cases of different data structures across different institutes. In addition, Bogowicz et al.[116] presented the approach of a distributed learning radiomics model training.This study showed comparable centralized and distributed learning results of a radiomics model for the prediction of two year overall survival and human papillomavirus (HPV) status, including six different head and neck cancer patients' cohorts of more than 1000 CT scans.

Reliable and responsible radiomics research requires standardized data collection, evaluation criteria and reporting guidelines. Large-scale data sharing is necessary for the validation and full potential that radiomics can present to mature as a clinical discipline. The combination of radiomics and FAIR data principles can accelerate the clinical translation of radiomics only with the development of initiatives and combination of knowledge among the different experts and stakeholders of the radiation oncology world. Especially nowadays, we are facing the new era of AI in radiation oncology where the explosion of medical imaging data creates an ideal environment for ML and data science. However, the implementation of FAIR data in radiomics and radiation oncology requires strong efforts and initiatives to prove the clinical value and benefit for the clinicians. With our study, we suggest the knowledge exchange and the adoption of FAIR-inclusion criteria for a better understanding, representation and commissioning of the radiomics models.

2.3 A Practical Roadmap for Radiomics Research in Radiation Oncology

The goal of this section is to synthesize and distill key points from this chapter and provide researchers a practical guide of the steps and tools required for a responsible, reliable, and reproducible radiomics study. The starting point is the clear definition of the clinical question. Furthermore, suggestions and recommendations are provided regarding the next steps of the radiomics pipeline such as the data selection, image preprocessing and model development.

2.3.1 Finding the clinical question and hypothesis

Prior to devoting the substantial resources, time, and labor needed for radiomics projects, it is crucial to clearly define the clinical question that one seeks to answer and the unmet need that radiomics could address in the particular clinical situation. Ideally, one should envision that successful use of the radiomics model will result in improvement of patient care and medical decision-making. Beyond this, the following must be defined:

- ***Is this a suitable problem for radiomics?***
    - There are several areas in which radiomics may be appropriate, including: deciphering underlying pathologic or molecular features of tumors, predicting clinical endpoints (survival, progression, tumor response), host-tumor interface (e.g. immune infiltration, necrosis, edema), predicting toxicity prediction of radiotherapy plans. Complementary to this question, one should ask *is radiomics necessary for this?* If the current problem is already adequately addressed by qualitative or simpler radiographic features, one should pause at the need for more complex machine learning based approaches.
- ***What are the study endpoints?***
    - Primary and secondary project endpoints must be clearly defined and objectively evaluable. For instance, will this model predict a rate of 1-year recurrence? A probability of genomic mutation?
- ***How will success be defined and measured?***
    - *Choose your metrics:* commonly used values to evaluate radiomic study endpoints for classification models are: receiver operating characteristic curves (ROC) with area under the curve (AUC), sensitivity (1-false negative rate), specificity (1-false positive rate), raw accuracy, F1 score, positive predictive value, and negative predictive value.
    - Based on initial classification probabilities, risk-groups can be defined, and survival analyses can be performed for clinical endpoints
    - Establish the benchmark performance for the current “standard”
    - Propose an alternate hypothesis for the study model’s performance, and calculate the sample size needed to power your study to ensure the primary endpoint is evaluable. If a benchmark is not available, it may be reasonable to proceed without one, but in this case, it is incumbent on the

researcher to defend the level of performance that would be deemed "success."

- Delineate clinical implementation prior to initiation of research: how does use of the model impact decision-making and clinical care? Decrease treatment toxicity? Improve patient cure rates? Save time and resources? These endpoints can then be evaluated in parallel or subsequent to the development phase of the model.

2.3.2. Database selection and curation

Database selection and curation is one of the most important and labor-intensive parts of radiomics research. Data curation, cleaning, and standardization is absolutely crucial to successful model building. Several specific questions must be addressed here:

- ***Choosing the ideal data source for the project***
  Determine the imaging modality or modalities best suited as *input* data for the project, given the clinical context and scientific rationale. The standard diagnostic oncologic imaging modalities consist of CT, MRI, and PET (and other radiopharmaceutical-based scans). In radiation oncology, simulation-planning scans, RT structure sets and dose information, and on-treatment image guidance can all be utilized as possible data sources depending on the application.

- ***Labeling, annotating, and establishing ground-truth***
  Equally important to identifying a robust data source for model input, is developing an accurately annotated set of *output* data for model training. If pathologic information is being predicted, quality of pathology review and molecular studies should be reviewed. If clinical outcomes are being predicted, clinical record abstraction should be conducted in a standardized way, with a dedicated instrument, and endpoints, particularly progression endpoints should be clearly defined and follow-up times recorded accurately. For retrospective toxicity measurement, standardized instruments should be used. Any uncertainty, heterogeneity, or missing data among annotations should be documented and considered during model training. Independent quality checks of data labels should be conducted, if possible.

2.3.3 Image preprocessing

Imaging information comes with significant inter- and intra- modality and institutional heterogeneity. As the goal of radiomics is to use machine learning to correlate quantified imaging features with specific endpoints, great lengths should be taken to both suppress unwanted distortions in data and enhance features hypothesized to correlate with the outcome of interest. Image preprocessing serves to "de-noise" your data and maximize the likelihood of identifying valid, biologically-driven radiomic signatures. Specific preprocessing tasks will be project and modality-specific, though these general principles apply:

- ***Choose a platform***: most preprocessing can be performed on any modern desktop or laptop CPU, and the most popular computing languages for medical image processing include Python (with packages such as Pyradiomics (pyradiomics.io), MatLab, and R.
- ***Image Acquisition:*** Information regarding scanner model, reconstruction algorithm, pixel resolution, slice thickness, and use of contrast must be obtained. For MRI, coil characteristics, sequence characteristics (echo time, repetition time, acceleration, bandwidth) are necessary. Post-acquisition processing algorithms should be determined, along with overall scan quality, and artifact due to implants or motion.
- ***Region of Interest (ROI) Segmentation and Extraction:*** Generally, radiomics studies require segmentation of the region of interest within an image. In the context of radiation oncology, this is often a tumor or particular organ of interest. Segmentation can be performed manually or with automated processing scripts and/or separate machine learning algorithms. Clinical and technical expertise should guide the most appropriate way to segment an ROI. For instance, sometimes important information may lie outside of the tumor boundaries, in which case inclusion in the ROI would be helpful for model development. It is recommended to remove additional image information outside the ROI.
- ***Image Normalization:*** Once the ROI has been defined, pixel/voxel values must be normalized across cases. Image intensity discretization (e.g. binning) should be performed. For CT, Hounsfield units should be converted to raw pixel values and interpolated to uniform x, y, and z spacing. For MRI, binning to a defined number of grey levels is necessary. For PET, standardized uptake values (SUV) should be calculated from the DICOM metadata, when available. Skipping any of these steps may result in unstable radiomic feature analyses.

2.3.4 Feature selection

A fundamental requirement for clinical utility of radiomic biomarkers is that their values internally consistent and also stable across various testing scenarios. While various methods to ensure reproducibility exist, it is recommended to follow the guidelines of the Imagel Biomarker Standardization Initiative (IBSI) (ibsi.readthedocs.io). There have been hundreds of radiomic features described and an important step in model building is determining which, and how many features to use. While as a starting point, it may be feasible to utilize all features in a complex machine learning model, often selecting the most important features and discarding others can minimize noise, collinearity, and dimensionality, and ultimately improve performance.

- **Filter methods:** Filter methods are feature-ranking methods that rank features based on their importance and redundancy in model prediction performance. These can be univariate and multivariate methods, such as Wilcoxon test, Gini index, Fisher score, and others. As a rule of thumb, several methods should be applied on the development dataset to determine the best approach, prior to model validation and testing.

- **Principal component analysis:** Principal component analysis reduces dimensionality by creating new combinations of radiomic features that contribute to the most prediction variance.
- **Prior knowledge:** In some cases, pre-existing literature has described certain radiomic features known to correlate with a particular outcome. In these cases, features hypothesized to be important should be included a priori.

2.3.5 Machine learning method

After identifying features which serve as inputs for a model one must decide the most appropriate machine learning technique to employ. We recommend attempting multiple machine learning methods to best identify the technique is best suited to model the radiomic features. As the use of machine learning increases in quantitative image analysis as do the number of available machine learning methods. The best technique to employ is not always clearly defined, but a few aspects of the data can help to guide the decision process.

-**Data Set Size**: Data set size is likely the largest driver for which machine learning technique to use. If the dataset of interest is small (n<100 samples), then we would likely recommend first attempting to model your selected features using generalized linear models. For particularly large datasets, generalized linear models may also be appropriate, but more complex machine learning algorithms such as random forest, deep learning, Bayesian networks may be better.

-**Outcome Variable:** The nature of your outcome variable will likely help define the most appropriate machine learning method to model your radiomic data. Certain machine learning algorithms like support vector machines and logistic regression are well suited for binary outcomes (for eg. Presence of tumor). Other machine learning techniques like deep learning and random forest are well suited for multi-class classification tasks and those tasks which may have class imbalances.

-**Feature Number:** As the number of radiomic features increases, more exotic machine learning techniques (deep learning, random forest) may be more useful to mitigate bias associated with excessive collinearity with similar radiomic features. Generalized linear models techniques may be employed with large amounts of features, however adjustment for excessive co-linear features using interaction terms may become cumbersome. If one is particularly interested in using a generalized linear model, we recommend using dimensionality reduction to reduce the number of correlated features first.

-**Computational Resources:** Certain machine learning methods are computationally expensive and difficult to complete with limited computer resources. Specifically, deep learning algorithms often require significant computational resources to train. If one is limited in their computational resources, generalized linear models and support vector machines are often more easily employed.

2.3.6 Training, validation, and benchmarking

Once a model has been selected, the training process begins. Splitting your data into separate training and validation sets (and sometimes a third independent test set) is crucial prior to initiating the training process, as to not overfit your model. A portion of your data should be set aside and not seen or explored until a final model has been selected from the training set. Within the confines of the training set, one can be comfortable tweaking training heuristics, model hyperparameters, feature selection techniques, and even preprocessing steps to achieve maximum performance. Once a final model has been developed and selected, the model should be locked and saved. Following this, it should be tested blindly on the held-out ***internal validation*** set. Furthermore, it should be tested on an ***external validation*** set that represents data from heterogeneous institutions and patient populations. Finally, ***benchmarking*** should be performed against the current standard of care for the application of interest. Depending on the application, this could consist of comparison to human experts (often radiologists or oncologists), traditional diagnostic criteria, staging systems, or clinical factors.

2.3.7 Stability

After a radiomic model has been developed and validated, continued stability assessment is still necessary. Often overlooked following the validation of a model, examination of model behavior in different conditions can often lead to more interesting findings for future study. We recommend the following post-validation assessments.

- **Feature Importance:** Identifying and examining radiomic features of importance may help better understand the clinical phenomenon which is being studied. Isolating particular radiomic features which are highly correlated with an outcome of interest can help guide further study in the lab and clinic.
- **Generalizability**: Testing a model across different patient populations will assure generalizability of a proposed radiomic model. Given the proclivity for machine learning algorithms to overfit training data and maintain bias present within training/validation datasets, it is import to continue to test radiomic models across different patient groups. If overfitting does appear when testing the model among a different group of patients, regularization techniques may help improve generalizability.
- **Stability Testing:** Ensuring stability of radiomic features and model predictions across patients is important to assure the model is truly modeling the clinical outcome of interest. Stability can be measured by examining differences in radiomic features and model predictions across patients with similar clinical characteristics.
- **Failure Examination:** Study of model failures is an important method to characterize potential pitfalls in your radiomic model. Specifically, identifying phenotypes of patients which a model consistently underperforms is a clinically useful endeavor. Examination of failures can identify unique sub-populations of patients which were previously unidentified or potentially uncover a bias in the data used to train the radiomic model.

2.4 Translation of radiomics into the clinic

In this section we will summarize the importance of understanding the relevance of the radiomics models for the physicians regarding the machine learning tools used. Interpretability and validation of the ML radiomics models enhance the introduction of these models to the clinical routine combined with their deployment and commissioning. Moreover, the importance of the outcome of the radiomics prediction model is compared with the operational excellence required for the model development.

2.4.1 Black box vs interpretability

Radiomic models can be considered "black-box" tools where it is not obvious what procedures are undertaken to generate the output from the inputs; though this does not necessarily have to be the case. This lack of transparency and interpretability can pose a challenge with regard to adoption and deployment of such models into routine clinical use from the aspects of both 1) if and how a physician will use such a tool as clinical decision support to improve patient care/outcomes and 2) what are the normative best practices for efficacious and ethical deployment of such tools at the institutional and societal level. The difficulty in addressing these challenges are at two levels: 1) regarding the nature of quantitative image derived radiomic features and 2) regarding the procedural operation of machine learning models.

With regard to imaging biomarkers, quantitative image derived radiomic features are contrasted against qualitative features (or "semantic features). Semantic imaging features typically have an intuitive interpretation related to pathophysiology (e.g. edema, contrast enhancement etc.); whereas quantitative imaging features (e.g. texture, statistics, shape etc.) typically do not have obvious interpretation or connection to the underlying biology of the disease process[117]. It has been proposed that as radiomic signatures are further studied and validated, meaningfulness, and thereby interpretability, will accrue as correlative associations with known genomic, cellular and metabolic oncogenic processes are uncovered[118,119]. It is also hoped strengthening connections and combining models with semantic features will improve meaningfulness and interpretability[119]. The validation and translational roadmap for such radiomic models is otherwise not dissimilar to that of imaging biomarkers generally: assay validation, biological and clinical validation, cost effectiveness assessment, and an end goal of large prospective trials[17]. Having followed a similar path, genomic biomarkers have achieved translational validity and, perhaps, some meaningfulness, to the point of routine use by oncologists in the clinical setting (e.g. OncotypeDx); that too despite lack of transparency[120,121]. This is likely due to the emphasis on empiricism in modern clinical oncology research and practice[122–125], radiation oncology practice in particular has had a strong tradition of proceeding on such basis[126,127].

Aside from the non-intuitiveness of radiomic features, concerns regarding black-box nature of machine learning algorithms (which are trained rather than specified[128]) used to develop such models remain, particularly with increasing size and complexity (e.g. when convolutional neural networks are used i.e. "deep radiomics"[3]). It is argued that there is an inherent trade-off between model performance and the ability to interrogate and understand the model and that this causes a conundrum as prioritizing performance would be at odds with the moral responsibility of a clinician to scrutinize, justify, and demonstrate sound decision-making[129]. How can one do so if one is unable to understand the model's working? As has been stated in a recent editorial regarding the application of such algorithms in healthcare: “[. . .] accuracy is not sufficient to engender trust. An understanding of why decisions are made by the algorithm, the rigour of evaluation, the accreditation of the algorithm, and when and why errors might occur are all points to consider before any algorithm is used in practice”[15]. An additional aspect to consider is the fact that in implementation, self-learning models are dynamically updated, are “plastic,” and could propagate errors at scale which adds additional complexity to validation, safety, and soundness[130]. There is the real possibility of harm. The ability to troubleshoot and debug a radiomic/machine learning model's implementation is critical. There are evolving efforts to tackle the legal aspects surrounding the development and implementation of such models such as liability (malpractice), intellectual property (corporate misincentives toward secrecy via opaque algorithms), and regulation (FDA, EU, “right to explainability”)[130–132]. Much work is progressing in outcomes prediction and prognostication using radiomic models in radiation oncology[133]. As work evolves toward impacting various levels of treatment-planning decision making, the stakes will increase and factors relating to interpretability will be of greater concern in deployment.

Given these issues, the question remains regarding the way forward: “Should machine learning models be pushed toward those implementations that are more interpretable, more mechanistically modeled, and ultimately aimed at increased understanding, or should acceptable models include fundamentally black box algorithms that are practically useful but provide little scientific insight?”[131] This continues to be a matter of debate. Some argue that it is imperative to use the most performant algorithm for patient-care, even if that necessitates a trade-off with transparency, and additionally argue that such black box nature is not too different than the black box aspects of conventional medical decision making[129]. However, it has also been argued that such a performance vs. transparency tradeoff is not axiomatic (and may not even be evidenced in real world implementations) and that given the high stakes with healthcare decisions, interpretability must be prioritized[134]. There is much debate on what makes a model interpretable[135,136]. Post-hoc “explainability” of a model has been posited as a way to address this, e.g. with visualization algorithms[137] ;however, there is skepticism as to the reliability and utility of such an approach and whether it will suffice to address the aforementioned issues. In its stead, a rigorous focus on developing inherently interpretable machine learning algorithms (designed as such from the outset) has been suggested. Though more burdensome and challenging (both computationally and intellectually), generating such a model that is just as performant as fully black box ones is thought to be achievable[134]. Interpretability remains a goal and subject worthy of further elaboration, especially as the foundational infrastructure expectations of the field continue to be established.

As radiomic models find routine clinical implementation and use by physicians, the way such models are incorporated into decision-making will likely require iteration and standards that enhance trust and transparency. To some extent this will be due to increasing comfort with radiomic models over time through empiric clinical use of those with prospective validation and validation against underlying biology. However, it is likely that legal/ethical implications, institutional/regulatory standards, and stakeholder expectations will demand as much understanding, clarity, and interpretability as reasonably achievable; this area will likely continue to evolve in an interdisciplinary fashion.

2.4.2 Deployment/implementation/commissioning of radiomics models

As more studies demonstrate the efficacy of radiomics within clinical practice, there is a growing focus on implementing and deploying radiomic models within the clinic. Although there is no formalized commissioning of clinical radiomic models, there are a considerable number of steps which can be taken to assure a candidate radiomic model is fit for clinical use.

-**Generalizability Testing**: We recommend testing radiomic models in various settings to assure generalizability. Specifically, it is recommended to test across different patient populations which are representative of the clinical population for which the proposed model may be used. We also recommend similar generalizability testing across different imaging protocols and scanners.

-**Benchmarking:** It is important to identify benchmarking standards of radiomic model performance which can be assessed frequently after the model is deployed.

-**Clinical Impact Measurement:** Following implementation, it is necessary to assess the potential impact a radiomic model has on clinical practice. Specifically, this can be achieved by retrospective review of clinical decision making prior to implementation and following implementation. Survey studies of clinicians may aid in understanding barriers to implementation of radiomic models within clinical practice.

2.4.3 Operational excellence vs outcome prediction

The introduction of radiomics has established the principles of personalized medicine in radiation oncology. Due to the numerous and continuous efforts of the radiomics research community to enhance the reproducibility of radiomics studies and enrich the knowledge, new clinically relevant questions have been raised. The main goal of these efforts is the operational excellence of radiomics methodology. Taking into account the significance and influence of the different parameters that are included on the radiomics pipeline for final outcome prediction, we have proposed potential solutions for the standardization of the workflow.

From the patients' perspective, overall survival constitutes the most common endpoint of interest. This specific endpoint depends on tumour features and clinical parameters of the patients such as toxicities, demographics and treatment. The complex task of the prognosis of overall survival presents higher accuracy in aggressive types of cancer such as glioblastoma [138]and NSCLC[139] while on the contrary the study of Ger et al.[140] reported the failure of the improvement of prediction of the overall survival of large cohorts of head and neck cancer patients. Due to the complexity of the prediction of the abovementioned endpoint, several endpoints can be implemented to the clinical routine such as the tumour regression or toxicities.

The cost of the prediction in the clinic constitutes a significant factor for the acceptance and implementation of the prognostic models. Generally, the predictive models are designed with equal cost of false results. For instance, for head and neck cancer patients the cost of the false-positive p16 result is significantly higher than potentially would be translated into a de-escalation of the treatment when a more aggressive approach is optimal. For this reason, the clinically acceptable accuracy of a model is a crucial factor for the introduction of radiomics models into the clinic. In any case, every prognostic model needs reliably labeled and quality data for accurate performance regardless of the statistical method or algorithm used for the model building[102].

Concluding, the application of models based on radiomics features beyond proof of concept needs the cooperation and knowledge exchange of each mutual field by the different experts. Especially nowadays, we are facing the new era of the demonstration of AI in several clinical studies as multisource datasets are used to improve the prediction of patients' treatment outcomes. The crucial point regarding the generalization and validation of the models relies on the usage of different databases consisting of different patients' demographic details, as the validation process may present difficulties and limitations.

2.5 Conclusion

In this chapter, we emphasized the fundamental principles that are part of a radiomics pipeline. We gave specific recommendations regarding the standardization of the radiomics workflow, with the aim to provide a useful guide for clinicians with a high interest in the radiomics field. Furthermore, we presented a roadmap with all the necessary steps that should be taken into account to produce more reproducible and reliable radiomics prediction models. Moreover, we described the general approach related to interpretable radiomics models, as well as the potential solutions regarding the implementation and commissioning of the models. Although there are significant barriers and challenges regarding the acceleration of the introduction of radiomics into the clinic, increasing efforts are made by the radiomics community to enhance the potential of radiomics and AI in radiation oncology. The multiple barriers and solutions highlighted in this chapter constitute an opportunity for knowledge exchange between the different professions involved in the radiomics field, which hopefully could lead to a broader acceptance of radiomics by the clinicians.

Bibliography

1. Fass L. Imaging and cancer: A review. Molecular Oncology. 2008;2(2):115-152. doi:10.1016/j.molonc.2008.04.001
2. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036
3. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181
4. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magnetic Resonance Imaging. 2012;30(9):1234-1248. doi:10.1016/j.mri.2012.06.010
5. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):4006. doi:10.1038/ncomms5006
6. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. arXiv:161207003 [cs, eess]. Published online October 23, 2019. Accessed November 27, 2019. http://arxiv.org/abs/1612.07003
7. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328-338. doi:10.1148/radiol.2020191145
8. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol. 2016;61(13):R150-R166. doi:10.1088/0031-9155/61/13/R150
9. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC, For the Group. A Collaborative Enterprise for Multi-Stakeholder Participation in the Advancement of Quantitative Imaging. Radiology. 2011;258(3):906-914. doi:10.1148/radiol.10100799
10. Varghese BA, Cen SY, Hwang DH, Duddalwar VA. Texture Analysis of Imaging: What Radiologists Need to Know. American Journal of Roentgenology. 2019;212(3):520-528. doi:10.2214/AJR.18.20624
11. Shafiq-ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44(3):1050-1062. doi:10.1002/mp.12123
12. Shafiq-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep. 2018;8(1):10545. doi:10.1038/s41598-018-28895-9
13. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 1998;17(1):87-97. doi:10.1109/42.668698
14. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging. 2010;29(6):1310-1320. doi:10.1109/TMI.2010.2046908
15. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magnetic Resonance Imaging. 2004;22(1):81-91. doi:10.1016/j.mri.2003.09.001
16. Parker JA, Kenyon RV, Troxel DE. Comparison of Interpolating Methods for Image

Resampling. IEEE Trans Med Imaging. 1983;2(1):31-39. doi:10.1109/TMI.1983.4307610

17. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169-186. doi:10.1038/nrclinonc.2016.162
18. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
19. Braman N, Prasanna P, Whitney J, et al. Association of Peritumoral Radiomics With Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2 (ERBB2) –Positive Breast Cancer. JAMA Netw Open. 2019;2(4):e192561. doi:10.1001/jamanetworkopen.2019.2561
20. Dou TH, Coroller TP, van Griethuysen JJM, Mak RH, Aerts HJWL. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. Lee H-S, ed. PLoS ONE. 2018;13(11):e0206108. doi:10.1371/journal.pone.0206108
21. Shan Q, Hu H, Feng S, et al. CT-based peritumoral radiomics signatures to predict early recurrence in hepatocellular carcinoma after curative tumor resection or ablation. Cancer Imaging. 2019;19(1):11. doi:10.1186/s40644-019-0197-5
22. Sun Q, Lin X, Zhao Y, et al. Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region. Front Oncol. 2020;10:53. doi:10.3389/fonc.2020.00053
23. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. Acta Oncologica. 2013;52(7):1391-1397. doi:10.3109/0284186X.2013.812798
24. Kocak B, Durmaz ES, Kaya OK, Ates E, Kilickesmez O. Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. American Journal of Roentgenology. 2019;213(2):377-383. doi:10.2214/AJR.19.21212
25. Traverso A, Kazmierski M, Welch ML, et al. Sensitivity of radiomic features to interobserver variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients. Radiotherapy and Oncology. Published online August 30, 2019. doi:10.1016/j.radonc.2019.08.008
26. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol Imaging Biol. 2016;18(5):788-795. doi:10.1007/s11307-016-0940-2
27. Shi Z, Foley KG, Pablo de Mey J, et al. External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients. Front Oncol. 2019;9:1411. doi:10.3389/fonc.2019.01411
28. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation. Woloschak GE, ed. PLoS ONE. 2014;9(7):e102107. doi:10.1371/journal.pone.0102107
29. Chen W, Liu B, Peng S, Sun J, Qiao X. Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. International Journal of Biomedical Imaging. 2018;2018:1-11. doi:10.1155/2018/2512037
30. Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: an automatic decision tree-based

learning algorithm for advanced image segmentation in positron emission tomography. Phys Med Biol. 2016;61(13):4855-4869. doi:10.1088/0031-9155/61/13/4855

31. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis. 2017;36:61-78. doi:10.1016/j.media.2016.10.004

32. Wang S, Zhou M, Liu Z, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Medical Image Analysis. 2017;40:172-183. doi:10.1016/j.media.2017.06.014

33. Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. Phys Med Biol. 2016;61(17):6553-6569. doi:10.1088/0031-9155/61/17/6553

34. Jiangdian Song, Caiyun Yang, Li Fan, et al. Lung Lesion Extraction Using a Toboggan Based Growing Automatic Segmentation Approach. IEEE Trans Med Imaging. 2016;35(1):337-353. doi:10.1109/TMI.2015.2474119

35. Just N. Improving tumour heterogeneity MRI assessment with histograms. Br J Cancer. 2014;111(12):2205-2213. doi:10.1038/bjc.2014.512

36. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-577. doi:10.1148/radiol.2015151169

37. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJR. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. European Journal of Nuclear Medicine and Molecular Imaging. 2013;40(1):133-140. doi:10.1007/s00259-012-2247-0

38. Parekh V, Jacobs MA. Radiomics: a new application from established techniques. Expert Review of Precision Medicine and Drug Development. 2016;1(2):207-226. doi:10.1080/23808993.2016.1164013

39. Scalco E, Rizzo G. Texture analysis of medical images for radiotherapy applications. BJR. 2017;90(1070):20160642. doi:10.1259/bjr.20160642

40. Banerjee J, Moelker A, Niessen WJ, van Walsum T. 3D LBP-Based Rotationally Invariant Region Description. In: Park J-I, Kim J, eds. Computer Vision - ACCV 2012 Workshops. Vol 7728. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013:26-37. doi:10.1007/978-3-642-37410-4_3

41. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Machine Intell. 2002;24(7):971-987. doi:10.1109/TPAMI.2002.1017623

42. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q. Deep learning for image-based cancer detection and diagnosis − A survey. Pattern Recognition. 2018;83:134-149. doi:10.1016/j.patcog.2018.05.014

43. Rios Velazquez E, Parmar C, Liu Y, et al. Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. Cancer Res. 2017;77(14):3922-3930. doi:10.1158/0008-5472.CAN-17-0122

44. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60-88. doi:10.1016/j.media.2017.07.005

45. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res.

2018;24(6):1248-1259. doi:10.1158/1078-0432.CCR-17-0853

46. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. Sci Rep. 2017;7(1):5467. doi:10.1038/s41598-017-05848-2
47. Ypsilantis P-P, Siddique M, Sohn H-M, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. Anto RJ, ed. PLoS ONE. 2015;10(9):e0137036. doi:10.1371/journal.pone.0137036
48. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of Clinical Epidemiology. 2016;71:76-85. doi:10.1016/j.jclinepi.2015.10.002
49. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. Physica Medica. 2019;61:44-51. doi:10.1016/j.ejmp.2019.04.009
50. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology. 2014;7(1):72-87. doi:10.1593/tlo.13844
51. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
52. Liu Z, Wang Y, Liu X, et al. Radiomics analysis allows for precise prediction of epilepsy in patients with low-grade gliomas. NeuroImage: Clinical. 2018;19:271-278. doi:10.1016/j.nicl.2018.04.024
53. Guo J, Liu Z, Shen C, et al. MR-based radiomics signature in differentiating ocular adnexal lymphoma from idiopathic orbital inflammation. Eur Radiol. 2018;28(9):3872-3881. doi:10.1007/s00330-018-5381-7
54. Tang Z, Liu Z, Li R, et al. Identifying the white matter impairments among ART-naïve HIV patients: a multivariate pattern analysis of DTI data. Eur Radiol. 2017;27(10):4153-4162. doi:10.1007/s00330-017-4820-1
55. Shen C, Liu Z, Guan M, et al. 2D and 3D CT Radiomics Features Prognostic Performance Comparison in Non-Small Cell Lung Cancer. Translational Oncology. 2017;10(6):886-894. doi:10.1016/j.tranon.2017.08.007
56. Shen C, Liu Z, Wang Z, et al. Building CT Radiomics Based Nomogram for Preoperative Esophageal Cancer Patients Lymph Node Metastasis Prediction. Translational Oncology. 2018;11(3):815-824. doi:10.1016/j.tranon.2018.04.005
57. Liu Z, Zhang X-Y, Shi Y-J, et al. Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. Clin Cancer Res. 2017;23(23):7253-7262. doi:10.1158/1078-0432.CCR-17-1038
58. Huang Y, Liang C, He L, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. JCO. 2016;34(18):2157-2164. doi:10.1200/JCO.2015.65.9128
59. Kakushadze Z, Yu W. *K-means and cluster models for cancer signatures. Biomolecular Detection and Quantification. 2017;13:7-31. doi:10.1016/j.bdq.2017.07.001
60. Clark MC, Hall LO, Goldgof DB, Velthuizen R, Murtagh FR, Silbiger MS. Automatic

tumor segmentation using knowledge-based techniques. IEEE Trans Med Imaging. 1998;17(2):187-201. doi:10.1109/42.700731

61. Itakura H, Achrol AS, Mitchell LA, et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med. 2015;7(303):303ra138-303ra138. doi:10.1126/scitranslmed.aaa7582
62. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*: Critical Care Medicine. 2007;35(9):2052-2056. doi:10.1097/01.CCM.0000275267.64078.B0
63. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. Sci Rep. 2016;6(1):34921. doi:10.1038/srep34921
64. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics. 2018;102(4):1143-1158. doi:10.1016/j.ijrobp.2018.05.053
65. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. The British Journal of Radiology. 2017;90(1070):20160665. doi:10.1259/bjr.20160665
66. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. Tian J, ed. PLoS ONE. 2016;11(12):e0166550. doi:10.1371/journal.pone.0166550
67. Zhao B, Tan Y, Tsai W-Y, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6(1):23428. doi:10.1038/srep23428
68. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. Translational Oncology. 2014;7(1):88-93. doi:10.1593/tlo.13865
69. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncologica. 2017;56(11):1544-1553. doi:10.1080/0284186X.2017.1351624
70. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. Tian J, ed. PLoS ONE. 2017;12(9):e0178524. doi:10.1371/journal.pone.0178524
71. Mackin D, Ger R, Dodge C, et al. Effect of tube current on computed tomography radiomic features. Sci Rep. 2018;8(1):2354. doi:10.1038/s41598-018-20713-6
72. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and Translational Radiation Oncology. 2019;19:33-38. doi:10.1016/j.ctro.2019.07.003
73. Kalendralis P, Traverso A, Shi Z, et al. Multicenter CT phantoms public dataset for radiomics reproducibility tests. Medical Physics. 2019;46(3):1512-1518. doi:10.1002/mp.13385
74. Mackin D, Fave X, Zhang L, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features: Investigative Radiology. 2015;50(11):757-765. doi:10.1097/RLI.0000000000000180
75. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features

in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncologica. 2010;49(7):1012-1016. doi:10.3109/0284186X.2010.498437

76. Yan J, Chu-Shern JL, Loi HY, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. Journal of Nuclear Medicine. 2015;56(11):1667-1673. doi:10.2967/jnumed.115.156927

77. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. Translational Oncology. 2015;8(6):524-534. doi:10.1016/j.tranon.2015.11.013

78. Grootjans W, Tixier F, van der Vos CS, et al. 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. 2016;57(11):1692-1698. doi:10.2967/jnumed.116.173112

79. Shankar LK, Hoffman JM, Bacharach S, et al. Consensus recommendations for the use of 18F-FDG PET as an indicator of therapeutic response in patients in National Cancer Institute Trials. J Nucl Med. 2006;47(6):1059-1066.

80. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42(2):328-354. doi:10.1007/s00259-014-2961-x

81. Lovat E, Siddique M, Goh V, Ferner RE, Cook GJR, Warbey VS. The effect of post-injection 18F-FDG PET scanning time on texture analysis of peripheral nerve sheath tumours in neurofibromatosis-1. EJNMMI Res. 2017;7(1):35. doi:10.1186/s13550-017-0282-3

82. Brooks FJ, Grigsby PW. The Effect of Small Tumor Volumes on Studies of Intratumoral Heterogeneity of Tracer Uptake. Journal of Nuclear Medicine. 2014;55(1):37-42. doi:10.2967/jnumed.112.116715

83. Hatt M, Majdoub M, Vallieres M, 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. 2015;56(1):38-44. doi:10.2967/jnumed.114.144055

84. Herlidou-Même S, Constans JM, Carsin B, et al. MRI texture analysis on texture test objects, normal brain and intracranial tumors. Magnetic Resonance Imaging. 2003;21(9):989-993. doi:10.1016/S0730-725X(03)00212-1

85. Jirák D, Dezortová M, Hájek M. Phantoms for texture analysis of MR images. Long-term and multi-center study. Med Phys. 2004;31(3):616-622. doi:10.1118/1.1646231

86. Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study: Effects of MRI acquisition parameters on texture analysis. Med Phys. 2009;36(4):1236-1243. doi:10.1118/1.3081408

87. Savio SJ, Harrison LC, Luukkaala T, et al. Effect of slice thickness on brain magnetic resonance image texture analysis. BioMed Eng OnLine. 2010;9(1):60. doi:10.1186/1475-925X-9-60

88. Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground

truth. Physica Medica. 2018;50:26-36. doi:10.1016/j.ejmp.2018.05.017

89. Waugh SA, Lerski RA, Bidaut L, Thompson AM. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms: Influence of different MRI protocols on texture analysis. Med Phys. 2011;38(9):5058-5066. doi:10.1118/1.3622605

90. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. 2011;54(3):2033-2044. doi:10.1016/j.neuroimage.2010.09.025

91. Shinohara RT, Sweeney EM, Goldsmith J, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical. 2014;6:9-19. doi:10.1016/j.nicl.2014.08.008

92. Velazquez ER, Parmar C, Jermoumi M, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. Sci Rep. 2013;3(1):3529. doi:10.1038/srep03529

93. Kalpathy-Cramer J, Zhao B, Goldgof D, et al. A Comparison of Lung Nodule Segmentation Algorithms: Methods and Results from a Multi-institutional Study. J Digit Imaging. 2016;29(4):476-487. doi:10.1007/s10278-016-9859-z

94. Pfaehler E, Zwanenburg A, de Jong JR, Boellaard R. RaCaT: An open source and easy to use radiomics calculator tool. Wang Y, ed. PLoS ONE. 2019;14(2):e0212223. doi:10.1371/journal.pone.0212223

95. Shi Z, Traverso A, Soest J van, Dekker A, Wee L. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). Medical Physics. 2019;46(12):5677-5684. doi:10.1002/mp.13844

96. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339

97. Nioche C, Orlhac F, Boughdad S, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res. 2018;78(16):4786-4789. doi:10.1158/0008-5472.CAN-18-0125

98. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42(3):1341-1353. doi:10.1118/1.4908210

99. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiotherapy and Oncology. 2018;127(3):370-373. doi:10.1016/j.radonc.2018.03.004

100. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350(jan07 4):g7594-g7594. doi:10.1136/bmj.g7594

101. Chalkidou A, O'Doherty MJ, Marsden PK. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. Rubin DL, ed. PLOS ONE. 2015;10(5):e0124165. doi:10.1371/journal.pone.0124165

102. Deist TM, Dankers FJWM, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys. 2018;45(7):3449-3459. doi:10.1002/mp.12967

103. Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med. 2018;59(2):189-193. doi:10.2967/jnumed.117.200501

104. Ibrahim A, Vallières M, Woodruff H, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. Seminars in Nuclear Medicine. 2019;49(5):438-449. doi:10.1053/j.semnuclmed.2019.06.005

105. Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics. Med Phys. 2020;47(5). doi:10.1002/mp.13678

106. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141

107. Beyan O, Choudhury A, van Soest J, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intelligence. 2020;2(1-2):96-107. doi:10.1162/dint_a_00032

108. Jochems A, Deist TM, El Naqa I, 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. 2017;99(2):344-352. doi:10.1016/j.ijrobp.2017.04.021

109. Jochems A, Deist TM, van Soest J, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiotherapy and Oncology. 2016;121(3):459-467. doi:10.1016/j.radonc.2016.10.002

110. Deist TM, Jochems A, van Soest J, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and Translational Radiation Oncology. 2017;4:24-31. doi:10.1016/j.ctro.2016.12.004

111. Deist TM, Dankers FJWM, Ojha P, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiotherapy and Oncology. 2020;144:189-200. doi:10.1016/j.radonc.2019.11.019

112. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18

113. Radiomics Ontology - Summary | NCBO BioPortal. Accessed August 16, 2019. https://bioportal.bioontology.org/ontologies/RO

114. Shi Z, Zhovannik I, Traverso A, et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Sci Data. 2019;6(1):218. doi:10.1038/s41597-019-0241-0

115. Wee, L., Aerts, H. J.L., Kalendralis, P., & Dekker, A. (2019). Data from NSCLC-Radiomics-Interobserver1 [Data Set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.cwvlpd26

116. Bogowicz M, Jochems A, Deist TM, et al. Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. Sci Rep. 2020;10(1):4542. doi:10.1038/s41598-020-61297-4

117. Liu Z, Wang S, Dong D, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. Theranostics. 2019;9(5):1303-1322. doi:10.7150/thno.30309

118. Limkin EJ, Sun R, Dercle L, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. Annals of Oncology. 2017;28(6):1191-1206. doi:10.1093/annonc/mdx034

119. Morin O, Vallières M, Jochems A, et al. 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. 2018;102(4):1074-1082. doi:10.1016/j.ijrobp.2018.08.032
120. Andre F, Ismaila N, Henry NL, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update—Integration of Results From TAILORx. JCO. 2019;37(22):1956-1964. doi:10.1200/JCO.19.00945
121. Schildgen V, Warm M, Brockmann M, Schildgen O. Oncotype DX Breast Cancer recurrence score resists inter-assay reproducibility with RT2-Profiler Multiplex RT-PCR. Sci Rep. 2019;9(1):20266. doi:10.1038/s41598-019-56910-0
122. Prasad V. Regarding Empiricism and Rationalism in Medicine and 2 Medical Worldviews. Mayo Clinic Proceedings. 2014;89(1):137. doi:10.1016/j.mayocp.2013.10.019
123. Prasad V. Why Randomized Controlled Trials Are Needed to Accept New Practices: 2 Medical Worldviews. Mayo Clinic Proceedings. 2013;88(10):1046-1050. doi:10.1016/j.mayocp.2013.04.026
124. Coveney PV, Dougherty ER, Highfield RR. Big data need big theory too. Phil Trans R Soc A. 2016;374(2080):20160153. doi:10.1098/rsta.2016.0153
125. Ponder BAJ, Waring MJ, eds. The Science of Cancer Treatment. Vol 2. Springer Netherlands; 1990. doi:10.1007/978-94-009-0709-6
126. Symonds P, Jones D. Advances in Clinical Radiobiology. Clinical Oncology. 2013;25(10):567-568. doi:10.1016/j.clon.2013.07.001
127. Halperin E. Chapter 1: The Discipline of Radiation Oncology. In: Perez & Brady's Principles And Practice Of Radiation Oncology. 7th Ed. Philadelphia, Baltimore, New York, London, Beunos Aires, Hong Kong, Sydney, Tokyo: Wolders Kluwer; 2018. Lippincott, Williams & Wilkins; 2019.
128. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society. 2016;3(1):205395171562251. doi:10.1177/2053951715622512
129. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report. 2019;49(1):15-21. doi:10.1002/hast.973
130. Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018;10(471):eaao5333. doi:10.1126/scitranslmed.aao5333
131. Price II WN. Black-Box Medicine by W. Nicholson Price II. Vol SSRN. 28.; 2014.
132. W. N. Price II, Regulating Black-Box Medicine, 116 Mich. L. Rev. 421 (2017).
133. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Computers in Biology and Medicine. 2018;98:126-146. doi:10.1016/j.compbiomed.2018.05.018
134. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
135. Lipton ZC. The Mythos of Model Interpretability. arXiv:160603490 [cs, stat]. Published online March 6, 2017. Accessed May 25, 2020. http://arxiv.org/abs/1606.03490

136. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:170208608 [cs, stat]. Published online March 2, 2017. Accessed May 25, 2020. http://arxiv.org/abs/1702.08608

137. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7

138. Lao J, Chen Y, Li Z-C, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Sci Rep. 2017;7(1):10353. doi:10.1038/s41598-017-10649-8

139. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and Oncology. 2017;123(3):363-369. doi:10.1016/j.radonc.2017.04.016

140. Ger RB, Zhou S, Elgohari B, et al. 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. Hsieh JC-H, ed. PLoS ONE. 2019;14(9):e0222509. doi:10.1371/journal.pone.0222509

Chapter 3

Multicenter CT phantoms public dataset for radiomics reproducibility tests

Adapted from: “Multicenter CT phantoms public dataset for radiomics reproducibility tests”

Petros Kalendralis, Alberto Traverso, Zhenwei Shi, Ivan Zhovannik, Rene Monshouwer, Martijn P. A. Starmans, Stefan Klein, Elisabeth Pfaehler, Ronald Boellaard, Andre Dekker, Leonard Wee

Medical physics 46.3 (2019): 1512-1518, https://doi.org/10.1002/mp.13385,

Contribution: First authorship

Abstract

Purpose: The aim of this paper is to describe a public open-access (Computed Tomography) CT phantoms image set acquired at three centers and collected especially for radiomics reproducibility research. The dataset is useful to test radiomics features reproducibility with respect to various parameters, such as acquisition settings, scanners and reconstruction algorithms.

Acquisition and validation methods: Three phantoms were scanned in three independent institutions. Images of the following phantoms were acquired: Catphan 700 and COPDGene Phantom II (Phantom Laboratory, Greenwich, NY, USA), and the Triple-modality 3D Abdominal Phantom (CIRS, Norfolk, Virginia, USA). Data were collected at three Dutch medical centers: MAASTRO Clinic (Maastricht, NL), Radboud University Medical Center (Nijmegen, NL), and University Medical Center Groningen (Groningen, NL) with scanners from two different manufacturers Siemens Healthcare and Philips Healthcare. The following acquisition parameter were varied in the phantom scans: slice thickness, reconstruction kernel~~s~~, and tube current.

Data Format and usage notes: We made the dataset publicly available on the Dutch instance of “Extensible Neuroimaging Archive Toolkit-XNAT” (https://xnat.bmia.nl). The dataset is freely available and reusable with attribution (Creative Commons 3.0 license).

Potential applications: Our goal was to provide a findable, open access, annotated and reusable CT phantom dataset for radiomics reproducibility studies. Reproducibility testing and harmonization are fundamental requirements for wide generalizability of radiomics-based clinical prediction models. It is highly desirable to include only reproducible features into models, to be more assured of external validity across hitherto unseen contexts. In this view, phantom data from different centers represent a valuable source of information to exclude CT radiomics features that may already be unstable with respect to simplified structures and tightly controlled scan settings. The intended extension of our shared dataset is to include other modalities and phantoms with more realistic lesion simulations.

3.1 Introduction

Computer-aided analysis of clinical radiological images offers a data-at-large-scale approach towards personalized medicine[1] wherein tumor phenotype may be inferred using images of the entire tumor instead of selective sample biopsies. On the premise that phenotypic variability affects clinical outcome[2], medical imaging offers an efficient and non-invasive method to determine prognosis.

This approach has immense potential to support clinical decision-making in the personalized medicine paradigm[3], i.e. which would be a superior choice of treatment for a given person. Studies in the active field of image-derived markers (i.e. "radiomics") strongly suggest that tomographic images do indeed embed more prognostic information than may be seen by an unassisted human eye[4–8]. In order to be widely generalizable and have meaningful clinical use, it is essential that reproducibility of features can be tested in phantoms[9-10], in addition to validating models in human subjects across different settings and multiple independent institutions [11–13].

Studies have shown that feature reproducibility may be affected by differences in image acquisition parameters, such as slice thickness and reconstruction algorithm[14–17]. Since clinical image acquisition protocols are one of the major sources of variation among different hospitals, phantoms allow testing, comparison and harmonization of radiomics features in similar vein to diagnostic imaging quality assurance. We hypothesize that even simplified phantoms allow us to test for radiomic features that may already become unstable even under tightly constrained conditions.

In this data publication, we offer Computed Tomography (CT) scans of simple phantoms across three Dutch academic medical centres for open access. We chose to start with CT since this modality is readily available in many centres and is a workhorse imaging modality for radiotherapy intervention planning. In many clinics, CT scanners are mature technology with well-established protocols for calibration, quality assurance and routine maintenance.

3.2 Acquisition and Validation methods

3.2.1 Phantoms

Catphan 700

To obtain a baseline for overall CT scanner performance, we scanned a Catphan 700 phantom (Phantom Laboratory, Greenwich, NY, USA) that had been designed specifically for routine quality assurance on CT scanners. It is only suitable for use in CT, and contains test modules for contrast, geometric accuracy and spatial resolution [18-19].

COPDGene Phantom II

The COPDGene phantom II (Phantom Laboratory, Greenwich, NY, USA) was designed for thoracic CT quality assurance in prospective clinical trials (specifically asthma and chronic obstructive pulmonary disorder) with guidance from the Quantitative Image Biomarker Alliance Technical Committee. We used the CCT162 version, which included the standard version CTP698 with two additional supports and acrylic end-plates for stabilization of the phantom during the scanning. An outer polyurethane ring simulated tissue attenuation while an internal oval body (15 cm x 25 cm) simulated lung attenuation. The inner oval held a number of cylindrical cavities for foam, acrylic and water[20-21], as well as a number of internal structures simulating different-sized bronchi.

Triple modality 3D Abdominal Phantom

A 3D multimodality Abdominal Phantom (CIRS, Norfolk, Virginia, USA) measuring 26 cm x 12.5 cm x 19 cm[22] was designed to be used for liver biopsy training under guidance by CT, magnetic resonance imaging or ultrasonography. We scanned Model 057A that simulated the abdomen of a small adult. The materials encased within the phantom represented liver, the portal vein, kidneys, bottom of lung, the abdominal aorta, the vena cava, lumbar spine and the six lowest ribs.

3.2.2 Image acquisition

The images used in our study were acquired using three different CT scanners at independent Dutch centres: MAASTRO Clinic (Maastricht), Radboud University Medical Center (Nijmegen), and University Medical Center Groningen (Groningen). The standard clinical operating procedures for thoracic and abdominal radiotherapy planning CT scans at each of the three centers were used to generate a baseline scan of each phantom. These baseline parameters are stated in Table 1A and 1B, for the Phantom Laboratory and CIRS phantoms, respectively.

We subsequently applied perturbations to imaging settings of the baseline scan. We adjusted the following parameters strictly one at a time and saved each scan: slice thickness (1mm, 3mm and 5mm), reconstruction kernels (between 3 and 5 settings depending on the scanner) and current-exposure product (50 mAs, 150 mAs and 300 mAs). The individual setting for each scan is given in Table 2A and 2B, for the Phantom Laboratory and CIRS phantoms, respectively.

3.2.3 Image annotations

CatPhan700 images were only used for image quality assessment of the baseline scans between participating centres, therefore no annotations were added to the scans.

Regions of interest (ROIs) on the COPDGene and Abdominal phantoms were manually delineated in MIRADA DBx (version 1.2.0.59, Mirada Medical, Oxford, United Kingdom). In the COPD phantom, we delineated four distinct spherical ROIs within two of the insert cavities. In the multimodality phantom, we delineated two different ROIs corresponding to two of the simulated liver lesions, one large and one small (as shown in Figure 1). The delineations were performed by one medical physicist at MAASTRO Clinic. All images and annotations were then exported as DICOM (Digital Imaging and Communications in Medicine)-RT(Radiotherapy) objects.

3.2.4 Data format and usage notes

Our scans are made open access via an instance of the Extensible Neuroimaging Archive Toolkit (XNAT) hosted within Dutch national research infrastructure (TraIT, www.ctmm-trait.nl)[23]. XNAT is an open source platform for imaging-based research and clinical investigations, which manages access to different datasets compartmentalized into separate projects (i.e. collections). Within each collection, XNAT permits browsing of individual cases. The platform supports direct uploading of DICOM images and DICOM-RT objects (plan, structure set and dose grid) with *http* file transfer[24]. Studies in XNAT can be queried and retrieved by means of an API (Application Programming Interface) in the Python programming language by installing the *xnat* library (https://pypi.org/project/xnat/).

The Phantom Laboratory images have been uploaded to the XNAT collection STW-STRATEGY-Phantom_Series1 : (https://xnat.bmia.nl/data/projects/stwstrategyps1).

The CIRS multimodality Abdominal Phantom images have been uploaded to the SNAT collection STW-STRATEGY-Phantom_Series2 : (https://xnat.bmia.nl/data/projects/stwstrategyps2).

In each of the above collections, the subject identifier matches exactly the names shown in the leftmost column of Tables 2A and 2B. DICOM-formatted images and the annotations as DICOM-RTStruct objects are nested under the subject level. A python script for downloading an entire collection is available here: (https://github.com/PetrosKalendralis/Download-XNAT-collections-script).

3.3 Discussion

We have made publically available multi-center phantom CT scans to support investigations in radiomics repeatability and reproducibility, specifically to identify features that may be unstable with respect to image acquisition settings in simplified geometry.

Radiomics reproducibility may be investigated as a function of: scanner manufacturer/scanner type, slice thickness~~es~~, tube current (i.e. signal to noise ratio), and reconstruction algorithm~~s~~. We invite the radiomics community to utilise our dataset for research by extracting radiomics features with their own processing pipelines and comparing the results with other investigators. We also invite the community to contact us in order to share the results of their computations. For the next steps, we intend to host the computed features set from the open source library *pyradiomics v2* (https://github.com/Radiomics/pyradiomics)[25] as well as the associated DICOM image metadata on a public open access website (www.radiomics.org).

This is a fundamental step towards improving benchmarking and standardization of the radiomics field of study. This is in support of valuable harmonization projects such as the IBSI (International Biomarker Standardization Initiative)[26]. The features and metadata will be made available as linked Resource Descriptor Format (RDF) objects labelled with a dedicated radiomic-specific semantic web ontology [https://bioportal.bioontology.org/ontologies/RO], such that the data can be queried through the SPARQL language. To assist the radiomics community with data sharing, a standard tabular template and conversion script to RDF will also be provided at www.radiomics.org.

A number of key limitations in the data must be noted at the present time. First, as explicitly declared by the phantom manufacturers, the phantoms used in this study had not been designed with the specific aim of simulating standard radiomic features. It is not presently not fully understood exactly what should be used as a canonical set of imaging features.

Secondly, we posit that the so-called “test lesions” within the current phantoms represent over-simplified geometries and relatively uniformly-dense material. Complex texture patterns and shape features are not well represented in such simple phantoms. However, these phantoms do present a preliminary opportunity for investigating reproducibility of radiomic features, thus we may be able to test for certain features that already unstable in simplified conditions. We would assert that a feature that is not reproducible in such a constrained setting might be unlikely to be highly reproducible in multi-institutional human studies. To improve on the current situation, the data set might be expanded by scans of more phantoms that contain more realistic tumor-mimicking inserts. These may prove to be more suitable for selecting stable features for inclusion in radiomics investigations.

Lastly, while we have started with CT as the most commonly available imaging modality in our field, we intend to expand this collection to include PET (Positron Emission Tomography) and MRI (Magnetic Resonance Imaging).

In addition to making available multi-center and multimodality phantoms for radiomics reproducibility studies, future work in this field should make publicly accessible DICOM metadata and image pre-processing steps, so as to make radiomics studies as FAIR (Findable, Accessible, Interoperable, Reusable) as possible. To this end, image metadata needs to be linked to the features using publicly available Semantic DICOM (SEDI) ontology[27] and the Radiomics ontology needs to extended to cover image pre-processing.

3.4 Conclusion

We offer a publicly accessible multi-center CT phantom dataset with carefully controlled image acquisition parameters to support reproducibility research in the field of radiomics. The dataset is hosted in a well-established and publicly funded XNAT instance. The data is shared under a Creative Commons Attribution 3.0 License (free to browse, download and use at no cost for scientific and educational purposes). The dataset is offered to the radiomics community to compare simple features extracted with different software pipelines as well as to identify features that may not be stable with respect to image acquisition conditions even under highly simplified conditions. Our unique contribution to the field is to investigate the robustness of each radiomics feature with respect to different scanning acquisition parameters.

3.5 Acknowledgments

The authors thank the in-kind contribution of the commercial vendors, Computerized Imaging Reference Systems (CIRS) and The Phantom Laboratory, that supported our study with the loan of the abovementioned phantoms. The commercial partners have made no direct contribution to the writing of this article.

This work has been carried out as part of the Dutch STW-Perspectief research program (STRaTeGy grant numbers 14929 and 14930). The national research infrastructure TraIT is being financially supported by the Dutch Cancer Society.

The authors declare no Conflict of Interests pertaining to the above scientific work.

3.6 Supplementary material

CATPHAN 700/ COPDGENE PHANTOM II BASELINE SCAN PARAMETERS

PARAMETERS	*DICOM tags*	*MAASTRO Clinic (MAAS)*	*Radboud University Medical Center (RADB)*	*University Medical Center Groningen (UMCG)*
MANUFACTURER	(0008,0070)	Siemens	Phillips	Siemens
MODEL	(0008,1090)	Biograph 40	Brilliance Big Bore	Biograph 64
SOFTWARE VERSION	(0018,1020)	syngo CT 2006A	3.6.6	VG60A
SLICE THICKNESS (MM)	(0018, 0050)	3	3	3
TUBE VOLTAGE (KV)	(0018, 0060)	120	120	80
RECONSTRUCTION DIAMETER (MM)	(0018, 1100)	500	255	239
TUBE CURRENT (MA)	(0018, 1151)	39	134	149
EXPOSURE (MAS)	(0018, 1152)	24	124	53
CONVOLUTION KERNEL	(0018, 1210)	B31f	B	I30f
ROWS	(0028, 0010)	512	1024	512
COLUMNS	(0028, 0011)	512	1024	512
PIXEL SPACING	(0028, 0030)	0.98	0.25	0.46

BITS STORED	(0028, 0101)	12	12	12
HIGH BIT	(0028, 0102)	11	11	11
RESCALE OFFSET	(0028, 1052)	-1024	-1024	-1024
RESCALE SLOPE	(0028, 1053)	1	1	1

Table 1A. CT scanner details and image acquisition parameters for baseline scans of the Catphan700 and COPDGene Phantom II in each of the participating clinics.

TRIPLE MODALITY 3D ABDOMINAL PHANTOM BASELINE SCAN PARAMETERS

PARAMETERS	*DICOM tags*	*MAASTRO Clinic (MAAS)*	*Radboud University Medical Center (RADB)*	*University Medical Center Groningen (UMCG)*
MANUFACTURER	(0008,0070)	Siemens	Phillips	Siemens
MODEL	(0008,1090)	Biograph 40	Brilliance Big Bore	Biograph 64
SOFTWARE VERSION	(0018,1020)	syngo CT 2006A	3.6.6	VG60A
MANUFACTURER	(0008,0070)	Siemens	Phillips	Siemens
TUBE VOLTAGE (KV)	(0018, 0060)	120	120	80
RECONSTRUCTION DIAMETER (MM)	(0018, 1100)	500	255	239
TUBE CURRENT (MA)	(0018, 1151)	118	190	18
EXPOSURE (MAS)	(0018, 1152)	73	175	9
CONVOLUTION KERNEL	(0018, 1210)	B30f	B	I30f
ROWS	(0028, 0010)	512	512	512
COLUMNS	(0028, 0011)	512	512	512
PIXEL SPACING	(0028, 0030)	0.98	0.75	0.59
BITS STORED	(0028, 0101)	12	12	12
HIGH BIT	(0028, 0102)	11	11	11
RESCALE OFFSET	(0028, 1052)	-1024	-1024	-1024
RESCALE SLOPE	(0028, 1053)	1	1	1

Table 1B. CT scanner details and image acquisition parameters for baseline scans of the multimodality CIRS Abdominal Phantom in each of the participating clinics.

COLLECTION : SERIES 1 – CATPHAN 700 AND COPD II INDIVIDUAL SUBJECT SCAN SETTINGS

SUBJECT	*Institution*	*Slice thickness (mm)*	*Voltage (kvp)*	*Current (mA)*	*Exposure (mAs)*	*Convolution kernel*
CATPHAN-01-MAAS	MAASTRO	3	120	39	24	B31f
CATPHAN-01-RADB	Radboud	3	120	134	124	B
CATPHAN-01-UMCG	Groningen	3	80	165.5	58.5	I30f
COPD-001-MAAS	MAASTRO	3	120	130	80.5	B31f
COPD-001-RADB	Radboud	3	120	210	194	B
COPD-001-UMCG	Groningen	3	120	191	68	I30f
COPD-002-MAAS	MAASTRO	1	120	112.5	69.5	B31f
COPD-002-RADB	Radboud	1	120	210	194	B
COPD-002-UMCG	Groningen	1	120	205	73	I30f
COPD-003-MAAS	MAASTRO	5	120	106.5	66	B31f
COPD-003-RADB	Radboud	5	120	210	194	B
COPD-003-UMCG	Groningen	5	120	195	69	I30f
COPD-004-MAAS	MAASTRO	3	120	91	56	B31f
COPD-004-RADB	Radboud	3	120	54	50	B
COPD-004-UMCG	Groningen	3	120	140	50	I30f
COPD-005-MAAS	MAASTRO	3	120	80	50	B31f
COPD-005-RADB	Radboud	3	120	108	100	B
COPD-005-UMCG	Groningen	3	120	280	100	I30f
COPD-006-MAAS	MAASTRO	3	120	130	80.5	B41f

COPD-006-RADB	Radboud	3	120	325	300	B
COPD-006-UMCG	Groningen	3	120	660	300	I30f
COPD-007-MAAS	MAASTRO	3	120	130	80.5	B41f
COPD-007-RADB	Radboud	3	120	210	194	A
COPD-007-UMCG	Groningen	3	100	230	104	I40f
COPD-008-MAAS	MAASTRO	3	120	130	80.5	B75f
COPD-008-RADB	Radboud	3	120	210	194	C
COPD-008-UMCG	Groningen	3	100	231	104	I44f
COPD-009-MAAS	MAASTRO	3	120	130	80.5	B60f
COPD-009-RADB	Radboud	3	120	210	194	E
COPD-009-UMCG	Groningen	3	100	236	107	I49f
COPD-010-MAAS	MAASTRO	3	120	130	80.5	B80f
COPD-010-RADB	Radboud	3	120	210	194	L
COPD-010-UMCG	Groningen	3	100	232	105	I50f
COPD-011-UMCG	Groningen	3	100	238	108	I70f
COPD-012-UMCG	Groningen	3	100	236	107	B30f

Table 2A. The individual scan settings for the Catphan 700 and COPD II phantoms from the participating different Dutch clinics.

COLLECTION : SERIES 2 – CIRS MULTIMODALITY PHANTOM INDIVIDUAL SUBJECT SCAN SETTINGS

SUBJECT	*Institution*	*Slice thickness (mm)*	*Voltage (kvp)*	*Current (mA)*	*Exposure (mAs)*	*Convolution kernel*
CIRS-AB-001-MAAS	MAASTRO	3	120	118	73	B30f
CIRS-AB-001-RADB	Radboud	3	120	190	175	B
CIRS-AB-001-UMCG	Groningen	3	100	100	50	I30f
CIRS-AB-002-MAAS	MAASTRO	1	120	133	83	B30f
CIRS-AB-002-RADB	Radboud	1	120	190	175	B
CIRS-AB-002-UMCG	Groningen	1	100	95	47	I30f
CIRS-AB-003-MAAS	MAASTRO	5	120	136	85	B30f
CIRS-AB-003-RADB	Radboud	5	120	190	175	B
CIRS-AB-003-UMCG	Groningen	5	100	98	49	I30f
CIRS-AB-004A-UMCG	Groningen	3	120	100	50	I30f
CIRS-AB-004B-UMCG	Groningen	3	120	100	50	I30f
CIRS-AB-004-MAAS	MAASTRO	3	120	141	88	B30f
CIRS-AB-004-RADB	Radboud	3	120	54	50	B
CIRS-AB-005A-UMCG	Groningen	3	120	200	100	I30f
CIRS-AB-005B-UMCG	Groningen	3	120	200	100	I30f
CIRS-AB-005-MAAS	MAASTRO	1	120	137	85	B30f
CIRS-AB-005-RADB	Radboud	3	120	108	100	B
CIRS-AB-006-MAAS	MAASTRO	5	120	137.5	85.5	B30f
CIRS-AB-006-RADB	Radboud	3	120	325	300	B
CIRS-AB-006-UMCG	Groningen	3	120	600	300	I30f
CIRS-AB-007-RADB	Radboud	3	120	190	175	A
CIRS-AB-007-UMCG	Groningen	3	100	98	49	I40f
CIRS-AB-008-RADB	Radboud	3	120	190	175	C
CIRS-AB-008-UMCG	Groningen	3	100	98	49	I44f

CIRS-AB-009-RADB	Radboud	3	120	190	175	D
CIRS-AB-009-UMCG	Groningen	3	100	96	48	I49f
CIRS-AB-010-UMCG	Groningen	3	100	97	48	I50f
CIRS-AB-011-UMCG	Groningen	3	100	98	49	I70f
CIRS-AB-012-UMCG	Groningen	3	100	97	48	B30f

Table 2B. The individual settings of the Triple modality 3D abdominal phantoms from the three participating Dutch clinics.

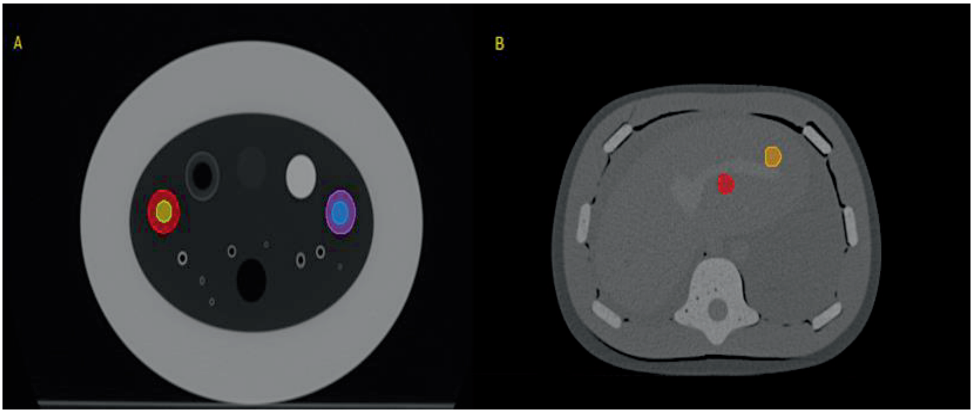


Figure 1. The delineated spherical ROIs within two of the inserts cavities for the COPD and Triple modality 3D abdominal phantoms are presented in the figure 1A and 1B respectively.

Bibliography

1. L. Fass, “Imaging and cancer: A review,” Mol. Oncol., vol. 2, no. 2, pp. 115–152, Aug. 2008.
2. S. Chicklore, V. Goh, M. Siddique, A. Roy, P. K. Marsden, and G. J. R. Cook, “Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis,” Eur. J. Nucl. Med. Mol. Imaging, vol. 40, no. 1, pp. 133–140, Jan. 2013.
3. P. Lambin et al., “Radiomics: Extracting more information from medical images using advanced feature analysis,” Eur. J. Cancer, vol. 48, no. 4, pp. 441–446, Mar. 2012.
4. A. K. Das, M. H. Bell, C. S. Nirodi, M. D. Story, and J. D. Minna, “Radiogenomics Predicting Tumor Responses to Radiotherapy in Lung Cancer,” Semin. Radiat. Oncol., vol. 20, no. 3, pp. 149–155, Jul. 2010.
5. R. T. H. M. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. van Elmpt, “Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures,” Br. J. Radiol., vol. 90, no. 1070, p. 20160665, Feb. 2017.
6. V. Kumar et al., “Radiomics: the process and the challenges,” Magn. Reson. Imaging, vol. 30, no. 9, pp. 1234–1248, Nov. 2012.
7. K. A. Miles, “How to use CT texture analysis for prognostication of non-small cell lung cancer,” Cancer Imaging, vol. 16, no. 1, Dec. 2016.
8. S. S. F. Yip and H. J. W. L. Aerts, “Applications and limitations of radiomics,” Phys. Med. Biol., vol. 61, no. 13, pp. R150–R166, Jul. 2016.
9. M. J. Nyflot, F. Yang, D. Byrd, S. R. Bowen, G. A. Sandison, and P. E. Kinahan, “Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards,” J. Med. Imaging, vol. 2, no. 4, p. 041002, Aug. 2015.
10. X. Fave et al., “Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?: Can radiomics features be measured from CBCT images?,” Med. Phys., vol. 42, no. 12, pp. 6784–6797, Nov. 2015.
11. G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement,” BMJ, vol. 350, no. jan07 4, pp. g7594–g7594, Jan. 2015.
12. A. Traverso, L. Wee, A. Dekker, and R. Gillies, “Repeatability and reproducibility of radiomic features: A systematic review,” Int. J. Radiat. Oncol., Jun. 2018.
13. A. Zwanenburg and S. Löck, “Why validation of prognostic models matters?,” Radiother. Oncol., vol. 127, no. 3, pp. 370–373, Jun. 2018.
14. P. Hu et al., “Reproducibility with repeat CT in radiomics study for rectal cancer,” Oncotarget, vol. 7, no. 44, Nov. 2016.
15. Y. Balagurunathan et al., “Reproducibility and Prognosis of Quantitative Features Extracted from CT Images,” Transl. Oncol., vol. 7, no. 1, pp. 72–87, Feb. 2014.
16. A. Midya, J. Chakraborty, M. Gönen, R. K. G. Do, and A. L. Simpson, “Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility,” J. Med. Imaging, vol. 5, no. 01, p. 1, Feb. 2018.

17. B. Zhao, Y. Tan, W. Y. Tsai, L. H. Schwartz, and L. Lu, “Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study,” Transl. Oncol., vol. 7, no. 1, pp. 88–93, Feb. 2014.
18. “Catphan® 700 Manual.” The Phantom Laboratory, 2016.
19. “Catphan® 700 Data Sheet.” The Phantom Laboratory.
20. “CTP698 and CCT162 COPDGene Lung Phantom II Data Sheet.” The Phantom Laboratory.
21. “CTP698 and CCT162 Lung Phantom II Manual.” The Phantom Laboratory, 2014.
22. “Triple modality 3D Abdominal Phantom Data Sheet.” Computerized Imaging Reference Systems (CIRS).
23. D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, “The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data,” Neuroinformatics, vol. 5, no. 1, pp. 11–34, 2007.
24. K. A. Archie and D. S. Marcus, “DicomBrowser: Software for Viewing and Modifying DICOM Metadata,” J. Digit. Imaging, vol. 25, no. 5, pp. 635–645, Oct. 2012.
25. J. J. M. van Griethuysen et al., “Computational Radiomics System to Decode the Radiographic Phenotype,” Cancer Res., vol. 77, no. 21, pp. e104–e107, Nov. 2017.
26. A. Zwanenburg, S. Leger, M. Vallières, S. Löck, and for the I. B. S. Initiative, “Image biomarker standardisation initiative,” ArXiv161207003 Cs, Dec. 2016.
27. V. S. Johan et al., “Towards a semantic PACS: Using Semantic Web technology to represent imaging data,” Stud. Health Technol. Inform., pp. 166–170, 2014.

Chapter 4

FAIR-compliant clinical, radiomics and DICOM metadata of public collections on The Cancer Imaging Archive (TCIA)

Adapted from: "FAIR-compliant clinical, radiomics and DICOM metadata of public collections on The Cancer Imaging Archive (TCIA)"

Petros Kalendralis, Zhenwei Shi , Alberto Traverso, Ananya Choudhury, Matthijs Sloep, Ivan Zhovannik, Martijn P.A. Starmans, Detlef Grittner, Peter Feltens, Rene Monshouwer, Stefan Klein, Rianne Fijten, Hugo Aerts, Andre Dekker, Johan van Soest, and Leonard Wee

Medical Physics 47.11 (2020): 5931-5940, https://doi.org/10.1002/mp.14322,

Contribution: First authorship

Abstract

Purpose: 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. 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 re-usable (FAIR) data management principles.

Acquisition and validation methods: Overall survival time intervals were updated using a national citizens registry after internal ethics board approval. Spatial offsets of the Primary Gross Tumor Volume (GTV) regions of interest (ROIs) associated with the Lung1 CT series were improved on The Cancer Imaging Archive (TCIA). GTV radiomics features were extracted using the open-source ontology-guided radiomics 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). DICOM metadata was extracted using a research version of Semantic DICOM (SOHARD, GmbH, Fuerth; Germany). Subjects' clinical data was described with metadata using the Radiation Oncology Ontology. All of the above were published in Resource Descriptor Format (RDF), i.e. 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.

Data format: The accumulated RDF data is 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.

Potential applications: 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 is supported by referencing existing public ontologies. The RDF data is 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.

4.1 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[1]. Radiomics refers to high throughput automated characterization of the tumor phenotype by analyzing quantitative features derived from a radiological image[2]. 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[4–7]. Recent studies have suggested that harmonization of radiomics features across multiple institutions and different scanner parameters may be needed to realize its full potential[8–11].

CT images for some frequently-cited studies[3,12], in the Digital Imaging and Communications in Medicine (DICOM) format, have been made available via The Cancer Imaging Archive (TCIA)[12–16]. The DICOM standard incorporates metadata about image acquisition settings and it extends to regions-of-interest 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[17–20]. Software libraries are available that easily change from DICOM to NIfTI/NRRD[21], but in keeping with FAIR (Findable, Accessible, Interoperable and Reusable) data stewardship principles[22], 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[23] 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 re-uses the Radiation Oncology Ontology (ROO)[24], Semantic DICOM ontology (SeDI)[25] and the Radiomics Ontology (RO)[26]. These ontologies themselves re-use existing terminologies and thesauri, such as the Image Biomarker Standardization Initiative (IBSI)[27], National Cancer Institute Thesaurus (NCIT)[28], the Units Of Measurement Ontology (UO)[29], and the DICOM data dictionary[30], 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 re-usability of radiomics features from multiple datasets. The core collection being made publicly available here consists of (i) improvements to the four clinical imaging datasets described in the seminal radiomics publication by Aerts et al.[3] (ii) extracted radiomics features described in line with IBSI recommendations[27,31] and (iii) updates to the subject clinical data associated with the aforementioned image collections.

4.2 Acquisition and validation methods

4.2A 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 3.0[12]), in DICOM format on TCIA and has been previously investigated by Aerts et al.[3]. These collections are described in detail elsewhere; a brief recapitulation is given in Table 1.

In each of these collections, primary Gross Tumor Volumes (GTVs) had been delineated by experienced radiation oncologists; regions of interest (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 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 has been made available with the imaging collections on TCIA.

Table 1. Overall representation of the datasets previously investigated by Aerts et al[3]. The name of each dataset is accompanied with a URL of the TCIA collection and a brief summary of the dataset.

Collection	Description
RIDER Lung CT (link)	This collection was prepared by Zhao et al.[12] to evaluate the variability of tumor unidimensional, bidimensional, and volumetric measurements across "test-retest" CT scans taken at an internal of about 15 minutes (e.g. a "coffee break") with the same image acquisition settings. This has been re-used for radiomics repeatability and segmentation studies. The associated ROIs denoted *GTVp_test_man* and *GTVp_retest_man* refer to manual delineations in the test and retest series, respectively. The ROIs denoted *GTVp_test_auto* and *GTVp_retest_auto* were initially generated by a semi-automated segmentation algorithm[32] in the test and retest series, respectively, and manually edited.
NSCLC-Radiomics-Interobserver1 (link)	This collection consists of radiotherapy dosimetry planning CT scans of 22 NSCLC subjects treated by conventionally fractionated external beam radiotherapy at a single Dutch center. The ROIs denoted were manually drawn by 5 experts working independently. The same procedure was repeated after an initial delineation by the above mentioned semi-automatic segmentation algorithm.
NSCLC-Radiomics (link)	This collection consists of radiotherapy dosimetry planning CT scans of 422 NSCLC subjects treated by conventionally fractionated (chemo)-radiotherapy at a single Dutch center. The ROI called *GTV-1* denotes the primary tumor.
Head-Neck-Radiomics-HN1 (link)	This collection consists of radiotherapy dosimetry planning CT scans of 137 subjects with either laryngeal or oropharyngeal cancer treated by conventionally fractionated (chemo)-radiotherapy at a single Dutch center. The ROI called *GTV-1* denotes the primary tumor.

4.2B Data format and usage notes

The workflow of the conversion of clinical data, DICOM metadata and radiomics features to RDF triples is represented in Figure 1.

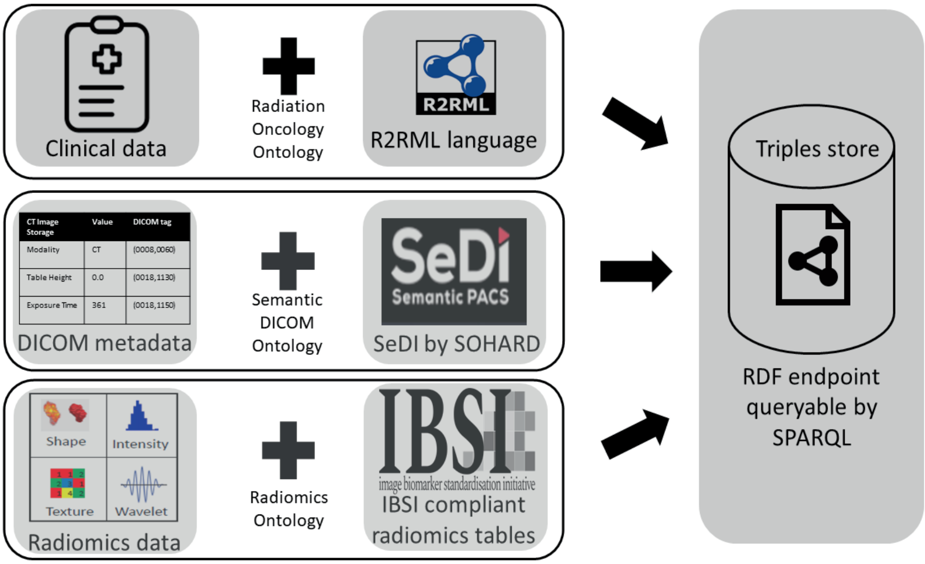


Figure 1. Representation of the conversion of the clinical data, DICOM headers and radiomics features to RDF. The procedures are outlined in the text in sections 2B2, 2B3 and 2B4. The RDF triples can be queried from a publicly accessible endpoint using the SPARQL language.

4.2B2 Clinical metadata as RDF

Clinical tables (in CSV format) from TCIA were imported as standard relational databases (e.g. in PostGreSQL[33] and then converted into RDF triples using a serializing scripting language such as R2RML[34]. 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.[24] .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.

4.2B3 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[35].

4.2B4. 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)[36], a PyRadiomics[37] -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[38]. Radiomics RDF triples were saved to the same aforementioned SPARQL endpoint.

4.2C 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[39] 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.

4.2D 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.[40], using a single radiomics feature that is known by its IBSI text label "Fmorph.vol". We have set up the example query in Text Box 1. In brief, a SPARQL query consists of :

i. Shorthand prefixes for namespaces referring to data, schema, syntax and ontologies that are needed;
ii. SELECT and FILTER commands that allow us to shape the contents to be returned; and,
iii. a sequence of pattern matching rules that allow us to link patients to radiomics features and overall survival outcome.

The contents of Text 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 10 subjects can be seen in Figure 2.

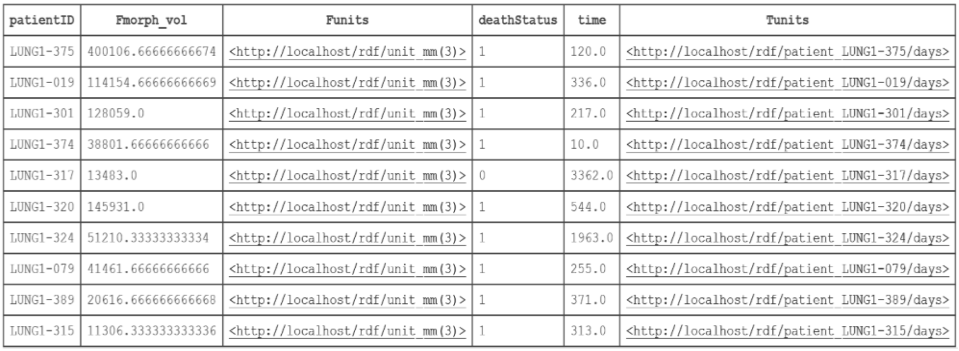
As another purely radiomics-based example, we may examine if distinct radiomics intensity discretization algorithms had been used during the extraction of a radiomics feature. If one were to execute the example query in Text Box 2, it would be seen that the specific radiomics feature labelled as RO:Y1RO[41] 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 Text 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 ob-
jects in the radiomics domain

  ?featureObj roo:100042 ?Fmorph_vol; roo:100027 ?Funits FILTER con-
tains(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 col-
lection
}
LIMIT 10 #purely for the example, we have limited the number of rows of output
to 10
```

Text box 1. 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 Figure 2.



patientID	Fmorph_vol	Funits	deathStatus	time	Tunits
LUNG1-375	400106.66666666674	<http://localhost/rdf/unit_mm(3)>	1	120.0	<http://localhost/rdf/patient_LUNG1-375/days>
LUNG1-019	114154.66666666669	<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	38801.66666666666	<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	51210.33333333334	<http://localhost/rdf/unit_mm(3)>	1	1963.0	<http://localhost/rdf/patient_LUNG1-324/days>
LUNG1-079	41461.66666666666	<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 2. The result of ten patients' cases of the example query given in Text Box 1. We can see the research study IDs of patients from the public TCIA 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).

```
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/>
```

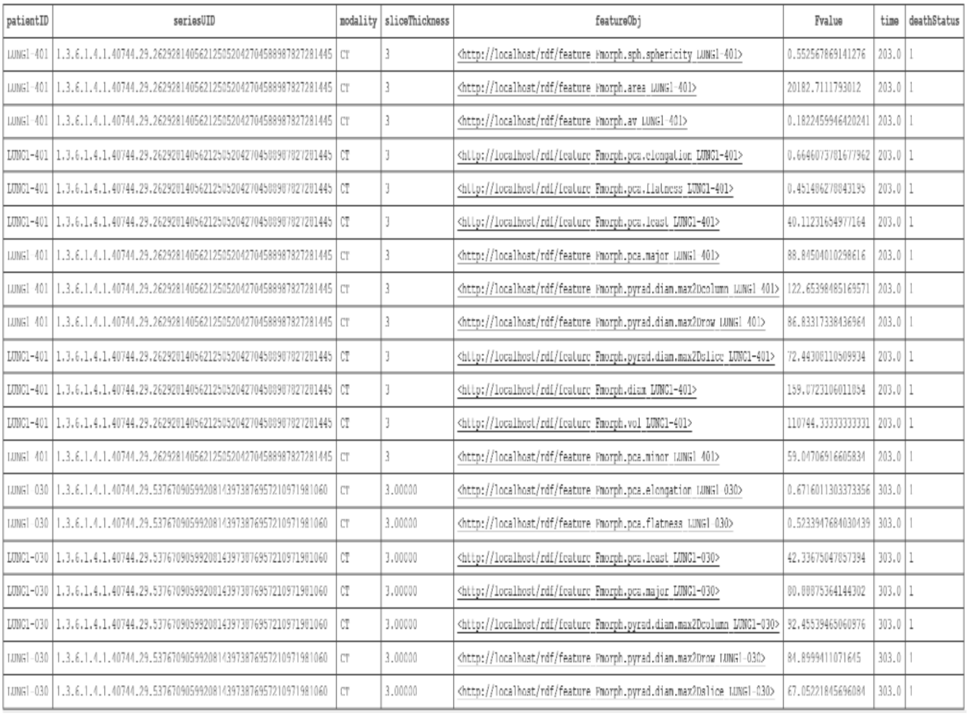
```
SELECT DISTINCT ?paramspace ?discretisationparam ?discretisationAlgorithm
WHERE{
  ?patient a ncit:C16960.
  ?patient roo:P100042 ?patientID.
  ?patient ro:P00088 ?featureObj.

  ?featureObj rdf:type ro:Y1RO.
  #the Radiomics Ontology defines "ro:Y1RO" as a grey-level size zone matrix tex-
tural feature, specifically grey-level nonuniformity normalized
  # i.e. https://bioportal.bioontology.org/ontologies/RO/?p=classes&concep-
tid=http%3A%2F%2Fwww.radiomics.org%2FRO%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 algo-
rithm was used

  FILTER regex(?patientID, "^HN1067").                            #purely for this ex-
ample, we arbitrarily selected one subject to examine
}
```

Text box 2. 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 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 ?survivaldayssinceRT.
   ?survivaldayssinceRT 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
```

Text box 3. 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 Figure 3.



patientID	seriesUID	modality	sliceThickness	featureObj	Fvalue	time	deathStatus
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.sph.sphericity_LUNG1-401>	0.552567869141276	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.area_LUNG1-401>	20182.7111793012	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.av_LUNG1-401>	0.1822459946420241	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pca.elongation_LUNG1-401>	0.6646073781677962	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pca.flatness_LUNG1-401>	0.451486278843195	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pca.least_LUNG1-401>	40.11231654977164	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pca.major_LUNG1-401>	88.84504010298616	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_LUNG1-401>	122.65398485169571	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Drow_LUNG1-401>	86.83317338436964	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dslice_LUNG1-401>	72.44308110509934	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.diam_LUNG1-401>	159.0723106011854	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.vol_LUNG1-401>	110744.33333333331	203.0	1
LUNG1-401	1.3.6.1.4.1.40744.29.262928140562125052042704588987827281445	CT	3	<http://localhost/rdf/feature_Fmorph.pca.minor_LUNG1-401>	59.04706916605834	203.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pca.elongation_LUNG1-030>	0.6716011303373356	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pca.flatness_LUNG1-030>	0.5233947684030439	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pca.least_LUNG1-030>	42.33675047857394	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pca.major_LUNG1-030>	80.88875364144302	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dcolumn_LUNG1-030>	92.45539465060976	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Drow_LUNG1-030>	84.8999411071645	303.0	1
LUNG1-030	1.3.6.1.4.1.40744.29.53767090599208143973876957210971981060	CT	3.00000	<http://localhost/rdf/feature_Fmorph.pyrad.diam.max2Dslice_LUNG1-030>	67.05221845696084	303.0	1

Figure 3. A partial snapshot of the example query given in Text Box 3. Given as a result of the query are : the subject research ID, the CT 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 DICOM and radiomics data are cross-referenced to the vital status and survival time interval as per the example in Text Box 1.

4.3 Discussion

4.3.1 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.

4.3.2 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 this 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[42–44].

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 multi-center studies. 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 re-used in future.

4.3.3 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 is therefore not static, so it is intended to be improved and refined together with developing methodology over time.

4.3.4 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 towards 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 towards a target ontology, the top-down approach requires data elements and a data structure to be rigidly defined first of all before the data is 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 towards 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 re-used 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.

4.4 Conclusion

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 multicenter research into reproducibility of radiomics features across multiple datasets.

Bibliography

1. Fass L. Imaging and cancer: A review. Molecular Oncology. 2008;2(2):115-152. doi:10.1016/j.molonc.2008.04.001
2. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036
3. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):4006. doi:10.1038/ncomms5006
4. Huang Y, Liu Z, He L, et al. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. Radiology. 2016;281(3):947-957. doi:10.1148/radiol.2016152234
5. Yang B, Guo L, Lu G, Shan W, Duan L, Duan S. Radiomic signature: a non-invasive biomarker for discriminating invasive and non-invasive cases of lung adenocarcinoma. CMAR. 2019;Volume 11:7825-7834. doi:10.2147/CMAR.S217887
6. Wu W, Ye J, Wang Q, Luo J, Xu S. CT-Based Radiomics Signature for the Preoperative Discrimination Between Head and Neck Squamous Cell Carcinoma Grades. Front Oncol. 2019;9:821. doi:10.3389/fonc.2019.00821
7. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol. 2019;20(7):1124. doi:10.3348/kjr.2018.0070
8. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics. 2018;102(4):1143-1158. doi:10.1016/j.ijrobp.2018.05.053
9. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? Tomography. 2016;2(4):361-365. doi:10.18383/j.tom.2016.00208
10. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. The British Journal of Radiology. 2017;90(1070):20160665. doi:10.1259/bjr.20160665
11. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol. 2016;61(13):R150-R166. doi:10.1088/0031-9155/61/13/R150
12. Zhao, Binsheng, Schwartz, Lawrence H, & Kris, Mark G. (2015). Data From RIDER_Lung CT. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.U1X8A5NR.
13. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

14. Wee, L., Aerts, H. J.L., Kalendralis, P., & Dekker, A. (2019). Data from NSCLC-Radiomics-Interobserver1 [Data Set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.cwvlpd26.
15. Wee, L., & Dekker, A. (2019). Data from Head-Neck-Radiomics-HN1 [Data Set]. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.8kap372n.
16. Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., … Lambin, P. (2019). Data From NSCLC-Radiomics [Data Set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.
17. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and Translational Radiation Oncology. 2019;19:33-38. doi:10.1016/j.ctro.2019.07.003
18. Mackin D, Fave X, Zhang L, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. Tian J, ed. PLoS ONE. 2017;12(9):e0178524. doi:10.1371/journal.pone.0178524
19. Shafiq-ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep. 2018;8(1):10545. doi:10.1038/s41598-018-28895-9
20. Fave X, Cook M, Frederick A, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. Computerized Medical Imaging and Graphics. 2015;44:54-61. doi:10.1016/j.compmedimag.2015.04.006
21. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods. 2016;264:47-56. doi:10.1016/j.jneumeth.2016.03.001
22. Jochems A, Deist TM, El Naqa I, 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. 2017;99(2):344-352. doi:10.1016/j.ijrobp.2017.04.021
23. Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. 1993;5(2):199-220. doi:10.1006/knac.1993.1008
24. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879
25. Grittner D, van Soest J, Lustberg T, Marshall MS, Feltens P, Dekker A. Semantic DICOM Ontology. http://bioportal.bioontology.org/ontologies/SEDI.
26. Radiomics Ontology - Summary | NCBO BioPortal. https://bioportal.bioontology.org/ontologies/RO. Accessed August 16, 2019.
27. Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS. Image biomarker standardisation initiative. arXiv:161207003 [cs]. December 2016. http://arxiv.org/abs/1612.07003. Accessed June 8, 2018.
28. National Cancer Institute Thesaurus Ontology. https://bioportal.bioontology.org/ontologies/NCIT.

29. Units of Measurement Ontology. https://bioportal.bioontology.org/ontologies/UO.
30. DICOM data dictionary. http://dicom.nema.org/medical/dicom/current/output/html/part06.html.
31. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high throughput image-based phenotyping. Radiological Society of North America (RSNA). 2020. http://orca.cf.ac.uk/id/eprint/128432.
32. Meldolesi E, van Soest J, Alitto AR, et al. VATE: VAlidation of high TEchnology based on large database analysis by learning machine. Colorectal Cancer. 2014;3(5):435-450. doi:10.2217/crc.14.34
33. PostGreSQL. https://www.postgresql.org/.
34. R2RML language. https://www.w3.org/ns/r2rml.
35. Johan VS, Tim L, Detlef G, et al. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. Studies in Health Technology and Informatics. 2014:166–170. doi:10.3233/978-1-61499-432-9-166
36. Shi Z, Traverso A, Soest J, Dekker A, Wee L. Technical Note: Ontology-guided radiomics analysis workflow (O-RAW). Med Phys. October 2019:mp.13844. doi:10.1002/mp.13844
37. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339
38. A.Traverso et al. The Radiomics Ontology (RO): standardizing radiomic studies following FAIR principles-Manuscript in preparation. 2020.
39. Blazegraph application. https://blazegraph.com/.
40. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. Radiotherapy and Oncology. 2019;130:2-9. doi:10.1016/j.radonc.2018.10.027
41. radiomics feature. https://bioportal.bioontology.org/ontologies/RO/?p=classes&conceptid=http%3A%2F%2Fwww.radiomics.org%2FRO%2FY1RO.
42. Shi Z, Zhovannik I, Traverso A, et al. Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Sci Data. 2019;6(1):218. doi:10.1038/s41597-019-0241-0
43. Deist TM, Jochems A, van Soest J, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and Translational Radiation Oncology. 2017;4:24-31. doi:10.1016/j.ctro.2016.12.004
44. Shi Z, Foley KG, Pablo de Mey J, et al. External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients. Front Oncol. 2019;9:1411. doi:10.3389/fonc.2019.01411

Chapter 5

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"

Zhenwei Shi, Ivan Zhovannik, Alberto Traverso, Frank J.W.M. Dankers, Timo M. Deist, **Petros Kalendralis**, René Monshouwer, Johan Bussink, Rianne Fijten, Hugo JWL Aerts, Andre Dekker, Leonard Wee

Scientific data 6.1 (2019): 1-8, https://doi.org/10.1038/s41597-019-0241-0,

Contribution: Data preparation for the study

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.

5.1 Introduction

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

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

A privacy-preserving distributed learning infrastructure based on World Wide Web Consortium "Semantic Web" data sharing standards[19,20], known as Personal Health Train (PHT; https://vimeo.com/143245835)[21] has been successfully used to develop and validate models on non-image clinical data[22-24]. 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)[26] 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)[27].

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") 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.

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

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

Outcomes		
Median follow-up in days	546	595
Median survival time in days	478	500
2-year overall survival rate	40%	41%

5.2 Results

Cohort summary information was exchanged through private discussion between the collaborating investigators, prior to performing this study. This was to confirm that general characteristics were comparable between the updated cohorts. This is shown in **Table 1**. None of the information contained in Table 1 was used in the model. There was a slightly higher proportion of patients with metastatic disease in Lung2 (10% vs 1%) compared to Lung1. The most common histology types in Lung1 were large-cell and squamous-cell carcinomas, whereas adenocarcinoma and squamous-cell carcinoma were most common in Lung2. The median follow-up time, the median survival time and the overall 2-year survival rate were similar in both cohorts.

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

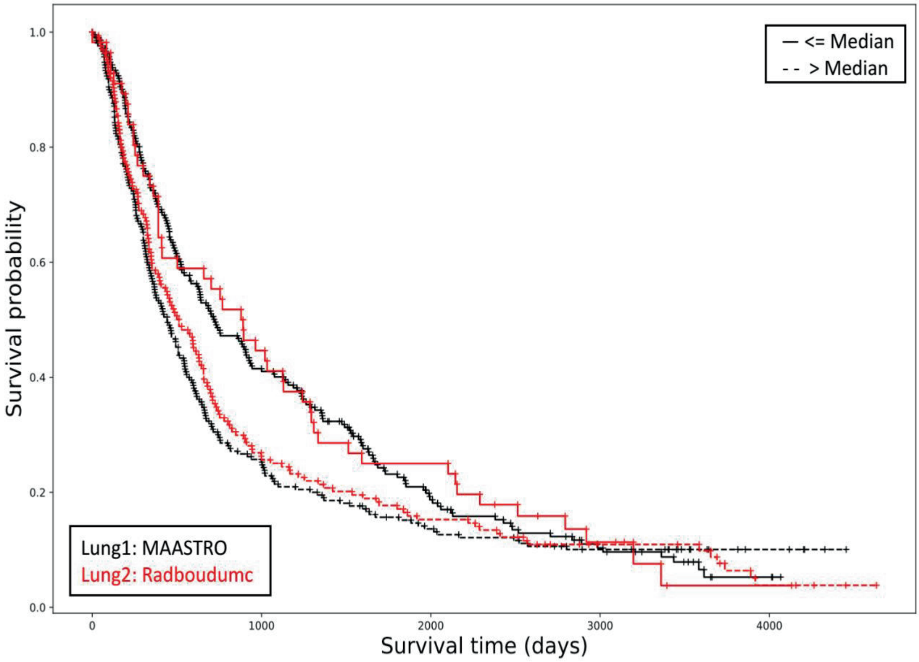


Figure 1. The performance of radiomic signature ZS2019 according to Kaplan-Meier survival analysis. The signature was developed in Lung1 (MAASTRO; black line) and then distributedly validated in Lung2 (Radboudumc; red line). The upper and lower survival curves were split according to the median of the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

Distributed learning code for Cox regression in MATLAB (MATLAB 2016a, Mathworks, Natick MA, USA) was deployed via the PHT infrastructure connecting MAASTRO Clinic and Radboudumc. We retrieved anonymous event timepoints and thus compiled Kaplan-Meier curves for overall survival in each of the training and validation cohorts (in **Figure 1**). Within each cohort, the subjects were stratified into two risk groups, based on the median of the risk score distribution in Lung1. 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.

5.3 Discussion

In this paper, a model (ZS2019) derived from radiomic features and overall survival locally within one institution was able to be exchanged interoperably with an external institution, without mandating any transfer of either images, feature values or clinical outcomes at the individual subject level. This is an essential and unique contribution to radiomic investigations, because we hereby demonstrate the concept for carrying out multi-centre radiomic studies with fully decentralized data. The results obtained with decentralized data were the same as if all the data had been brought into the same location. However, the unique advantage of our approach is that no one party needs to risk breaking patient confidentiality by exposing the original data to another party. Each institutional data owner retains complete control over their privacy-sensitive patient data, and decides what they wish to share for a collaborative project.

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

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

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

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[28]. Development of an optimal radiomic signature for NSCLC overall survival would require a detailed re-examination of features and feature selection in the updated datasets, which is not the primary objective of the present study.

The PHT approach utilises existing data to answer key questions in personalised healthcare, preventive medicine and value-based healthcare. PHT is one of a number of innovative approaches (DataSHIELD[29] and WebDISCO[30]) 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)[29], WebDisco (https://omictools.com/webdisco-tool)[30] and ppDLI (https://distributedlearning.ai/blog) may also be used for distributed radiomics. The differences between the present study and the original study may be traced to : (i) the original Matlab code is commercial confidential and not available to the authors, so we used PyRadiomics developed by Aerts et al.[2] as a practical alternative and (ii) we tried our best to replicate the original method using the documented steps in the original manuscript, but we also improved the survival follow-up such that many right-censored events were now confirmed deaths.

5.4 Conclusion

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

5.5 Methods

Patients

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

Tumour delineations

Radiotherapy treatment planning DICOM CT images and physician-delineated primary NSCLC tumours as RT structure sets were used. From 422 available, 34 cases were found to have a reference frame translation between the image and delineation due to incorrect coding of the treatment couch height offset in the planning system; these have been rectified for the TCIA collection. Only 1 patient was post-operative radiotherapy, so this case was excluded from any further analysis, leading to 421 eligible cases in Lung1 for model training.

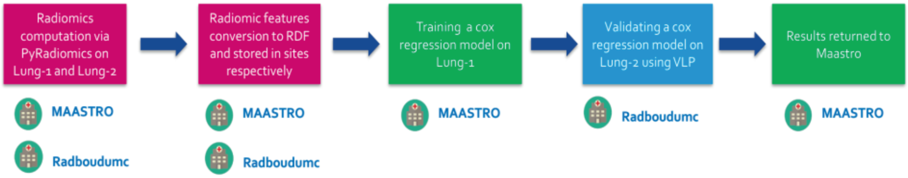

Radiomics computation via PyRadiomics on Lung-1 and Lung-2
MAASTRO
Radboudumc
Radiomic features conversion to RDF and stored in sites respectively
MAASTRO
Radboudumc
Training a cox regression model on Lung-1
MAASTRO
Validating a cox regression model on Lung-2 using VLP
Radboudumc
Results returned to Maastro
MAASTRO

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

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

Outcomes

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

Data processing

The study steps are shown schematically in Figure 2 for MAASTRO and Radboudumc. The core of the radiomic feature extraction process utilizes free and open-source PyRadiomics[15] (v1.3) libraries. Software wrapper extensions collectively known as O-RAW (https://gitlab.com/UM-CDS/o-raw) were used to convert DICOM objects into numerical arrays as inputs for PyRadiomics; these were based on the SimpleITK (v1.0.1)[31] toolkit.

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

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

For the development of ZS2019 we did not select new radiomic features, and instead used the four features with names corresponding to those described previously in the supplementary material[28] that accompanied the original publication:

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

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

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

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

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

Semantic web ontologies

Semantic Web technologies and ontologies play a key role in distributed learning by enabling semantic interoperability between data from multi-centres. In this study, radiomic features and clinical data were defined by a Radiomics Ontology v1.3 (https://bioportal.bioontology.org/ontologies/RO) and a Radiation Oncology Ontology[32] , 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[20] on feature graphs to also probe for similarities in imaging setting, pre-processing methods, and suchlike. We hypothesise that the data graph based approach is more scalable than pairwise cross-referencing of multiple dictionaries of feature definitions.

Distributed approach

The VLP distributed learning architecture has been described in deep detail elsewhere[22-24] . In brief, VLP consists of (i) a global web-based clinical learning environment that spans across any number of participating institutes for a given learning project, and (ii) a local connector application that runs exclusively inside the IT firewall of each institute. The former coordinates access permission, asynchronous messaging, web security and site privacy protocols across the learning network, while the latter hosts a local FAIR data repository. Radiomic feature values were hosted in the respective VLP local connector application (v2.0.1) as RDF.

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

Model training

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

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

Cox model evaluation

Anonymous timepoints for Kaplan-Meier survival curves[33] 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)[34] implemented using the python lifelines package (v0.14.4) was used to quantify discrimination performance using the retrieved timepoints. The log-rank method[35] was used to calculate a chi-squared test statistic and p-value for the significance of the discrimination. To assess if the survival model had any value beyond random discrimination (null hypothesis: c-index = 0.5), we used a two-sided Wilcoxon test with a bootstrap approach on 100 repeated sub-samples of 100 patients per repetition from Lung2.

2-year overall survival

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

Code availability

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

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

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

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

5.6 Data Availability

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

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

5.7 Acknowledgements

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

5.8 Competing interests

MAASTRO Clinic receives institutional research support from Varian Medical Systems.

Andre Dekker receives speaking and consultancy honoraria from Varian Medical Systems.

Andre Dekker holds a patent on radiomics (US Patent 9721340 B2).

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

Bibliography

1. McKnight, J., Babineau, B. & Gahm, J. North American Health Care Provider Information Market Size & Forecast. ESG-Enterprise Strategy Group (2011).
2. Aerts, H. J. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5, 4006 (2014).
3. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. Radiology 278, 563-577 (2015).
4. Kumar, V. et al. Radiomics: the process and the challenges. Magnetic resonance imaging 30, 1234-1248 (2012).
5. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14, 749 (2017).
6. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48, 441-446 (2012).
7. Coroller, T. P. et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiotherapy and Oncology 114, 345-350 (2015).
8. Huang, Y.-q. et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. Journal of Clinical Oncology 34, 2157-2164 (2016).
9. Parmar, C. et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Scientific reports 5, 11044 (2015).
10. Nie, K. et al. Rectal cancer: assessment of neoadjuvant chemo-radiation outcome based on radiomics of multi-parametric MRI. Clinical cancer research 22.21, 5256-5264 (2016).
11. Zhang, B. et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clinical Cancer Research 23.15, 4259-4269 (2017).
12. Foley, K. G. et al. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. European radiology 28, 428-436 (2018).
13. Leijenaar, R. T. et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta oncologica 52, 1391-1397 (2013).
14. Apte, A. P. et al. Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. Medical physics 45.8 3713-3720 (2018).
15. van Griethuysen, J. J. et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer research 77, e104-e107 (2017).
16. Zhang, L. et al. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Medical physics 42, 1341-1353 (2015).
17. Nioche, C. et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer research 78, 4786-4789 (2018).
18. Clark, K. et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging 26, 1045-1057 (2013).

19. Berners-Lee, T., Hendler, J. & Lassila, O. The semantic web. Scientific american 284, 28-37 (2001).
20. Prud'hommeaux, E. & Seaborne, A. SPARQL query language for RDF, https://www.w3.org/TR/rdf-sparql-query (2008).
21. van Soest, J. et al. Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data. Studies in health technology and informatics 247, 581-585 (2018).
22. Jochems, A. et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. International Journal of Radiation Oncology• Biology• Physics 99, 344-352 (2017).
23. Jochems, A. et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–A real life proof of concept. Radiotherapy and Oncology 121, 459-467 (2016).
24. Deist, T. M. et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and translational radiation oncology 4, 24-31 (2017).
25. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3 (2016).
26. Traverso, A. Radiomics Ontology, https://bioportal.bioontology.org/ontologies/RO
27. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative-feature definitions. Preprint at arXiv: https://arxiv.org/abs/1612.07003
28. Aerts, H. J. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. (Supplementary). Nature communications 5, 4006 (2014).
29. Wolfson, M. et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. International journal of epidemiology 39, 1372-1382 (2010).
30. Lu, C.-L. et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association 22, 1212-1219 (2015).
31. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleITK. Frontiers in neuroinformatics 7, 45 (2013).
32. Traverso, A., van Soest, J., Wee, L. & Dekker, A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Medical physics 45.10, e854-e862 (2018).
33. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. Journal of the American statistical association 53, 457-481 (1958).
34. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine 15, 361-387 (1996).
35. Peto, R. & Peto, J. Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society: Series A (General) 135, 185-198 (1972).

Chapter 6

Independent validation of dysphagia NTCP models for the selection of head and neck cancer patients for proton therapy in the Netherlands

Adopted from: "Independent validation of dysphagia NTCP models for the selection of head and neck cancer patients for proton therapy in the Netherlands"

Petros Kalendralis, Matthijs Sloep, Nibin Moni George, Joeri Veugen, Martijn Veening, Johannes A. Langendijk, Andre Dekker, Johan van Soes, Rianne Fijten

Submission in preparation, Contribution: First authorship

Abstract

Purpose/Background: We externally validated the Normal Tissue Complication Probability (NTCP) Grade ≥ 2 at 6months dysphagia model for head and neck cancer patients included in the Dutch National Indication Protocol for Proton Therapy (NIPP) using an independent patient cohort treated with (chemo)radiotherapy in MAASTRO clinic.

Materials/Methods: We used 277 head and neck cancer patients treated with (chemo)radiotherapy in MAASTRO clinic between 2019-2021. For the evaluation of the model discrimination we used statistics metrics such as the sensitivity, specificity and the area under the receiver operating characteristic (ROC) curve.

After validation we evaluated if the NTCP model can be improved using a closed testing procedure (CTP). Specifically, we used calibration curves to graphically assess the i) Original model, ii) Recalibration in the large, iii) Recalibration and iv) Model revision).

The code used for the implementation of CTP and the creation of calibration curves was written in the statistical analysis software RStudio.

Results: The performance of the original NTCP model for dysphagia grade ≥ 2 at 6 months was good in the independent cohort of MAASTRO clinic (AUC=0.80) but according to its calibration curve it was underestimating the risk of the head and neck patients to develop dysphagia. Therefore, we implemented the CTP. The CTP indicated that the model had to be updated and selected a revised model with updated predictor coefficients as an updated model. The revised model had also a satisfactory discrimination in MAASTRO's cohort (AUC=0.83) with an improved calibration of predicted and observed NTCP values.

Conclusion: The validation of the NIPP NTCP model for dysphagia Grade ≥ 2 was successful in our independent validation cohort but can be improved using the CTP. Future steps include the participation of more independent radiotherapy centres for the validation of the NIPP NTCP models through a federated learning approach of the personal health train (PHT) infrastructure.

6.1 Introduction

Head and neck cancer (HNC) constitutes one of the most common cancer types worldwide. It is estimated that over 400.000 deaths are caused by HNC malignancies annually[1]. In Europe specifically, HNC accounts for 4 percent of the cancer incidence with more than 60.000 deaths annually [2]. Radiotherapy (RT) is one of the main treatment modalities for HNC cancer and is generally prescribed alone or in combination with surgery and/or chemotherapy. During the last years, several novel photon-based RT techniques have been implemented in clinical practice such as intensity-modulated radiation therapy (IMRT) and the Volumetric Modulated Arc Therapy (VMAT). The main goal of these RT techniques is to deliver the optimal radiation dose to the treatment target while minimising the radiation dose to the nearby healthy tissues and organs at risk (OARs) and therefore reducing acute and late radiation-induced toxicities[3]. This is especially relevant to HNC patients due to the close proximity of several important organs within a small area of the body.

In addition, proton-based RT is used increasingly in clinical practice to treat HNC due to its potential to further reduce the dose delivered to healthy normal tissues and OARs compared to photon-based treatments. Proton particles have a particular physical property called "Bragg peak"[4] where they deposit the maximum amount of their energy in the end of their "pathway" to the anatomical tumour target with a very low radiation dose after this peak. Therefore, PT techniques such as the Intensity-Modulated Proton Therapy (IMPT) can potentially benefit HNC patients treated for palliative or curative purposes[5].

This benefit is mainly based on the improvement of the local tumour control and the dose sparing of the normal tissues reducing early and late toxicities after PT such as dysphagia. This is one of the main RT-induced complications in HNC patients and is characterised by difficulty in the swallowing process in the bolus from the oral cavity to the stomach. Dysphagia can greatly reduce quality of life and cause other late RT induced late effects such as nutritional implications and feeding tube dependence[6].

In the Netherlands it is estimated that approximately 2000 HNC patients are treated with RT annually[7]. The first Proton therapy (PT) treatments took place in the Netherlands in 2018 and currently more than 2000 patients have been treated collectively by the three PT centres of the country (MAASTRO clinic in Maastricht, University Medical Centre of Groningen (UMCG) and the Holland Proton Therapy Centre (PTC) in Delft). Before the start of the first PT treatments, a systematic effort was initiated to develop data-driven selection and qualification of patients that will benefit most from PT called the "model-based approach" (MBA)[8]. The MBA works by using machine learning (ML), a subdivision of artificial intelligence, to compare different normal tissue complication probability (NTCP) profiles between the most optimal photon and proton RT treatment plans. These insights then enable clinicians to select those patients for PT that will have a clinical benefit in terms of reduced radiation-induced toxicity rates after the RT treatment translated in the difference between the proton and photon NTCP profiles estimation (ΔNTCP).

The different dose parameters of the different OARs as well as other clinical variables such as the baseline toxicity scores according to Patient Reported Outcome (PROMs) questionnaires or physician-rated scores (CTC-AE) or the tumour location are included in these NTCP profiles. The first National Indication Protocol for Proton therapy (NIPP) was clinically implemented in University Medical Centre Groningen (UMCG) in the Netherlands in 2019[7]. The NIPP, approved by the National Health Care Institute (Zorginstituut Nederland; ZIN) in the Netherlands, is the official patient selection mechanism for treatment with PT in the Netherlands. As a result, treatment of patients selected by the NIPP MBA is fully reimbursed by Dutch healthcare insurance.

However, to continue the support and ensure accurate selection via the MBA, standardised registration of high quality patient data is required. The ProTRAIT initiative (PROton Therapy ReseArch regIsTry)[9] was initiated in 2019 in the Dutch PT centres. The ProTRAIT goal is to systematically register patients data from different tumour groups including demographic data[10] that can support the MBA. Furthemore, the data are transformed in a FAIR data[11] format in each participating PT centre so that the different NTCP statistical profiles can be validated in a privacy preserving manner using the Personal Health Train (PHT) infrastructure[12]. In this study, we aimed to assess the accuracy and robustness of part of the current NIPP[13] MBA. To this end, we validated the logistic regression-based NTCP model for Dysphagia grade ≥ 2 at 6 months (primary setting) as described by the NIPP[7] using data from photon and proton-based RT treatment plans of patients treated in the PT centre of MAASTRO clinic in the Netherlands.

6.2 Materials/Methods

6.2.1 Developed NTCP model

The NTCP dysphagia logistic regression model as described in the National Protocol for model-based selection for PT by Langendijk et al.[13]. The model's goal is to predict the NTCP values of patient candidates for PT to develop greater than second grade dysphagia six months after the end of their RT treatment. Dysphagia was graded by physicians according to the Common Toxicity Criteria for Adverse Events version 4.0 (CTCAEv4.0). The model consists of six different predictors, i) the mean radiation delivered dose in Gray (Gy) to the oral cavity, ii-iv) the mean radiation delivered dose in Gy to the superior, medium and superior pharyngeal constrictor muscle (PCM), v) the baseline dysphagia score in the start of RT and vi) the tumour location. The model was developed using 813 patients treated with RT in UMCG and it is currently clinically integrated in the Dutch PT centres for the selection of patients for PT according to the MBA[8,13]. The NTCP dysphagia model in the primary setting is described in equation 1.

Equation 1

$$NTCP\ dyspghagia\ in\ six\ months\ = \frac{1}{1 + e^{-LP}}$$

$$LP = \ -4.0536 + Dmean\ Oral\ Cavity * 0.0300 + Dmean\ Superior\ PCM * 0.0236$$

$$Dmean\ Medium\ \ PCM * 0.0095 + Dmean\ Inferior\ PCM * 0.0133$$
$$+\ Baseline\ dysphagia\ score$$

$$+Tumour\ location$$

$$Baseline\ dysphagia\ score = 0.000\ \ for\ dysphagia\ score = 0 - 1$$

$$Baseline\ dysphagia\ score = 0.9382\ \ for\ dysphagia\ score = 2$$

$$Baseline\ dysphagia\ score = 1.2900\ for\ dysphagia\ score\ \ \geq\ 3$$

$$Tumour\ location = 0.000\ \ for\ oral\ cavity\ tumours$$

$$Tumour\ location = -0.6281\ for\ pharynx\ \ tumours$$

$$Tumour\ location = -0.7711\ \ for\ larynx\ \ tumours$$

$$Abbreviations: LP = Linear\ Predictor, Dmean = mean\ dose,$$
$$PCM = \ \ Pharyngeal\ Constrictor\ Muscle$$

6.2.2 External validation cohort

For the external validation of the NTCP logistic regression model described in equation 1, we used an independent dataset of 277 patients treated with primary (chemo-)RT in MAASTRO clinic between 2019 and 2021. The patients were diagnosed with malignancies of the pharynx and larynx and were treated using photon (263 patients) and proton-based (14 patients) RT techniques. The demographic, clinical and OARs dosimetric characteristics are presented in Table 1.

Table 1: Patients' cohort characteristics (n=277) that was used for the validation of the NTCP ≥ 2 grade six months dysphagia model	
Demographic characteristics	
Gender	N (%)
Men	194 (70)
Women	83 (30)
Age groups in years	N (%)
≤ 60 years old	90(32.5)
> 60 years old	187(67.5)
Clinical characteristics	
Clinical T stage 8th edition	N (%)
T1-T2	122(44)
T3-T4	142(51.2)
Tis	2(0.7)
Tx	11(4.1)
Clinical N stage 8th edition	N (%)
≤N2	250(90.2)
≥N3	18(6.5)
Nx	9(3.3)
Clinical M stage 8th edition	N (%)
M0	253(91.3)

M1	2(0.7)
Mx	22(8)
Tumour location	N (%)
Pharynx	188(67.9)
Larynx	89(32.1)
Dosimetric characteristics-predictors of the NTCP model for dysphagia grade ≥ 2 at 6 months (Gy)	
Dmean oral cavity	24.6
Dmean PCM superior	36.2
Dmean PCM medium	41.8
Dmean PCM inferior	37.6
Dysphagia	
Baseline dysphagia CTCAEv4.0	N (%)
< grade 2	233(84.1)
>= grade 2	44(15.9)
Six months dysphagia CTCAEv4.0	N (%)
< grade 2	190(68.5)
>= grade 2	87(31.5)

Abbreviations: NTCP= Normal Tissue Complication Probability, Dmean= Mean radiation dose, PCM= Pharyngeal Constrictor Muscle, CTCAEv4.0=Common Toxicity Criteria for Adverse Events version 4.0

6.2.2 Statistical analysis

We used the closed testing procedure (CTP) as described and implemented by Vergouwe et al.[14] to validate the NTCP model (equation 1) and to examine whether the model needs updating. CTP follows a three levels calibration hierarchy comparing the updated calibrated models against the original using likelihood ratio tests by testing the statistical significance of them (ie. p value <0.05) . Following the CTP methodology, we examined four different logistic regression NTCP models. The first one included the calculation of the NTCP values according to the original ≥2 grade dysphagia model. For the second model, a new intercept was estimated for the NTCP model of the equation 1 after setting its coefficient equal to 1 (**“**Recalibration in the large**”**). For the third model, a new updated coefficient of the original NTCP model’s linear predictor was estimated (ie. slope) as well as with the intercept of the model (“Logistic Recalibration”). For the fourth model, we used the complete set of predictor variables used in the original NTCP model, to estimate their respective coefficients (“Model revision/update”). Table 2 presents all models described above and includes the names used to describe the models in the remainder of the manuscript. The code used to execute these four aforementioned models was written in the open-source statistical analysis software tool RStudio[15]. The selected final model was chosen according to the CTP function of Vergouwe et al.[14]**.** The RStudio-based[15] code used for the CTP implementation is publicly available in the Github repository (ProTRAIT/CTP_dysphagia_NTCP.R at main · MaastrichtU-CDS/ProTRAIT (github.com))

Table 2:Definition of the different models according to the closed testing procedure (CTP)

Model name	Definition	Estimated parameters
Model 0	Original NTCP ≥2 grade dysphagia model.	No parameters
Model 1	Recalibration in the large	Intercept
Model 2	Logistic Recalibration	Intercept and slope
Model 3	Model revision/update	Logistic regression coefficients

6.2.3 Model performance

For model performance, Brier Scores (scale 0 to 1, with the lower values indicating a higher accuracy of the model) were calculated to evaluate the four different models as suggested by Steyeberg et al.[16]. Moreover, we performed a graphical assessment of the calibration of the four different models of the CTP to evaluate the correctness of the predicted compared to the observed probabilities of the NTCP values. The four different models were graphically assessed using the maximum and average difference between the predicted and calibrated probabilities (Emax and Eavg). For the creation of the calibration curves we used the function "calPlot2" from the RStudio[15] package "ModelGood"[17]. For the discrimination evaluation of the four different models, the sensitivity, specificity and the area under the receiver-operating characteristic curve (AUC) were calculated.

6.3 Results

6.3.1 Model performance

A summary of the validation performance with the independent MAASTRO cohort of the four different models developed according to the CTP methodology as well as the different calibration metrics are listed in table 3. The original ≥2 grade dysphagia model (model 0) included in the NIPP[13] presented acceptable discrimination in the validation dataset (AUC=0.80, sensitivity=0.71, specificity=1, operating point=0.50) while the revised model with new updated coefficients (model 3) presented excellent discrimination (AUC=0.83, sensitivity=0.80, specificity=0.67, operating point=0.50). The ROC curves of the original (model 0) and revised (model 3) NTCP model for grade dysphagia are both presented in figure 1. The Brier scores also indicated that the accuracy of model 0 was not as high as the other calibrated models in the validation cohort. The lowest Brier score was reported for model 3 and therefore the highest accuracy was observed. Furthermore, model 0 presented the highest difference between the predicted and calibrated probabilities according to the average absolute difference in predicted and calibrated probabilities (Eavg). Therefore, according to the CTP, a model update was needed for the independent validation model 0 in MAASTRO's head and neck patients cohort as it was underestimating the risk of dysphagia ≧2nd grade at six months after the end of RT (figure 2). The CTP selected model 3 as the ideal updated model. The revised equation belonging to this model (model 3) is displayed in equation 2.

Table 3: Performance of the of the original NTCP and the calibrated models in MAASTRO's patients cohort (n=386)

Models	**Original NTCP model (model 0)**	**Re-calibration in the large (model 1)**	**Logistic recalibration (model 2)**	***Model revision/update (model 3)***
Performance measure	**Discrimination**			
AUC (95% CI)	0.80(0.75-0.85)	0.80(0.75-0.85	0.80(0.75-0.85	**0.83(0.78-0.88)**
Sensitivity	0.71	0.76	0.78	**0.80**
Specificity	1	0.66	0.63	**0.67**
Calibration evaluation	**Calibration**			
Calibration intercept	0	1.11	1.41	-
Calibration slope	1	1	1.18	-
Brier	0.20	0.16	0.16	**0.15**
Emax	0.30	0.06	0.08	**0.12**
Eavg	0.16	0.02	0.02	**0.03**
E90	0.27	0.04	0.03	**0.06**

Abbreviations: 95% CI:confidence interval with a 95% confidence level, AUC:the area under the receiver-operating characteristic curve, Brier: Brier score (average squared difference in predicted and actual probabilities), Emax/E90/Eavg: Maximum/90th quantile, average absolute difference in predicted and calibrated probabilities.

Equation 2

$$NTCP\ dyspghagia\ in\ six\ months\ = \frac{1}{1+e^{-LP}}$$

$$LP = \ -6.9939 + Dmean\ Oral\ Cavity * 0.0141 + Dmean\ Superior\ PCM * 0.0649 +$$

$$Dmean\ Medium\ \ PCM * (-0.0118) \quad + Dmean\ Inferior\ PCM * 0.0052$$

$$+ Baseline\ dysphagia\ score * 2.1734 + Tumour\ location * (-4.7274)$$

$$Baseline\ dysphagia\ score = 0.000\ \ for\ dysphagia\ score = 0 - 1$$

$$Baseline\ dysphagia\ score = 0.9382\ \ for\ dysphagia\ score = 2$$

$$Baseline\ dysphagia\ score = 1.2900\ for\ dysphagia\ score\ \ \geq\ 3$$

$$Tumour\ location = 0.000\ \ for\ oral\ cavity\ tumours$$

$$Tumour\ location = -0.6281\ for\ pharynx\ \ tumours$$

$$Tumour\ location = -0.7711\ \ for\ larynx\ \ tumours$$

$$Abbreviations: LP = Linear\ Predictor, Dmean = mean\ dose,$$
$$PCM = \ \ Pharyngeal\ Constrictor\ Muscle$$

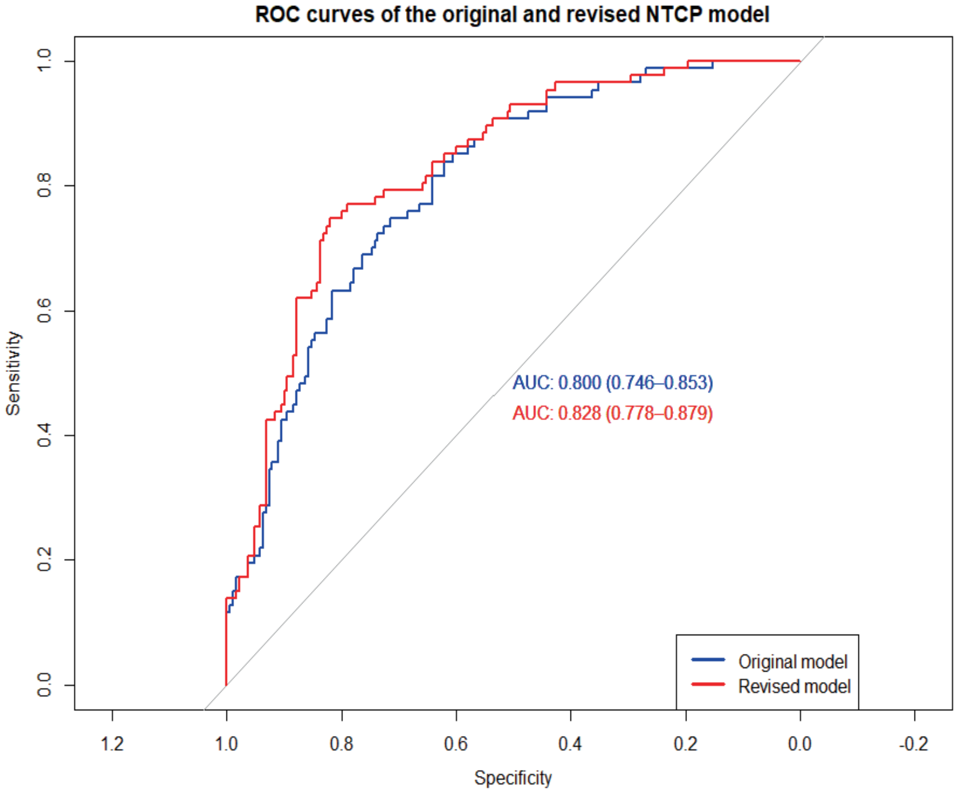


Figure 1: ROC curves of the original NTCP ≥ 2 grade six months dysphagia model(model 0)[13] and the revised model (equation 2, model 3) as selected by the CTP, showing good discriminative performance in MAASTRO's cohort as indicated by the AUC values (>0.75).

Abbreviations: ROC = receiver operating characteristic; NTCP = normal-tissue complication probability; AUC = area under the curve.

The four different levels of calibration of the i) original (model 0) ii) recalibrated in the large (model 1), iii)logistic recalibrated (model 2) and iv) revised models (model 3) can be visually assessed in the calibration plots presented in figure 2. The figure shows that model 0 underestimated the risk of dysphagia ≧2nd grade in the time-point of six months after the end of the RT treatment. Furthemore, the three calibration levels of models 1,2 and 3 significantly improved the agreement between the predicted and observed NTCP risks according to figure 2. The individual calibration curves for each calibrated NTCP ≥ 2 grade dysphagia model including the non-parametric estimate of the calibration relationship between the actual and predicted NTCP values can be found in the supplementary material.

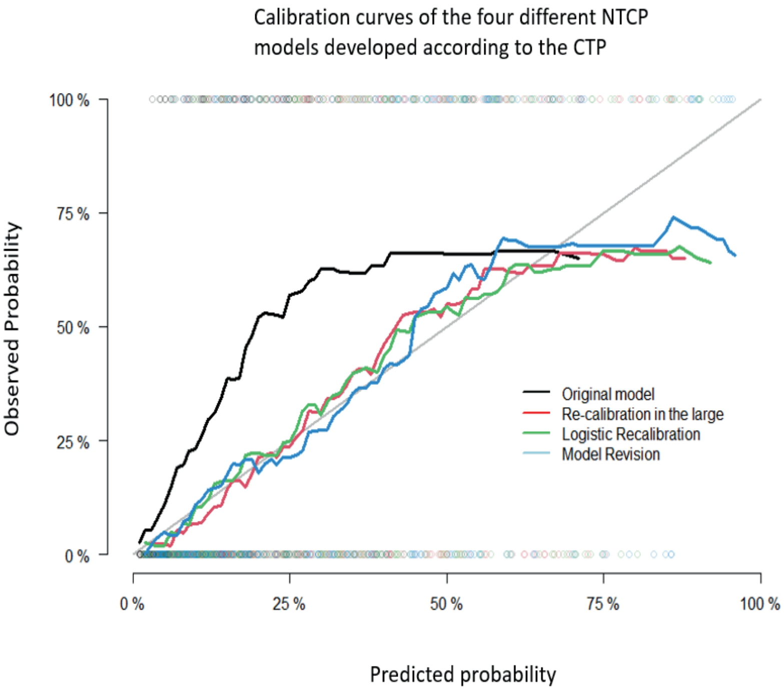


Figure 2: Calibration curves of the different NTCP ≥ 2 grade six months dysphagia models as indicated by the CTP (i) original NTCP model, ii) Re-calibration in the large, iii) Logistic recalibration iv) Model revision.

Abbreviations: NTCP = normal-tissue complication probability; CTP= Closed Testing Procedure

6.4 Discussion

The innovative methodology of the MBA[8] constitutes the main clinically integrated tool in the Netherlands for the selection of patients for PT. Multiple logistic regression NTCP models that predict the radiation-induced toxicity rates of xerostomia and dysphagia for head and neck cancer patients are included. However, it is highly important to externally validate the above-mentioned NTCP models using independent patients' cohorts from external centres and ensure that they can be transferable in other head and neck patients' cohorts[16,18]. Several factors of model transferability and reproducibility can be taken into consideration for external validation studies such as geographical location (location of the hospital/patients) or methodological (RT treatment protocol used) transferability. To account for all these factors we performed an external and independent validation of the NTCP logistic regression model for grade ≥ 2 dysphagia withiin six months after the end of RT which is integrated in the NIPP[13], in an independent head and neck patients cohort of MAASTRO clinic treated with RT.

In this study we followed the methodology of the CTP described by Vergouwe et al.[14] to evaluate the performance of the model. This approach includes three different levels of model recalibration (adjustment of the NTCP model coefficients for the correction of the predicted event probabilities/rates). In the original NIPP[13] model (model 0) we observed moderate calibration as the model underestimated the risk dysphagia and its discrimination of the original model in MAASTRO's cohort was satisfactory (AUC=0.80, sensitivity=0.71 and specificity=1.00). After implementing the "Recalibration in the large" (re-estimation of the intercept of the original NTCP model (model 0)) in the NTCP ≥ 2 grade dysphagia model, the calibration curve improved (figure 2). The calibration curves also graphically improved after implementing the other two calibration levels of the CTP procedure (recalibration (model 2) and model revision (model 3)).

The ideal scenario in the case of the external validation of a prediction model in an independent cohort includes its high performance in terms of statistical metrics such as sensitivity, specificity and the area under the ROC curve. According to Van Calster et al.[19] this high performance can be in other words called "strong calibration" and implies that a model is totally correct in the validation dataset. However, according to the same study, the "strong calibration" can be unrealistic in real word data. Therefore, the external validation of NTCP models in independent cohorts may require a specific adjustment/update mechanism that takes into account the different factors that make the external validation of NTCP models unsuccessful [20,21].

The CTP recommends a revised version (model 3) of the original ≥ 2 grade dysphagia NTCP model as the ideal update method. Despite our initial goal to externally validate the NTCP using an independent patients' cohort by assessing its transferability, there are some discrepancies between the methods used in this study to assess the performance of the original NTCP dysphagia model and the methodologies proposed from other studies[19,22]. Therefore some limitations should be taken into account. First, as stated by the NIPP publication[13] **,** in the validation datasets of the original NTCP model multiple imputation was performed for the computation of missing values of the logistic regression model predictors. In our case, we included only complete cases of and did not perform any imputation method to account for missing values. This is possibly one of the reasons that model 0 was not selected by the CTP and its performance was not as high as model 3 which was selected by the CTP. Secondly, according to Van Calster et al.[19] it is recommended that at least 200 events and 200 non-events are required for the development of flexible calibration curves. In our dataset consisting of 277 patients we included 87 patients who developed ≥ 2 grade dysphagia (events) in six months after RT and 190 patients who did not (non-events) for creating and assessing graphically the calibration plots of the different levels of calibrations according to the CTP. Moreover, according to Van de Bosch et al.[22] an external validation of the updated model is recommended in the case of a selection of the revised model by the CTP. In our case, the model selected by the CTP (model 3) was not validated by another external and independent dataset and so are prone to overfitting and overoptimistic performance. The aforementioned reported limitations of our study have to be taken into account in the case of a potential independent validation of the revised model by other external centres. Therefore, we encourage the independent external validation by other RT institutions (inter)nationally of the revised model selected by the CTP for its transferability and generalisability assessment.

Another factor that can influence the performance of a NTCP model containing dosimetric predictor OARs variables is the delineation method used for the OARs contours. In our study we included patients with manual OARs delineations for the dosimetric OARs NTCP predictor variables. The last few years, several studies proposed the implementation of AI-based techniques for the automation of the delineation procedure for head and neck cancer patients[23,24]. For instance, Deep learning (DL)-based delineations have the potential to establish a standardised delineations framework for head and neck patients, decreasing the clinical burden[25]. Interobserver variability among different clinicians for head and neck patients is a common phenomenon [26] that can impact the quality of dosimetric data included in a prediction model and therefore the performance of it in different independent patients' cohorts. DL assisted delineation is expected to reduce this variability and is recommended for future studies.

The CTP was adopted in other European patient cohorts such as the Danish Head and Neck Cancer Group (DAHANCA) implementing a similar model to the Dutch NIPP based dysphagia model with the inclusion of different OARs dosimetric variables in the NTCP model such as the supraglottic larynx[27]. In the Danish cohort, the NTCP dysphagia model by Christianen et al.[28] was externally validated using 588 patients of the DAHANCA 19 trial. The CTP selected in the Danish case the "recalibration in the large" (re-estimation of the intercept) as the best updated model to the external validation Danish dataset used implementing a a five-fold validation type of the CTP, which was not the case in our study.

The implementation of AI-based techniques in the treatment planning procedure may not only reduce the interobserver variability in the delineation procedure for example, but it can potentially reduce the time and resources needed for it. For instance, the MBA is based on the continuous RT treatment planning comparisons of patients candidates for PT. AI-based automated procedures such as the automatic OARs delineation or the radiation dose optimisation[24,29] can significantly contribute to the acceleration of the MBA and evaluation of more patients based on it. The first experience with the MBA in UMCG for head and neck patients as published by Tambas et al.[30] indicated also the need for the inclusion of AI techniques in the treatment planning procedure and therefore the planning comparison for the MBA, that can potentially reduce the resources needed for it.

As a next step, we aim to implement federated learning techniques adhering to the FAIR principles[11] for the privacy-preserving external validation of the NIPP NTCP ≥ 2 grade dysphagia model[13] in the Dutch PT centres. Using the Personal Health Train (PHT) infrastructure [12] we aim to exchange statistical algorithms that can use the CTP approach in a privacy-preserving manner (ie. without the exchange of patients data). With this federated approach we aim to include larger patient cohorts for the development and validation of the NIPP[13] based NTCP models including patients who are treated with different RT treatment protocols for head and neck cancer.

6.5 Conclusion

In conclusion, with this study we performed an independent validation of the NTCP ≥ 2 grade dysphagia model (primary setting) which is used in the Netherlands for the selection of patients for PT according to the NIPP[13]. We concluded that the performance of the model in the independent and external MAASTRO patient cohort was good. There was still room for improvement, however, as the distribution of the observed compared to the predicted probabilities of the model according to the calibration plot generated was not ideal. Following the CTP methodology, it was indicated that the model should be updated and calibrated. We therefore, based on the CTP, selected the revised version of the original model with updated intercept and predictor coefficients for further development. The revised version of the model had a high discrimination in the internal validation MAASTRO cohort, but an additional external and independent validation from other RT centres is needed to further evaluate its robustness and transferability.

6.6 Supplementary material

For the creation of the calibration curves of the different models presented in table 2 of the main text of the study we used the package “val.prob” of the RStudio library “rms”[1]. In the different figures, the dashed curve represents the non-parametic estimate of the calibration probabilities between the predicted and observed/actual values. The grey diagonal line presents the ideal probability distribution (intercept=0 and slope=1).

Table 2:Definition of the different models according to the closed testing procedure (CTP)

Model	Definition	Estimated parameters
Model 0	Original NTCP ≥2 grade dysphagia model.	No parameters
Model 1	Recalibration in the large	Intercept
Model 2	Recalibration	Intercept and slope
Model 3	Model revision	Logistic regression coefficients

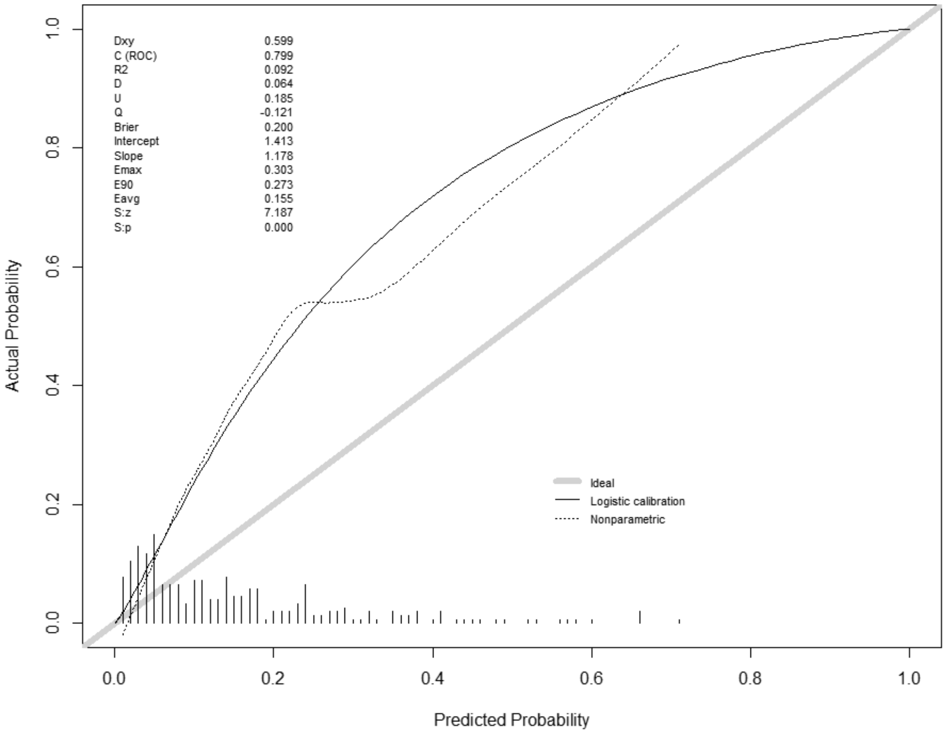


Figure S1: The calibration curve of the original NIPP[2] NTCP grade ≥ 2 dysphagia model (model 0) . As it is shown according to the distribution of the predicted and actual probabilities, model 0 underestimates the risk of the head and neck MAASTRO patients (N=277) to develop grade ≥ 2 dysphagia six months after the end of the RT (calibration curve above the diagonal line). **Abbreviations:** Dxy: Somer's rank correlation, C(ROC): Area Under the Curve for discrimination assessment, R2: Nagelkerke-Cox-Snell-Maddala-Magee R-squared index, D: discrimination index, U: unreliability index, Q: quality index, Brier: Brier score (average squared difference in predicted and actual probabilities), Emax/E90/Eavg: Maximum/90th quantile, average absolute difference in predicted and smoothed calibrated probabilities, S:z/S:p the z and two sided p-value of the Spiegelhalter test for calibration accuracy.

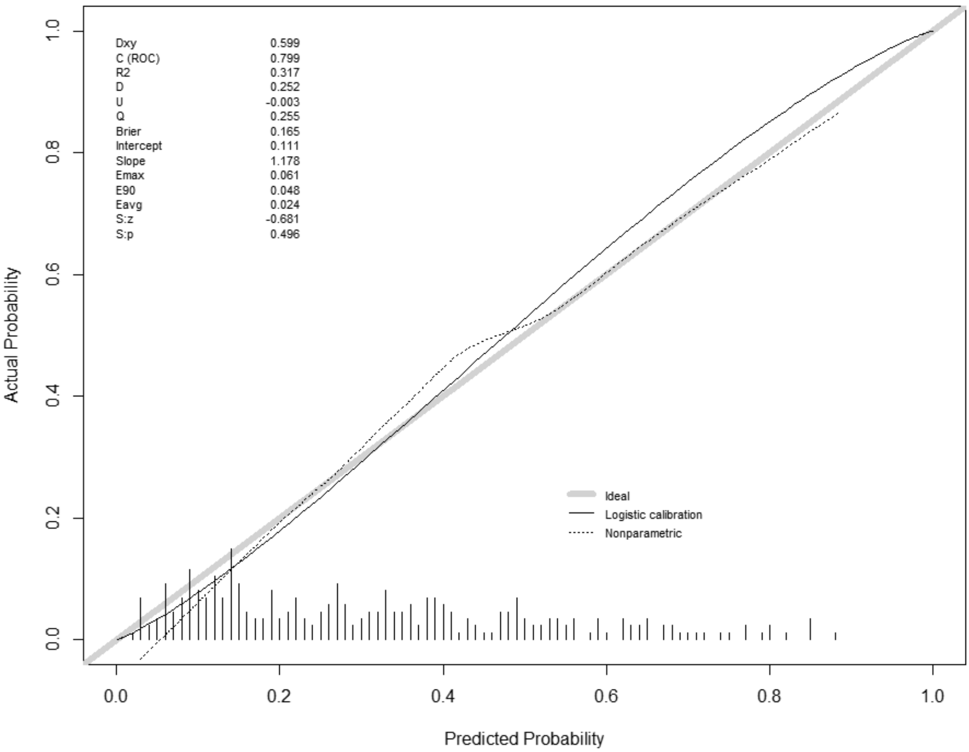


Figure S2: The calibration curve of model 1 (estimation of an updated intercept of model 0) . As it is shown, there is an improvement of the distribution of the predicted and actual probabilities indicated by the brier score and the average absolute difference in predicted and smoothed calibrated probabilities (Eavg). **Abbreviations:** Dxy: Somer's rank correlation, C(ROC): Area Under the Curve for discrimination assessment, R2: Nagelkerke-Cox-Snell-Maddala-Magee R-squared index, D: discrimination index, U: unreliability index, Q: quality index, Brier: Brier score (average squared difference in predicted and actual probabilities), Emax/E90/Eavg: Maximum/90th quantile, average absolute difference in predicted and smoothed calibrated probabilities, S:z/S:p the z and two sided p-value of the Spiegelhalter test for calibration accuracy.

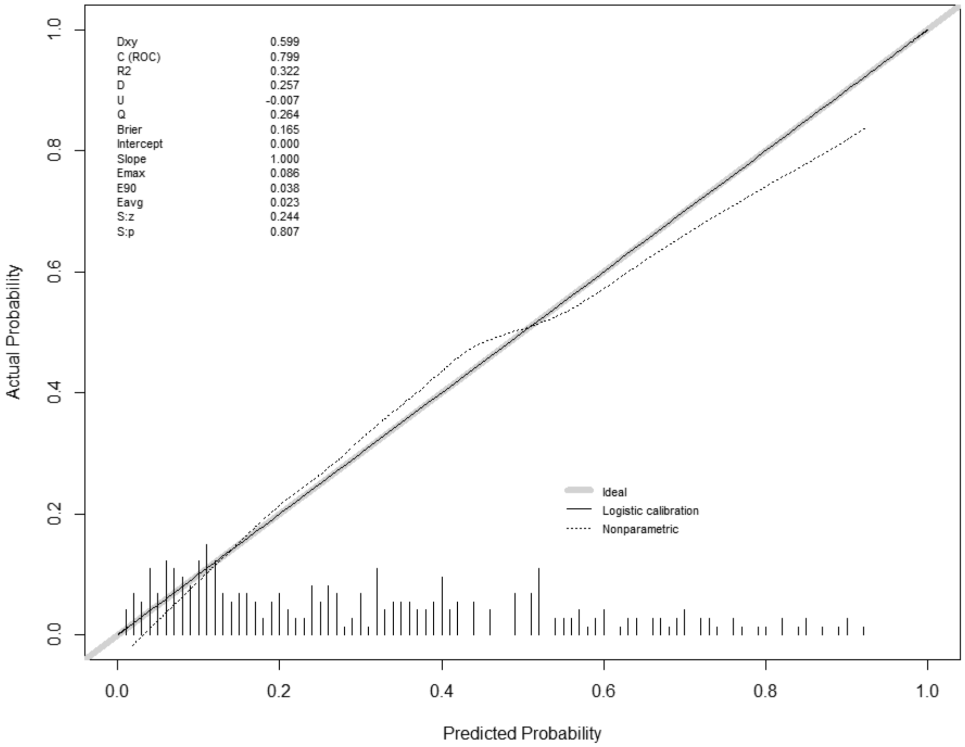


Figure S3: The calibration curve of model 2 (estimation of an updated intercept and slope of model 0) . As it is shown, there is an improvement of the distribution of the predicted and actual probabilities indicated by the brier score and the average absolute difference in predicted and smoothed calibrated probabilities (Eavg) **Abbreviations:** Dxy: Somer's rank correlation, C(ROC): Area Under the Curve for discrimination assessment, R2: Nagelkerke-Cox-Snell-Maddala-Magee R-squared index, D: discrimination index, U: unreliability index, Q: quality index, Brier: Brier score (average squared difference in predicted and actual probabilities), Emax/E90/Eavg: Maximum/90th quantile, average absolute difference in predicted and smoothed calibrated probabilities, S:z/S:p the z and two sided p-value of the Spiegelhalter test for calibration accuracy.

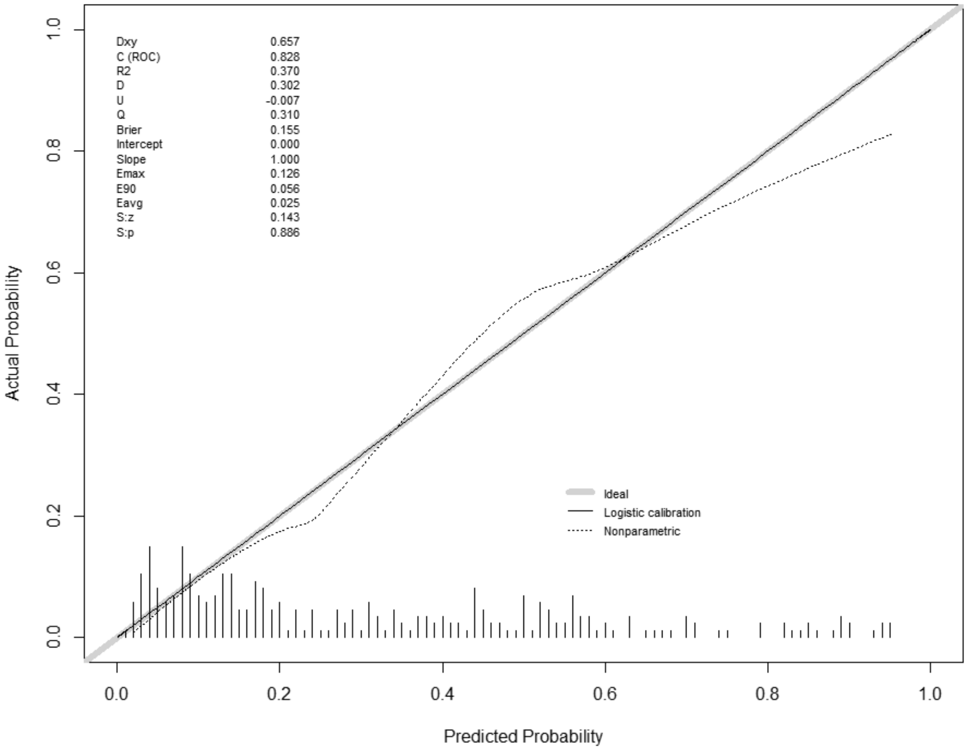


Figure S4: The calibration curve of model 3 (estimation of an updated intercept and slope of model 0) . As it is shown, there is an improvement of the distribution of the predicted and actual probabilities indicated by the brier score and the C(ROC). **Abbreviations:** Dxy: Somer's rank correlation, C(ROC): Area Under the Curve for discrimination assessment, R2: Nagelkerke-Cox-Snell-Maddala-Magee R-squared index, D: discrimination index, U: unreliability index, Q: quality index, Brier: Brier score (average squared difference in predicted and actual probabilities), Emax/E90/Eavg: Maximum/90th quantile, average absolute difference in predicted and smoothed calibrated probabilities, S:z/S:p the z and two sided p-value of the Spiegelhalter test for calibration accuracy.

Bibliography

1. Ragin CCR, Modugno F, Gollin SM. The Epidemiology and Risk Factors of Head and Neck Cancer: a Focus on Human Papillomavirus. *J Dent Res*. 2007;86(2):104-114. doi:10.1177/154405910708600202
2. Gatta G, Botta L, Sánchez MJ, et al. Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: The EUROCARE-5 population-based study. *European Journal of Cancer*. 2015;51(15):2130-2143. doi:10.1016/j.ejca.2015.07.043
3. Langendijk JA, Doornaert P, Verdonck-de Leeuw IM, Leemans CR, Aaronson NK, Slotman BJ. Impact of Late Treatment-Related Toxicity on Quality of Life Among Patients With Head and Neck Cancer Treated With Radiotherapy. *JCO*. 2008;26(22):3770-3776. doi:10.1200/JCO.2007.14.6647
4. *National Indication Protocol for Proton Therapy in the Netherlands*. https://nvro.nl/images/documenten/rapporten/2019-08-15__Landelijk_Indicatie-protocol_Protonentherapie_Hoofdhals_v2.2.pdf
5. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. *Radiotherapy and Oncology*. 2013;107(3):267-273. doi:10.1016/j.radonc.2013.05.007
6. Wilson RR. Radiological Use of Fast Protons. *Radiology*. 1946;47(5):487-491. doi:10.1148/47.5.487
7. Moreno AC, Frank SJ, Garden AS, et al. Intensity modulated proton therapy (IMPT) – The future of IMRT for head and neck cancer. *Oral Oncology*. 2019;88:66-74. doi:10.1016/j.oraloncology.2018.11.015
8. Murphy BA, Gilbert J. Dysphagia in Head and Neck Cancer Patients Treated With Radiation: Assessment, Sequelae, and Rehabilitation. *Seminars in Radiation Oncology*. 2009;19(1):35-42. doi:10.1016/j.semradonc.2008.09.007
9. *ProTRAIT (PROton Therapy ReseArch RegIsTry)*. www.protrait.nl
10. Sloep M, Kalendralis P, Choudhury A, et al. A knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry. *Clinical and Translational Radiation Oncology*. 2021;31:93-96. doi:10.1016/j.ctro.2021.10.001
11. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3:160018. doi:10.1038/sdata.2016.18
12. Beyan O, Choudhury A, van Soest J, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. *Data Intellegence*. 2020;2(1-2):96-107. doi:10.1162/dint_a_00032
13. Langendijk JA, Hoebers FJP, de Jong MA, et al. National Protocol for Model-Based Selection for Proton Therapy in Head and Neck Cancer. *International Journal of Particle Therapy*. 2021;8(1):354-365. doi:10.14338/IJPT-20-00089.1
14. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2017;36(28):4529-4539. doi:10.1002/sim.7179

15. *RStudio*. https://www.rstudio.com/
16. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
17. ModelGood: Validation of risk prediction models.
18. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*. 2015;68(3):279-289. doi:10.1016/j.jclinepi.2014.06.018
19. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005
20. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statist Med*. 2004;23(16):2567-2586. doi:10.1002/sim.1844
21. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. *Can J Anesth/J Can Anesth*. 2009;56(3):194-201. doi:10.1007/s12630-009-9041-x
22. Van den Bosch L, Schuit E, van der Laan HP, et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. *Radiotherapy and Oncology*. 2020;148:151-156. doi:10.1016/j.radonc.2020.04.012
23. van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology*. 2019;138:68-74. doi:10.1016/j.radonc.2019.05.010
24. Dai X, Lei Y, Wang T, et al. Automated delineation of head and neck organs at risk using synthetic MRI-aided mask scoring regional convolutional neural network. *Med Phys*. 2021;48(10):5862-5873. doi:10.1002/mp.15146
25. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. *Radiotherapy and Oncology*. 2020;142:115-123. doi:10.1016/j.radonc.2019.09.022
26. van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiat Oncol*. 2021;16(1):120. doi:10.1186/s13014-020-01677-2
27. Hansen CR, Friborg J, Jensen K, et al. NTCP model validation method for DAHANCA patient selection of protons versus photons in head and neck cancer radiotherapy. *Acta Oncologica*. 2019;58(10):1410-1415. doi:10.1080/0284186X.2019.1654129
28. Christianen MEMC, Schilstra C, Beetz I, et al. Predictive modelling for swallowing dysfunction after primary (chemo)radiation: Results of a prospective observational study. *Radiotherapy and Oncology*. 2012;105(1):107-114. doi:10.1016/j.radonc.2011.08.009
29. Kierkels RGJ, Fredriksson A, Both S, Langendijk JA, Scandurra D, Korevaar EW. Automated Robust Proton Planning Using Dose-Volume Histogram-Based Mimicking

of the Photon Reference Dose and Reducing Organ at Risk Dose Optimization. *International Journal of Radiation Oncology*Biology*Physics*. 2019;103(1):251-258. doi:10.1016/j.ijrobp.2018.08.023

30. Tambas M, Steenbakkers RJHM, van der Laan HP, et al. First experience with model-based selection of head and neck cancer patients for proton therapy. *Radiotherapy and Oncology*. 2020;151:206-213. doi:10.1016/j.radonc.2020.07.056

Chapter 7

External validation of a Bayesian network for error detection in radiotherapy plans

Adopted from: “External validation of a Bayesian network for error detection in radiotherapy plans”

Petros Kalendralis, Denis Eyssen, Richard Canters, Samuel M.H. Luk, Alan M. Kalet, Wouter van Elmpt, Rianne Fijten, Andre Dekker, Catharina M.L. Zegers, Inigo Bermejo

IEEE Transactions on Radiation and Plasma Medical Sciences (2021), https://doi.org/10.1109/TRPMS.2021.3070656,

Contribution: First authorship

Abstract

Artificial intelligence (AI) applications have recently been proposed to detect errors in radiotherapy plans. External validation of such systems is essential to assess their performance and safety before applying them to clinical practice. We collected data from 5,238 patients treated at Maastro Clinic, and introduced a range of common radiotherapy plan errors for the model to detect. We estimated the model's discrimination calculating the area under the receiver-operating characteristic curve (AUC). We also assessed its clinical usefulness as an alert system that could reduce the need for manual checks by calculating the percentage of values flagged as errors and the positive predictive value (PPV) for a range of high sensitivities (95% to 99%) and error prevalence. The AUC when considering all variables was 67.8% (95% CI, 65.6%-69.9%). The AUC varied widely for different types of errors (from 90.4% for table angle errors to 54.5% for Planning Tumor Volume-PTV dose errors). The percentage of flagged values ranged from 84% to 90% for sensitivities between 95% and 99% and the PPV was only slightly higher than the prevalence of the errors. The model's performance in the external validation was significantly worse than that in its original setting (AUC of 68% versus 89%). Its usefulness as alert system to reduce the need for manual checks is questionable due to the low PPV and high percentage of values flagged as potential errors to achieve a high sensitivity. We analyzed the apparent limitations of the model and we proposed actions to overcome them.

Keywords—Artificial intelligence, Bayesian network, Radiotherapy, Treatment planning

7.1 Introduction

Over the past decades, radiotherapy has constituted a fundamental treatment modality for cancer patients along with other treatment options such as surgery, chemotherapy and immunotherapy[1]. Radiotherapy's cost-effectiveness (5% of the total cost of oncological care)[2] as well as the number of patients that are treated with it (50% of cancer patients)[1,3] and its potentially curative nature[4], stress the need for an accurate treatment plan construction and delivery. Recent advancements in the field of artificial intelligence (AI) have contributed to a significant progress regarding the automation of the treatment planning process such as the automatic delineation of the clinical target volumes (CTV) or organs at risk (OAR)[5-6] and the automatic dosimetric evaluation of treatment planning[7].

Radiotherapy treatment planning is a complex procedure that requires a coordinated team effort by an interdisciplinary group that consists of radiation oncologists, medical physicists, radiation technologists and dosimetrists. The objective of radiotherapy treatment planning is to safely and efficiently prescribe the optimal dose to the anatomical target volume of the patients. Mistakes made during this process can cause serious risks during the treatment planning execution. In the past, several organizations such as the world health organization (WHO), the American association of physicists in Medicine (AAPM) and the European society for therapeutic radiation oncology (ESTRO) have published recommendation guidelines for the elimination of the radiotherapy errors[8-10]. Generally, the radiotherapy treatment plan errors can be subdivided into operational or system errors. For instance, malfunction of the multileaf collimators (MLCs) system of the linear accelerator (LINAC) in a case of intensity modulated radiation therapy (IMRT) or differences between the prescribed dose and the dose per radiotherapy fraction due to adjustments of the reference points are some of the potential errors. These errors can lead to serious accidents with extremely severe consequences for both patients and clinical professionals[11-12].

Increased automation, supported by AI techniques and combined with human expertise, could reduce the time needed for the development and execution of a radiotherapy treatment plan. Furthermore, the implementation of AI methods can potentially contribute to the early detection of plan errors and the reduction of the time needed for their detection[13-14].

Currently, we are entering a new challenging and promising era in radiotherapy where AI has started to manifest its potential with several applications. For example, several studies introduced automated treatment plan verification for the detection or errors during radiotherapy[15-19]. Moreover, with the development of the automated pipelines for the validation and quality assurance (QA) of the radiotherapy plans, objections raised regarding their accuracy and implementation such as the requirement of expertise knowledge of the manual planning (i.e. human intervention) and reproducibility issues[20].

To address these limitations, Luk *et al.*[21] proposed a model to detect radiotherapy errors using an AI-based approach. Their Bayesian network (BN) model can flag anomalies in 29 variables related to diagnostic, prescription, plan and setup level parameters to assist clinical physicists and clinicians on the time-consuming and error-prone radiotherapy treatment planning procedure.

Bayesian networks (BNs) are the most popular type of probabilistic graphical models (PGMs), which emerged during the 1980s and rose to prominence in the next decade[22]. PGMs use graphs to represent the probabilistic dependencies between the variables in a model. Bayesian networks, for example, use directed acyclic graphs (DAGs) where each variable is represented by a node and links between variables imply causality. In addition, the conditional probability distribution (CPD) of each variable is defined as a function of its parents in the graph (i.e. the set of nodes that have links pointing at one particular node). The structure of the graph of the BN and the CPDs can be either defined based on expert knowledge or learnt from data using machine learning algorithms[23]. Probabilistic reasoning in BNs allows for different types of queries, such as the probability distribution of one or more target variables given a set of findings (e.g. what is the probability of rain given the grass is wet), or the probability of a set of findings (e.g. what is the probability of rain and dry grass). A set of such findings is referred to as evidence. The intuitiveness of the probabilistic reasoning in BNs thanks to their graphical structure in contrast to black box algorithms prominent in AI has led to a wide adoption in healthcare[24].

Luk *et al.*[21] defined the DAG based on expert knowledge and learnt the CPDs based on historical data from their institution. Consequently, they showed that they could detect anomalies in radiotherapy plans assigning the values of a given radiotherapy plan to the variables of the BN and calculating the probability of the evidence, because radiotherapy plans with errors will generally result in a lower probability.

We hypothesized that such a model is clinically relevant and can provide significant added value, reducing the need for manual checks and detecting errors that would otherwise could go unnoticed. An external validation is an empirical evaluation in a dataset that was not used to develop the model and they are essential before considering whether to use a clinical prediction model[25]. Therefore, we performed an external validation of the model using data from Maastro clinic (The Netherlands), with the aim to assess the generalizability of the model.

7.2 Materials and methods

7.2.1 Data acquisition

We used data from 5238 patients (19054 treatment plans) for this study, collected at the Maastro radiation oncology clinic (Maastricht, The Netherlands) between 2012 and 2020. The patients were treated with external beam radiotherapy using electrons or photons with IMRT and volumetric modulated arc therapy (VMAT) in seven different Truebeam LINACs of Varian medical systems. Patients treated with protons were excluded from the dataset as the original model by Luk *et al.*[21] did not include them. The radiotherapy elements were extracted and collected from the Varian Eclipse (version 11 and 15) treatment planning system database and amended with information from the electronic patient dossier (EPD). A description of all the variables used as well as with representative examples can be found in table 1.

<table>
<tr><td colspan="4">Table 1:Description and examples of the Maastro clinic’s variables used</td></tr>
<tr><td colspan="4"><b>Diagnostic variables</b></td></tr>
<tr><td><b>Variable name</b></td><td><b>Description</b></td><td><b>States number</b></td><td><b>Examples</b></td></tr>
<tr><td>Diagnose</td><td>Anatomic tumor location</td><td>226</td><td>”prostaat”, ”long”</td></tr>
<tr><td>cT</td><td>Clinical T stage</td><td>29</td><td>0,1,1a</td></tr>
<tr><td>cN</td><td>Clinical N stage</td><td>17</td><td>1b,1c,2a</td></tr>
<tr><td>cM</td><td>Clinical M stage</td><td>8</td><td>1,1a,1b</td></tr>
<tr><td colspan="4"><b>Prescription variables</b></td></tr>
<tr><td><b>Variable name</b></td><td><b>Description</b></td><td><b>States number</b></td><td><b>Examples</b></td></tr>
<tr><td>Treatment_ Intent</td><td>Treatment Intent</td><td>3</td><td>”Radicaal”, ”Palliatief”</td></tr>
<tr><td>NumberOf Rxs</td><td>Number of prescriptions</td><td>14</td><td>1,2,3</td></tr>
<tr><td>DosePer Fraction</td><td>Dose per fraction</td><td>61</td><td>2.75,2,4</td></tr>
<tr><td>PTVDoseRx</td><td>Total dose</td><td>210</td><td>15,20,48</td></tr>
<tr><td>Total Fractions</td><td>Fractions</td><td>35</td><td>4,5,8</td></tr>
<tr><td>RxRadiationType</td><td>Radiation Type</td><td>7</td><td>6X,10X</td></tr>
<tr><td colspan="4"><b>Plan/Beam variables</b></td></tr>
<tr><td><b>Variable name</b></td><td><b>Description</b></td><td><b>States number</b></td><td><b>Examples</b></td></tr>
<tr><td>Plan Technique</td><td>Planning technique</td><td>5</td><td>”ARC”, ”STATIC”</td></tr>
<tr><td>TableAngle</td><td>Table angle</td><td>50</td><td>0,10,355</td></tr>
<tr><td>NumberOf Beams</td><td>Number of beams</td><td>14</td><td>3,4,5</td></tr>
<tr><td>Wedge</td><td>Wedge position</td><td>1</td><td>0,00%</td></tr>
<tr><td>ControlPoints</td><td>Control points</td><td>714</td><td>6,8,10</td></tr>
<tr><td>SSD</td><td>Source to surface distance</td><td>15478</td><td>70.1,72.2,73.6</td></tr>
<tr><td>Bolus</td><td>Presence/</td><td>4</td><td>N,Y,</td></tr>
</table>

	type of bolus		”MULTI-VALUE”
GantryAngle	Gantry angle	1329	103.8,104,104.1
Collimator Angle	Collimator angle	725	347.5,348.3,354.9
BeamEnergy	Beam energy	6	2,3,4
Setup variables			
Variable name	**Description**	**States number**	**Examples**
Orientation	Patient scan orientation	2	”Head First-Supine”, ”HeadFirst-Decubitus Right”
CouchLat	Lateral couch position	5681	17.1,326.3,-1.70
CouchLong	Longitudional couch position	5743	103.4,108.7,112.9
CouchVert	Vertical couch position	5707	-10.5, -5.4
Tolerance	Setup tolerance table	1	”Console RUIM”

7.2.2 Variable mapping

The numerical variables of the dataset were mapped to the nearest value in the corresponding variable from of Luk *et al.* For the categorical variables, such as anatomic tumor location, we mapped the values from our dataset to the matching values in the corresponding variable. If there was more than one matching value (e.g., the variable T_stage contains the values 1a, 1A, T1a), we selected the one with the highest marginal probability in the model (i.e., the most common occurrence in the original training dataset).

7.2.3 Errors

Reports of errors and near-misses that happened in Maastro clinic (The Netherlands) related to radiotherapy were collected and validated from the Prevention and Recovery Information System for Monitoring and Analysis (PRISMA) Prisma database[26]. After the assessment of the 19054 treatment plans of the 5238 patients we encountered 5 radiotherapy treatment plan errors reported that were checked manually. One of the errors was related to a wrong table angle, two errors were related to an incorrect PTV dose and the remaining two errors were related to the usage of the bolus. Since our goal is to replace or support these manual and time consuming checks with the introduction of BNs, we simulated errors in 3% of the plans following instructions of experts in the area.

These errors can be categorized into four main types: patient positioning, prescription level, LINAC mechanical and general radiotherapy plan errors. The patient positioning error category consisted of the LINAC table rotation errors simulation errors with a values bigger than 10 degrees. In the category of prescription level errors, differences between the prescribed dose to the Planning Tumor Volume (PTV) and the dose per fraction were evaluated. Specifically, we simulated errors with values bigger than 100cGy planned dose to the PTV on VMAT and IMRT plans of 15 and 20 fractions. Errors regarding the LINAC collimator angle were simulated and included into the LINAC mechanical errors. In this category, the simulated errors collimator angle values were increased by 10-15 degrees. Under the category of the generic radiotherapy plan errors, we simulated errors for whether the usage of bolus or not was included. In table 2 you can find different categories and the description of the errors. The selection of the above mentioned simulated errors was based on the reported and manually checked errors of the PRISMA database (Table rotation, Incorrect PTV dose and Bolus usage) and the suggestions of manually checked errors (Collimator angle) from the radiotherapy technologists (RTTs) of Maastro Clinic.

Table 2: Errors simulation overview

Errors category	**Errors description**	**Errors specification**
Patient positioning	Table rotation errors	Table rotation values bigger than 10 degrees from the planned value
Prescription level	Prescribed dose to the PTV is not equal to the fractionation	PTV dose values increased by values bigger than 100cGy for VMAT and IMRT plans
LINAC mechanical	Collimator angle errors	Collimator angle values increased by 10-15 degrees
General radiotherapy plan errors	Bolus usage	Bolus involvement to the plans that bolus was not prescribed and bolus absence to the plans involved the prescription of bolus

7.2.4. Evaluation

We used the Java API (application programming interface) of Hugin Researcher 7.4[27] to load the network provided by the authors and calculate the relevant probabilities. Following the instructions in the original article, for each case we instantiated the variables Anatomic_tumor_loc, T_Stage, M_Stage and N_Stage and Treatment_Intent and calculated the probabilities of the rest of the variables. Each probability P was compared against a threshold T that designated whether that parameter should be flagged as correct or as an error. Setup_Device variables were excluded, since these were not available in our database.

In order to compare the performance of the model reported by its authors with its performance in our dataset, we plotted the receiver operating characteristic curve (ROC) and calculated the area under the curve (AUC), which provides an estimate of the discriminative power of the model. We plotted the ROC and calculated the AUC of the whole dataset (i.e. all variables combined) as well as for each of the variables where we simulated errors: collimator angle, table angle, gantry angle, PTV dose, and bolus. We used the ROC and calculated the AUC and its confidence intervals (CIs) using the R language (version 3.6.1) and the 'classifierplots' package.

We also performed an analysis to assess the usefulness of the model in a clinic as an alert system that helps reduce the need for manual checks. As such, it would be only of added value if it could detect almost all errors (i.e., sensitivity ≥ 95%) with a reasonable positive predictive value (PPV, i.e., the probability that an instance flagged as an error is actually an error). Therefore, we undertook scenario analyses to calculate the model's PPV for different sensitivities and different prevalence of errors (since the PPV depends on how frequently errors occur in clinical practice and the prevalence is unknown). We did not assess calibration because the model's output is not meant to be interpreted as a probability.

The source code of our analysis is available at https://gitlab.com/UM-CDS/projects/ext-val-bn-rt-plan-qa.

7.3 Results

Figure 1shows the ROC curve for all the variables used in the external validation. The model achieved an AUC of 67.8% (95% CI, 65.6% – 69.9%) when considering all variables together.

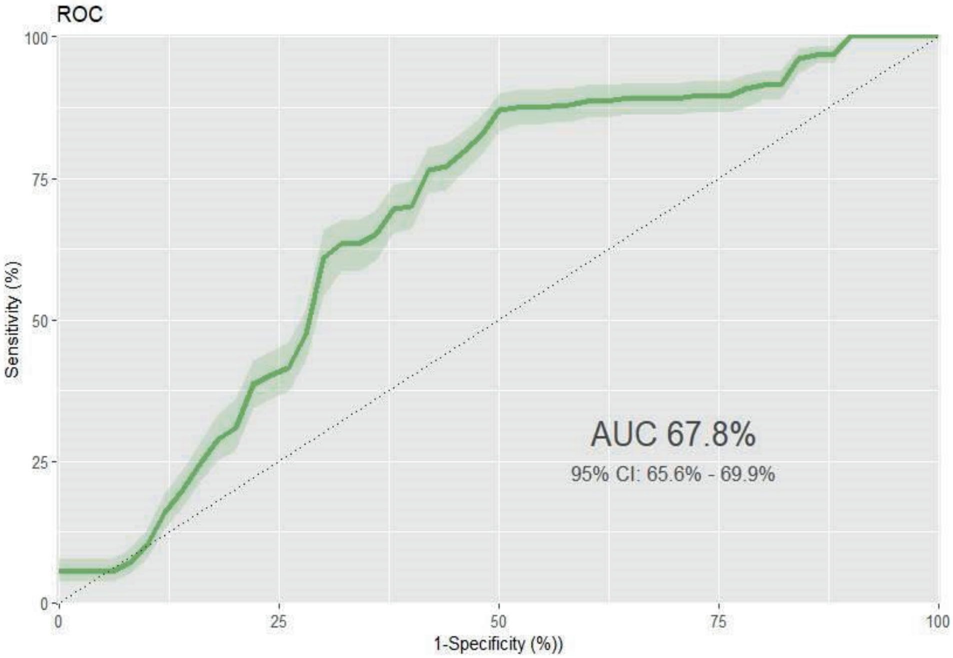


Figure 1: ROC curve for the external validation dataset

Table 3 shows the AUCs for the six types of simulated errors. The discriminative performance of the model is very high for the table rotation errors (“Table Angle” variable) achieving an AUC of 90.4% (95% CI, 87.1%-93.5%). For the category of the simulated errors related to the bolus, gantry angle and the collimator angle, the model performs worse with AUCs of 75.6 % (95% CI, 71.3%-79.9%), 67% (61%-72.7%) and 69.6% (66.3%-73.1%) respectively. However, the Bayesian network fails to detect the errors comprising a difference between the prescribed dose to the PTV and the dose per fraction, resulting in an AUC of 54.5% (49.3%-59.4%).

Table 3: AUCs for different types of errors

Type of error	**Mean**	**95% CI**
Bolus	75.6	71.3 - 79.9
Collimator angle	69.6	66.3 - 73.1
Table angle	90.4	87.1 - 93.5
PTV dose	54.5	49.3 - 59.4
Gantry angle	67.0	61.0 - 72.7
Overall	**67.8**	**65.6 – 69.9**

The results of our analysis regarding the usefulness of the model as an alert system are shown in Table 4, which includes the probability threshold at which different levels of high sensitivities are achieved and the resulting percentage of values flagged as errors and PPVs. According to our analyses, the model would flag as possible errors 84%, 89% and 90% of the values in order to detect 95%, 97% and 99% of errors, respectively. This implies that human technicians would still need to manually review almost all values to check whether they are correct. For these high sensitivity levels, the PPV, or the probability that a value flagged as an error is actually an error, was not significantly higher than the error prevalence itself.

Table 4: Percentage of flagged values and PPV for different combinations of sensitivity and prevalence of errors

Sensitivity	**Percentage of flagged values**	**PPV (%)**		
		Prevalence of errors		
		0.1%	**1%**	**3%**
95%	68%	0.14	1.39	4.15
97%	78%	0.13	1.25	3.73

99%	79%	0.13	1.26	3.75

Table 5 includes some of the cases from the external validation dataset where the model missed and detected errors. We selected the missed errors from those plans containing errors for which the model estimated a probability higher than the median probability in the test set for the variable that contained the error. Detected error were selected from those plans containing errors for which the model estimated a probability lower than the 3rd percentile probability in the test set for the variable that contained the error. The analysis of patterns in the cases where the model succeeded and failed could potentially lead to insights to guide re-training and fine-tuning the process in the future.

Table 5: Selection of missed errors (estimated probability higher than the median) and detected errors (estimated probability lower than the 3rd percentile)

Missed errors					
Anatomic tumor location	**TNM Stage**			**Error**	**Erroneous value**
PROSTA-TE GLAND	T2a	N0	M0	Bolus should be present	None
CERVIX	T1b1	N0	M0	Gantry angle should start at 170	182
SKIN	T2	N0	M0	PTV dose should be 4000 cGy	3600
LUNG	T1c	N0	M0	Table angle should have been 10	0
HEAD /FACE/ NECK	T1	N0	M0	Radia-tion type should be 10X	6X
Detected errors					
Anatomic tumor location	**TNM Stage**			**Error**	**Erroneous value**
LUNG	T4	N3	M1c	There should be no bolus	*custom
BREAST FEMALE	T2	N0	M0	Gantry angle should start at 168	179
BREAST FEMALE	T4d	N1	M0	PTV dose should be 4500 cGy	4005
PROSTA-TE GLAND	NULL	NULL	M0	Table angle should have been 0	5
ABDO-MEN	T3	N2	M1	Radia-tion type should be 6X	10X

7.4 Discussion

We have performed an external validation of a Bayesian network for error detection in radiotherapy plans described in Luk *et al.*[20] using data routinely collected at Maastro clinic. The results show that the model's performance is significantly deteriorated when using it outside of the environment it was developed in. We have also shown that the performance of the model varies heavily for different types of errors. We undertook an analysis that shows that in order to achieve a high sensitivity, the model needs to flag almost all values as potential errors, which reduces its usefulness as an alert system.

The deterioration in the performance of the model in our external validation might be caused by differences in radiotherapy practices between the two clinics and limitations in the implementation of the original model. For example, the institution from which the data to train the model originated uses Elekta's MOSAIQ oncology information system, while Maastro uses Varian's ARIA (Eclipse treatment planning system). On the other hand, the poor performance of the model detecting PTV dose errors could be caused by differences in institutional preferences on dose prescriptions and fractionation schedules. For example, in our institute hypofractionation (>2 Gy per fraction) is frequently applied in prostate cancer patients, while in the original dataset used to train the model, a more conventional treatment schedule was used. The model flagging fractionation schedules different to those in the original institution as errors is likely to be a consequence of training a model in a single institution. However, it is arguable to which extent such models need to be generalizable (e.g. able to accept different fractionation schedules) and to which extent they should be adjusted to the implementing clinic (e.g. to deliberately flag as errors fractionation schedules different to the clinic's) through a commissioning process [28].

Another potential source of model performance deterioration are limitations in the model's development. For example, some categorical variables in the model contain redundant values (e.g. the variable T_Stage contains the values "1a", "1A" and "T1a") and numerical variables often contain a high number of values (e.g. more than 200 states). This in turn led to conditional probability tables (CPTs) with a high number of parameters, as the number of probabilities in a CPT grows exponentially with the number of states in each variable (e.g., the CPT for Number_of_Rxs contains more than 20 million probabilities). Since these parameters need to be estimated from data, the higher the number of parameters, the higher the number of samples required to learn these parameters. Options to alleviate the issue by reducing the number of values in each variable include removing redundant values, discretizing numeric variables and grouping values that are similar or equivalent when considering the task at hand. In addition, the evaluation of the network as proposed by Luk *et al.*[21], considers the probability of each plan parameter independently, conditioned on the diagnostic variables and the treatment intent. This prevents the model from being able to detect erroneous combinations of plan parameters, such as a wrong value for Total_Fraction given a particular Dose_Per_Fraction.

It is worth assessing whether using a single probability threshold to determine whether to flag a value as an error or not is ideal. There is high variance in the number of states or categories across different variables and probabilities tend to be lower the higher the number of states. Therefore, adjusting the threshold per variable to reflect this could lead to improved performance.

There is also room for improvement in the handling of missing data. Many variables contain a special value to reflect missing data (e.g. 'NULL'). This approach has been shown to lead to suboptimal results and is unnecessary in this case given that the algorithm used to learn the probabilities, the Expected Maximization (EM) algorithm, is especially suited to handle missing data[29]. Moreover, BNs are well capable of dealing with missing data when queried for probabilities (i.e., inference).

Moreover, the model was trained using data from a single institution. This is a common practice, but one that leads to models that often do not generalize well outside of the environment in which they were developed. Our results offer yet another example of the importance of using data from multiple sources (e.g. different clinics across the world) when training and testing models to achieve generalizability. This is not easy to achieve because ethical and legal barriers prevent sharing of privacy-sensitive data. However, the recently proposed federated learning paradigm[30] and related initiatives such as the Personal Health Train[31] aim to provide a framework where learning from multiple sites becomes straightforward. Another barrier to combining data from multiple institutions are the differences in the way different institutions encode the data. The FAIR (Findable, Accessible, Interoperable, Reusable) data[32] principles establish a series of guidelines to make data interoperable, specifically by using publicly available ontologies for the creation of a semantic web model. Such ontologies already exist for radiation oncology and radiotherapy[33-36].

Our external validation suffers from a number of limitations. The most important limitation is that while the information about the plans used in the validation are real, the errors are simulated. As explained in the methods section, after analyzing the database used to log misses and near-misses, we only found five errors related to radiotherapy planning. This is likely because technicians check every plan manually before and correct it before the plans are approved for treatment execution or because some errors go undetected. As a consequence, we were forced to simulate errors. Considering how much the model's performance varies across different types of errors, differences between the simulated and actual error distributions could lead to biased overall performance estimates. We mitigated this risk by simulating the errors partly based on the errors encountered in the database, and partly also by simulating the kind of errors that are manually corrected according to experienced technicians' feedback. Another limitation of our external validation is that our dataset was missing the information about the set-up or immobilization devices (e.g. breast board, head-rest) used during radiotherapy. As a consequence, we could not validate the performance of the model detecting errors in these variables. Finally, we did not assess the model's ability to detect errors that might have gone unnoticed in the clinic. In principle, by sacrificing sensitivity, one could use the model to try to flag a few errors that could otherwise go unnoticed with high specificity. However, this could be potentially dangerous because the existence of such a system could give a false sense of security to technicians unaware that by sacrificing sensitivity, most errors would go undetected.

The above mentioned limitations in combination with the different radiotherapy treatment planning software between the two clinics (Mosaiq in Washington vs ARIA Eclipse in Maastro) and the LINAC models (Elekta in Washington vs Varian in Maastro) contributed to the relatively low performance of the model in the external validation. To further investigate the root cause of the low performance of the model in the validation cohort, we aim to address the limitations mentioned in the discussion and train the model in Maastro clinic as a next step of a future study.

The findings of our external validation suggest that the model is not yet ready to be useful in clinical practice in institutions different to its original. However, we believe that if the limitations identified in this external validation are successfully addressed, such a model could lead to reduction in cost of radiotherapy planning and increase its safety.

7.5 Conclusion

We have performed an external validation of a Bayesian network for error detection in radiotherapy plans proposed by Luk *et al.* [21], by testing the performance of the model in actual plans delivered in Maastro clinic with simulated errors. The results show that the performance of the model proposed by Luk *et al.*[21] significantly deteriorated when applied in an environment different to the source institution where it was developed (AUC of 65% versus 89%). The performance of the model varied widely for different types of errors (from 99.5% for table angle errors to 39.2% for PTV dose errors). This result shows the importance of external validations and the advantages of developing models using data from more than one institution. We analyzed the apparent limitations of the model (data preprocessing, handling of missing data, model evaluation) and we have proposed actions to overcome them.

7.6 Acknowledgments

The authors wish to thank Nico Lustberg (Business Intelligence Developer), for helping with the extraction of the data from ARIA oncology information system at Maastro.

Bibliography

1. Delaney G, Jacob S, Featherstone C, Barton M. The role of radiotherapy in cancer treatment: Estimating optimal utilization from a review of evidence-based clinical guidelines. Cancer. 2005;104(6):1129-1137. doi:10.1002/cncr.21324
2. Ringborg U, Bergqvist D, Brorsson B, et al. The Swedish Council on Technology Assessment in Health Care (SBU) Systematic Overview of Radiotherapy for Cancer including a Prospective Survey of Radiotherapy Practice in Sweden 2001--Summary and Conclusions. Acta Oncologica. 2003;42(5-6):357-365. doi:10.1080/02841860310010826
3. Begg AC, Stewart FA, Vens C. Strategies to improve radiotherapy with targeted drugs. Nat Rev Cancer. 2011;11(4):239-253. doi:10.1038/nrc3007
4. Barnett GC, West CML, Dunning AM, et al. Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype. Nat Rev Cancer. 2009;9(2):134-142. doi:10.1038/nrc2587
5. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiotherapy and Oncology. 2018;126(2):312-317. doi:10.1016/j.radonc.2017.11.012
6. Kim N, Chang JS, Kim YB, Kim JS. Atlas-based auto-segmentation for postoperative radiotherapy planning in endometrial and cervical cancers. Radiat Oncol. 2020;15(1):106. doi:10.1186/s13014-020-01562-y
7. Cilla S, Ianiro A, Romano C, et al. Template-based automation of treatment planning in advanced radiotherapy: a comprehensive dosimetric and clinical evaluation. Sci Rep. 2020;10(1):423. doi:10.1038/s41598-019-56966-y
8. WHO Radiotherapy Risk Profile-Technical manual. https://www.who.int/patientsafety/activities/technical/radiotherapy_risk_profile.pdf?ua=1
9. Fraass B, Doppke K, Hunt M, et al. American Association of Physicists in Medicine Radiation Therapy Committee Task Group 53: Quality assurance for clinical radiotherapy treatment planning. Med Phys. 1998;25(10):1773-1829. doi:10.1118/1.598373
10. Thwaites D, Scalliet P, Leer JW, Overgaard J. Quality assurance in radiotherapy. Radiotherapy and Oncology. 1995;35(1):61-73. doi:10.1016/0167-8140(95)01549-V
11. Radiotherapy Errors and Near Misses Data Report (December 2013 to November 2015) Report No. 4. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549847/radiotherapy_errors_and_near_misses_data_report.pdf
12. Unintended Overexposure of a Patient during Radiotherapy Treatment at the Edinburgh Cancer Centre, in September 2015. https://www.gov.scot/publications/unintended-overexposure-patient-during-radiotherapy-treatment-edinburgh-cancer-centre-september/

13. Wolfs CJA, Canters RAM, Verhaegen F. Identification of treatment error types for lung cancer patients using convolutional neural networks and EPID dosimetry. Radiotherapy and Oncology. 2020;153:243-249. doi:10.1016/j.radonc.2020.09.048
14. Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, van Elmpt W. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol. 2020 Dec;153:55-66. doi: 10.1016/j.radonc.2020.09.008. Epub 2020 Sep 10. PMID: 32920005.
15. Yang D, Moore KL. Automated radiotherapy treatment plan integrity verification: Plan checking using PINNACLE scripts. Med Phys. 2012;39(3):1542-1551. doi:10.1118/1.3683646
16. Sun B, Rangaraj D, Palaniswaamy G, et al. Initial experience with TrueBeam trajectory log files for radiation therapy delivery verification. Practical Radiation Oncology. 2013;3(4):e199-e208. doi:10.1016/j.prro.2012.11.013

17. Xia J, Mart C, Bayouth J. A computer aided treatment event recognition system in radiation therapy: Error detection in radiation therapy. Med Phys. 2013;41(1):011713. doi:10.1118/1.4852895
18. Holdsworth C, Kukluk J, Molodowitch C, et al. Computerized System for Safety Verification of External Beam Radiation Therapy Planning. International Journal of Radiation Oncology*Biology*Physics. 2017;98(3):691-698. doi:10.1016/j.ijrobp.2017.03.001
19. Halabi T, Lu H. Automating checks of plan check automation. Journal of Applied Clinical Medical Physics. 2014;15(4):1-8. doi:10.1120/jacmp.v15i4.4889
20. Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations. BJR. 2018;91(1092):20180270. doi:10.1259/bjr.20180270
21. Luk SMH, Meyer J, Young LA, et al. Characterization of a Bayesian network-based radiotherapy plan verification model. Med Phys. 2019;46(5):2006-2014. doi:10.1002/mp.13515
22. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Rev. 2. print., 12. [Dr.]. Kaufmann; 2008.
23. Neapolitan RE. Learning Bayesian Networks. Pearson Prentice Hall; 2004.
24. Kyrimi E, McLachlan S, Dube K, Neves MR, Fahmi A, Fenton N. A Comprehensive Scoping Review of Bayesian Networks in Healthcare: Past, Present and Future. arXiv:200208627 [cs]. Published online February 28, 2020. Accessed November 5, 2020. http://arxiv.org/abs/2002.08627
25. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40. doi:10.1186/1471-2288-14-40

26. W. Vuuren, van TW Schaaf, van der. The Development of an Incident Analysis Tool for the Medical Field. https://research.tue.nl/nl/publications/the-development-of-an-incident-analysis-tool-for-the-medical-fiel
27. Andersen SK, Olesen KG, Jensen FV, Jensen F, Shafer G, Pearl (Eds.) J. Hugin – A shell for building belief universes for expert systems. In: Reading in Uncertainty. ; 1990:332-337.
28. Mahadevaiah G, Rv P, Bermejo I, Jaffray D, Dekker A, Wee L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. Med Phys. 2020;47(5). doi:10.1002/mp.13562
29. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977;39(1):1-22. doi:10.1111/j.2517-6161.1977.tb01600.x
30. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. doi:10.1038/s41598-020-69250-1
31. Beyan O, Choudhury A, van Soest J, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intellegence. 2020;2(1-2):96-107. doi:10.1162/dint_a_00032
32. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. doi:10.1038/sdata.2016.18
33. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879
34. National Cancer Institute Thesaurus Ontology. https://bioportal.bioontology.org/ontologies/NCIT
35. Radiation Oncology Ontology (ROO). https://bioportal.bioontology.org/ontologies/ROO
36. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017

Chapter 8

Making radiotherapy more efficient with FAIR data

Adopted from: "Making radiotherapy more efficient with FAIR data"

Petros Kalendralis, Matthijs Sloep, Johan van Soest, Andre Dekker, Rianne Fijten

Physica Medica 82 (2021): 158-162,

https://doi.org/10.1016/j.ejmp.2021.01.083,

Contribution: First authorship

Abstract

Given the rapid growth of artificial intelligence (AI) applications in radiotherapy and the related transformations toward the data-driven healthcare domain, this article summarizes the need and usage of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles in radiotherapy. This work introduces the FAIR data concept, presents practical and relevant use cases and the future role of the different parties involved. The goal of this article is to provide guidance and potential applications of FAIR to various radiotherapy stakeholders, focusing on the central role of medical physicists.

Keywords: Radiotherapy, FAIR data, Artificial Intelligence

Numerous data science advancements have been made in the radiotherapy domain, such as the extended incorporation of patients' imaging data for treatment purposes and the development of outcome prediction models. These were made possible by including high quality data consisting of information of patients, their treatment and follow-up.

With the rapid introduction of new technologies in radiotherapy, such as artificial intelligence (AI) and machine learning (ML), data driven approaches could influence the way patients are treated. Image-based, biological, dosimetric and clinical variables can be combined with ML techniques to predict radiotherapy tumor outcomes[1-3] and toxicities[4]. However, radiotherapy data are highly complex and thus require clear definition-terminologies to ensure accessibility and interoperability (i.e. understandable by both machines and humans). Datasets without detailed, formal, standardized and applicable terminologies that enforce relationships within the data elements (i.e. ontologies) cannot be interpreted by others without expert knowledge of that specific dataset. As a result, many existing radiotherapy datasets are not reusable due to the absence of these ontological items.

This problem stems from barriers in the exchange of health data due to administrative, political, ethical and technical issues. For instance, inconsistencies in labeling and nomenclature of anatomical structure names in radiotherapy structures sets (RTSTRUCTs) are common. For example, when validating a published radiomics model, researchers discover a variety of label names for the Gross Tumor Volume (GTV), making the correct GTV selection problematic.

Aiming also to overcome problems like these, we need to implement the FAIR (Findable, Accessible, Interoperable, Reusable) data principles [5]. With this commentary article we would like to present our educational opinion based in our research findings and experience with the implementation of the FAIR data principles, rather than giving an exhaustive view of the FAIR principles. Furthermore, this manuscript is intended to provide an overview of the challenges and opportunities of implementing the FAIR principles in radiotherapy, highlighting the medical physicists' role. It ends with a suggested framework to develop responsible FAIR radiotherapy research.

FAIR stands for Findable, Accessible, Interoperable and Reusable (Figure 1). Since its first publication[5] in 2016, the FAIR principles have been adopted by many institutes and research organizations worldwide[6-7] and applied in a variety of disciplines[8-9].

Findability of data refers to a detailed description of metadata, indexed in a searchable source. Each dataset should be assigned a unique and persistent identifier for their unequivocal reference and citation. For instance, the publicly available NCSLC-Radiomics dataset[10] can be cited and referenced with its own unique identifier such as the Uniform Resource Identifier (URI).

Accessible data means that data are readable by both humans and computers with the appropriate authorization. Data should be stored in a trusted repository with an open protocol. Examples of trusted public repository for radiotherapy datasets case are the Cancer Imaging Archive (TCIA)[11],the Extensible Neuroimaging Archive Toolkit (XNAT)[12] hosted within the Dutch national research infrastructure[13], Dryad[14] and Zenodo[15]. It is important to mention that accessible data are not open data without constraint, it means that humans and machines may have access to them by respecting clear rules.

Interoperability refers to the use of formal, universal and broadly applicable languages for knowledge sharing and representation, such as public ontologies. Ontologies are terminologies that give a meaning to the essential relationships between different data concepts. Typical examples of these languages are the Radiation Oncology Ontology (ROO)[16-17], specifically designed for the radiotherapy domain, the Radiation Oncology Structures (ROS)[18] ontology and the Biomedical Imaging Methods (FBbi)[19] ontology. Similar interoperability-enhancing initiatives have been and continue to be undertaken by professional societies[20].

For data to be Reusable, researchers need to include detailed documentation and rich metadata. Publicly available reports about the acquisition, processing and origin of the data combined with a detailed description of technical details such as statistical analysis methods in a format of publicly available codes/algorithms, are highly recommended.

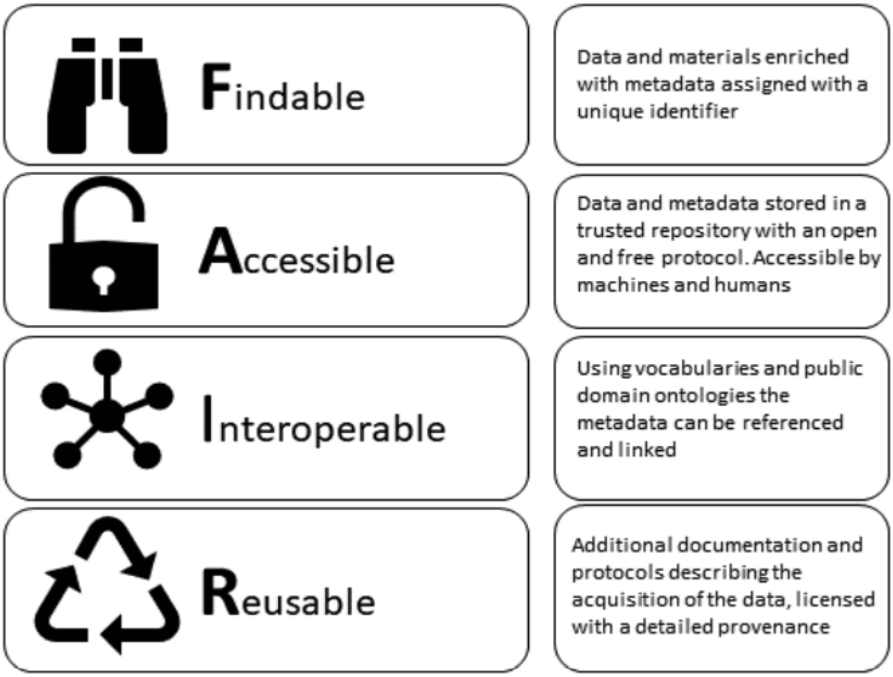

Findable
Data and materials enriched with metadata assigned with a unique identifier
Accessible
Data and metadata stored in a trusted repository with an open and free protocol. Accessible by machines and humans
Interoperable
Using vocabularies and public domain ontologies the metadata can be referenced and linked
Reusable
Additional documentation and protocols describing the acquisition of the data, licensed with a detailed provenance

Figure 1: Schematic representation and description of the FAIR data principles.

Despite its advantages, the FAIR approach has not yet been widely adopted in the radiotherapy domain due to various barriers. First, data preparation is costly, both monetary and labor intensive. There is an emerging need for novel IT solutions that bridge the gap between the need to “FAIRify” data for research purposes and every-day use of Electronic

Health Record (EHR) systems, treatment planning systems and medical image storage archives. The attachment of ontologies for each data element recorded in the EHR systems to clearly establish their definition is an example.

Because each hospital uses various data sources at the same time, there is a vast amount of multi-source data that are mostly unstructured and not integrated. Its transformation into a valuable source of knowledge is therefore problematic due to the lack of interoperability between the different data sources. This, in combination with the unstructured nature of multiple sources, underlies the urgent need to introduce the FAIR data principles.

Moreover, as the radiotherapy data encrypt “sensitive” personal patients’ information, there are ethical and security barriers regarding their use and property. There is an urgent need for a higher level of security that has to be taken into account as it is fundamental to ensure that patient data are protected. The different national regulations of each institute for data use decelerate the FAIR concept. Furthermore, the implementation of the FAIR data principles in compliance with the mandatory General Data Protection Regulation (GDPR)[21] is challenging. Specifically, the GDPR[21] includes the right of data transfer between social networks that prerequisite FAIR data. Contrarily, the GDPR does not allow data sharing between different institutions without a definitive purpose for research or other enterprise activities.

Additionally, the “FAIRification” process requires technical programming skills, which not all clinicians are familiar or proficient with. As a consequence, its adoption and implementation can be difficult as the short- and long-term return or reward are obscure to many.

Several other challenges and barriers for the implementation of the FAIR principles that are not mentioned in our study are described extensively by (inter)national guidelines such as the final report and action plan on FAIR data from the European Commission[22], academic publications such as the study of Jacobsen et al.[23] and non profit organisations initiatives like the Go FAIR foundation[24].

The FAIR principles have the potential to tackle the interoperability and reusability issues, using publicly available ontologies for radiation oncology and radiotherapy [25] such as the ROO [16], [17] and other semantic technologies[26-27], such as the Resource Description Framework (RDF)[28]. Combining these two components results in a semantic data model. Semantic models are used to represent relationships between multiple concepts in the data. RDF represents the data in something called triples. Triples are composed of a subject (“patient”), a predicate (“has biological sex”) and an object(“gender”) that link the relationships between the data items of a dataset. Each of these elements of a triple need to be defined by a publicly available ontological identifier.

Figure 2 displays the relationship between a patient and their gender in triple form.

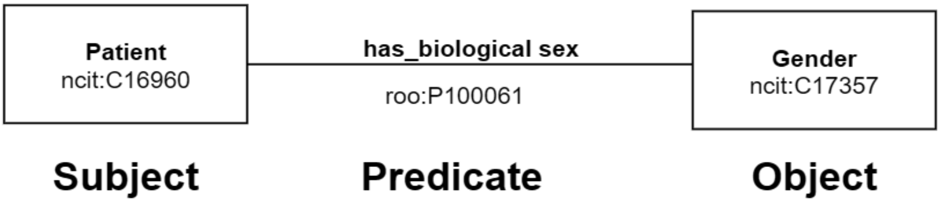


Figure 2. The triples RDF concept represented for the patient and gender data items, connected with the predicates from the ROO[16-17] and NCIT[29] ontology codes available in the Bioportal[30].

By applying the FAIR principles to radiotherapy data, algorithms can assist the users in processing and manipulating data. As a consequence, AI/ML implementation could be facilitated in numerous applications with large potentials in sparing resources in time-consuming repetitive procedures. A relevant example concerns the automatic delineation of the anatomical structures for treatment planning.

Furthermore, the implementation and adoption of the FAIR data principles ensure that research results and outputs can be exchanged and shared among different institutions. One of the potential applications that FAIR data enable in the radiotherapy domain, is the radiotherapy outcomes prediction modelling (exchange and validation) among different institutes exchanging model's parameters instead of patients' data[31].

FAIR compliant datasets have the potential to tackle the interoperability issues between the researchers. A representative example is the set of collections hosted by the above mentioned TCIA provided by Kalendralis et al.[32]. This study provided the clinical metadata, quantitative imaging features and DICOM metadata from four radiotherapy datasets mainly used for radiomics studies[33-34], in RDF[28] format using public ontologies[16-17,29].

Besides the benefits of the establishment of collaborations between different institutes/clinics, one of the basic and valuable achievements of the FAIR principles for radiotherapy researchers and clinicians is the increase of the scientific outcome and of its impact on the community thanks to the largely improved generalizability and usability of the published results: and this is expected to increase the visibility of their work. Of note, publications associated with FAIR compliant data sharing are cited 69% more frequently[35].

The incorrect implementation of the FAIR principles might have significant risks such as the "abuse" of data use by private interests. These risks can be prevented by the implementation of the FAIR principles as the data usage as the FAIR data users can make them accessible ensuring authentication and authorisation steps.

The generation of multisource data, with the goal to make advantage of it by reusing the data besides the immediate care of the patient, requires the inclusion of different stakeholders in the radiotherapy domain. It is important to identify and define the different stakeholders and their action items, establishing a shared responsibility relationship.

These stakeholders in the FAIR radiotherapy domain are radiation oncologists, radiologists (and other clinicians experts in imaging applications), medical physicists, researchers, radiotherapy enterprises (such as EHR or radiotherapy treatment planning systems companies), patients, IT personnel and managerial boards.

High-quality FAIR-compliant data will have a positive impact on the robustness of the clinical decisions for the patients with the prerequisite of a shared vision between the stakeholders. Our suggestion for their roles and responsibilities are shown in Table 1

Table 1: Stakeholders of FAIR data in radiotherapy with the action items for them

Stakeholders	**Role definition**
Clinicians	●Influence the hospital boards to provide resources/funding for the FAIR principles implementation with their clinical knowledge
Medical Physicists	●Data managers ●Data collection and curation tasks ●Commissioning of FAIR data tools
Researchers	●Clinical orientated FAIR research ●Collaborate with doctors and physicists during data curing and conceptualizations of data usage
Companies in radiotherapy market	●FAIR-friendly tools ●Professionalize FAIR data infrastructure tools
IT personnel	●Closer collaboration with the clinicians, investigating the incorporation of the FAIR principles into the daily practice ● Adapting/tailoring infrastructures
Patients	●Understand the benefits of FAIR data ●Give their consent to use their data
Hospital board members	●Data governance plan ●Legal interoperability ●Resources/funding

The definition of valuable clinical questions from the clinicians is crucial for the development and implementation of beneficial FAIR data driven decisions into the clinic. Toxicity reduction to the organs at risk (OAR) during radiotherapy, improvement of the patients' cure rates, time and resources optimization are some of the interesting endpoints that can be evaluated prior to the development of a FAIR data tool in the clinic. Using their clinical knowledge, clinicians should take the role of ambassadors to influence the hospital board members to invest resources/funding to FAIR data driven clinical decisions for a personalized treatment approach of the patients.

Medical physicists can be considered as "data managers" as they usually have the database privileges and rights to acquire and extract the data in collaboration with the IT experts. Their role of facilitator is highly crucial and will be focused on later. Furthermore, they can commission new technologies for use in the clinic. As a result, they are typically assigned to the labor-intensive task of data curation and selection for patient care.

Researchers are assigned with the task of presenting the value of FAIR data by executing clinically oriented research. In most of the cases they may overlap and collaborate with the medical physicists and doctors involved in the process of data curation and conceptualizations regarding how to use these data for the development of new valuable knowledge. In the field of translational research, the involvement of both clinical and basic researchers will contribute a significant value in radiotherapy.

During the past years, various funding agencies and organizations globally encouraged researchers to submit proposals including the implementation of the FAIR data principles[6-7,36].

A crucial bridge between the research and clinic are the radiotherapy market companies that can introduce products and applications related to FAIR data principles. The professionalization of FAIR data infrastructure tools should in future be included in the portfolios of the radiotherapy vendors. FAIR-inclusion criteria for patients cohorts discovery, public repositories including radiotherapy outcomes prediction models, annotated applications with the visualization/registration of patients characteristics[37] and platforms that contain different imaging/treatment protocols from different centers are some of the examples of FAIR compliant applications that can be adopted by vendors in the radiotherapy market.

The need for a closer collaboration between the clinicians and IT professionals in the clinic is emerging for the introduction of the FAIR principles in daily practice. Important knowledge of data structures, data models and clinical databases schemas is held by IT personnel of the clinics and the implementation of a standardized data acquisition and usage plan should be established together with them. Moreover, IT professionals should be assigned the role of adapting and tailoring data infrastructures to fit the needs of a particular center.

Additionally, protocols regarding data reuse requests including approval from the institutional review board (IRB) of each institute should include clear steps describing the action points that should be taken from the different parties that would like to use patients' data and the further procedures that have to be followed for the acquisition of the data in the right format. This should include describing data access principles and protocols for outside parties to gain access to the data. Specifically, being informed about the benefits of FAIR data, the patients' role includes the task of giving their consent/approval to use their data.

The implementation of the FAIR principles should include leadership of each department and data governance policy should be established by such management/board members. Data lifecycle and stewardship plans that include the FAIR principles need to be as compulsory for research proposals as the data analysis protocol. Furthermore, assessment frameworks that include the certification of FAIR services provided should be implemented to the radiotherapy departments in combination with data science and data stewardship training curriculums/programs. Moreover, in the radiotherapy domain we deal with patients' sensitive information so there is a higher level of security that has to be taken into account as it is fundamental to ensure that patient data are protected.

As it is mentioned above, the medical physicists have the role of facilitator as they have the reputation of introducing and implementing novel technologies such as AI into the clinic. Although a leading and active role in the AI field is beneficial for the future career of medical physicists, they need to be included in a working group of multi-discipline experts. Closer collaboration with data scientists and computer scientists is more than necessary for the safe and sufficient development of AI technologies such as the FAIR concept, enabling access to high quality curated datasets needed for AI applications[38]. For their effective and constructive contribution of the medical physicists to this new emerging field in radiotherapy advanced data science skills (AI, data analysis, statistics) should be included in their educational curriculums and training schemes [38-39].

In this manuscript, we made an introduction of the FAIR principles presenting the challenges and opportunities arising from their implementation to the radiotherapy data. Specific efforts and actions points are summarized and underlined from the stakeholders' perspective. In order to continue providing high quality personalized radiotherapy services, the stakeholders of the radiotherapy community should develop a close collaboration with each other trying to overcome and cover technological barriers. To further support that, it is desirable to include the FAIR data principles related topics in the educational schemes of the international medical physicists associations. Medical physics' new trend is correlated with the challenging AI field and the related professions should cope with this including establishing a clear data governance framework in radiotherapy departments.

Bibliography

1. Chao H-H, Valdes G, Luna JM, et al. Exploratory analysis using machine learning to predict for chest wall pain in patients with stage I non-small-cell lung cancer treated with stereotactic body radiation therapy. J Appl Clin Med Phys. 2018;19(5):539-546. doi:10.1002/acm2.12415
2. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. Phys Med Biol. 2016;61(16):6105-6120. doi:10.1088/0031-9155/61/16/6105
3. Gennatas ED, Wu A, Braunstein SE, et al. Preoperative and postoperative prediction of long-term meningioma outcomes. PLoS ONE. 2018;13(9):e0204161. doi:10.1371/journal.pone.0204161
4. Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine Learning-Based Models for Prediction of Toxicity Outcomes in Radiotherapy. Front Oncol. 2020;10:790. doi:10.3389/fonc.2020.00790
5. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18
6. European Commission. Horizon 2020, Work Programme 2018– 2020, Health, demographic change and wellbeing, https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-health_en.pdf [accessed 26 January 2021]
7. Notice Announcing Funding Opportunity Issued for the NIH Data Commons Pilot Phase. Published June 23, 2017, https://grants.nih.gov/grants/guide/notice-files/NOT-RM-17-031.html [accessed 26 January 2021]
8. Lannom L, Koureas D, Hardisty AR. FAIR Data and Services in Biodiversity Science and Geoscience. Data Intellegence. 2020;2(1-2):122-130. doi:10.1162/dint_a_00034
9. Hiebel G, Goldenberg G, Grutsch C, Hanke K, Staudt M. FAIR data for prehistoric mining archaeology. Int J Digit Libr. Published online January 23, 2020. doi:10.1007/s00799-020-00282-8
10. [NSCLC-Radiomics] Aerts, H. J. W. L., Wee, L., Rios Velazquez, E., Leijenaar, R. T. H., Parmar, C., Grossmann, P., … Lambin, P. (2019). Data From NSCLC-Radiomics [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
11. 11. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
12. Marcus, D.S., Olsen, T.R., Ramaratnam, M. et al. The extensible neuroimaging archive toolkit. Neuroinform 5, 11–33 (2007). https://doi.org/10.1385/NI:5:1:11
13. Dutch national research infrastructure TraIT, www.ctmm-trait.nl [accessed 26 January 2021]
14. Dryad repository, https://datadryad.org/stash [accessed 26 January 2021]
15. Zenodo repository, https://zenodo.org/ [accessed 26 January 2021]

16. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879
17. Radiation Oncology Ontology (ROO), https://bioportal.bioontology.org/ontologies/ROO [accessed 26 January 2021]
18. Radiation Oncology Structures (ROS) Ontology, https://bioportal.bioontology.org/ontologies/ROS [accessed 26 January 2021]
19. Biomedical Imaging Methods (FBbi) Ontology, https://bioportal.bioontology.org/ontologies/FBbi [accessed 26 January 2021]
20. Mayo CS, Moran JM, Bosch W, et al. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. International Journal of Radiation Oncology*Biology*Physics. 2018;100(4):1057-1066. doi:10.1016/j.ijrobp.2017.12.013
21. General Data Protection Regulation (GDPR), https://gdpr-info.eu/ [accessed 26 January 2021]
22. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, “Turning FAIR into reality.” https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf [accessed 26 January 2021]
23. Jacobsen A, de Miranda Azevedo R, Juty N, et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intellegence. 2020;2(1-2):10-29. doi:10.1162/dint_r_00024
24. Go FAIR non profit organisation, https://www.go-fair.org/[accessed 26 January 2021]
25. Min H, Manion FJ, Goralczyk E, Wong Y-N, Ross E, Beck JR. Integration of prostate cancer clinical data using an ontology. Journal of Biomedical Informatics. 2009;42(6):1035-1045. doi:10.1016/j.jbi.2009.05.007
26. Van Soest J, Lustberg T, Grittner D, et al. Towards a semantic PACS: Using Semantic Web technology to represent imaging data. Stud Health Technol Inform. 2014;205:166-170.
27. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. Computer Methods and Programs in Biomedicine. 2015;118(3):322-329. doi:10.1016/j.cmpb.2015.01.003
28. Resource Description Framework (RDF), https://www.w3.org/RDF/ [accessed 26 January 2021]
29. National Cancer Institute Thesaurus Ontology, https://bioportal.bioontology.org/ontologies/NCIT [accessed 26 January 2021]
30. Bioportal website, https://bioportal.bioontology.org/ [accessed 26 January 2021]
31. Deist TM, Dankers FJWM, Ojha P, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiotherapy and Oncology. 2020;144:189-200. doi:10.1016/j.radonc.2019.11.019
32. Kalendralis P, Shi Z, Traverso A, et al. FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. Med Phys. Published online June 27, 2020:mp.14322. doi:10.1002/mp.14322

33. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):4006. doi:10.1038/ncomms5006

34. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141

35. Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. Ioannidis J, ed. PLoS ONE. 2007;2(3):e308. doi:10.1371/journal.pone.0000308

36. 2016 National Research Infrastructure, https://www.dese.gov.au/resources/2016-national-research-infrastructure-roadmap [accessed 26 January 2021]

37. Gopinath, D., Agrawal, M., Murray, L., Horng, S., Karger, D., & Sontag, D. (2020). Fast, Structured Clinical Documentation via Contextual Autocomplete. arXiv e-prints, arXiv-2007.

38. Fiorino C, Jeraj R, Clark CH, et al. Grand challenges for medical physics in radiation oncology. Radiotherapy and Oncology. 2020;153:7-14. doi:10.1016/j.radonc.2020.10.001

39. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? Radiotherapy and Oncology. 2018;129(3):421-426. doi:10.1016/j.radonc.2018.05.030

Chapter 9

A Knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry.

Adapted from: “A Knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry”

Petros Kalendralis, Matthijs Sloep, Ananya Choudhury, Lerau Seyben, Jasper Snel, Nibin Moni George, Martijn Veening, Johannes A. Langendijk, Andre Dekker, Johan van Soest, Rianne Fijten

Clinical and translational radiation oncology 31 (2021): 93-96, https://doi.org/10.1016/j.ctro.2021.10.001,
Contribution: Shared first authorship

Abstract

Cancer registries collect multisource data and provide valuable information that can lead to unique research opportunities. In the Netherlands, a registry and model-based approach (MBA) are used for the selection of patients that are eligible for proton therapy. We collected baseline characteristics including demographic, clinical, tumour and treatment information. These data were transformed into a machine readable format using the FAIR (Findable, Accessible, Interoperable, Reusable) data principles and resulted in a knowledge graph with baseline characteristics of proton therapy patients. With this approach, we enable the possibility of linking external data sources and optimal flexibility to easily adapt the data structure of the existing knowledge graph to the needs of the clinic.

9.1 Introduction

Proton therapy has emerged as a novel treatment modality that has the potential to reduce toxicity rates and further improve tumour control due to the depth-dose characteristics of the proton particle[1]. Because of scarcity and costs, the Netherlands has initiated the development of a model-based approach (MBA) to select those patients for proton therapy that will benefit the most[2]. By comparing the dose difference in the organs at risk (OARs) in delta Normal Tissue Complication Probability (ΔNTCP) models for photon and proton plans and the resulting 3D radiation dose, the MBA estimates the potential benefit for an individual patient.

To ensure the MBA remains valid, one needs to continuously update and validate these ΔNTCP models. In the Netherlands, the ProTRAIT (**PRO**ton **T**herapy **R**ese**A**rch reg**I**s**T**ry) initiative has set up a national registry that collects real world data from patients previously treated with proton or photon radiotherapy. The initiative's aim is to systematically and automatically register these data and its ultimate goals are to minimise radiation-induced toxicities in the healthy tissues, to improve quality of life, and to escalate the dose to target (tumor) cells.

In order to fulfill the ProTRAIT initiative's aim, it is important to develop an architecture that can handle the semi-structured nature of radiotherapy data and adheres to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The authors who first published the principles point out the need to improve infrastructures to support the reuse of (scholarly) data and created guidelines to facilitate this[3]. Taking into account the semi-structured and multisource nature of radiotherapy data (imaging, biological and clinical data), as well as the heterogeneous clinical workflows and data analysis pipelines between individual centres, the implementation of the FAIR principles can enable a standardised framework for data management and processing. Furthermore, the accessibility, interoperability and reusability aspect of FAIR will enable a quicker external and independent validation of research findings[3].

FAIR data are often collected in a semantic data model and represented in a knowledge graph[4]. Literature defines a (biomedical) knowledge graph as "a resource that integrates one or more expert-derived sources of information into a graph where nodes represent biomedical entities and edges represent relationships between those entities"[5]. In biomedical science, knowledge graphs are often built based upon an existing database. In our case, the ProTRAIT data registry graph had to be built manually, however inspired by previous work on the Radiation Oncology Ontology[6].

Up until this moment the data needed for the ProTRAIT registry are manually collected in each individual centre and then entered to a centralised electronic data capture (EDC) system.An alternative approach was suggested by Zapletal et al. automating the integration of radiotherapy related data items such as the prescribed dose and the Dose Volume Histogram (DVH) parameters7 using an i2b2-based clinical data warehouse. Although the automatic registration of this approach based on the i2b2 standard has the same goal as our approach, the proton therapy data items and the data model of our case are not fully compliant with the i2b2 standard for instance the i2b2 data model lack many items needed in the proTRAIT registry. Furthermore, the approach of Zapletal et al. is based on a relational data model, where our approach is based on a graph-based data model, which enforces standardized terminologies and is flexible in addition of concepts and structures and is preferable over a relational model as mentioned by the review of Gamal et al.8.A vital part of the upload infrastructure is choosing an interoperable data model; this means choosing a data structure. In this paper we present our choice for a domain specific knowledge graph that stores relevant observational patient data into an interoperable, machine readable format. The knowledge graph specification is publicly available (DOI:10.5281/zenodo.50600699) but the data itself is not due to privacy and legal requirements associated with patient data; access to these data will be formalized in the near future by the ProTRAIT consortium.

9.2 Methods and Results

The clinical items listed in the registry and graph were selected by domain experts based on established clinical workflows and were subsequently reviewed by the relevant national expert community: part of the “Nederlandse Vereniging voor Radiotherapie en Oncologie.” - the Dutch national association for radiotherapy and oncology. This resulted in a flat data model, a list of items with little to no relation between the individual elements. The structure of this flat data model will be discussed in depth in the next paragraphs. The items in this list and others can be found on the Github repository, including the definition and required data element type (integer, string, date etc.) in .xlsx format.

In our case the construction of the graph and ontology classes was a coordinated effort by medical physicists, physicians and computer scientists. Their combined expertise was used to define nodes and edges in the graph. The knowledge graph was modeled using the Resource Description Framework (RDF), a World Wide Web Consortium data standard (W3C). It is originally designed for metadata but also used for knowledge management applications[5]. The RDF format is based on the representation of the data in triples format (Subject-Predicate-Object, eg. Patient-has Disease-Neoplasm). The structure of the knowledge graph was constructed with the patient class at the centre with connections to different sections of clinical information, such as the age and biological sex, and information regarding the baseline treatment (eg. date of first radiotherapy course).

The list of clinical items were represented in the R2RML mapping language10 to describe the graph structure, and to facilitate the data conversion process. In this graph structure, several publicly available ontologies related to the radiation oncology field, bundled in the Radiation Oncology Ontology (ROO)6, were reused. The importance of ontologies specific to the radiotherapy domain has been highlighted by several studies such as the publications of Phillips et al.11 and Bibault et al.12. The use of ontologies enhances the data interoperability and reusability with a clear definition of the different data classes including knowledge representation.

Figure 1 shows the baseline characteristics in our knowledge graph with all the nodes and edges designed by domain experts. Moreover, table 1 presents an overview of the data items used for the creation of the knowledge graph with their definition and description. The different data classes can be grouped in three different categories; (I) baseline characteristics and demographic information, (II) baseline tumour, and (III) radiotherapy planning information. In this visualisation of the graph, we present the relation between variables connected to a patient that all together make up our generic list. Furthermore, we would like to underline that additional data can be linked to this graph easily. The structure of the knowledge graph is open source and the R2RML mapping files can be found on http://www.protrait.nl (licence: CC-BY).

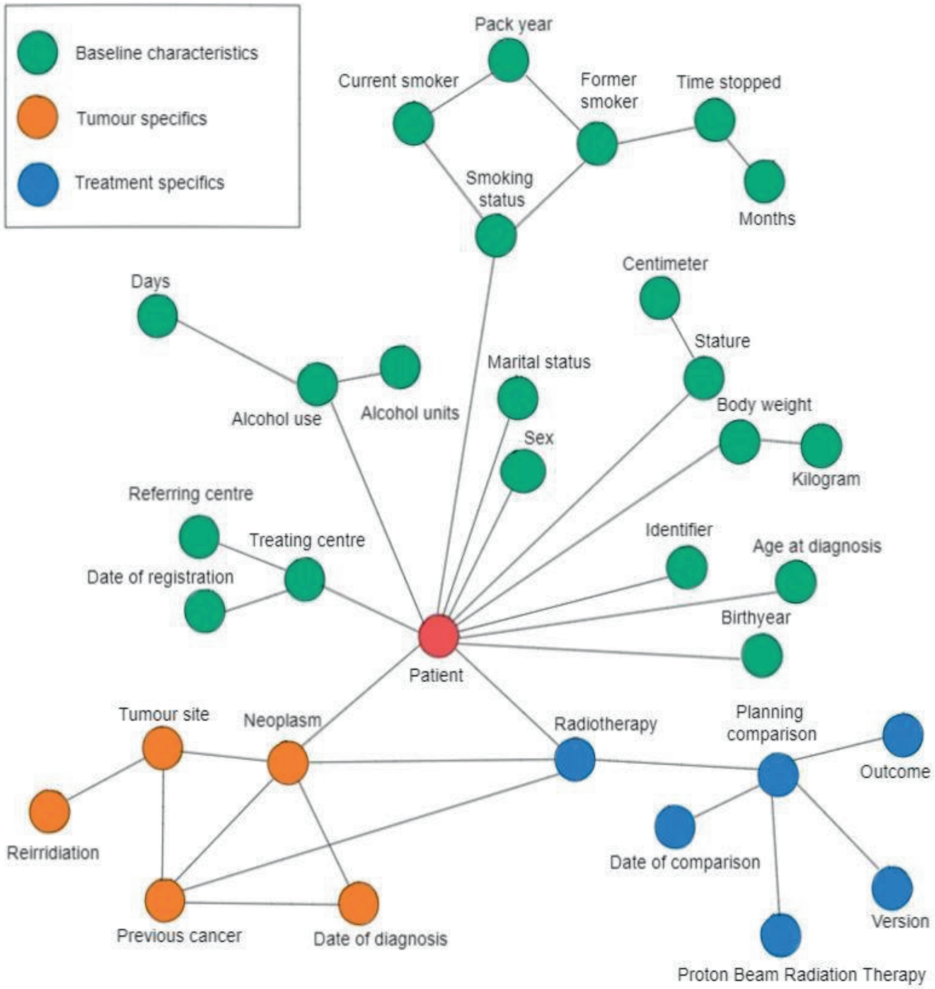


Figure 1: Visualisation of the knowledge graph created with the baseline characteristics (green) tumour specific variables (orange) and treatment variables (blue) of the patients

eligible for proton radiotherapy. For readability purposes we have excluded the predicates between the different instances and classes.

Table 1: Overview of the different data items used in the knowledge graph with their definition and description

Generic list data items	Definition-description
Patient	Patient that has been diagnosed with cancer
Identifier	Patient identifier
Age at diagnosis	Age at diagnosis of the patient
Birthyear	Birth year of the patient
Treating centre	Particle treating centre
Referring centre	Referring centre
Date of registration	Date of registration (first visit in the radiotherapy department)
Neoplasm	Neoplasm
Tumour site	Tumour site
Previous cancer	Previous cancer
Re-irridation	Re-irridation in the case of a previous cancer
Date of diagnosis	Date of first diagnosis (first pathology)
Radiotherapy	Radiotherapy
Planning comparison	Planning comparison performed
Date of comparison	Date of planning comparison (if it was performed)
Outcome	Outcome of the planning comparison
Version	Version planning comparison (version LIPP* protocol)
Proton beam radiation therapy	Proton radiotherapy
Alcohol use	Current alcohol use
Alcohol units	Alcohol units
Days	If the patient is a current alcohol user, number of days per month where ≥ 1 alcohol unit is consumed
Marital status	Marital status
Sex	Biological sex
Body weight	Body weight
Kilogram	Body weight in Kilograms
Stature	Height

Centimeter	Height in Centimetres
Smoking status	Smoking status
Former smoker	Former smoker
Pack year	If patient is a current/former smoker, number of pack years
Current smoker	Current smoker
Former smoker	Former smoker
Time stopped	If patient is a past smoker, number of stopped months
Months	Months(unit of Time stopped)
*LIPP=landelijk indicatie protocol protonen	

9.3 Discussion

In this study, we developed a knowledge graph to store clinical patient characteristics for the proton therapy registry. We chose this data model because of the specific characteristics of the ProTRAIT project. Our knowledge graph was based on the RDF[13] data model using the R2RML[10] mapping language and publicly available ontologies[6,11,14]as it facilitates the interoperability and reusability of data.

Proton therapy is a relatively new treatment and its indications and application are likely to change when new insights develop. Hence, the data model and structure must be designed with flexibility and a transient practical application in mind: as proton therapy gains salience in The Netherlands, new clinical applications will appear and there will be a shift in the threshold of which patients can undergo treatment because of the limited treatment capacity. New models are developed to tackle this shift and for this reason, our architecture must be able to easily adapt to new data elements and model transformations. The semantic data model is flexible because we can define and add ontology classes and their definitions are shared and accessible to others in line with the FAIR principles, which gained traction in the radiotherapy world[15], while still keeping them backward compatible. The semantic data model was developed with machine readability and thus exchange and interoperability in mind. The flexibility further shows in adding new variables to the data model and the cardinality limitations that relational databases have. There is no need to create new tables to tackle {one/many}-to-many relations; for instance, additional treatments can exist in the same graph as additional instances of the treatment class. Finally, the ease to which multiple datasets/data sources can be queried (eg. third parties datasets) in a single unified query is an additional point that makes the knowledge graph an advantageous data format over a relational database[6].

Knowledge graphs are not mainstream in clinical data capture systems. Indeed, relational databases still are widely used for clinical data storage. However, knowledge graphs have significant advantages over relational databases, such as flexibility and the ease with which semantic data may be enriched. Most important, perhaps, is that the ProTRAIT data model stands out in interoperability. As hospitals generally use local syntaxes for data registration, their relational databases are not interoperable.

An alternative to our RDF based ProTRAIT approach could have been the Observational Medical Outcomes Partnership (OMOP)[16]. OMOP is a common data model and technical architecture, which works in a similar manner to our FAIR approach initiative to collect observational data using relational databases. OMOP supports population exchange, but not with the flexibility that FAIR or RDF representations have. Semantic integration adds context, uncertainty and detail to the data annotation on a level that OMOP cannot[4]. Furthermore, OMOP is not designed for handling detailed procedural information in a structured and standardized manner. Health Level 7 (HL7) is the clinical standard that describes data formats and elements is a relevant standardisation initiative; the latest version, Fast Healthcare Interoperability Resources (FHIR), focuses on communication and information exchange[13]. FHIR is designed for electronic health record (EHR) based sharing of data from individual patients between institutions and is broadly supported[14]. However, since FHIR has limited functionality in exchanging population level data we opted for the semantic data model.

If the field of radiation oncology wants to make a quick translation from technological advancements to patient care improvements it needs a flexible data system. A system that can incorporate standardised structured data and common data elements as easily as new input. Standardising takes up a lot of time and effort; for the identification of a set of clinical and genomic data elements the OSIRIS group needed a year of weekly multidisciplinary meetings[18]. For this reason standardising will always lag behind innovation and research. Thus a flexible system that combines both is needed especially in technology heavy disciplines like radiation oncology.

The metadata that ontologies add to the knowledge graph not only make the data adhere to the FAIR principles but also enrich the data and serve another practical purpose. By using domain specific ontologies in our knowledge graph, the original real world data can co-exist in the same graph together with the project specific categories and numerical values. In the analysis there is the potential to allow algorithms to infer indirect knowledge from the graph, which is not possible in a flat relational database. In other words, the clinical expertise in the creation of the ontology and knowledge graph means that the metadata is enhancing the instance data, and that inferencing could potentially improve AI/ML analysis of the data. For example, identification of similar patient groups, depending on their characteristics, may enable a personalised approach for prognostic studies.

In the future the central registry may become substituted by a federated/distributed analysis of data using the Personal Health Train (PHT)[19]. The requirements set by the Semantic Web technologies allows machines to understand and interpret the data element classes. Furthermore, ML applications can be implemented, such as the validation and exchange of prediction models using the privacy preserving PHT infrastructure[19]. Moreover, the knowledge graph format may efficiently serve data handling and storage of (distributed) large-scale datasets (“big data”).

9.4 Conclusion

With this study we present our knowledge graph; a database solution for a clinical and research repository that ensures a high degree of flexibility which is needed in a new and advancing field. Our research repository promotes adherence to the FAIR principles. This will facilitate re-use of the data for instance by linking the data to other data sets or incorporating the PHT infrastructure for federated learning analysis. Lastly, the knowledge graph enhances the data and creates opportunities for improved Machine Learning (ML)/Artificial Intelligence (AI) analysis. Future plans are to link sets of tumour specific items that contain data elements related to the treatment, patient reported outcome measures and radiotherapy dose information in order to allow for the design and validation of ΔNTCP models needed in the MBA proton patients’ selection.

9.5 Acknowledgements

We want to thank the Dutch Cancer Society KWF 30943405N for funding the development of the proTRAIT infrastructure and subsequent proton therapy research repository.

Bibliography

1. Mohan R, Grosshans D. Proton therapy – Present and future. Advanced Drug Delivery Reviews. 2017;109:26-44. doi:10.1016/j.addr.2016.11.006
2. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. Radiotherapy and Oncology. 2013;107(3):267-273. doi:10.1016/j.radonc.2013.05.007
3. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3:160018. doi:10.1038/sdata.2016.18
4. Bona JP, Prior FW, Zozus MN, Brochhausen M. Enhancing Clinical Data and Clinical Research Data with Biomedical Ontologies - Insights from the Knowledge Representation Perspective. Yearb Med Inform. 2019;28(01):140-151. doi:10.1055/s-0039-1677912
5. Miller E. An Introduction to the Resource Description Framework. Bul Am Soc Info Sci Tech. 2005;25(1):15-19. doi:10.1002/bult.105
6. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879
7. Zapletal E, Bibault J-E, Giraud P, Burgun A. Integrating Multimodal Radiation Therapy Data into i2b2. Appl Clin Inform. 2018;09(02):377-390. doi:10.1055/s-0038-1651497
8. Gamal A, Barakat S, Rezk A. Standardized electronic health record data modeling and persistence: A comparative review. Journal of Biomedical Informatics. 2021;114:103670. doi:10.1016/j.jbi.2020.103670
9. Kalendralis P, Sloep M, Fijten R, Dekker A. Version 1.0.0 of the proTRAIT r2rml mappings. Published online July 2, 2021. doi:10.5281/ZENODO.5060069
10. R2RML language. https://www.w3.org/ns/r2rml
11. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017
12. Bibault J-E, Zapletal E, Rance B, Giraud P, Burgun A. Labeling for Big Data in radiation oncology: The Radiation Oncology Structures ontology. Amendola R, ed. PLoS ONE. 2018;13(1):e0191263. doi:10.1371/journal.pone.0191263
13. Resource Description Framework (RDF). https://www.w3.org/RDF/
14. National Cancer Institute Thesaurus Ontology. https://bioportal.bioontology.org/ontologies/NCIT
15. Kalendralis P, Sloep M, van Soest J, Dekker A, Fijten R. Making radiotherapy more efficient with FAIR data. Physica Medica. 2021;82:158-162. doi:10.1016/j.ejmp.2021.01.083

16. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the Science for Active Surveillance:

Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153(9):600. doi:10.7326/0003-4819-153-9-201011020-00010

17. HL7 Fast Healthcare Interoperability Resources Specification (FHIR®), DSTU Release 1. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=343
18. Guérin J, Laizet Y, Le Texier V, et al. OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology. JCO Clinical Cancer Informatics. 2021;(5):256-265. doi:10.1200/CCI.20.00094
19. Beyan O, Choudhury A, van Soest J, et al. Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intellegence. 2020;2(1-2):96-107. doi:10.1162/dint_a_00032

Chapter 10

A knowledge graph approach to registering tumour specific data of patients-candidates for proton therapy in the Netherlands

Adopted from: “A knowledge graph approach to registering tumour specific data of patients-candidates for proton therapy in the Netherlands”

Petros Kalendralis, Matthijs Sloep, Ananya Choudhury, Lerau Seyben, Jasper Snel, Nibin Moni George, Martijn Veening, Johannes A. Langendijk, Andre Dekker, Johan van Soest, Rianne Fijten

Submission in preparation, Contribution: First authorship

Abstract

Purpose

The registration of multi-source radiation oncology data is a time consuming and labour intensive procedure. The standardisation of data collection offers the possibility for the acquisition of quality data for research and clinical purposes. With this study we present an overview of the different tumour group data lists in the Dutch national proton therapy registry. Our goal is to provide the radiotherapy community with a flexible and interoperable data model for data exchange between centres. We highlight data variables that are needed for models used in the model-based approach (MBA), which ensures a fair selection of patients that will benefit most from proton therapy.

Methods

As a representative example of the workings of these different tumour specific knowledge graphs, we present the FAIR (Findable, Accessible, Interoperable, Reusable) data principles-compliant knowledge graph approach describing the head and neck tumour variables using radiotherapy domain ontologies and semantic web technologies. We used dosimetric and clinical variables included in the standardised head and neck tumour group items list (protrait.nl) that are used for the selection of patients candidates for proton therapy.

Results

We successfully implemented the creation of the knowledge graph using data items from the head and neck tumour list. Furthermore, we presented the structure of an interoperable data model based on the usage of publicly available ontologies and semantic web technologies.

Conclusion

With this study we provide a synopsis of the different data items lists of the ProTRAIT registry, focusing on a particular knowledge graph with data items included in the head and neck tumour group list. We also highlight the importance of the FAIR data principles that can establish a standardised framework of data reusability in radiotherapy.

10.1 Introduction

Recently, significant efforts have taken place for the digital transformation of radiation oncology[1]. These efforts are targeted toward gaining and exploiting new knowledge from the data acquired for and used in daily practice for more efficient and personalised treatments[2,3]. The vast majority of these "big-data"[2] is multisource and usually labelled with a hospital's specific terminology. The storage of these data in multiple sources within the local hospitals systems leads to a lack of standardised clinical data curation and inconsistencies in labelling. As a result, the exploitation of this valuable information is very difficult and time consuming, especially when combining data from multiple centres. When these data are in fact usable, artificial intelligence (AI) algorithms are able to predict clinically relevant patient outcomes and subsequently improve care for patients and possibly produce new knowledge[3].

In order to make exploitation of these multi-source and multi-center data possible, we need to use tools that can guarantee a harmonised and uniform data registration pipeline. This can only be achieved by implementing tools that adhere to the FAIR (Findable, Accessible, Interoperable. Reusable) data principles[4]. For example, taking into account the rapid development of the Fast Healthcare Interoperability Resources (FHIR)[5] in radiotherapy, labelling data with publicly available ontologies is beneficial as the interoperability is reassured so that the data output can be re-used in a wider perspective. Furthermore, with the transformation of the data elements into a machine readable format such as the Resource Description Framework (RDF)[6] format, the data can be more easily used as an input for AI-based analysis.

The harmonisation of radiotherapy real-world data according to the FAIR principles was recently achieved within the ProTRAIT (**PRO**ton **T**herapy **R**ese**A**rch reg**I**s**T**ry)[7] initiative, which had as a main goal the standardised registration of patients candidates for proton therapy in the Netherlands. This registration includes different data elements categorised in different tumour specific groups that are collected at a multi-institutional level, including radiation dose and toxicity related data items as well as demographic information [7]. These data elements are transformed in a FAIR format using semantic web technologies such as the resource description framework (RDF)[6] standard and radiotherapy specific ontologies[9–11]**.** Moreover, the use of the FAIR principles within the standardised data collection of proton therapy patients will enable multi-centre interoperability for proton therapy data.

In this manuscript, we present the structure of the data model we used for the creation of the knowledge graphs used in the ProTrait initiative and will focus on the head and neck tumour section, and more specifically the subsection of data elements that are included in the national indication protocol for proton therapy (NIPP)[12] for the computation of the Normal Tissue Complication Probability (NTCP) dysphagia and xerostomia models using photon and proton dose variables. These models are used for the selection of patients for proton therapy in the Netherlands according to the model-based approach (MBA)[13].

10.2 Materials and methods

Knowledge graph creation

Generally, the data elements defined in the tumour specific lists of the ProTRAIT data registry contain for example tumour specific, diagnostics, treatment planning information, acute and late toxicity, and patient-reported outcome measures. The different data elements registered for the head and neck cancer registry were selected by a coordinated effort led by physicians, medical physicists and computer scientists who collected and consistently defined the necessary clinical items needed. They are publicly available in .xlsx format in the GitHub repository with detailed definitions and explanations such as the type of the element (character or integer) or the exact clinical definition of each data element.

The data elements of the registry were transformed into a FAIR format in using publicly available ontologies[9] needed for transforming the data into RDF[6]. Specifically, the RDF format transforms the data in triples, ensuring that the data are machine actionable. Two radiation oncology focused ontologies were used: the Radiation Oncology Ontology (ROO)[11] and the Radiation Oncology Structures (ROS)[14] ontology. In addition, the more generic National Cancer Institute Thesaurus (NCIT)[10] was used. Each ontology used in the creation of these knowledge graphs is publicly available via Bioportal[15]. Furthermore, for the transformation of the relational databases to RDF format the mapping language R2RML[16] was used. The R2RML language allows the representation of a relational database as an equivalent graph data object using ontologies which can be used and controlled by a mapping syntax file. The knowledge graph data model was chosen because it adds the flexibility that the real world data registry needs[17].

As an example, we present the knowledge graph created with the variables used for the NTCP dysphagia model of the NIPP[12] that predicts the NTCP values for the head and neck cancer patients candidates for proton therapy to develop greater than second grade dysphagia six months after the end of their radiotherapy treatment. The different prognostic variables in the NTCP model as described by the NIPP[12] are: i) the baseline dysphagia score as rated by physicians in the start of the treatment, ii) the mean photon-proton dose to the oral cavity and iii) the mean photon/proton dose to the superior pharyngeal constrictor muscle (PCM). For the creation of the knowledge graph presenting the prognostic variables of the aforementioned NTCP model, we used 368 head and neck cancer patients treated with photon and proton based radiotherapy between 2019 and 2021 in MAASTRO clinic, Maastricht, The Netherlands.

10.3 Results

An overview of the different tumour groups in the data registry and the number of the different items in the respective knowledge graphs for each tumour site are shown in Table 1. The different ProTRAIT data element lists can be found in the GitHub repository. These lists contain different data elements related to the radiotherapy treatment planning such as the mean dose to organs at risk (OARs), toxicity rates at different timepoints like the six months xerostomia rates and clinical data items such as the TNM stage for each specific tumour group. Table 1 enumerates and categorises the data items of four specific tumour group lists with several examples.

Table 1: Overview of the data items number with examples categorised in different groups

Item lists	**Head and neck**	**Lung**	**Oesophagus**	**Breast**	**Clinical items Examples**	**Radiotherapy specific items examples**	**Toxicity items examples**
Number of clinical items	57	58	55	123	Clinical TNM stage, Tumour location, Date of diagnosis, Histology, Surgery	Mean Dose Thyroid, V20 Heart, Gross Target Volume GTV, Planning Tumour Volume (PTV), Mean dose - Brainstem,Maximal dose - Spinal cord	Dermatitis, Dysphagia, Xerostomia, Feeding Tube dependency, Heart failure,Liver disease
Number of Radiotherapy specific items	221	158	166	261			
Number of toxicity items	16	20	29	26			

Examples of the different knowledge graphs that can potentially be created are shown in the figures S.1-S.4 of the supplementary material. The four knowledge graphs show some representative data items for each tumour group as there are some common data elements among them such as the clinical TNM stage or the Planning Target Volume (PTV). Specifically, we would like to focus on the section of our knowledge graph containing the variables used for the head and neck NTCP model development according to the indication protocol[12], as shown in figure 1.

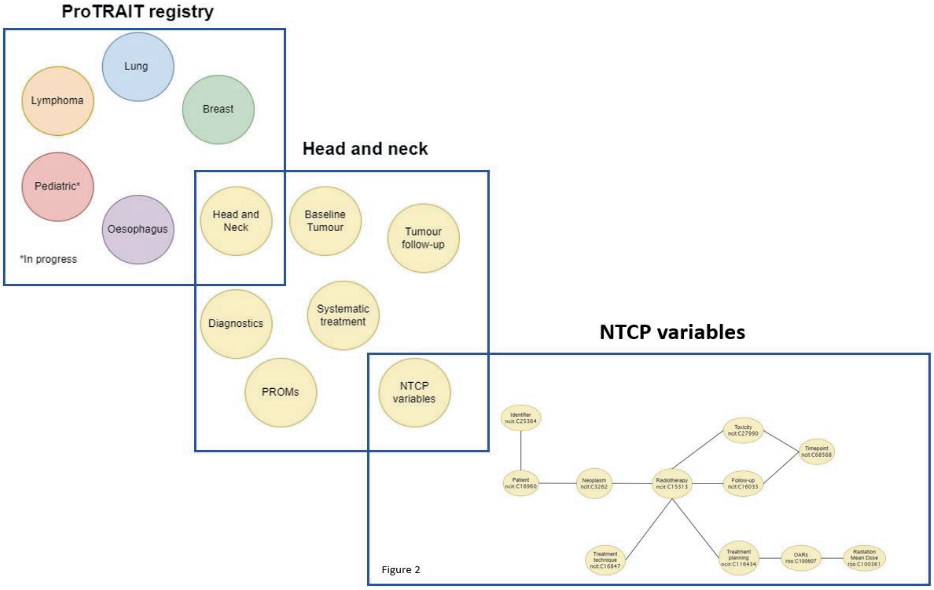


Figure 1: Schematic representation of the NTCP variables graph as a subsection of the overall ProTRAIT registry knowledge graph. Figure 2 is presented in an enlarged version below. For the creation of the knowledge graphs the Radiation Oncology Ontology (ROO)[11] and the National Cancer Institute Thesaurus (NCIT)[10] ontology were used. The ROO and NCIT codes are included in the knowledge graph nodes.

Abbreviations: PROMs= Patient Reported Outcome Measures, NTCP= Normal Tissue Complication Probability

Displaying the data elements used for the computation of the xerostomia and dysphagia NTCP models according to the NIPP[12], we present the knowledge graph subsection in figure 2. The different concepts are connected based on clinically relevant relationships with each other in a logical representation in their graph[18]. Dosimetric photon and proton data elements are connected with the planning comparison data item. The planning comparison between photon and proton treatment plans using the NTCP models is the main part of the MBA correctly predicting the clinical benefit of the proton therapy to the patients in terms of reduced toxicity rates.

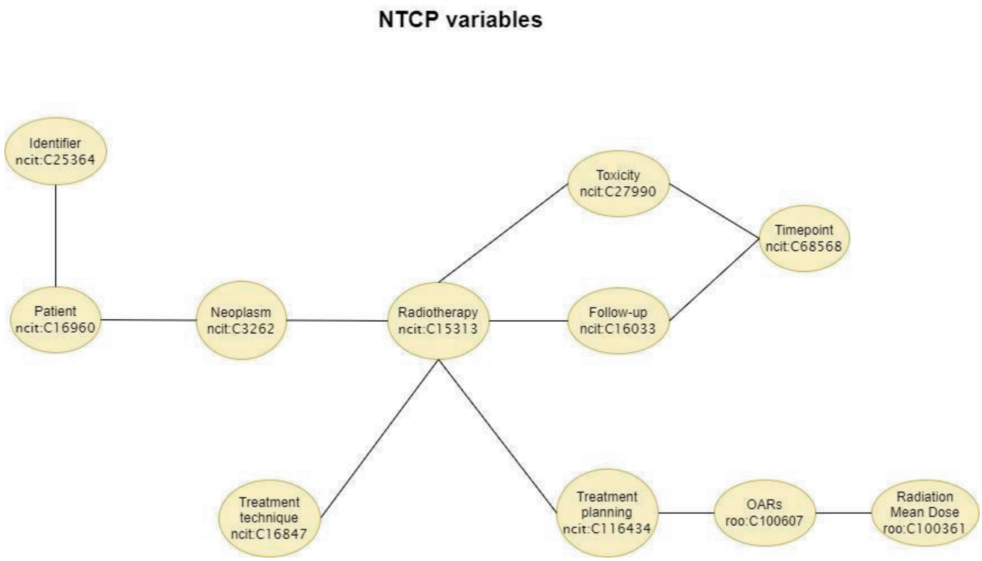


Figure 2: Schematic overview of the NTCP subsection of the knowledge graph created with data items included in the head and neck data item list and used in the NIPP[12] NTCP models for xerostomia and dysphagia of the Dutch indication protocol for proton therapy. For space economy and readability purposes the predicates between the different instances and data classes are not present in the figure.For the creation of the knowledge graphs the Radiation Oncology Ontology (ROO)[11] and the National Cancer Institute Thesaurus (NCIT)[10] ontology were used. The ROO and NCIT codes are included in the knowledge graph nodes.

Abbreviations:OARs= Organs at Risk

An example of a NTCP model included in the national protocol[12] is the logistic regression based model that predicts the probabilities of patients to develop grade 2 or more dysphagia six months after the radiotherapy treatment. Using the variables included in the NTCP model (as described in the results section) we created a sub-graph by selecting a subsection of the figure 2 knowledge graph as shown in figure 3.

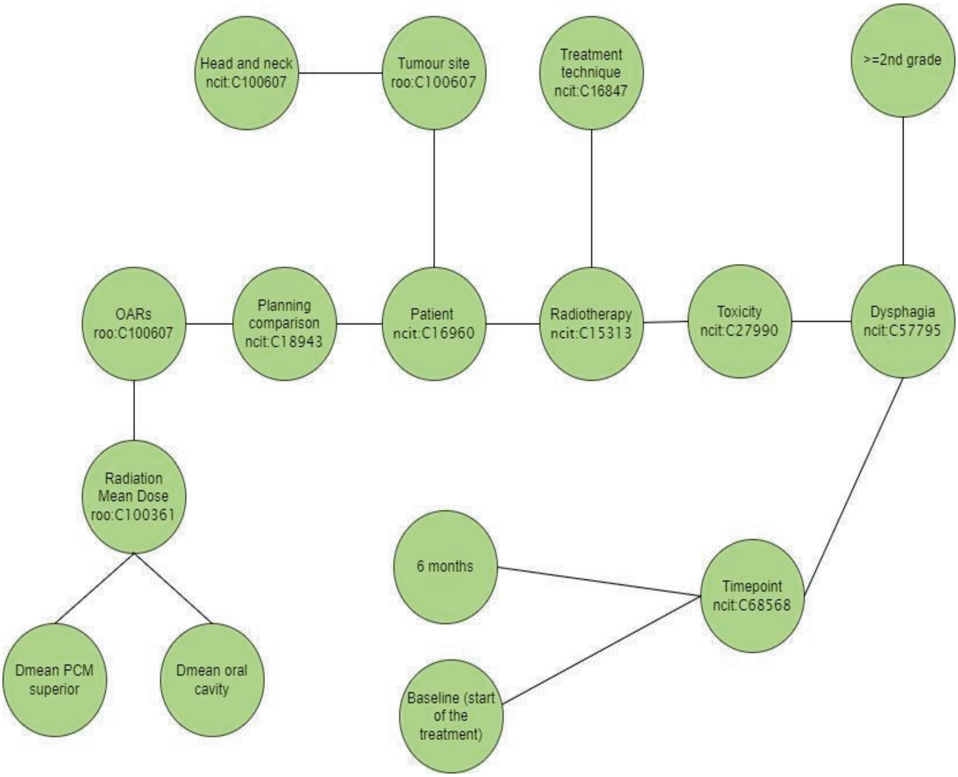


Figure 3: Schematic overview of the knowledge graph created using the variables of the NIPP[12] NTCP grade 2+. For space economy and readability purposes the predicates between the different instances are not present in the figure.For the creation of the knowledge graphs the Radiation Oncology Ontology (ROO)[11] and the National Cancer Institute Thesaurus (NCIT)[10] ontology were used. The ROO and NCIT codes are included in the knowledge graph nodes.

Abbreviations: Dmean=Mean delivered radiation Dose, PCM=pharyngeal constrictor muscle

After development of the knowledge graphs we used data included in the dysphagia NTCP model of the NIPP[19] from 368 patients registered in the ProTRAIT registry to verify its use with real-world data. 95.7% of these patients were treated with photons while 4.3% were treated with protons (Intensity modulated proton therapy-IMPT). The percentage of patients who developed dysphagia equal or greater than second grade in the start of the radiotherapy treatment was 22%. This percentage increased by 11% to 33% for the time-point of six months after the end of the radiotherapy as presented in figure 4.

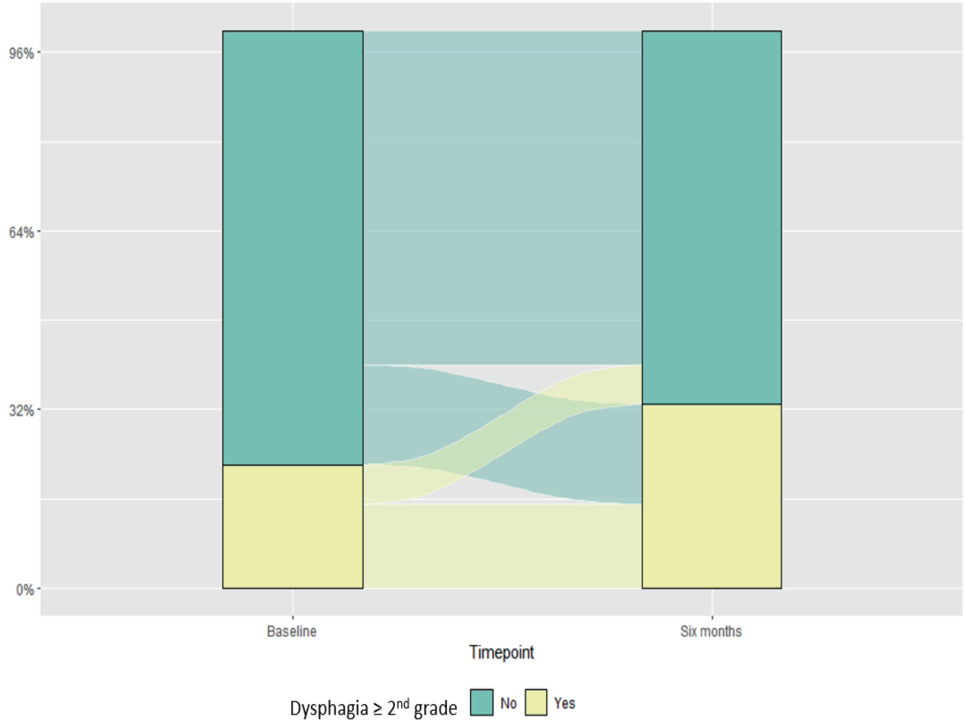


Figure 4: Flowchart with the comparison of the proportion of patients that developed equal or greater than second grade dysphagia in the start of the treatment and six months after it. The percentage of patients who developed second grade dysphagia increased by 11% in the time-point of six months compared to the start of the radiotherapy treatment.

10.4 Discussion

In this technical note, we provided an overview of the tumour-specific ProTRAIT registry knowledge graphs, with a special focus on the head and neck data elements list. We presented a knowledge graph that represents the different data items used for the computation of the NTCP dysphagia and xerostomia models utilised in the MBA for the selection of patients for proton therapy in the Netherlands. For the creation of the RDF-based knowledge graph, publicly available radiation oncology-related ontologies[10,11] were used in combination with the relational database to RDF R2RML mapping language [16].

The transformation of the relational data warehouses of the participating centres to a semantic RDF data model enables interoperability between different centres and a level of flexibility, something that a relational database structure cannot easily replicate. Furthermore, the RDF standard makes the data machine actionable which is key for the implementation of federated learning studies. One of the main intentions for which the ProTRAIT knowledge graph is designed is the Personal Health Train (PHT) approach where statistical analysis can be exchanged between different centres in a privacy preserving manner. In our case, an envisioned purpose is the creation and validation of the head and neck NTCP dysphagia and xerostomia models in the different Dutch proton therapy centres without the exchange of patients' data using the above mentioned knowledge graph data model and PHT.

Furthermore, we would like to underline the importance of the ontologies used for the creation of interoperable data. Ontologies offer the possibility to combine radiotherapy data with different data groups such as the different data items lists of ProTRAIT and numerical or categorical values, inferring indirect knowledge from the created knowledge graph. Thus, the implementation of AI and machine learning (ML) techniques could be potentially improved due to the inference offered by the semantic data model. For instance, the inclusion of patients specific dose parameters or tumour specific data items such as the TNM stage can be used for the exchange of statistical prognostic models in a distributed learning manner.

Therefore, it is of paramount importance for the different radiotherapy centres to systematically collect data in a standardised way. The creation of a national radiotherapy registry of clinical data requires the systematic and continuous collaboration of different professional disciplines. Clinicians, radiotherapy technicians, data and computer scientists should collaborate to create standardised data item lists and privacy-preserving Information Technology (IT) infrastructures in order for a registry to be established. Problems with the data structure and terminologies of each hospital or the free text fields in the electronic health record (EHR) systems make the registration and use of data a labour intensive and time consuming task. As a result, additional resources in terms of personnel are still necessary such as data managers who will undertake the task of manual data registration.

Radiation oncology relies on accurate data to ensure patients receive the best care. However, a significant amount of data is stored in different sources and often in unstructured text format such as medical history reports and clinicians' notes captured in the electronic health record EHR systems of the hospitals. Usually, deploying a labour-intensive process is required to extract information needed for research or clinical purposes. Because of the significance of acquiring structured and quality data, the different care providers dedicate highly skilled professionals such as clinical data managers to manually review and extract clinical insights, which is a time-consuming, expensive and error prone procedure. In the ProTRAIT case, without automation in accordance with the FAIR principles, the data managers and data engineers of each participating centre need to continue manual registration of clinical information necessary for the Dutch national registry. The implementation of clinical natural language processing (NLP) technologies potentially someday provide an alternative and comprehensive, automated and cost effective solution to manual clinical data extraction[20,21]. In our radiotherapy case the acceleration of the data extraction from the different EHR systems will enable a rapid and quality data input for semantic web and FAIR compliant technologies like our knowledge graph approach. It is worth mentioning that several health data FAIRfication approaches have been proposed based on the fast healthcare interoperability resources (FHIR) standard[22–24]. These could turn out to be viable alternatives to our knowledge graphs. Currently, FHIR-based profiles are not clinically implemented and adopted yet in the RT domain and were therefore not implemented yet in our registry.

10.5 Conclusion

With this work we provide an overview of the different data items lists of the ProTRAIT registry, presenting in particular a subsection of our knowledge graph with data items included in the head and neck tumour group list. Specifically, these head and neck data items are included in the NIPP[12] which is used for the selection of patients for proton therapy in the Netherlands. The creation of the data items lists by experienced radiotherapy professionals in combination with the inclusion of semantic data scientists for the transformation of the different data items into a machine readable format, facilitates the interoperability and flexibility that a semantic data model offers. Furthermore, adhering to the FAIR principles in the creation of our knowledge graph enables the link of external data sources as well as with the reusability of data. In conclusion, we would like to underline the potential of the knowledge graph approach regarding the utilisation of federated AI/ML analysis in radiotherapy.

10.6 Acknowledgments

This publication is part of the “ProTRAIT” project with project number 30943405N funded by the Dutch Cancer Society KWF for the development of the ProTRAIT infrastructure and subsequent proton therapy research repository.

This publication is also part of the project “Making radiotherapy sustainable” with project number 10070012010002 of the Highly Specialised Care & Research programme (TZO programme) which is (partly) financed by the Netherlands Organisation for Health Research and Development (ZonMw).

Bibliography

1. Aznar MC, Bacchus C, Coppes RP, et al. Radiation oncology in the new virtual and digital-era. Radiotherapy and Oncology. 2021;154:A1-A4. doi:10.1016/j.radonc.2020.12.031

2. Lustberg T, van Soest J, Jochems A, et al. Big Data in radiation therapy: challenges and opportunities. BJR. 2017;90(1069):20160689. doi:10.1259/bjr.20160689

3. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Affairs. 2014;33(7):1163-1170. doi:10.1377/hlthaff.2014.0053

4. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. doi:10.1038/sdata.2016.18

5. HL7 Fast Healthcare Interoperability Resources Specification (FHIR®), DSTU Release 1. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=343

6. Resource Description Framework (RDF). https://www.w3.org/RDF/

7. ProTRAIT (PROton Therapy ReseArch RegIsTry). www.protrait.nl

8. Sloep M, Kalendralis P, Choudhury A, et al. A Knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry. arXiv:210702482 [cs]. Published online July 6, 2021. Accessed July 15, 2021. http://arxiv.org/abs/2107.02482

9. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017

10. National Cancer Institute Thesaurus Ontology. https://bioportal.bioontology.org/ontologies/NCIT

11. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879

12. Langendijk JA, Hoebers FJP, de Jong MA, et al. National Protocol for Model-Based Selection for Proton Therapy in Head and Neck Cancer. International Journal of Particle Therapy. 2021;8(1):354-365. doi:10.14338/IJPT-20-00089.1

13. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: The model-based approach. Radiotherapy and Oncology. 2013;107(3):267-273. doi:10.1016/j.radonc.2013.05.007

14. Bibault JE, Zapletal E, Rance B, Giraud P, Burgun A. Labeling for Big Data in radiation oncology: The Radiation Oncology Structures ontology. Amendola R, ed. PLoS ONE. 2018;13(1):e0191263. doi:10.1371/journal.pone.0191263

15. Bioportal. https://bioportal.bioontology.org/

16. R2RML language. https://www.w3.org/ns/r2rml

17. Gamal A, Barakat S, Rezk A. Standardized electronic health record data modeling and persistence: A comparative review. Journal of Biomedical Informatics. 2021;114:103670. doi:10.1016/j.jbi.2020.103670

18. Hasan SMS, Rivera D, Wu XC, Durbin EB, Christian JB, Tourassi G. Knowledge Graph-Enabled Cancer Data Analytics. IEEE J Biomed Health Inform. 2020;24(7):1952-1967. doi:10.1109/JBHI.2020.2990797

19. National Indication Protocol for Proton Therapy in the Netherlands. https://nvro.nl/images/documenten/rapporten/2019-08-15__Landelijk_Indicatieprotocol_Protonentherapie_Hoofdhals_v2.2.pdf

20. Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. Natural Language Processing in Dutch Free Text Radiology Reports: Challenges in a Small Language Area Staging Pulmonary Oncology. J Digit Imaging. Published online February 19, 2020. doi:10.1007/s10278-020-00327-z

21. Wang Y, Tafti A, Sohn S, Zhang R. Applications of Natural Language Processing in Clinical Research and Practice. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics; 2019:22-25. doi:10.18653/v1/N19-5006

22. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020;59(S 01):e21-e32. doi:10.1055/s-0040-1713684

23. Choudhury A, van Soest J, Nayak S, Dekker A. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee A, Borgohain SKr, Soni B, Verma G, Gao XZ, eds. Machine Learning, Image Processing, Network Security and Data Sciences. Vol 1240. Communications in Computer and Information Science. Springer Singapore; 2020:85-95. doi:10.1007/978-981-15-6315-7_7

24. Phillips M, Halasz L. Radiation Oncology Needs to Adopt a Comprehensive Standard for Data Transfer: The Case for HL7 FHIR. International Journal of Radiation Oncology*Biology*Physics. 2017;99(5):1073-1075. doi:10.1016/j.ijrobp.2017.08.007

10.7 Supplementary material

Using publicly available radiation oncology related ontologies[1–3] and the Resource Description Framework[4] (RDF) standard we present characteristic examples of the knowledge graphs (figures S.1-S.4) that can be created using the data elements of the four different tumour group standardised lists of ProTRAIT (**PRO**ton **T**herapy **R**ese**A**rch reg**I**s**T**ry).

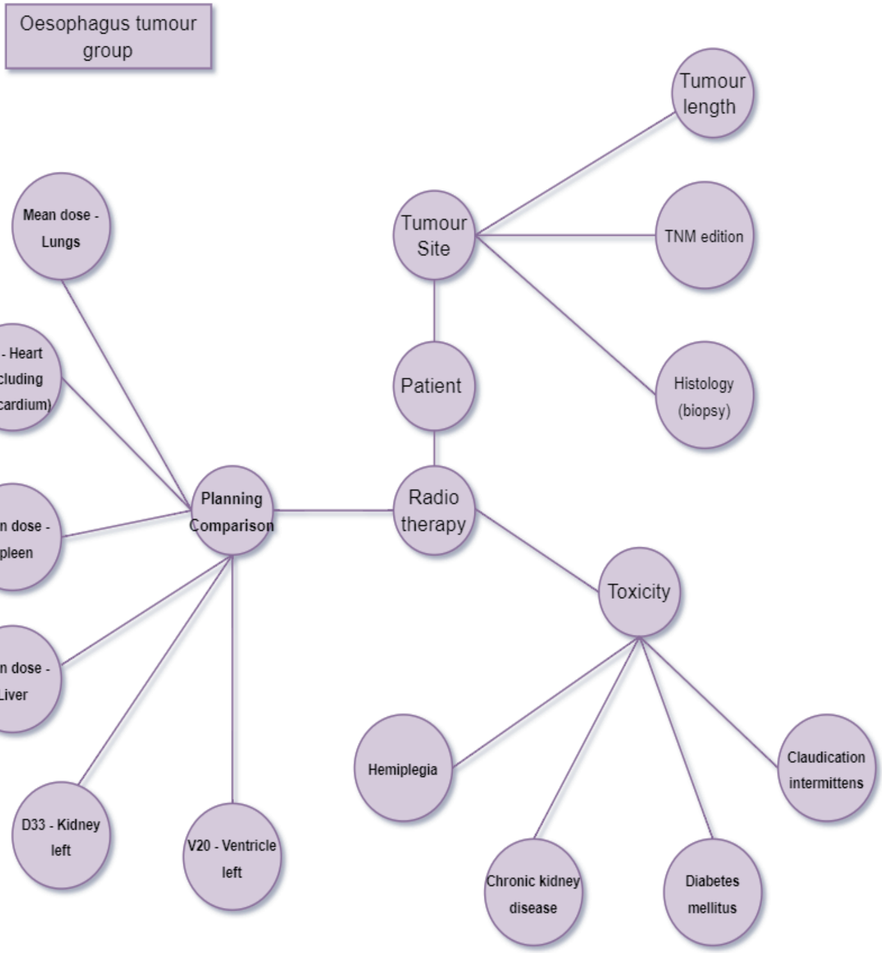


Figure S.1: Knowledge graph containing different characteristic clinical, toxicity and radiotherapy treatment variables from the standardised ProTRAIT oesophagus tumour group list (Item-lists/Lists/Definitieve CRF's OpenClinica/Oesophagus at main · ProTraitInfra/Item-lists · GitHub).

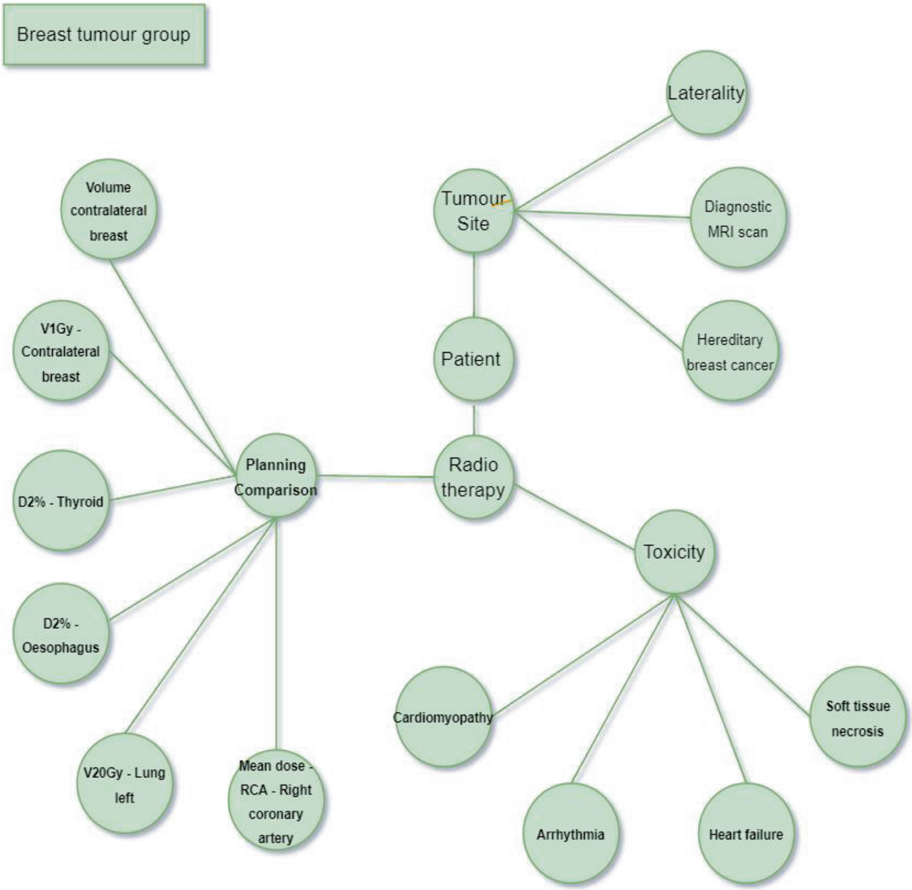


Figure S.2: Knowledge graph containing different characteristic clinical, toxicity and radiotherapy treatment variables from the standardised ProTRAIT breast tumour group list (Item-lists/Lists/Definitieve CRF's OpenClinica/Breast at main · ProTraitInfra/Item-lists · GitHub)

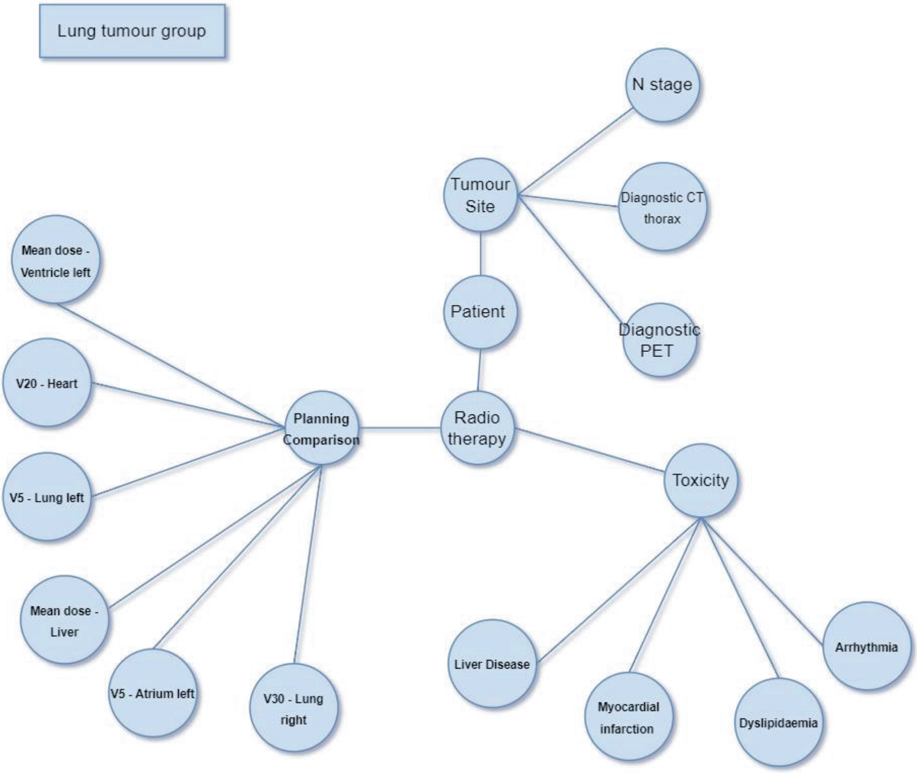


Figure S.3: Knowledge graph containing different characteristic clinical, toxicity and radiotherapy treatment variables from the standardised ProTRAIT lung tumour group list (Item-lists/Lists/Definitieve CRF's OpenClinica/Lung at main · ProTraitInfra/Item-lists · GitHub)

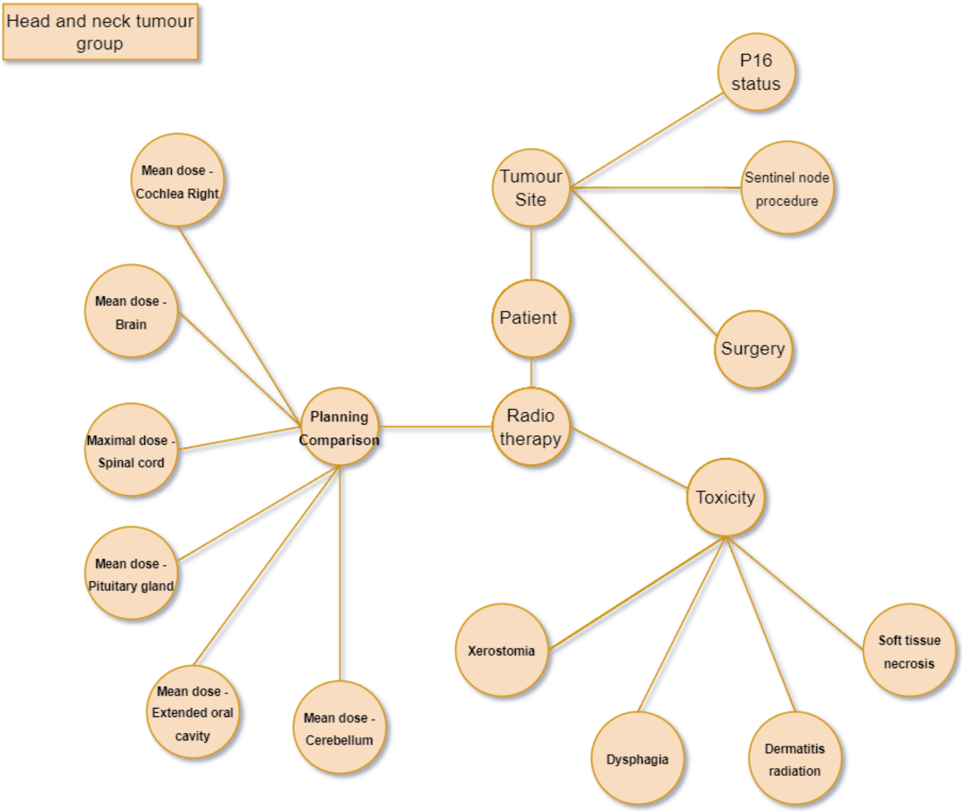


Figure S.4: Knowledge graph containing different characteristic clinical, toxicity and radiotherapy treatment variables from the standardised ProTRAIT head and neck tumour group list (Item-lists/Lists/Definitieve CRF's OpenClinica/Head and Neck at main · ProTraitInfra/Item-lists · GitHub)

Supplementary material bibliography

1. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017

2. National Cancer Institute Thesaurus Ontology. https://bioportal.bioontology.org/ontologies/NCIT

3. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys. 2018;45(10):e854-e862. doi:10.1002/mp.12879

4. Resource Description Framework (RDF). https://www.w3.org/RDF/

Chapter 11
Discussion, future perspectives and research impact

In this thesis we have investigated the current status of artificial intelligence (AI) applications in radiotherapy (RT) and have identified the pitfalls and barriers for their respective implementation in three RT domains; i) medical imaging (**Chapters 2-5**), ii) prediction modelling (**Chapter 6**) and iii) RT treatment planning quality assurance (QA) checks (**Chapters 7-8**). We focused on potential solutions that can play the role of “accelerators'' in the responsible and cautious implementation of AI in the RT clinical routine procedures. Of specific interest and focus has been the introduction of the Findable, Accessible, Interoperable and Reusable (FAIR) data principles[1] in the RT domain (**Chapter 8**). The FAIR data principles have the potential to establish a standardised framework that can tackle data-sharing difficulties that come with privacy-sensitive RT data. The basic components for the transformation of the multi-source RT data in a FAIR format are the radiation oncology related ontologies[2] and semantic web techniques[3]. These two components have the robustness to integrate and implement AI techniques as they can transform data toa machine-readable format, enabling interoperability and flexibility in RT data structures (**Chapters 9-10**).

As an extension of the thesis, this discussion chapter will provide an overview of five important points: 1) the different barriers and methodological pitfalls in the implementation of AI and particularly machine learning (ML) techniques using quantitative imaging features for the prediction of RT treatment related outcomes and providing potential solutions, 2) the importance of external validation of ML-based prediction models in RT, 3) the role of AI in the daily RT treatment planning QA checks, 4) FAIR data principles in RT and 5) the future perspectives for the significant step of the clinical integration of AI in the three different RT domains investigated in this thesis.

11.1 Artificial intelligence in radiotherapy clinical routine

In the last decade, AI has had some impressive breakthroughs that have solved challenging problems that existed for decades. The most impressive achievements of AI are in the fields of image recognition, voice transcription and translation between languages. These very impressive milestones have led to a justified excitement about the impact of those developments in the real world. The AI pipeline very naturally starts with the cognitive task assigned to the computers by humans. The collection process for input data to solve the learning problem follows, accompanied with statistics and the selected algorithm. The input data should be representative of the conditions that the AI system will encounter in the future. The final step is the model deployment after optimization steps to achieve the best performance.

These AI-based technologies will transform many different disciplines. Since AI tools and frameworks have been improved significantly and accurate predictions for well defined problems in a research environment are achieved, there should be a shift in focus toward the operationalisation of AI for its integration in clinical procedures. Because of the significant research and efforts in the RT domain[4], the clinical implementation of AI-based solutions in this domain has already been successful in some cases, such as the introduction of the automated RT treatment planning[5–7].

However there are important obstacles and technical challenges related to the robustness of AI. These challenges include the effects of biassed, incomplete and shifting input data, unknown dependencies in the input data that cannot be considered in the prediction model building and the context that underlie the input data. For example AI algorithms can be antagonised by similar AI algorithms resulting in opposing or different outcomes. An important obstacle regarding the robustness of AI techniques is its inability to change effectively between tasks and its inability to improvise, a task at which humans are very good. This difference between humans and machines is related to the fact that humans are good at transferring intuition from some learning tasks that they have already solved to some other new learning tasks that they do not have a lot of experience with. The application of AI algorithms trained in a specific source domain to a relatively different target domain has been investigated by the domain adaptation field in computer science[8]. Moreover, humans do not need a lot of training data to develop a new cognitive task, contrariwise computers need a lot of training data to transfer learning from a cognitive task they perform well to novel tasks that they do not have experience with.

A second important issue underlying the robustness of AI is causal inference. According to the personal opinion of the author of this thesis, understanding a clinical question-problem is not about being able to make predictions about it, but about the causal mechanisms that link different variables that underlie the clinical problem and are either observable or unobservable. A lot of emphasis has been placed on the ability of algorithms to make accurate predictions but there is no understanding on whether humans are even able to understand the causal mechanisms underlying the input data used in AI-based technologies. Undoubtedly, the clinical translation of AI also raises privacy and ethical issues that we need to understand establishing an intimate interaction between computer, data scientists and RT domain experts as discussed in **Chapter 8**. Taking into account the aforementioned obstacles, and specifically the emerging need to first understand the different data items variables used in AI techniques before implementing an AI algorithm, we established a Findable Accessible Interoperable and Reusable (FAIR) data principles framework using multi-source RT data (**Chapters 4-5,9-10**). The goal of our FAIR-based research is not only focused on the interoperability and reusability of RT data, but also on the need to fully understand the clear definitions, relationships and connections between the different data items used as an input of AI algorithms.

11.2 Modelling radiotherapy treatment outcomes using imaging features - Pitfalls and solutions

Medical imaging is one of the potential applications that AI can contribute to and revolutionise. The various imaging modalities used in combination with the rapid advances of the imaging technologies in RT result in large amounts of available data for research. This plethora of available medical images in the RT departments attracted the attention of researchers and clinicians and contributed to the introduction of the radiomics concept as a novel field of interest in the RT (AI) research community[9,10]. Radiomics refers to the comprehensive quantification of tumour phenotypes by extracting a large number of quantitative imaging derived features[10]. Radiomics has many potential uses, including predicting outcomes such as toxicities and overall survival of patients using ML algorithms[11] for personalised cancer treatment. Currently, the goal of the clinical translation of radiomics-based models has not yet been achieved. In the first part of this thesis, we presented the basic radiomics principles and described the barriers and methodological pitfalls that decelerate radiomics integration in the decision making process of clinicians. Specifically, in **Chapter 2** we investigated uncertainties related to the i) imaging data acquisition settings, ii) methodological ML-based pitfalls with respect to imaging features extraction and iii) the standardised radiomics infrastructure/pipeline that can be reproduced by multiple RT centres. For each of the above-mentioned uncertainty categories we provide a practical roadmap and standardisation guidelines as potential solutions that will accelerate the reproducibility and repeatability of radiomics studies.

11.2.1 Data acquisition during the imaging procedure

The majority of radiomics studies are based on retrospective imaging data analysis from clinical trials without a need for quantification of the measured features. There was therefore no need for standardisation in terms of imaging acquisition parameters as these images were mainly used to determine the size and staging of the tumours. Due to the still persisting lack of standardisation between various data generating scanners, the inclusion of retrospective imaging data from clinical trials in radiomics studies with the acquisition of prospective new datasets requires the detailed and adequate reporting of the imaging acquisition parameters.

Hence, the imaging data acquired from different scanner manufacturers in combination with different acquisition parameters is an important barrier in radiomics studies in terms of radiomics findings reproducibility. Consequently, data acquisition protocols, reconstruction algorithms and scanners vendor names are some of the settings that should be reported in every radiomics study. The appropriate selection, labelling and annotation of the source data can accelerate the translation of radiomics in a clinical environment. Taking into account the need for efficient reporting of the imaging acquisition settings in radiomics studies we provided a set of useful and practical principles regarding the image acquisition settings in **Chapter 2.**

11.2.2 Standardisation of imaging features extraction

Radiomics features are computed based on mathematical equations available in different sources in the literature[12]. This means that theoretically, the definition of each radiomics feature is mathematically defined. However, various open source and licensed features extraction radiomics software tools implement things differently, resulting in variations and inconsistencies in feature names, the selection of specific values for certain feature parameters and even some feature equations. Consequently, it is a reality that radiomics researchers have to choose among different open-source and licenced commercial software that have different feature settings and naming lexicons from images acquired from different imaging devices. Those variations have an impact in the comparison and reproducibility of radiomics studies when the feature definitions and the extraction parameters settings are not adequately reported, especially in a case of images acquired from different scanner vendors with different technical settings. To tackle these problems, we proposed in **Chapter 2** guidelines of radiomics features computation reporting accompanied with the radiomics computation workflow defined by the Image Biomarker Standardisation Initiative (IBSI)[13].The IBSI's effort has as a main goal to establish a standardised lexicon and definition of radiomics features with reporting guidelines advancing the reproducibility and external validation of radiomics based models. It is important to highlight that despite the significant efforts of the IBSI in the radiomics field, the reproducibility of radiomics studies is not ensured by using extraction software tested by the IBSI. Specifically, the different RT institutes with a radiomics research profile cannot fully harmonise and adapt an IBSI recommended radiomics pipeline due to the fact that their main goal is and always will be tumour detection and size determination for medical treatment. Nonetheless, the IBSI initiative constitutes an ideal solution for radiomics studies using prospective data. The real challenges are related to the harmonisation of specific settings that do not necessarily follow the default settings of the software used and the consistency of the maintenance of the different software versions that can possibly change in the future.

11.3 The role of publicly available datasets

Focusing on the problematic reproducibility of radiomics features[14] due to the barriers described in **Chapter 2** and taking into account the difficulties of data sharing across the different centres[15], we initiated a multi-centre data sharing study suitable for radiomics researchers with publicly available datasets in **Chapter 3.** Specifically, we provided computed tomography (CT) scans of two different phantoms scanned in three different RT centres in the Netherlands with varying imaging parameters such as the slice thickness of the CT images and the reconstruction algorithms. Our goal was to provide an annotated multi-centre dataset that enables reproducibility and standardisation of radiomics features which are fundamental requirements for the generalisability of radiomics-based models. Moreover, with this chapter we encourage the radiomics community to investigate the repeatability and reproducibility of imaging derived features identifying unstable feature values with respect to image acquisition settings. Expanding the work undertaken in **Chapter 3,** we introduced the concept of combining semantic web technologies and radiation oncology related ontologies with publicly available datasets in **Chapter 4**. Semantic web-based ontologies benefit the radiomics field by providing definitions of radiomics features and describing their calculation procedure. For instance, specific radiomics orientated ontologies such as the radiomics ontology (RO)[16] are mainly designed with the goal to introduce standardised reporting of radiomics features and parts of the technical parameters of the radiomics workflow.

In **Chapter 4,** our goal was to establish a publicly available semantic web-based interoperability framework applied to multi-source public radiomics related datasets. Moreover, with the application of semantic web technologies such as the Resource Description Framework (RDF) the data were transformed in a machine-readable format, which is a fundamental component for the implementation of ML techniques. Implementing the two above-mentioned components we offered to the radiomics community a set of four publicly available FAIR (Findable Accessible Interoperable and Reusable)[1] compliant radiomics datasets highlighting the benefits and flexibility of transforming and storing radiomics data in a FAIR principles-based data model, which enforces standardised terminologies. With the implementation of these FAIR compliant frameworks, federated learning studies are enabled as well as external validation of radiomics-based models.

In conclusion, a lot of effort from the radiomics research community is still necessary for proper clinical translation and usability of radiomics data and models. Besides the challenges presented in this thesis, there is an emerging need to understand the biological meaning of radiomics features and changes in features. As radiomics is mainly a data-driven technology, there are no insights regarding the impact of biological processes on the radiomics features and subsequently the model outputs. Therefore, we believe that fundamental research is necessary where not only radiomics is investigated, but also other (biological) data types such as genetic, clinical and histologic data, which could result in discovery of correlations between radiomics features and biological processes in the human body. These correlations will further prove the validity of radiomics as a tool for clinical use. Moreover, we would like to underline that the radiomics approach is not totally "cost-free" despite the large amount of seemingly readily available medical images in the RT departments. The input data acquisition, preparation and pre-processing steps for the data quality improvement are resource and labour intensive tasks assigned to the data engineers or data managers of the RT departments. Data sharing can be a potential solution addressing this barrier including the time consuming task of legal, ethical and administrative approvals that medical data sharing requires as less effort is required for the generation of new data.

11.4 Artificial Intelligence-based prediction modelling in radiotherapy - The need for external validation

During the previous years, a surge in the development of prediction model studies was observed in the RT domain, which is not limited to radiomics alone. The main goal of these studies was to clinically integrate prediction models that can be used for personalised RT, decision making, toxicity rate prediction and risk classification of patients. However, despite the significant amount of studies related to prediction models in RT during the last years (6098 studies between 2010-2020 according to Pubmed) and their potential and promising perspectives in patients' care, their clinical integration is not successful in most of the cases. One of the main reasons is the small number of externally validated prediction models in RT (295 studies between 2010-2020 according to PubMed). The term "external validation" is defined as the procedure of validating the original prediction model in an independent external patients cohort to determine whether the original model performs sufficiently. This external cohort ideally is a different population compared to the original in terms of geographical originality and treatment time period. External validation is needed to ensure reproducibility and repeatability of models due to ML's (especially supervised ML techniques) nature to perform perfectly in a development cohort, also known as overfitting. This results in a lack of generalizability as no cohort is ever a perfect representation of the true population[17]**.** This lack of external validation according to the scientific literature is one of the factors that leads to a lack of clinical integration of these prediction models that often do have clinical relevance. The complexity of AI algorithms is also a factor that retrogresses the external validation of AI-based prediction models as the AI algorithms might not be understandable by some RT professionals and considered as a "black-box" [18].

Ideally, a successful external validation of a prediction model is followed by a coordinated effort by physicians and researchers, to further investigate whether the prediction model can assist them in the daily clinical practice by improving patients' RT related outcomes or quality of life. For the second part of this thesis, in **Chapter 6** we performed an external validation study to highlight the significance of testing AI-based prediction models before their clinical integration. We investigated whether the Normal Tissue Complication Probability (NTCP) model for dysphagia performs well in an external validation dataset. The logistic regression-based NTCP model is used for the selection of head and neck cancer patients that will benefit more from proton therapy in the Netherlands according to the national protocol for proton therapy in the Netherlands. We showed that the assessment of AI-based models needs additional validation steps for the acquisition of reliable risk estimates even if an algorithm has a good discrimination (the ability of prediction algorithms to give a higher estimated risk to the patients with the event rather than the patients without the event).

Taking into account that some logistic regression based clinical prediction models may perform poorly when validated in external and independent cohorts compared to their training cohorts, they might require to be updated. Furthermore, due to the fact that different AI algorithms can be used in models that are poorly calibrated[19], we decided to follow a different validation approach as described by the study of Vergouwe et al.[20]. This study described the "closed testing procedure" consisting of three different levels of calibration of logistic regression models in the case of an external validation. The term calibration can be defined as the risk estimates accuracy on the agreement between the number of estimated and observed events. Inadequate or poor calibration can lead to unreliable predictions and a less clinically useful model[21,22]. A model can be recalibrated as follows: i) the re-estimation of the logistic regression model intercept (recalibration in the large), ii) the re-estimation of the logistic regression model intercept and slope (recalibration) and iii) the re-estimation of the logistic regression model predictors coefficients (model revision).

Using the model performance measure of likelihood ratio test, the closed testing procedure selects the most suitable method of calibration for the external validation cohort. It is suggested that in the case of a model revision an additional external validation dataset should be used for the model's generalisability evaluation while in the case of recalibration there is not a common consent[23]. In the case of recalibration in the large selection based on the closed testing procedure, there is no indication for using an additional external validation dataset[23]. Based on the above-mentioned indications, the implementation of the closed testing procedure becomes problematic for limited sample size datasets where model revision requires an additional patient cohort. Furthermore, there should be robust evidence on the effects of the updated predictors regarding the improved model performance before selecting the model revision as an updated method. Moreover, the model revision update method cancels the prognostic value of the original logistic regression model predictors leading to less robust predictions in the external validation cohort with an over-fitting model. In the case that the predicted risk of the logistic regression model is not fully represented by the predictor variables of the original model, the recalibration in the large method is suitable especially where there is a disagreement between the predicted and observed risks. In **Chapter 6,** the significance of external validation of AI-based models is highlighted and specifically the multivariable logistic regression models used for the selection of patients for proton therapy, we performed additional model update steps which are most of the times ignored[19] although it is recommended by the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) recommendations for prediction modelling studies[24].

11.5 Artificial Intelligence-assisted treatment planning quality assurance

Quality assurance (QA) checks are one of the most significant parts of the RT workflow for the efficient and safe delivery of a treatment plan to the patients. QA is a time consuming procedure for the medical physicists and radiation technologists. The detection of RT treatment planning errors before treatment takes place, is a labour intensive procedure that can cause delays in the daily clinical routine workflow of a RT department. In the third section of this thesis, in **Chapter 7**, we focused on the introduction of AI algorithms, specifically in Bayesian networks, in the treatment planning QA procedure. As a first step, in **Chapter 7,** we wanted to highlight the significance of the external validation of AI prediction models as a prerequisite for their clinical translation. In this chapter, we attempted to externally validate an existing Bayesian network model for the early detection of RT treatment planning errors that can assist RT professionals in their daily clinical routine. We used an independent Dutch cohort of patients, but were unsuccessful in externally validating the existing model. We concluded that some additional data pre-processing steps were needed in combination with a structure update of the Bayesian network developed in Washington in the United States of America (USA). As a next step of this study, we aim to establish a collaboration with the creators of the original model in the USA to develop and provide to the RT community with an externally validated and reproducible QA model that could predict treatment planning errors. Specifically, we aim to enrich the Bayesian network structure with dosimetric treatment planning variables that can improve the performance of the probabilistic model making it reusable by other RT centres.

Currently, the introduction of AI in the QA checks in the RT departments is hesitant as the majority of the RT professionals lack the basic data science or AI knowledge for the basic understanding of AI algorithms, although significant studies have demonstrated promising AI-based QA solutions[4,25]. This hesitancy stems from distrust due to the opaque nature of many of these models, which are generally described as "black boxes" by some of RT professionals (such as medical physicists) despite their high interest in AI projects in general[26]. The absence of basic data and computer science knowledge in the medical physicists education curricula is one of the main reasons for the lag in implementation of AI in the QA procedures and the absence of commercial AI-based solutions. Furthermore, the effort and time required for the acquisition of high quality datasets ("cleaned" data in the language of data scientists) in combination with the privacy regulations regarding the usage of RT data with patients' sensitive information are important barriers for testing and implementing the AI algorithms in multiple RT centres. All the aforementioned points highlight the importance of the strong collaboration between the different disciplines within a RT department. Data managers/stewards, computer scientists, medical physicists and clinicians should together establish a "round table" of discussion and collaboration for the robust implementation and development of AI methods in the QA workload.

11.6 The significance of knowledge exchange in radiotherapy studies

As we have mentioned in the **Introduction chapter,** there are important legal and administrative barriers[27] that hamper data exchange between different RT centres. For example, in many cases hospitals are not allowed or willing to share data due to concerns about patient privacy. Additionally, much of the data collected in a routine clinical care setting is not stored in a structured way, but generally in free text format. A result of these barriers is that it is difficult to receive and extract data from data archiving systems, which hampers the building and validation of AI models for relevant clinical outcomes and therefore knowledge exchange. These data sharing and usage difficulties led to the need to invest in privacy preserving technologies. A promising form of privacy-preserving technology for data sharing is federated learning, which constitutes a novel approach in the healthcare domain that has the potential to enable the exchange of statistical algorithms on a multi-centric level without the exchange of patients' data[28,29]. Specifically, secure local "data-stations" are used in each participating centre for the federated training of statistical ML-based algorithms without the exchange of patients' data, enabling multi-centre collaborations. **In Chapter 5,** we implemented and expanded the federated learning approach in the radiomics field using the infrastructure presented in **Chapter 4** with the transformation of radiomics data in a FAIR compliant format. We exchanged statistical models predicting the two-years overall survival as an outcome between two different Dutch RT centres following the privacy and security guidelines. Having as a goal to establish and promote a proof-of-concept framework for federated learning studies in the radiomics field rather than developing a novel radiomics model, we successfully validated the radiomics signature published by Aerts et al.[9] in an independent RT centre. Moreover, we verified and concluded that the federated radiomics model validation approach has indistinguishable results to the centralised approach without exchanging sensitive patient data.

11.7 The role of the FAIR data principles

In this thesis we provided an overview of the AI applications in three different RT domains (imaging, prediction modelling and treatment planning QA), investigating pitfalls and providing useful solutions. Across all these different domains, one challenge arises everywhere, namely the difficulty of exchanging data across different centres, which hampers implementation of these useful technological innovations.Furthermore, taking into account the need to primarily "understand" the data and the relationships between them before implementing an AI technique, we stressed the importance of implementing the FAIR data principles in RT research **Chapters 4-5, 9-10.** Based on the experience gained during the research executed in this thesis, regarding the data acquisition from the different data archive hospitals systems and the data structures that can support the FAIR data principles, we provided a FAIR data transformation practical guide presenting use cases that can be implemented in RT data in **Chapter 9**.This guide includes the different action points from the RT stakeholders for the responsible and coordinated clinical integration of AI methods in RT enabling the privacy preserving multi-centre collaborations.

The future of RT is strictly dependent on the coordinated and systematic efforts of the different professional parties involved, for the utilisation of the knowledge that multi-source RT data offer when adhering to the FAIR data principles. This transformation for the new era of RT research requires time and significant resources for the integration of this new technology to the hospitals data archive systems. Regarding resources, doing research that has the potential to be integrated into the clinic in most of the cases is not a reality without economic or technical resources. During the last years scientific journals[30] and research funding organisations[31] initiated requirements for FAIR data management plans in space science publications and grant applications respectively. The RT domain requires specific data manipulations in terms of data management plans and data usage policies by the different hospitals due to the privacy sensitivity of data. Although the definition of the FAIR data principles does not include a detailed guide regarding technical details of data usage and publication policy, it is necessary to adapt research policies in an institutional level that are aligned with the FAIR principles. An ideal propulsion for the implementation of the FAIR principles by the different data usage policies in an institutional level stems from the aforementioned requirements from the funding agencies. Therefore, it is more than necessary to rethink carefully the data management plans concerning the research ethical and sensitive data usage and their FAIR compliant transformation starting from the base of initiating a research project to its translation into the clinic.

The reusability of data constitutes one of the main aims of the FAIR principles. In the RT domain data can be reused for external parties in a case of an external validation multi-centre study. In this case, it is necessary (from a legal and ethical perspective) to establish data sharing agreements if multiple hospitals are involved in a research project. Data sharing agreements ensure the adequate documentation of compliance aspects of a multi-centre collaboration pr_oviding a common framework regarding the ethical and data protection requirements. Based on the different technical “FAIRification” infrastructures and the different national legal requirements the field should include common technical standards specifying the different data stages of the data “FAIRification” pipeline and the responsibilities of each party on the data sharing agreements. Industrial parties that may support technically a research project should be convinced to start from a standardised FAIR compliant framework that ensures the commercial protection of sensitive data.

It should be highlighted that the FAIR data principles are not a data standard as stated by the original publication[1]. On the contrary, they describe a set of fundamentals of data resources that have the potential to utilise data knowledge via data discovery and reusability. The principles are compatible with various approaches that have the potential to transform different data sources in a FAIR compatible format. In other words, different data standards can be developed based on the FAIR principles. One of the data standards that has the potential to represent individual patients records in a sustainable and flexible standard is the Fast Healthcare Interoperability Resources (FHIR) standard. It is a relatively new standard developed by the Health Level Seven International (HL7) organisation and constitutes a modern iteration of HL7 older versions (version 2.x and version 3.x.) and Clinical Document Architecture (CDA) protocols. The FHIR standard has uniform resource identifiers (URIs) based resources as basic elements and can be easily implemented as these resources are using the Hypertext Transfer Protocol (HTTP)-based Representational State Transfer (RESTful) protocol to communicate. Furthermore, the flexibility that the FHIR standard offers is reflected by the different formats of data representation supported, such as JavaScript Object Notation (JSON), extensible markup language (XML) or RDF. The initial aim of the FHIR standard is to facilitate interoperability between different healthcare electronic data capture (EDC) or electronic health record (EHR) systems, providing information to both healthcare providers and individual patients on different devices such as mobiles phones and computers. Currently, FHIR does not directly contain RT based resources as it lacks clinically validated profiles but there are significant ongoing efforts for modelling RT data in FHIR standard such as the federated learning privacy preserving approach of the Personal Health Train (PHT) using FAIR data[32]. The combination of the FAIR data principles and the FHIR data standard can establish a novel healthcare technology landscape in the future as interoperability between different health systems or devices is ensured using FHIR while technology becomes a valuable ally in discovering machine readable data by using the FAIR principles. Moreover, other technical architectures and data models such as the Observational Medical Outcomes Partnership (OMOP)[33] constitute an alternative to the RDF-based FAIR format approach proposed in this thesis. OMOP is mainly used for systematic analysis of observational databases using standardised terminologies and vocabularies. The RDF-based flexibility of the FAIR format in combination with its structured and standardised manner to collect observational data are some of the main factors that make the FAIR format suitable for RT studies.

11.8 Future perspectives

This thesis implemented and presented AI techniques in three RT subdomains (medical imaging, prediction modelling and treatment planning QA), discussing and identifying the barriers and gaps that hamper the clinical translation of AI-based clinical decision support systems. From the different chapters of this thesis, it is more than clear that multi-source data, including imaging, clinical, or demographics data, play a significant role in the development of AI applications in RT. The main challenge remains to fully and responsibly “exploit” the knowledge from multi-source data. Specific solutions such as the transformation of multi-source data in a FAIR format and the public availability of datasets are suggested from this thesis. The suggested FAIR transformation approach contains specific tools including publicly available radiation oncology ontologies in combination with rich metadata that efficiently describe the RT data.

In this thesis, we mainly used FAIRification tools that are exclusively focused on quantitative (imaging-clinical) data and metadata that have as a main goal to establish an interoperability framework enabling data reusability. However, there is an emerging need for the development of FAIRification tools that focus on qualitative data that potentially can support clinical or imaging data. For instance in the QA domain of RT delivery machines such as the LINACs there are numerous automatically registered parameters in the different manufacturer's log files that are mainly the cause of a machine breakdown. This amount of free text data can be potentially exploited and included in AI-based automatic QA checks in combination with an additional FAIRification step enabling interoperability among the different data users. Furthermore, taking into account the wide variety of applications where the FAIR data principles can be applied, it is worth highlighting that they need to be focused on clinically relevant RT applications such as the mapping of toxicities between different institutions or automated patients cohorts discovery implementing FAIR inclusion criteria that can improve the performance of prediction models.

Focusing on the first part of the thesis, the radiomics concept proved that it has the potential to be integrated into clinical practice regardless of the various gaps and pitfalls analysed by this work. Radiomics can be a useful cost-effective tool acting as a “clinician’s assistant” transforming the available medical images in every RT department into mineable knowledge using AI-based approaches. Based on my experience from the research conducted during the last three years, a significant gap exists between the radiomics researchers and clinicians. Although significant efforts have taken place in the radiomics research community for the standardisation and explainability of radiomics models, currently there is not a clinically integrated radiomics based model or a radiomics based clinical decision support system. It is still under negotiation in the RT community whether a robust and (externally) successfully validated radiomics model is robust enough for routine clinical use. According to my experience from the interaction with clinicians in AI-based research projects, they tend to prefer more transparent models as preliminary solutions (such as Bayesian networks) because more complex models are difficult to understand and validate. Therefore, the fact that an opaque model could statistically perform almost perfectly is not necessarily the most important factor for clinical use. Therefore, there is an emerging need to reconsider the whole radiomics concept and to foster a strong relationship between clinicians and all the other technical professionals involved by first understanding and defining the clinical question and requirements (such as model transparency). The introduction of basic data science/statistics educational initiatives focused on clinicians may be a possible starting point of this coordinated effort in the future. Moreover, the identification of systematic or statistical errors in the radiomics workflow is necessary for future radiomics studies. We, as a radiomics community, need to define a priority list of systematic errors that affect radiomics and subsequently prioritise the factors that can be controlled in radiomics analyses. For example, reducing or cataloguing scanner variability by using radiomics-specific phantoms to calibrate scanners and acquisition variability focusing on raw data (e.g. synograms) since most images are optimised for visual inspection, but not for automated analysis.

In conclusion, there is a need to re-think the way that AI-based prediction models are developed in RT. As mentioned in previous sections of this work, a significant percentage of clinicians consider AI as a "black-box" approach without understanding the basic technical and statistical concepts behind building a prediction model. Clinicians should have a leading role in the design of a prediction model as they fully understand the clinical question that has to be investigated with the involvement of technical/data science professionals due to their strong expertise in the field of statistics and mathematics. Future work in the field of prognostic modelling in RT should include the publication/availability of models in central repositories/archives adhering to the FAIR data principles having as a goal the continuous external validation of these models in different institutions and the investigation of patients cohorts or variables that fit to a specific clinical problem. This approach is significant especially for prediction models that are not published in the academic literature.

11.9 Conclusion

In conclusion, with this thesis we introduced and implemented AI-based techniques in three main domains in RT and the application of FAIR data principles in RT. There are still efforts needed from both clinicians and data science professionals for the clinical translation of AI for the improvement of patient care in the future. Every novel clinical solution design such as the implementation of the FAIR data principles in the clinic includes several challenges. For instance, institutional culture and resources distribution from the different stakeholders. Moreover, the FAIR initiative requires support from different AI sub-fields such as ML and Natural Language Processing (NLP) for data analysis or filtering.

Personally, as a data scientist with a clinical physics background, I would like to stress the significance of AI and its impact in our scientific approach as humans to improve patients' lives as a general final conclusion of this thesis. There is an ongoing discussion whether we, as scientists, made progress in understanding human intelligence at a higher level using AI by training machines to perform and automate tasks. By trying and succeeding to replicate a cognitive ability outside the human brain in some machines, humans understand that this particular cognitive task is not as hard as they thought. In the meantime, if humans fail to replicate this task with existing computational tools, there are insights given to them regarding the reasons that led to this failure. Specifically, computers sometimes are able to beat humans in some clinical tasks such as the delineation of anatomical targets in CT scans as mentioned in the introduction of this thesis. The delineation procedure is a hard and time consuming one.

However, the available mathematical and technical understanding we have in combination with hardware development are able to achieve better delineations than humans. This fact reveals that our mathematical understanding to perform an automatic tumour delineation in combination with computational power may give a solution and accelerate a clinical routine procedure. On the other hand, our inability to create algorithms that can transfer learning from one cognitive task to another and the “data hungriness” of the AI algorithms, which is much higher than that of humans who are very good to adapt their skills from one task to the other, show an important ability of the human brain that we do not know how to replicate outside of it. That is a natural target for us, as scientists, to investigate. In other words, our access to hardware and data and our ability to expedite this trial and error process trying to develop cognitive abilities outside the human brain, sheds light on which cognitive tasks are tractable and not tractable and require paradigm changes to “attack” them. This aforementioned fact progresses and develops the human brain by using and implementing AI and therefore improves our scientific way of solving a clinical problem as RT scientists.

Bibliography

1. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. doi:10.1038/sdata.2016.18
2. Phillips MH, Serra LM, Dekker A, et al. Ontologies in radiation oncology. Physica Medica. 2020;72:103-113. doi:10.1016/j.ejmp.2020.03.017
3. Soest J van, Choudhury A, Gaikwad N, Sloep M, Dumontier M, Dekker A. Annotation of Existing Databases using Semantic Web Technologies: Making Data more FAIR. In: CEUR-WS; 2019:94-101. http://ceur-ws.org/Vol-2849/#paper-11
4. Vandewinckele L, Claessens M, Dinkla A, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiotherapy and Oncology. 2020;153:55-66. doi:10.1016/j.radonc.2020.09.008
5. Han EY, Cardenas CE, Nguyen C, et al. Clinical implementation of automated treatment planning for whole-brain radiotherapy. J Appl Clin Med Phys. 2021;22(9):94-102. doi:10.1002/acm2.13350
6. Yoo S, Sheng Y, Blitzblau R, et al. Clinical Experience With Machine Learning-Based Automated Treatment Planning for Whole Breast Radiation Therapy. Advances in Radiation Oncology. 2021;6(2):100656. doi:10.1016/j.adro.2021.100656
7. Netherton TJ, Cardenas CE, Rhee DJ, Court LE, Beadle BM. The Emergence of Artificial Intelligence within Radiation Oncology Treatment Planning. Oncology. 2021;99(2):124-134. doi:10.1159/000512172
8. Redko I, Morvant E, Habrard A, Sebban M, Bennani Y. Advances in Domain Adaptation Theory. ISTE Press Ltd ; Elsevier Ltd; 2019.
9. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5(1):4006. doi:10.1038/ncomms5006
10. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141
11. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. Front Oncol. 2015;5. doi:10.3389/fonc.2015.00272
12. Depeursinge A, Foncubierta-Rodriguez A, Van De Ville D, Müller H. Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities. Medical Image Analysis. 2014;18(1):176-196. doi:10.1016/j.media.2013.10.005
13. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328-338. doi:10.1148/radiol.2020191145
14. Berenguer R, Pastor-Juan M del R, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407-415. doi:10.1148/radiol.2018172361

15. Packer M. Data sharing in medical research. BMJ. Published online February 14, 2018:k510. doi:10.1136/bmj.k510
16. A.Traverso et al. The Radiomics Ontology (RO): standardizing radiomic studies following FAIR principles-Manuscript in preparation. Published online 2020.
17. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology. 2015;68(1):25-34. doi:10.1016/j.jclinepi.2014.09.007
18. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. The Lancet. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6
19. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005
20. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. Stat Med. 2017;36(28):4529-4539. doi:10.1002/sim.7179
21. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004
22. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40. doi:10.1186/1471-2288-14-40
23. Van den Bosch L, Schuit E, van der Laan HP, et al. Key challenges in normal tissue complication probability model development and validation: towards a comprehensive strategy. Radiotherapy and Oncology. 2020;148:151-156. doi:10.1016/j.radonc.2020.04.012
24. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350(jan07 4):g7594-g7594. doi:10.1136/bmj.g7594
25. Kalet AM, Luk SMH, Phillips MH. Radiation Therapy Quality Assurance Tasks and Tools: The Many Roles of Machine Learning. Med Phys. 2020;47(5). doi:10.1002/mp.13445
26. Diaz O, Guidi G, Ivashchenko O, Colgan N, Zanca F. Artificial intelligence in the medical physics community: An international survey. Physica Medica. 2021;81:141-146. doi:10.1016/j.ejmp.2020.11.037
27. Annas GJ. HIPAA Regulations — A New Era of Medical-Record Privacy? N Engl J Med. 2003;348(15):1486-1490. doi:10.1056/NEJMlim035027
28. Zerka F, Barakat S, Walsh S, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clinical Cancer Informatics. 2020;(4):184-200. doi:10.1200/CCI.19.00047
29. Kirienko M, Sollini M, Ninatti G, et al. Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI. Eur J Nucl Med Mol Imaging. Published online April 13, 2021. doi:10.1007/s00259-021-05339-7
30. Announcement: FAIR data in Earth science. Nature. 2019;565(7738):134-134. doi:10.1038/d41586-019-00075-3

31. Notice Announcing Funding Opportunity Issued for the NIH Data Commons Pilot Phase. https://grants.nih.gov/grants/guide/notice-files/NOT-RM-17-031.html
32. Choudhury A, van Soest J, Nayak S, Dekker A. Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee A, Borgohain SKr, Soni B, Verma G, Gao XZ, eds. Machine Learning, Image Processing, Network Security and Data Sciences. Vol 1240. Communications in Computer and Information Science. Springer Singapore; 2020:85-95. doi:10.1007/978-981-15-6315-7_7
33. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153(9):600. doi:10.7326/0003-4819-153-9-201011020-00010

Thesis summary

In recent years, research on artificial intelligence (AI) in the radiotherapy (RT) domain has begun to monopolise the interest of different clinicians and researchers.This will likely make routine RT clinical procedures more reliable and effective with a direct benefit to the patients. However, the clinical integration of these AI-based applications is still lacking as several steps still need to be included in the AI clinical implementation roadmap starting from the research stage and ending up to the clinical integration. One of these crucial steps is the adoption of the FAIR (Findable, Accessible, Interoperable, Reusable) data principles among the different data structure/archive systems of the radiotherapy centres. The FAIR principles can establish an interoperability framework for the reusability of RT multisource data that can potentially decrypt valuable information for prognostic or diagnostic research and clinical purposes.

This thesis focuses on the introduction of AI techniques and the FAIR data principles in four RT subdomains:

i) medical imaging, ii) prediction modelling or RT related outcomes, iii) quality assurance (QA) of RT treatment planning and iv) the implementation of the FAIR data principles.

Medical imaging

In the first part of the thesis we focused on the medical imaging domain. We investigated how RT can be transformed into a personalised treatment modality using imaging derived features that can decrypt valuable information that is not visible by the human eyes using machine learning (ML) techniques (ie. radiomics). Specifically, in **chapter 2** we provided a roadmap for the clinical implementation of radiomics-based prediction models in the clinic, identifying the pitfalls and uncertainties encountered in the radiomics methodology/pipeline followed by the researchers. Furthermore, based on the pitfalls and uncertainties we presented, we proposed a standardisation framework with all the necessary technical aspects that should be taken into account in the design/development of a radiomics study.

One main take-away from **chapter 2** is that the reproducibility of radiomics studies/findings is one the main requirements for the standardisation and implementation of the radiomics concept in the clinic. Therefore, in **chapter 3** we provided a publicly available dataset consisting of Computer Tomography (CT) phantoms (suitable for radiomics studies) scans from three different Dutch RT centres, having as a goal to promote the reproducibility and interoperability of radiomics studies.

Prediction modelling or RT related outcomes

In the second part of the thesis we focused on the prediction modelling of RT outcomes using AI algorithms. Taking into account the patients data privacy regulations, in **chapter 5**, we implemented a radiomics-based federated decentralised multicentre study, using tNon Small Cell Lung Cancer (NSCLC) patients. Having as a base the FAIR transformed clinical and radiomics features-based data, we validated a radiomics signature that predicts the 2-years overall survival of NSCLC patients without the exchange of patients data, comparing the results of the centralised analysis. The study showed that the performance of the radiomics signature of the decentralised approach was not significantly different than the centralised one. Furthermore, with this study, the significance of the FAIR transformation of data for federated learning radiomics studies is underlined, implementing a privacy preserving infrastructure without the data exchange of patients.

In **chapter 6**, we performed an independent validation of the logistic regression-based **N**ormal **T**issue **C**omplication **P**robability (NTCP) model that predicts the six months ≥ 2nd grade dysphagia for head and neck cancer patients. This model is part of the Dutch **N**ational **I**ndication **P**rotocol for **P**roton therapy (NIPP) for the selection of patients candidates for proton therapy (PT). In this study, we showed that logistic regression models need a specific validation approach in independent cohorts, examining the potential update of the different components of the logistic regression models (intercept, slope and predictor coefficients) through the AI-based approach of the closed testing procedure (CTP). The CTP in combination with the graphical assessment of the calibration curves of the different updated models, indicated that the original dysphagia NIPP model needed an update and a new updated calibrated model was selected. However, it is important to perform a federated privacy preserving multicentre study using FAIR transformed datasets in the different Dutch PT centres that can robust the results of the CTP.

Quality assurance

The RT treatment planning is a complicated procedure that requires the coordinated efforts of clinical and technical RT professionals. For the third part of the thesis, In **chapter 7** we focused on the quality assurance (QA) of the RT treatment planning procedure by externally validating an AI-based method developed in the United States (US) using a Dutch independent patients cohort. This AI method using Bayesian Networks (BNs) has as a goal to early detect errors encountered in the verification phase of RT treatment planning and alert humans for possible erroneous variables included on it. The external validation using an independent Dutch patients cohort was not successful, due to the different technical characteristics of the treatment machines and software used in the different RT centres possibly. According to this study, further steps are required for the generalisability and reusability of AI-based systems focusing on the automatic error detection among different centres such as data preprocessing and the inclusion of more variables included in the treatment planning such as imaging-based data.

Implementation of FAIR principles

In **chapter 4,** expanding the work of **chapter 3,** we provided a set of four publicly available datasets that were used in a breakthrough publication in the radiomics community in 2014 in a machine-readable format. Specifically, using radiation oncology related ontologies and semantic web technologies, we transformed multisource (clinical, radiomics feature-based, and imaging) data in a FAIR format for enabling the automation of data processing by the machines with minimal human intervention.

For the last part of the thesis, we focused on the implementation of the FAIR principles in the RT domain. In **chapter 8**, having as a goal to introduce the FAIR concept in the RT community, we provided an overview of the action points required from the different RT stakeholders for the introduction and adoption of the FAIR principles in the different data archive systems of the hospitals. Some of the advantages that a FAIR data transformation offers are the flexibility to adapt databases and the ability of machines to "read" the different data for automated AI-based studies. In **chapters 9-10**, we provided FAIR data models structured in knowledge graphs using the data elements used for the Dutch national registry of patients candidates for PT. Using publicly available ontologies and semantic web technologies we underline the significance of the flexibility and interoperability of the FAIR format using routine clinical data.

Scriptie samenvatting

In de afgelopen jaren is de interesse van verschillende clinici en onderzoekers in onderzoek naar kunstmatige intelligentie (AI) in het gebied van radiotherapie (RT) aan het toenemen. Dit omdat het waarschijnlijk routinematige RT-klinische procedures betrouwbaarder en effectiever zal maken met een direct voordeel voor de patiënten. De klinische integratie van deze AI-gebaseerde applicaties ontbreekt echter nog, aangezien er nog verschillende stappen moeten worden genomen in de AI-roadmap voor klinische implementatie, beginnend bij de onderzoeksfase en eindigend met de klinische integratie. Een van deze cruciale stappen is de invoering van de FAIR (Findable, Accessible, Interoperable, Reusable) data principes binnen de verschillende datastructuur/archiefsystemen van de radiotherapiecentra. De FAIR-principes kunnen een interoperabiliteits kader tot stand brengen ten behoeve van de herbruikbaarheid van multi source RT gegevens die mogelijk waardevolle informatie kunnen ontsluiten voor prognostisch of diagnostisch onderzoek en klinische doeleinden.

Dit proefschrift richt zich op de introductie van AI-technieken en de FAIR-data principes in vier RT-subdomeinen:

i) medische beeldvorming, ii) voorspellen van RT-gerelateerde uitkomsten, iii) kwaliteitsborging (QA) van RT-behandelingsplanning en iv) de implementatie van de FAIR-gegevens principes.

Medische beeldvorming

In het eerste deel van het proefschrift hebben we ons gericht op het domein van de medische beeldvorming. We onderzochten hoe RT kan worden getransformeerd tot een gepersonaliseerde behandelingsmodaliteit met behulp van variabelen uit medische beelden die waardevolle informatie met behulp van machine learning (ML) technieken (dwz radiomics) kunnen ontsluiten die niet zichtbaar is voor het menselijk oog. In hoofdstuk 2 hebben we specifiek een routekaart gegeven voor de klinische implementatie van op radiomics gebaseerde voorspellingsmodellen in de kliniek, waarbij we de valkuilen en onzekerheden identificeren die we tegenkwamen in de radiomics methodologie/pijplijn. Daarnaast, hebben we op basis van de door ons gepresenteerde valkuilen en onzekerheden een standaardisatie raamwerk voorgesteld met alle noodzakelijke technische aspecten waarmee rekening moet worden gehouden bij het ontwerp/de ontwikkeling van radiomics-onderzoek.

Een belangrijke conclusie uit hoofdstuk 2 is dat de reproduceerbaarheid van radiomics-onderzoeken/bevindingen een van de belangrijkste vereisten is voor de standaardisatie en implementatie van het radiomics-concept in de kliniek. Daarom hebben we in hoofdstuk 3 een publiek beschikbare dataset gepresenteerd bestaande uit computertomografie (CT) fantomen (geschikt voor radiomics-onderzoeken) scans van drie verschillende Nederlandse RT-centra, met als doel de reproduceerbaarheid en interoperabiliteit van radiomics-onderzoeken te bevorderen.

Voorspellen van RT-gerelateerde uitkomsten

In het tweede deel van het proefschrift hebben we ons gericht op de voorspellen van RT-uitkomsten met behulp van AI-algoritmen. Rekening houdend met de privacyregelgeving met betrekking tot patiënten gegevens, hebben we in hoofdstuk 5 een op radiomics gebaseerde federatieve gedecentraliseerde multicenter onderzoek geïmplementeerd met Non Small Cell Lung Cancer (NSCLC) patiënten. Op basis van FAIR getransformeerde klinische en op radiomics gebaseerde gegevens, hebben we een radiomics-profiel gevalideerd.Dit profiel voorspelt de algemene overleving van 2 jaar van NSCLC-patiënten voorspelt zonder de uitwisseling van patiëntgegevens en hebben we deze resultaten vergeleken met een gecentraliseerde analyse. Deze studie toonde aan dat de prestatie van de radiomics-profiel van de gedecentraliseerde aanpak niet significant verschilde van die van de gecentraliseerde. Bovendien wordt met deze studie het belang van de FAIR-transformatie van gegevens voor federatief leren in radiomics-onderzoeken onderstreept, bij een privacy beschermende infrastructuur wordt geïmplementeerd zonder de gegevensuitwisseling van patiënten.

In hoofdstuk 6 hebben we een onafhankelijke validatie uitgevoerd van het Normale Weefsel Complicatie Probabiliteit (NTCP) model gebaseerd op logistische regressie dat de zes maanden ≥ 2e graad dysfagie voorspelt voor hoofd-halskankerpatiënten. Dit model maakt deel uit van de Nederlandse “Landelijk Indicatie Protocol Protonentherapie” (LIPP) voor de selectie van patiënten die in aanmerking komen voor protonentherapie (PT) . In deze studie hebben we aangetoond dat logistische regressiemodellen een specifieke validatie benadering nodig hebben in onafhankelijke cohorten, waarbij de mogelijke update van de verschillende componenten van de logistische regressie modellen (intercept, helling en voorspeller coëfficiënten) wordt onderzocht via de AI-gebaseerde benadering van de gesloten testprocedure (CTP). De CTP in combinatie met de grafische beoordeling van de kalibratiecurves van de verschillende bijgewerkte modellen, gaf aan dat het oorspronkelijke dysfagie-LIPP-model een update nodig had en dat er een nieuw bijgewerkt gekalibreerd model werd gekozen. Het is echter belangrijk om een federatief privacy behoudend multicenter onderzoek uit te voeren met behulp van FAIR-getransformeerde datasets in de verschillende Nederlandse PT centra die de resultaten van de CTP kunnen robuust maken.

Kwaliteitsverzekering

De planning van de RT-behandeling is een gecompliceerde procedure die de gecoördineerde inspanningen vereist van klinische en technische RT-professionals. Voor het derde deel van het proefschrift, in hoofdstuk 7 , hebben we ons gericht op de kwaliteitsborging (QA) van de procedure voor het plannen van RT-behandeling door extern een op AI gebaseerde methode te valideren die is ontwikkeld in de Verenigde Staten (VS) met behulp van een Nederlands onafhankelijk patiënten cohort. Deze AI-methode die gebruikmaakt van Bayesian Networks (BN's) heeft als doel fouten in de verificatiefase van de RT-behandelingsplanning vroegtijdig te detecteren en mensen te waarschuwen voor mogelijke foutieve variabelen die erop zijn opgenomen. De externe validatie met behulp van een onafhankelijk Nederlands patiënten cohort was niet succesvol, mogelijk vanwege de verschillende technische kenmerken van de behandel machines en software die in de verschillende RT-centra worden gebruikt. Volgens deze studie zijn verdere stappen nodig voor de generaliseerbaarheid en herbruikbaarheid van op AI gebaseerde systemen, waarbij de nadruk ligt op de automatische foutdetectie tussen verschillende centra, zoals gegevens voor verwerking en het opnemen van meer variabelen in de behandelplanning, zoals op beeldvorming gebaseerde gegevens.

Implementatie van FAIR-principes

In hoofdstuk 4, dat het werk van hoofdstuk 3 voortzetten, hebben we een set van vier publiek beschikbare datasets gepresenteerd die in 2014 in een invloedrijke publicatie in het radiomics gemeenschap werden gebruikt in een machineleesbaar formaat. In het bijzonder hebben we met behulp van RT specifieke ontologieën en semantische webtechnologieën gegevens uit meerdere bronnen (klinische, op radiomics gebaseerde functies en beeldvorming) getransformeerd in een FAIR formaat om de automatisering van gegevensverwerking door de machines mogelijk te maken met minimale menselijke tussenkomst.

Voor het laatste deel van het proefschrift hebben we ons gericht op de implementatie van de FAIR-principes in het RT domein. Met als doel om het FAIR-concept in de RT-gemeenschap te introduceren, hebben we in hoofdstuk 8 een overzicht gegeven van de actiepunten die nodig zijn van de verschillende RT-stakeholders voor de introductie en adoptie van de FAIR-principes in de verschillende data-archiefsystemen van de ziekenhuizen. Enkele voordelen die een FAIR transformatie biedt, zijn de flexibiliteit om databases aan te passen en het vermogen van machines om de verschillende gegevens te "lezen" voor geautomatiseerde, op AI gebaseerde onderzoeken. In de hoofdstukken 9-10 hebben we FAIR gegevensmodellen geleverd die zijn gestructureerd in “knowledge graphs” met behulp van de gegevenselementen die worden gebruikt voor de Nederlandse nationale registratie van patiënten die in aanmerking komen voor PT. Met behulp van openbaar beschikbare ontologieën en semantische webtechnologieën onderstrepen we het belang van de flexibiliteit en interoperabiliteit van het FAIR formaat met behulp van data verzameld in de dagelijkse klinische praktijk.

Research impact

Sociocultural impact

In this thesis we proposed and analysed AI solutions that can shorten the gap between research and clinical implementation of AI-based applications and identified specific barriers. These barriers significantly delay the clinical translation of AI research findings. This delay has a negative impact, not only to the RT clinicians, but also to the patients. The continuous challenge and goal of RT is to improve patient care by providing the best treatment option fitted to an individual's life and disease. The most important question of patients to the clinicians when they are diagnosed with a cancer related disease is whether they will survive or how much survival time they have. The new trend of personalised medicine having not only a curative intent but also aiming to decrease radiation induced toxicities/implication should be aligned with the modern technological AI advancements. The current evidence for treatment choices originates from clinical trials that investigate a well-defined subset of the larger patient population according to the requirements of the trial design. Generally, the insights gained from these clinical trials are only applicable to a small percentage of the patient populations with similar characteristics to those in the trial. The exchange of data from all patients rather than a subset and subsequent building of prediction models in a privacy preserving manner constitutes a solution of unlocking the potential of medical imaging and big data in RT. In this thesis, the exchange of biomarkers extracted from medical images enriched with clinical data FAIR compliant data is presented. This AI-application can potentially support clinicians in the decision-making process having a promising future impact in patient care.

Economical and technological impact

During the last decades due to technological and medicine related advancements, different novel treatment options, such as brachytherapy, immunotherapy, proton RT and FLASH RT(ultra high doses of radiation), have been made available to cancer patients. Clinicians have a plethora of available treatments in their hands to implement. However, many of these novel treatment options are costly and time consuming. This is a problem for every national health system in the world. Hospitals board members and clinicians that need to find an equilibrium between cost effectiveness and the most efficient treatment for patients, have difficulties in selecting the best treatment modality. In this thesis, we presented the model based approach (MBA) comparing the photon and proton dosimetric differences in terms of normal tissue complication probability (NTCP) for proton therapy in the Netherlands which is one of the high cost treatment options. In this way, the patients that benefit the most from proton therapy are selected while those who will respond well to conventional therapy are offered that. Through this selection, significant resources are saved for future proton treatments. Specifically, according to Peeters et al.[1] the construction cost of a proton therapy centre can be four times higher in terms of a capital investment while the operational cost can be three times bigger per fraction compared to photon therapy. The aforementioned numbers stress the significance and necessity of a "mechanism" such as the model based approach that can clinically benefit cancer patients.

Moreover, this thesis presented and implemented a framework for the adoption of the FAIR data principles in RT involving different stakeholders in the domain. The FAIR approach does not only have a significant impact on the way researchers store and transform the different data sources. There is an important economical impact in terms of data loss and data archive costs. According to the European Commission the annual cost of not transforming research data in a FAIR format reaches the “astronomic” amount of 10.2 billion euros[2]. In this cost, different parts of a research pipeline implementation are taken into account such as storage and licence costs. Although a FAIR compliant software is not developed in this thesis, we presented a FAIR compliant framework using RT multi-source data that can have a significant impact on the different RT stakeholders to implement the FAIR concept in the hospital systems changing the way that science and research is conducted.

Bibliography

1. Peeters A, Grutters JPC, Pijls-Johannesma M, et al. How costly is particle therapy? Cost analysis of external beam radiotherapy with carbon-ions, protons and photons. *Radiotherapy and Oncology*. 2010;95(1):45-53. doi:10.1016/j.radonc.2009.12.002
2. *Cost of Not Having FAIR Research Data*. http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

List of manuscripts

Published original research (*contributed equally)

1. **Kalendralis, P**., Traverso, A., Shi, Z., Zhovannik, I., Monshouwer, R., Starmans, M. P., ... & Wee, L. (2019). Multicenter CT phantoms public dataset for radiomics reproducibility tests. Medical physics, 46(3), 1512-1518.
2. **Kalendralis, P.,** Shi, Z., Traverso, A., Choudhury, A., Sloep, M., Zhovannik, I., ... & Wee, L. (2020). FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. Medical Physics, 47(11), 5931-5940.
3. **Kalendralis, P**., Sloep, M., van Soest, J., Dekker, A., & Fijten, R. (2021). Making radiotherapy more efficient with FAIR data. Physica Medica, 82, 158-162.
4. **Kalendralis, P**., Eyssen, D., Canters, R., Luk, S. M., Kalet, A. M., van Elmpt, W., ... & Bermejo, I. (2021). External validation of a Bayesian network for error detection in radiotherapy plans. IEEE Transactions on Radiation and Plasma Medical Sciences, 6(2), 200-206.
5. Sloep, M.*, **Kalendralis, P.***, Choudhury, A., Seyben, L., Snel, J., George, N. M., ... & Fijten, R. (2021). A knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry. Clinical and translational radiation oncology, 31, 93-96.
6. Foley, K. G., Shi, Z., Whybra, P., **Kalendralis, P.**, Larue, R., Berbee, M., ... & Spezi, E. (2019). External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiotherapy and Oncology, 133, 205-212.
7. Shi, Z., Zhang, C., Welch, M., **Kalendralis, P.**, Wee, L., & Dekker, A. (2019, April). CT-based Radiomics Predicting HPV Status in Head and Neck Squamous Cell Carcinoma. In Radiotherapy and Oncology (Vol. 133, pp. S515-S515).
8. Traverso, A., Kazmierski, M., Shi, Z., **Kalendralis, P.**, Welch, M., Nissen, H. D., ... & Wee, L. (2019). 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.
9. Shi, Z., Zhovannik, I., Traverso, A., Dankers, F. J., Deist, T. M., **Kalendralis, P**., ... & Wee, L. (2019). Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Scientific data, 6(1), 1-8.
10. Zhovannik, I., Bussink, J., Traverso, A., Shi, Z., **Kalendralis, P**., Wee, L., ... & Monshouwer, R. (2019). Learning from scanners: Bias reduction and feature correction in radiomics. Clinical and translational radiation oncology, 19, 33-38.
11. Zhai, T. T., Wesseling, F., Langendijk, J. A., Shi, Z., **Kalendralis, P.**, van Dijk, L. V., ... & Sijtsema, N. M. (2021). External validation of nodal failure prediction models including radiomics in head and neck cancer. Oral Oncology, 112, 105083.

12. Shi, Z., Zhang, C., **Kalendralis, P.**, Whybra, P., Parkinson, C., Berbee, M., ... & Foley, K. G. (2021). Prediction of lymph node metastases using pre-treatment PET radiomics of the primary tumour in esophageal adenocarcinoma: an external validation study. British Journal of Radiology, 94(1118).
13. Jha, A. K., Mithun, S., Sherkhane, U. B., Jaiswar, V., Shi, Z., **Kalendralis, P.**, ... & Dekker, A. (2021). Implementation of Big Imaging Data Pipeline Adhering to FAIR Principles for Federated Machine Learning in Oncology. IEEE Transactions on Radiation and Plasma Medical Sciences.

Submitted/In preparation

1. **Kalendralis, P.**, Sloep, M., Choudhury, A., Seyben, L., Snel, J., George, N. M, Veugen, Veening M., Langendijk H., van Soest, J., Dekker, A., & Fijten, R. (2022). A knowledge graph approach to registering tumour specific data of patients-candidates for proton therapy in the Netherlands., **Submitted to Medical Physics (https://osf.io/hynk2)**
2. **Kalendralis, P.**, Sloep, M., Choudhury, A., Seyben, L., Snel, J., George, N. M, Veugen, J., Hoebers, F., Wesseling, F., Unipan, M., Veening M., Langendijk H., van Soest, J., Dekker, A., & Fijten, R. (2022). Independent validation of dysphagia NTCP model for the selection of head and neck cancer patients for proton therapy in the Netherlands., **Submitted to Physics and Imaging in Radiation Oncology (phiRO) (https://osf.io/rt7bs)**
3. **Kalendralis, P.**, Vallières, M., Kann, B., Sanjay, A., Rashid, A., Dekker, A., & Fijten, R. Book chapter: Radiomics: “Unlocking the potential of medical images for precision radiation oncology” (2022) **Submitted to Elsevier**
4. **Kalendralis, P**., Eyssen, D., Canters, R., Luk, S. M., Kalet, A. M., van Elmpt, W., ... & Bermejo, I., Automatic quality assurance in radiotherapy treatment plans using Bayesian Networks-A multicentre study (In preparation)
5. Hasannejadasl, H., Roumen, C., van der Poel, H., Vanneste, B., **Kalendralis, P**., van Roermund, J., Aben, K., ... & Fijten, R. R. (2021). Predicting erectile dysfunction after treatment for localized prostate cancer. arXiv preprint arXiv:2110.00615.
6. Swart R., Fijten R., Boersma L., **Kalendralis, P**., Behrendt M., Ketelaars M., Cheryl Roumen C., Jacobs M. (2022). External validation of a prediction model for timely implementation of innovations in radiotherapy., **Submitted to Radiotherapy & Oncology.**

International Conferences abstracts and presentations (first authorship)

1. Report back from ESTRO mobility grants physics: Modelling Head and Neck Radiotherapy outcomes using radiomics biomarkers, ESTRO 38. **(Oral Presentation-Invite speaker)**
2. Multicenter CT phantoms public dataset for radiomics reproducibility studies, ESTRO 38.**(Poster)**
3. Public radiomics data collections in an open access Semantic Web (SPARQL) endpoint, ESTRO 2020.**(Poster)**

4. Modelling Head and Neck Radiotherapy outcomes using radiomics biomarkers, Young Scientists' Forum 2019 **(Oral Presentation)**
5. FAIRifying clinical data for proton radiotherapy patients, ESTRO physics workshop 2019**.(Oral Presentation)**
6. Bayesian network model for the early detection of errors in radiotherapy treatment plans, Winter Institute of Medical Physics 2021. **(Poster)**
7. A knowledge graph based on the proton therapy model based approach for head and neck patients, Semantic Web Applications and Tools for Health Care and Life Sciences conference 2022 **(Poster)**
8. Retraining of a Bayesian network for the detection of radiotherapy plan errors, ESTRO 2022**(Oral Presentation)**
9. A federated learning IT-infrastructure to support the Dutch model-based approach for proton therapy, ESTRO 2022**(Oral Presentation)**
10. A federated learning IT-infrastructure to support the Dutch model-based approach for proton therapy, European Conference of Medical Physics 2022 **(Poster)**
11. The Dutch initiative for the model-based selection of patients for proton therapy-A federated learning IT infrastructure, ASTRO 2022 **(Oral Presentation)**

Curriculum vitae

Petros Kalendralis was born in Greece on August 13th 1991. In 2009 he started his Bachelor of Science in Physics at the University of Ioannina in Greece. In 2015 he joined the 424 military hospital of Thessaloniki in Greece as a trainee in Medical Physics in combination with his military obligatory service. In 2016, he started his Master degree in Medical Physics at the University of Aberdeen in Scotland. During his master, he focused in radiotherapy and specifically, he showed interest in the investigation of different novel treatment techniques for breast cancer. In 2017, he joined the Clinical Data Science research group of MAASTRO clinic, Maastricht, The Netherlands as a clinical data engineer, where he had the chance to develop skills in statistical modelling based on imaging biomarkers (radiomics).

In 2019, he transitioned to a PhD trajectory within the Clinical Data Science research division of MAASTRO clinic and Maastricht University. His research focuses on the implementation of Artificial Intelligence applications in radiotherapy and the role of the FAIR principles in medical imaging, machine learning prediction models and quality assurance of radiotherapy treatment planning. During his PhD, he obtained two travel grants (European Society of Radiation Oncology-ESTRO and European Nuclear Education Network -ENEN awards) and participated to international conferences (such as ESTRO and European Conference of Medical Physics) presenting his research. He has established collaborations with international radiotherapy centres such as the University of Washington Medical Centre in Seattle and the radiotherapy department of Greater Poland Cancer Centre in Poznan Poland (research visit funded by ESTRO). Next to his research activities, since July 2020 for two years, he was a representative of the GROW PhD students in the GROW council aiming to promote the interests of the PhD students within the GROW institute. Furthermore, he was an active member of the Faculty PhD Committee (FPC) of the Faculty of Health, Medicine and Life Sciences (FHML) of Maastricht University. In September 2022, he starts his postdoctoral research on FAIR data infrastructures at the Clinical Data Science department of MAASTRO clinic and Maastricht University under the supervision of Prof. dr. Andre Dekker.

Acknowledgements

When I started writing this thesis, I was in the “work-room” of my house that was adapted according to the needs of a suitable working environment during the COVID-19 pandemic period. During this difficult and challenging time for the humanity, being isolated for long time periods, I learned how to deal with the new reality of virtual conferences and collaborations with colleagues. However, the most important experience I gained during this time was the opportunity I had to spend time for personal development and discover the serenity that writing offered to myself. After a long time period, having the feeling that all this difficult COVID-19 lockdowns period was similar to a really long day behind closed doors and face masks, the life returns back to a normal mode of “live” social interaction with colleagues and friends. It is remarkable how much I missed the feeling of talking and interacting with my colleagues during this period, understanding that I should be thankful to all the people who helped me during my PhD journey.

Andre, I would like to thank you for giving me the opportunity and honour to work with you all the past years. You are a true leader who gave me the chance to develop myself seeing the potential and ability I had, creating a vision when I started this academic journey. You gave me important lessons on how to handle difficult and challenging tasks that will be useful in my life.

Rianne, I could have spent a lot of paragraphs explaining the reasons that I feel grateful working with you all the past three years. You taught me a lot with your supervision. Especially, you taught me how to implement a clear project management flow during my PhD, executing open and high quality research. The most important, you motivated and inspired me to research in a systematic and efficient way.

Johan, I would like to thank you for all the support and help you gave me during my PhD. Your technical expertise helped me to understand and solve complicated problems giving me simple and quick feedback and solutions. It was a pleasure to work with you.

Leonard, thank you for being a valuable teacher. You taught me very important skills during my first steps within the Clinical Data Science group. I will never forget the scientific lessons you gave me during our visit in Poland.

I would like to thank all the administration team of MAASTRO clinic and Maastricht University (Sonia, Ellen, Audrey) for all the support and help you provided me to manage all the administrative tasks during my academic journey.

I would also like to thank all the members of the Clinical Data Science group for their help and collaboration in the different projects. Especially I would like to thank Alberto and Zhenwei for the help and guidance during my first academic steps.

I would like to thank all the people and colleagues I collaborated with during the external collaborations I had during my PhD. I would especially like to thank all the colleagues from the Greater Poland Cancer Centre in Poznan, where I had the opportunity to work in an international clinical environment. Moreover, I would like to thank my colleagues from the University of Washington Medical Center in Seattle and the University of Vermont Medical Center in Burlington, Vermont in the United states for the fruitful collaboration we had during the last years.

Special thanks to all my colleagues from the Physics group of MAASTRO clinic who helped me in different projects during the last years.

Greece and Scotland, I am really grateful for the education you provided me before I move and find my home and family in The Netherlands.

I would also like to thank Davide Magliacano for his approval to use a part of his amazing art as a cover of my thesis.

Finally, I would like to express the most sincere thanks and appreciations to my parents. Μαμά, μπαμπά, σας ευχαριστώ πολύ για όλα όσα μου έχετε προσφέρει όλα αυτά τα χρόνια. Δίχως εσάς δεν θα είχα καταφέρει όσα έχω καταφέρει μέχρι τώρα. Κάθε μέρα που περνάει θέλω να σας κάνω πιο πολύ υπερήφανους έχοντας σας ως παράδειγμα γονέων. Σας αγαπώ πολύ.