Learning from big imaging data to predict radiotherapy treatment outcomes and side-effects

Learning from Big Imaging Data to Predict Radiotherapy Treatment Outcomes and Side-effects

Zhenwei Shi

Propositions

Learning from Big Imaging Data to Predict Radiotherapy Treatment Outcomes and Side-effects

1. Radiomics: the bridge between medical imaging and personalized medicine. (Philippe Lambin et al.)
2. Medical imaging represents one of the largest segments of big cancer data and remains hitherto an under-utilized data resource (Chapter 2 and 3)
3. A prediction model based radiomics may not generalize to a wider population, which strengthens the necessity of external validation when assessing prediction models based on radiomics (Chapter 5 and 6)
4. Radiation induced toxicity of radiotherapy can be modelled based on the organ being irradiated, not solely on the type of primary tumor itself. (Chapter 7)
5. An imaging dataset can be made FAIR that allows federated machine learning without having to make data publicly open. (Chapter 8)
6. To increase the clinical applicability and wider generalizability of radiomics, appropriate methods for feature selection and signature compilation are much needed in the field (Chapter 9)
7. Validation is required to assess generalization of prognostic models to new, unseen data (Zwanenburg et al.)
8. If you cannot bring the data to the research, you could bring the research to the data. (Andre Dekker)
9. You have to believe that the dots will somehow connect in your future. (Steve Jobs)
10. Life is trying things to see if they work. (Ray Bradbury)

Learning from Big Imaging Data to Predict Radiotherapy Treatment Outcomes and Side-effects

Colofon

Layout and cover design: Zhenwei Shi
Printing: Proefschriftmaken, Proefschriftmaken.nl
ISBN: 978-94-6380-927-6

Learning from Big Imaging Data to Predict Radiotherapy Treatment Outcomes and Side-effects

Dissertation

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Rianne M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public on Tuesday 8 September 2020 at
10.00 hrs

by

Zhenwei Shi

石镇维

born on August 8, 1988
in Changchun, China

Supervisor:
Prof. dr. ir. Andre Dekker

Co-supervisor:
Dr. Leonard Wee

Assessment Committee
Prof. dr. Joachim Ernst Wildberger (chair)
Prof. dr. Bram van Ginneken (Radboud University, Nijmegen)
Dr. Henry Christian Woodruff
Prof. dr. Zhen Zhang (Fudan University, China)

CONTENT

Chapter 1

General introduction and outline of the thesis

Zhenwei Shi

A large stream of data is generated during routine cancer care in the continuum from diagnosis to treatment and follow-up. This includes a vast variety of data such as diagnostic images, radiologist reports, pathology images with reports, pre-treatment verification imaging, surgical notes, medication notes, billing/insurance data and so on.

In many developed countries, this data is often stored digitally which is already a major advantage, however there is the strong tendency for it to be scattered across different departments (radiology, radiation oncology, and surgery) that use multiple data storage systems (Electronic Medical Record (EMR) and Picture Archiving Communication System (PACS)) and a range of different formats (DICOM, ASCII and PDF) [1]. The situation is certainly much more challenging in rapidly developing countries, where a significantly smaller segment of healthcare data is digitized. The ever-increasing capacity of telecommunications infrastructure, connectivity with World Wide Web protocols and usability of "smart phones with apps" suggests a trend towards a global healthcare data revolution that looks increasingly decentralized, privacy-sensitive and patient-controlled.

Looking in the radiation oncology domain specifically, multiple types of data are routinely generated in clinical practice from multiple sources, which offers an opportunity to improve cancer care based on these data. An increasingly important source, in particular in terms of data volume, is the imaging data stream produced by diagnostic and treatment imaging modalities.

In oncology, the imaging modalities Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) are widely used. A recent study [2] estimated that around 140 million patients were diagnosed with cancer in about 100,000 hospitals globally in the last decade. If one patient has a data volume of 0.1–10 Gb, the total data volume of cancer patient worldwide of the past decade is approximately 14–1400 petabyte [2]. By the definition of the four V's of "Big Data", imaging data generated in the oncology domain can be classified as "Big Data", due to the properties: (1) the use of data-intensive imaging modalities (Volume), (2) the imaging archives are growing rapidly (Velocity), (3) there is an increasing amount of imaging and diagnostic modalities available (Variety), (4) and interpretation and quality differs between care providers (Veracity). This data is on a scale that is impossible for humans to manually process, but it is readily within the capabilities of (semi-)autonomous machine algorithms (for example, as early as 2008, Google servers alone were already capable of handling and processing 20PB per day).

Sixty years ago, a radiation oncologist only needed to consider a handful of key clinical parameters such as histological type, size/extent of the tumor and its relative position relative to radiosensitive organs (such as the heart and spinal cord). There were typically only a couple of treatment options in terms of radiotherapy technology and pharmaceutical agents, while the patient's perspective was rarely considered (if at all). The data needed for decision-making were based on averaged population effects measured in a few randomized trials sponsored by one or two of the eminent academic medical centers.

In this age, doctors consistently speak of decision fatigue, information overload and lack of an effective "filter" that allows them to focus on the information needed for a decision at hand. Doctors are becoming increasingly overwhelmed with scientific literature of variable quality, rapidly evolving and diverging options for cancer treatments and the exponentially increasing amount of imaging data that needs to be inspected.

To provide high-quality individualized cancer care, there is an urgent need to translate the explosive growth (in terms of quantity) of big data into actionable knowledge that could be used to support doctors for decision-making in routine operations. The angle that is increasingly being explored in research is how machine algorithms, of various types, could be harnessed to help with automated data processing and information extraction, such that the "mining" work does not need to be done by humans alone.

To use big cancer data more easily and efficiently, an approach is needed to generate Findable, Accessible, Interoperable, and Reusable (FAIR) [3] data (Figure 1). It is abundantly clear that FAIR data is not just a concept for humans to use data meaningfully, but it is also intended for machine algorithms to be able to automatically search and process the data to assist humans; that is why data needs to be imbued with additional metadata that incorporates machine-readable semantics (i.e., what the data means).

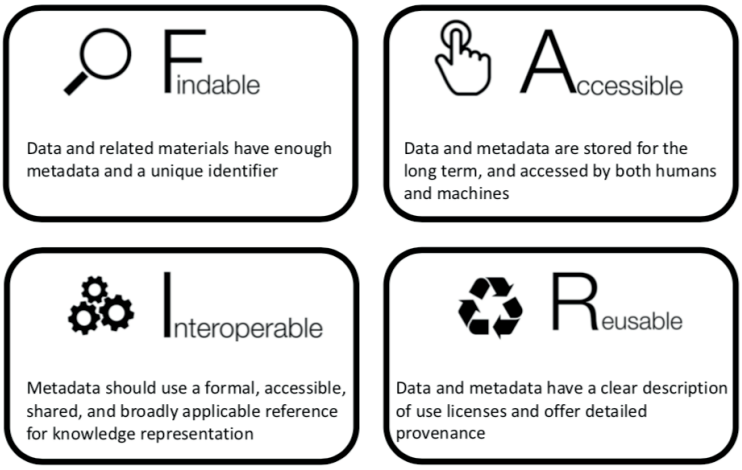


Figure 1: A brief description of what FAIR data is.

If we could prepare the data in a FAIR manner, artificial intelligence (AI), specifically machine learning, could have the ability to link the gap between this big data and useful knowledge, and we hypothesize this would improve clinical decision making for cancer care. Specifically in this thesis, we will explore how AI-based quantitative imaging techniques could be used towards this achieving this goal of better cancer care.

Quantitative Imaging Analysis

Multi-modality medical images consist of a large amount of valuable information reflecting the development and progression of a given oncological disease, such as lung cancer and esophageal cancer. With the rapid advances in AI, especially machine learning, it becomes possible to extract many quantitative features from an image of the tumor and translate these features into human-understandable knowledge.

This comprehensive approach used for image-based quantitative analysis is known as radiomics, that is able to quantify tumor heterogeneity, predict outcome, monitor treatment response and so on. The concept of radiomics was first proposed in 2012 [4, 5], and has since been applied to the field of translational clinical research. Radiomics aims to convert vast amounts of routine clinical imaging data into a mineable big data resource to promote research into better treatments and to derive actionable insights to guide medical decision-making. Therefore, radiomics has attracted the attention of researchers throughout the world [6] and is today an active topic of investigation.

Radiomics analysis is able to extract large amounts of high-dimensional quantitative features from multimodality medical images such as CT [7, 8], PET [9, 10], and MRI [11, 12]. The image-derived features have been found to correlate with the diagnosis and prognosis of cancer [8, 13], hence are able to decode information about the cancer and its phenotype, using subtle characteristics of pictures that are not easily quantifiable by an unaided human eye [14]. Previous studies have demonstrated that quantitative assessment of non-invasive biomarkers holds prognostic information in numerous cancer types in addition to molecular and clinical characteristics [8, 15, 16]. With complementary information from clinical reports, treatment responses, and genomic/proteomic assays, radiomics may reflect the global outlook of cancer [17]. Radiomics incorporates a series of computational technologies, and the methodologies used in radiomics are usually oriented for clinical problems. Substantial progress has been made within the field of radiomics regarding technology that meets clinical requirements, with advances towards cancer diagnosis and cancer treatment. Compared with previous methods that process medical images as pictures for visual inspection, radiomics has introduced a new way to mine for information in medical images.

Prediction modelling of treatment outcomes and side effects

Prediction modeling concerns the development and implementation of models or algorithms that are able to predict outcomes through statistical analyses. The developed models usually consist of one or more available predictors associated with the observed outcome. Prediction models can be used to calculate a probability of observing a certain outcome depending on the individual patient values for the model variables. In the radiation oncology domain, for instance, prediction models are used to estimate the probability of achieving tumor control or the risk of inducing side effects for a given treatment. Such models can be used to individualize treatments by optimizing the trade-off between tumor control and side effects. As another example, prediction models are currently used in the Netherlands to stratify patients who might benefit from proton or photon radiotherapy [18].

Federated learning

With the advances in AI technology, machine learning, especially deep learning, has achieved impressive progress recently especially in the domains such as face recognition [19] and object detection [20]. In principle, there are two main types of machine learning algorithms, related to the way data is stored, centralized learning and federated/distributed learning [21-23]. Traditional machine learning algorithms (centralized learning) are performed on a single data repository. In contrast, federated/distributed learning techniques allow machine learning algorithms to be performed on the data stored in different locations.

While AI in general already forms the base of a pervasive suite of applications in biology, clinical medicine, genomics and public health as surveyed in [24-27], deep learning applications in the medical imaging domain developed relatively slowly. One major challenge is that developing deep learning applications require large amounts of varied and high-quality training data [28, 29].

The popular image dataset ImageNet has over 14 million natural images in a central repository, which has tremendously supported the advance of AI specifically for centralized learning. Unfortunately, data is much more limited in the medical imaging domain, which are often data-poor settings with low sample sizes. Although healthcare practitioners are starting to build centralized, large, high-quality medical image repositories, it is still too limited to meet the rapid advance of AI. As a consequence, healthcare institutes have had to apply machine learning on their own data sources, which can be biased due to patient demographics, instruments, and clinical specializations.

The healthcare domain as a whole does possess the much-needed “big data” [30] for AI. The problem is that most healthcare data are locked in local data repositories inside hospitals. Due to the concerns of political, ethical, legal, privacy and technical natures, these data are often not easy to be share among centers, especially among international centers where HIPAA [31, 32] and GDPR [33, 34] regulations need to be observed [35]. These privacy concerns are important and worth to treat cautiously, but the negative aspect is that they have slowed down the healthcare domain from fully maximizing the benefits of AI.

To address the dispersal of big data in healthcare and leverage it for the advance of AI in a privacy–preserving manner, federated or distributed machine learning [21-23] might be a potential solution. Generally, federated or distributed learning can be considered to convey the same meaning, that is, training a machine learning algorithm across multiple decentralized data sources without directly exchanging original data samples. In this thesis, we will use them interchangeably. The fundamental theory of federated learning is to exchange learning parameters (e.g., weights or gradients of deep neural networks) between locally accessed data, rather than exchanging patient-level data directly. The idea of federated learning could be described by a saying – “bring the algorithm to the data, instead of the data to the algorithm” [23]. This approach enables multiple centers to collaborate on the development of AI-based

models without the need to share sensitive clinical data with each other, so that it is able to achieve the goal of preserving privacy.

Assessment of radiomics models

One recent criticism of radiomics is that prediction models based on radiomics will tend to achieve promising performance in training and even in internal validation datasets, but the external validation performance often drops dramatically. This phenomenon is generally grouped under the term "over-fitting", but this may not adequately express that lack of predictive performance in validation could arise from different sources.

One of the sources of difference that impacts radiomics is in terms of generalizability to different populations. Prediction models must not be limited to the same population of patients as the model development cohort, but should be able to predict outcomes in an unseen target population. Many radiomics studies to date have reported on internal validation or cross-validation, but only few reports on external validation in a non-randomly assigned group of unseen patients.

In many cases, internal validation appears to confirm the specific findings of a study which was discovered in the training process, but the findings are still limited in generalizability due to the selection bias of the training population. The reasons for this failure could be multiple between training and validation datasets, such as similarity of data distribution (age, disease and treatment), consistent measurement of outcome (if we define the outcome using a consistent measurement) and measurement of predictors (e.g., different approaches for calculation of radiomic features). Hence, it is indeed necessary to conduct external validation in other populations for prediction models based on radiomics.

Unfortunately, external validation datasets are not always available, mainly because of the concerns of political, ethical, legal, privacy and technical natures between centers. If we only consider the technical issue here, the concept of federated machine learning is once again a potential solution to solve the practical issue of external validation of a prediction model based on radiomics.

Furthermore, although there is overwhelming literature evidence on prediction models, the quality of reporting of prediction model studies has been found to be generally sub-optimal [7]. Radiomics, as for any other type of model, should be reported in a complete and transparent way, describing all the essential steps of developing the prediction model. Reporting guidelines such as Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [36] has been available for some years, but adherence amongst published radiomics studies still remains poor. The TRIPOD checklist was developed so that the possible usefulness of prediction models can be fully evaluated. Specifically in radiomics, the barriers towards fully transparent reporting still appears to be (1) images used for radiomics studies are not shared, (2) software used for radiomics feature calculation are not shared, (3) discrepancies

in the mathematical implementation of radiomic features in different software and (4) investigators not reporting in sufficient detail about pre-processing and image manipulation steps that were performed in their analysis.

Outline of the thesis

This thesis contains of five sections described in **Table-1**.

Table 1: Summary of the topics and characteristics of the studies presented in each chapter of this thesis.

Section	Chapter	Learning setting	Title
Introduction	Chapter 1		
Background	Chapter 2		Cancer registry and big data exchange
	Chapter 3		Data sharing and toxicity modelling – a vision of the near future
Software Development	Chapter 4		Ontology-guided Radiomics Analysis Workflow (O-RAW)
Application	Chapter 5	Centralized learning	Development and External Validation of a Prediction Model Incorporating PET Radiomics for Pathological Lymph Node Metastases in Esophageal Adenocarcinoma
	Chapter 6	Centralized learning	External validation of a prognostic model incorporating quantitative PET image features in esophageal cancer
	Chapter 7	Federated learning	External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients
	Chapter 8	Federated learning	Distributed radiomics as a signature validation study using the Personal Health train infrastructure
Discussion and Future Perspectives	Chapter 9		

After this introduction, the "Background" section, the basic principles of big data and treatment toxicity in radiation oncology are described. The importance of big data exchange (**Chapter 2**) and the generation of predictions for treatment outcomes and side-effects (**Chapter 3**) are clarified.

The next section presents the software development (**Chapter 4**), which makes the extraction of quantitative imaging features easier.

The “Application” Section contains the centralized learning (**Chapters 5-6**) and federated learning (**Chapters 7-8**) AI-based applications.

In the final section, the use of quantitative imaging features from big imaging data to predict radiotherapy treatment outcomes and side-effects are discussed. Then the challenges and potential solutions of federated learning are introduced. Finally, future perspectives are described.

References

1. Deng, J., *Big data in radiation oncology: challenges and opportunities.* Cancer Sci Res Open Access, 2014. **1**(2): p. 1-2.
2. Lustberg, T., et al., *Big Data in radiation therapy: challenges and opportunities.* The British journal of radiology, 2017. **90**(1069): p. 20160689.
3. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.
4. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* European journal of cancer, 2012. **48**(4): p. 441-446.
5. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic resonance imaging, 2012. **30**(9): p. 1234-1248.
6. Liu, Z., et al., *The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges.* Theranostics, 2019. **9**(5): p. 1303.
7. Traverso, A., et al., *Repeatability and reproducibility of radiomic features: a systematic review.* International Journal of Radiation Oncology* Biology* Physics, 2018. **102**(4): p. 1143-1158.
8. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nature communications, 2014. **5**: p. 4006.
9. Foley, K.G., et al., *External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer.* Radiotherapy and Oncology, 2019. **133**: p. 205-212.
10. 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, 2018. **28**(1): p. 428-436.
11. Traverso, A., 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**: p. 44-51.
12. Peerlings, J., et al., *Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial.* Scientific reports, 2019. **9**(1): p. 4800.
13. Gatenby, R.A., O. Grove, and R.J. Gillies, *Quantitative imaging in cancer evolution and ecology.* Radiology, 2013. **269**(1): p. 8-14.
14. Aerts, H.J., *The potential of radiomic-based phenotyping in precision medicine: a review.* JAMA oncology, 2016. **2**(12): p. 1636-1642.
15. Ou, D., et al., *Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status.* Oral oncology, 2017. **71**: p. 150-155.
16. Parmar, C., et al., *Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer.* Frontiers in oncology, 2015. **5**: p. 272.
17. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: images are more than pictures, they are data.* Radiology, 2015. **278**(2): p. 563-577.

18. Langendijk, J.A., et al., *Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach.* Radiotherapy and Oncology, 2013. **107**(3): p. 267-273.
19. Sun, Y., X. Wang, and X. Tang. *Deep learning face representation from predicting 10,000 classes*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
20. Zhao, Z.-Q., et al., *Object detection with deep learning: A review.* IEEE transactions on neural networks and learning systems, 2019. **30**(11): p. 3212-3232.
21. Sheller, M.J., et al. *Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation*. in *International MICCAI Brainlesion Workshop*. 2018. Springer.
22. McMahan, H.B., et al., *Communication-efficient learning of deep networks from decentralized data.* arXiv preprint arXiv:1602.05629, 2016.
23. Gaye, A., et al., *DataSHIELD: taking the analysis to the data, not the data to the analysis.* International journal of epidemiology, 2014. **43**(6): p. 1929-1944.
24. Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* Journal of The Royal Society Interface, 2018. **15**(141): p. 20170387.
25. Litjens, G., et al., *A survey on deep learning in medical image analysis.* Med Image Anal, 2017. **42**: p. 60-88.
26. Miotto, R., et al., *Deep learning for healthcare: review, opportunities and challenges.* Brief Bioinform, 2018. **19**(6): p. 1236-1246.
27. Shickel, B., et al., *Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis.* IEEE journal of biomedical and health informatics, 2017. **22**(5): p. 1589-1604.
28. Chen, X.-W. and X. Lin, *Big data deep learning: challenges and perspectives.* IEEE access, 2014. **2**: p. 514-525.
29. Sun, C., et al. *Revisiting unreasonable effectiveness of data in deep learning era*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
30. Obermeyer, Z. and E.J. Emanuel, *Predicting the future—big data, machine learning, and clinical medicine.* The New England journal of medicine, 2016. **375**(13): p. 1216.
31. Annas, G.J., *HIPAA regulations-a new era of medical-record privacy?* New England Journal of Medicine, 2003. **348**(15): p. 1486-1490.
32. Mercuri, R.T., *The HIPAA-potamus in health care data security.* Communications of the ACM, 2004. **47**(7): p. 25-28.
33. Tikkinen-Piri, C., A. Rohunen, and J. Markkula, *EU General Data Protection Regulation: Changes and implications for personal data collecting companies.* Computer Law & Security Review, 2018. **34**(1): p. 134-153.
34. Voigt, P. and A. Von dem Bussche, *The eu general data protection regulation (gdpr).* A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.
35. Fenton, S., et al., *Health information management: changing with time.* Yearbook of medical informatics, 2017. **26**(01): p. 72-77.
36. Collins, G.S., et al., Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. British Journal of Surgery, 2015. 102(3): p. 148-158.

Chapter 2

Cancer Registry and Big Data Exchange

Zhenwei Shi, Leonard Wee and Andre Dekker

Adapted from:
Shi, Zhenwei., et al. "Cancer registry and big data exchange." Big Data in Radiation Oncology (2019): 153.
DOI: http://dx.doi.org/10.1201/9781315207582-11

Introduction

A vast stream of data is generated by the routine operations of modern cancer diagnosis and oncologic treatments. Usually, the data is stored electronically but it tends to be scattered across different disciplines (e.g., radiation oncology, radiology, medical oncology and surgery), different data storage platforms (e.g., EMR and PACS) and in a wide variety of formats (e.g., DICOM, ASCII, and PDF) [1]. In addition, data in cancer care has the particular properties of large volume and complex dependencies between data elements, which is creating growing difficulties for conventional methods of data handling. By handling, we refer to collection, storage, update and exchange. Although the variety and volume of big data continues to grow exponentially within the field of oncology [2], it has not been easy to exploit this rich vein of data to improve patient safety and health outcomes [3].

Data, as a whole, in clinical routine is one of the most valuable, but most under-utilized, assets within radiotherapy and oncology studies [4]. Rapid Learning Health Care (RLHC) [5] envisions virtuous cycles of rapid knowledge generation and knowledge utilization within healthcare, by combining routine clinical practice with prospective research studies using a big data exchange paradigm. The resulting interconnected web of data repositories across departments and institutions becomes a global resource to mine for new knowledge, generate hypotheses for novel treatments, and test the effectiveness of interventions in real-world settings. RLHC also leads to a wide range of distributed software applications that can exploit FAIR (Findable, Accessible, Interoperable and Re-usable) data repositories [6], to create predictive outcome models for clinical decision support [7] and to discover predictive digital signatures in biomedical images (i.e., radiomics) [8].

To be effective in clinical decision support, predictive models must be able to estimate the probability of a given outcome over a range of clinical situations. Therefore, an approach that integrates data of many patients over different treatment settings is essential. Predictive outcome models have the potential to improve quality of life, identify patients at high (or low) risk, and to prolong the survival of patients with cancer [9-12]. Some predictive outcome models for various cancer sites can be found in (**http://www.predictcancer.org/**). These prediction models support the practice of personalized radiotherapy that is tailored to the individual risk profile of the patient. If multiple treatment possibilities exist where clinical evidence is equivocal, patient and physician preferences for certain outcomes may play a role. The collaborative consultation process between a patient and their treatment physician is known as shared decision making [13, 14]. Reliable model-based predictions of potential treatment responses are a prerequisite for information trade-offs between the risks of harm and strength of treatment in the shared decision-making paradigm. In keeping with RLHC, observational data on quality-of-life, patient-reported outcomes and decision regret should be put back into the web of clinical knowledge, so that continually updated models are better able to inform physicians and patients' decisions.

We have prefaced this chapter with a consideration of the wide-ranging opportunities for clinical innovation and improved outcomes that could be possible with comprehensively voluminous, multi-modality and multi-institutional data. Central to this ambition is the integration of all the data that presently remains under-utilized in closed, unconnected private repositories. The remainder of this chapter is organized as follows: first, we explore the

paradigm of data centralization in the form of a cancer registry (CR). This, whether large or small, represents the current orthodoxy for collection, storage, analysis, and distribution of oncology data. Second, as the multiplicity, volume and dimensionality of cancer data continues to grow rapidly, we shall argue that a new paradigm for data exchange is urgently needed. We thus describe practical and feasible big data architectures that can be applied to overcome the challenges of poorly connected data repositories in radiation oncology. Third, the remaining barriers impeding universal rollout of big data exchange architecture in radiation oncology field will be examined, as will the unique characteristics of certain data architectures that may be more naturally adapted to overcoming these barriers. The primary aim of this chapter is to support the understanding of readers about big data exchange architectures that facilitate rapid learning in clinical practice and scientific research in the field of radiation oncology.

Cancer Registry

A CR refers to architectures that are capable of systematically capturing, storing, analyzing, and reporting data on patients with cancer [15]. Data on cancer occurrence and properties is registered in CRs. A cancer registrar is a person with over-arching responsibility to capture complete, accurate and timely data on cancer patients. Participation in cancer registries is typically, but not solely, mandated by local legislations.

Data within a CR typically consists of demographic information, medical history, diagnostic investigations, outcomes of therapy, and some follow-up details of patients. CRs may have a variety of functionalities. For example, data within a CR can be used to:

1. Assess treatment outcomes (e.g., survival and toxicity),
2. Assess disease survivorship aspects such as quality of life,
3. Evaluate efficacy and/or economic impacts of diseases and their interventions,
4. Provide follow-up investigation and guidance,
5. Allocate health resources at regional or country level,
6. Compare quality of treatment and outcomes between care providers,
7. Report on cancer incidence.

Different Types of Cancer Registry

A CR may have different specific purposes and functionalities as described above, which primarily depend on local demands and requirements. Two primary types of CR are commonly used worldwide; hospital-based and population-based CR. **Table 1** shows a brief description of these two types of CR.

Table 1: Brief description of each type of CR

TYPES	HOSPITAL–BASED CR	POLULATION-BASED CR

Purposes	• Improvement of cancer care • Administration of cancer information • Clinical research • Training and education	• Cancer prevention and early detection • Determine cancer incidence and trends • Academic research • Assessment of population cancer outcomes
Details	• Maintains data on all cancer patients diagnosed and/or treated at a particular hospital • Provides medical audit-style assessment of outcomes within a particular hospital • Support institutional registries with common standard protocols and integrated data	• Records all new cases in a well defined population (e.g., geographic area) with an emphasis on epidemiology and public healthcare • Informs cancer agencies and organizations of cancer statistics in specific populations • Informs cancer researchers about an unbiased group of cases that can be selected for studies

Hospital-based CRs

Hospital-based CRs are established to record the information of cancer patients collected within a specific care setting, such as a major hospital or cancer clinic that aims to offer readily accessible, complete, accurate, and timely information of patients with cancer [16]. This may include, for example, data about diagnosis, treatments and outcomes. Hospital-based CRs generate reports on the number of cancers observed within a particular hospital per year by site, sex and age. These reports are very useful for clinical research by comparing the frequency of certain types of cancer within a single hospital to the total amount of cancer cases [15]. Furthermore, it can lead to several potential applications for epidemiological research by (1) providing information on approaches of diagnosis, stage distribution, outcomes to treatment, and overall survival at the hospital level and identifying potential drawbacks of treatments; and (2) predicting future needs for services, equipment and human resource within a particular cancer center.

Generally, the endpoints of a hospital-based CR are geared towards quality, management and caseload planning goals. However, many hospital-based CRs are required to submit data to a centralized disease-specific CR as well. For this purpose, the hospital-based CR usually has to collect data elements that are useful for the central CR but may not be immediately useful for the hospital. Therefore, data elements within a hospital-based CR often cover a wider range than data within a population-based CR [16].

Population-based CRs

Population-based CRs are concerned with collecting data on all new patients with cancer arising in a well-defined community [15, 17-19]. The primary purpose of population-based CRs is to report cancer incidence and produce analytic findings about cancer in a defined population. In addition, population-based CRs are a benefit of assessment and control of cancer care. Thus, the focus of population-based CRs is on epidemiology research and public healthcare.

Furthermore, population-based CRs can monitor the cancer occurrence and prevalence of diseases; thus they are of importance in planning and evaluating region-based or population-based programs on cancer control through:

1. Standardizing treatment priorities and predict resources needed in future
2. Examining the effectiveness and appropriateness of screening programs in the community
3. Comparing health care providers in terms of practice and quality
4. Evaluating cancer care population outcomes through survival statistics.

Because it is not always possible to strictly define a catchment population, a hospital-based CR is not necessarily able to provide assessment and statistics on the cancer occurrence in the defined population [16], which is a major difference between hospital- and population-based CRs.

Examples of Cancer Registries

There are many running CRs in the world. In this section, we will describe a small collection of them, which are from three countries: the United States, Italy and the Netherlands.

Surveillance, Epidemiology, and End Results (SEER)

The Surveillance, Epidemiology, and End Results (SEER) program is a source of information on cancer incidence and survival in the United States. The data is published on its Web site (**https://seer.cancer.gov/about/overview.html**). SEER currently captures and publishes cancer data on incidence and survival from the cancer data sources that cover around 28% of the American population.

The data captured by the SEER Program comprises demographics information, main tumor site, tumor morphology and cancer stage, treatment at first course, and follow-up details. The SEER program is the only program that collects the data on cancer stage at diagnosis and survival data of patients in the United States. The data is updated annually and published in reports as a public service. The U.S. Census Bureau provides periodical data on population for the SEER program to use to compute_cancer ratio. Many practitioners and members of the public have used the data reported by the SEER program.

Nation Radiation Oncology Registry (NROR)

The National Radiation Oncology Registry (NROR; **http://www.roinstitute.org/What-We-Do/NROR/Index.aspx**) is a collaborative initiative of the Radiation Oncology Institute (ROI) and the American Society of Radiation Oncology (ASTRO) through guidance and data support from other major stakeholders in oncology. The NROR captures related data on cancer patients' treatment delivery and outcomes; the data are used to improve cancer care. The overarching purposes of the NROR are to (1) compare cancer patients who have similar cancer states or profiles, (2) identify suitable treatment and possible drawbacks in cancer care, and (3) build a CR for health care in a defined population.

The national registry comprises standardized aggregated data on therapies for specific types of cancers. An analysis of the outcomes achieved would yield invaluable benchmarking measures, help define best practices, evaluate the comparative effectiveness of treatments, and identify gaps in quality. However, due to funding limitations, the NROR is no longer active.

Registro Tumori Ospedaliero (RTO)

The Italian Association of Cancer Registries (AIRTUM), established in Florence in 1997, aims to coordinate multiple cancer registries in Italy. For more details see (**http://www.registri-tumori.it/cms/en)**. The linkage created by the association supports research, editorial output, and methodological development of the various member registries. The association is connected to equivalent bodies in other countries at European and global levels. Statistics about the distribution of cancer in the areas covered by the member registries covers:

1. Incidence – the number of new cancer cases per year
2. Prevalence - the number of Italians have a particular cancer
3. Mortality - the various different causes of death for Italians registered in the CR
4. Trends - whether the number of cancer cases has been increasing or decreasing with respect to preceding years
5. Survival - how long do Italians survive after treatments for cancer
6. Comparison of registries - whether the impacts of cancer is uniform across Italy
7. International comparisons – how the situation in Italy compare with the rest of the world

Dutch Surgical Colorectal Audit (DSCA)

The Dutch Surgical Colorectal Audit (DSCA) [20] was established by the Association of Surgeons of the Netherlands (ASN) in 2009 and aims at surveillance, assessment, and improvement of colorectal cancer treatment. For more details see (**https://www.dica.nl/dlca**).

All Dutch hospitals that perform bowel cancer and rectal cancer surgery, participate in the Web-based quality registration. At present, more than 60,000 treatments have been registered. This registration allows quality benchmarking, where hospitals compare the quality of their cancer care with those of others. The comparisons are statistically corrected for differences in level of care (local case-mix) and random sampling, which renders the analysis meaningful and "fair". The system has been able to propose possible improvements to individual colorectal surgeons, while the national professional association facilitates and monitors these improvements.

One of the main contributions of DSCA is that it leads to effective multicenter surgical collaboration. Because the ASN has an important role in audits, all colorectal surgeons in the Netherlands have participated in the collaboration.

Data Sources of CRs

Data sources of CRs are external data resources available to the registry, that are used for the collection and verification of cancer-related information. According to the relation between the data elements and the goals of CRs, there are two types of data sources: primary and secondary data sources [21]. Primary data sources refer to the data collected for immediate goals of the CR. The data collection from primary data sources can promote elements of data quality such as completeness, validity, and reliability. This data collection is implemented via a standard protocol. The protocol is intended to enforce the same procedures and data format used in all CRs and patients, which ultimately benefits data analysis, tracking and integration. Due to the auditability of collected data, the entered data can typically be traced back to an individual patient. Finally, the quality of primary data sources is usually better than secondary data

sources because of automatic quality control procedures or follow-up checks made by data managers.

The initial purposes of secondary data sources are not for cancer registration (e.g., data generated in routine medical practice and insurance claim forms). The data in secondary data sources is usually stored electronically and can be accessed through appropriate permissions. [21].

To ensure few cancer cases are missed and the quality of data (i.e., dimensionality, completeness, accuracy, timeliness) remains high, CRs usually collect data of patients with cancer from multiple sources. For instance, data sources of a population-based CR often refer to cancer centers, general practitioners, screening programs, coroners' recording systems, health insurance companies, and other CRs. However, as a disadvantage, the use of multiple sources of information raises concerns about receiving multiple notifications of the same cancer patient. To avoid this problem, the data of the same patient existing within multiple data sources should be linkable, thereby eliminating duplicate registry.

It might be generally presumed that it is simple to maintain a population-based CR when sub-registries (e.g., hospital-based CRs) can openly transfer identifiable data such that the population-based CR only needs to integrate the incoming data. However, this is not always possible in real-world scenarios. In practice, the population-based CR still needs to collect overlapping data from numerous data sources. The reasons are twofold. First, patients with an eligible condition might never attend a contributing hospital, therefore the population-based CR needs to use multiple sources to prevent eligible cases being missed. Secondly, patients may attend more than one hospital over different parts of their treatment pathway. The use of multiple sources is then of benefit of identifying duplicate registrations or missing registrations of the same patient. Although it is not always possible to collect all data from all data sources in practice, the aim is still to use as many cancer data sources as possible. As described in [21], advantages and disadvantages of key sources are presented in **Table 2.**

Table 2: Advantages and disadvantages of key data sources.

DATA SOURCE	ADVANTAGES	DISADVANTAGES
Patient report	• Unique perspective based on patient experience of disease and treatment. • Information on treatments not necessarily prescribed by clinicians. • Obtaining information about intended compliance. • Useful when timing of follow-up may not be concordant with timing of clinical encounter. • Can capture patient and/or caregiver outcomes.	• Literacy, language, physician access or other barriers that may result in under-enrolment of some sub-populations. • Validated data collection equipment may need to be established. • Patients may refuse to participate in follow-up study. • Limited confidence on clinical information and utilization information.

Clinician report	• More specific information than available from coded data or medical record. • Tends to be more objective and focused on impacts on care delivery.	• Clinicians are highly conscious of administrative to burden. • Potential inconsistencies in capture of patient signs, symptoms, use of non-prescribed therapy.
Medical chart abstraction	• Information on routine medical care, with more clinical context than coded claims. • Potential for comprehensive view of patient medical and clinical history. • Use of abstraction and strict coding standards (including handling missing data) increases the quality and interpretation of data abstracted.	• The underlying information is not always collected in a systematic way. For example, a diagnosis of bacterial pneumonia by one physician may be based on a physical exam and patient report of symptoms, while another physician may record the diagnosis only in the presence of a confirmed laboratory test. • It is difficult to interpret missing data. For example, absence of a specific symptom in the visit record would not indicate whether the symptom was truly absent or that the physician did not actively inquire about this specific symptom or set of symptoms. • Data abstraction is more (technical) resource intensive. • Complete medical and clinical history may not be available (e.g., new patient to clinic).

Electronic health records (EHRs)	• Information on routine medical care and practice, with more clinical context than coded claims. • Potential for comprehensive view of patient medical and clinical history. • Effective access to medical and clinical data. • Use of data transfer and coding standards (including handling of missing data) will increase the quality of data abstracted.	• Underlying information from clinicians is not collected using uniform decision rules. (See example under "Medical chart abstraction.") • Consistency of data quality and breadth of data collected varies across sites. • Difficult to handle information uploaded as image files into the EHRs (e.g., scanned clinician reports) vs. direct entry into data fields. • Historical data capture may require manual chart abstraction prior to implementation date of medical records system. • Complete medical and clinical history may not be available (e.g., new patient to clinic). • EHR systems vary widely. If data come from multiple systems, the registry should plan to work with each system individually to understand the requirements of the transfer.
Institutional databases	• Diagnostic and treatment information (e.g., pharmacy, laboratory, blood bank, radiology). • Resource utilization data (e.g., days in hospital). • May incorporate cost data (e.g., billed and/or paid amounts from insurance claims submissions).	• Important to be knowledgeable about coding systems used in entering data into the original systems. • Institutional or organizational databases vary widely. The registry should plan to work with each system individually to understand the requirements of the transfer.

Administrative databases	• Useful for tracking health care resource utilization and cost-related information. • Range of data includes anything that is reimbursed by health insurance, generally including visits to physicians and allied health providers, most prescription drugs, many devices, hospitalization(s), if a lab test was performed, and in some cases, actual lab test results for selected tests (e.g., blood test results for cholesterol, diabetes). • In some cases, demographic information (e.g., gender, date of birth from billing files) can be uploaded. • Potential for efficient capture of large populations.	• Represents clinical cost drivers vs. complete clinical diagnostic and treatment information. • Important to be knowledgeable about the process and standards used in claims submission. For example, only primary diagnosis may be coded and secondary diagnoses not captured. In other situations, value-laden claims may not be used (e.g., an event may be coded as a "nonspecific gynecologic infection" rather than a "sexually transmitted disease"). • Important to be knowledgeable about data handling and coding systems used when incorporating the claims data into the administrative systems. • Can be difficult to gain the cooperation of partner groups, particularly in regard to receiving the submissions in a timely manner.
Death indexes	• Completeness—death reporting is mandated by law in many countries, such as the United States. • Dependable alternative source for mortality tracking (e.g., if a patient was lost to follow-up). • National Death Index (NDI) — centralized database of death records from State vital statistics offices; database updated annually. • NDI causes of death relatively reliable (93–96%) compared with State death certificates.	• Time delay—indexes depend on information from other data sources (e.g., State vital statistics offices), with delays of 12 to 18 months or longer (NDI). It is important to understand the frequency of updates of specific indexes that may be used. • Absence of information in death indexes does not necessarily indicate "alive" status at a given point in time.

	• Social Security Administration's (SSA) Death Master File—database of deaths reported to SSA; database updated weekly.	• Most data sources are country specific and thus do not include deaths that occurred outside of the country. • As of November 2011, Death Master File no longer includes protected State records.
Existing registries	• Can be merged with another data source to answer additional questions not considered in the original registry protocol or plan. • May include specific data not generally collected in routine medical practice. • Can provide historical comparison data. • Reduces data collection burden for sites, thereby encouraging participation	• Important to understand the existing registry protocol or plan to evaluate data collected for element definitions, timing, and format, as it may not be possible to merge data unless many of these aspects are similar. • Creates a reliance on the other registry. Other registry may end. • Other registry may change data elements (which highlights the need for regular communication). • Some sites may not participate in both. Must rely on the data quality of the other registry.

Receiving and Reporting Information in CRs

Because a CR is an organization for systematically processing data of patients with cancer, it is very important to determine (1) the data items that a CR will receive from various data sources, and (2) the data items that a CR will report to other organizations. Received information by CRs refers to the information (e.g., demographic information, medical history, diagnostic discoveries, therapeutic information, follow-up details) collected by CRs from multiple sources or reported by certain institutes. On the other hand, reported information by CRs is simply concerned with summary analyses generated by the CRs, which can be deemed the functional output of the CRs. Because CRs can be different from each other in terms of aims, functions, local needs and requirements, receiving and reporting information might also be different. The details of receiving and reporting information of CR are as follows.

Receiving information

The data elements collected by a CR are directly related to its aims and functionalities. The primary purpose of a hospital-based CR is administration of patients with cancer in a particular site. On the other hand, the main goal of a population-based CR is to produce statistics about cancer occurrence in a defined population. Thus, choosing data elements for a CR requires considering many aspects such as data reliability, the necessity in analyzing treatment responses, and even the cost of data collection [21].

Basic data elements

Although we have to determine the purposes and functionalities of a CR before specifying the data elements, some basic common elements exist in most CRs. The basic data elements that have been described in [21] are listed in **Table 3**, with the caveat that many directly identifying data elements will often be coded or partitioned in a secure part of the registry for privacy purposes.

Table 3: Examples of possible basic data elements.

Registrar information	• Registrar contact information • Contact information (e.g., address, telephone and email) of another individual who can be reached for follow-up
Patient information	• Patient identifiers (e.g., name, age, date of birth, place of birth, Social Security number) • Permission/consent • Source of enrolment (e.g., provider, institution, phone number, address, contact information) • Enrolment criteria • Sociodemographic characteristics, including race, gender, and age or date of birth • Education and/or economic status, insurance, etc. • Place of birth • Location of residence at enrolment • Source of information • Country, State, city, county, ZIP Code of residence.

Optional data items

Adding additional collection elements will increase the complexity and cost of registration [22]. Therefore, when designing registration forms and before performing registration, one should first consider whether a CR really needs certain data elements, and that it can sustain the cost associated with collecting these data element. Apart from the basic data elements, other data elements may be needed based on the specific design and purpose of a registry. **Table 4** shows a collection of possible data elements that are described in [21].

Table 4: Examples of optional data elements.

PRE-ENROLMENT HISTORY	
Medical history	• Morbidities/conditions • Onset/duration • Severity • Treatment history • Medications • Adherence • Health care resource utilization • Diagnostic tests and results • Procedures and outcomes • Emergency room visits, hospitalizations (including length of stay), long-term care, or stays in skilled nursing facilities • Genetic information • Comorbidities
Environmental exposures	• Places of residence • Hazardous occupations? • Exposure to occupational hazards?
Patient characteristics	• Development (pediatric/adolescent) • Functional status (including ability to perform tasks related to daily living), quality of life, symptoms • Health behaviors (alcohol, tobacco use, physical activity, diet) • Social history • Marital status • Family history • Work history • Employment, industry, job category • Social support networks • Economic status, income, living situation • Sexual history • Foreign travel, citizenship • Legal characteristics (e.g., incarceration, legal status) • Reproductive history • Health literacy • Social environment (e.g., community services) • Enrollment in clinical trials (if patients enrolled in clinical trials are eligible for the registry)

Provider/system characteristics	• Geographical coverage • Access barriers • Quality improvement programs • Disease management, case management • Compliance programs • Information technology use (e.g., computerized physician order entry, e-prescribing, electronic medical records)
Follow-up/ Outcomes	• Safety: adverse events (see Chapter 12) • Quality measurement/improvement: key selected measures at appropriate intervals • Effectiveness and value: intermediate and endpoint outcomes; health care resource use and hospitalizations, diagnostic tests and results. Particularly important are outcomes meaningful to patients, including survival, symptoms, function, and patient-reported outcomes, such as health-related quality-of-life measures. • Natural history: progression of disease severity; use of health care services; diagnostic tests, procedures, and results; quality of life; mortality; cause/date of death • Economic status • Social functioning
Other potentially important information	• Changes in medical status • Changes in patient characteristics • Changes in provider characteristics • Changes in financial status • Residence • Changes to, additions to, or discontinuation of exposures (medications, environment, behaviors, procedures) • Changes in health insurance coverage • Sources of care (e.g., where hospitalized) • Changes in individual attitudes, behaviors

Data element selection is primarily dependent on the goals of a CR, the approaches used for data collection, and the resources available. Many cancer CRs have failed as they attempted to capture too many data elements. The focus of the CR should be on the quality of data rather than the quantity. As we described in **section 2**, those successful and productive CRs only collected a limited amount of information for each patient.

Reporting information

The most important purpose of a CR is to perform statistics on cancer occurrence, treatment and outcomes in a particular region or population [23]. Therefore, collation, examination and explanation of the captured data are the main reporting tasks of a CR.

Information reported by CRs is often presented by means of cancer incidence reports, practice and treatment outcomes reports, and scientific publications. Results and conclusions are usually documented in reports and subsequently published to users. Generally, the reports

contain background information about registration, procedures of registration, population of covering, data quality (e.g., completeness and validity), and results of analysis. The population-based CR should perform basic statistics that are primarily about the distribution of the tumor in the community. The data and findings may be displayed in various types of format such as tabular and graphical forms, by which the readers can draw their own conclusions according to their interests.

Assessment of Data Quality within CRs

CRs have evolved beyond a data provider that reports cancer incidence within a well-defined population [19]. By linking sufficient resources, a CR is useful in many aspects of the cancer control domain, such as identification of causes of specific cancer, assessment of screening programs and improvement of cancer care [24, 25].

The functionalities of a modern CR and its capacity to perform cancer control activities are highly dependent on the quality of data within the CR. Three dimensions of data quality have been introduced in the earlier publication [26]: comparability, completeness and validity. As described in [27], timeliness is another key indicator of data quality for CRs.

In order to assure data quality, quality control plays as an important role. Theoretically, it is possible that a CR can collect very high-quality data (similar to clinical trials) without extensive quality control processes, but this is seldom the case in real world scenarios. There is no large-scale database that can be perfect in regard to completeness and validity. Therefore, routine quality control is the necessary step to identify the area needing improvement. The quality control can help with data interpretation and may further indicate the need for procedural changes [28].

From CR to Big Data in Radiation Oncology

Barriers of CRs

We have seen that CRs can act as a valuable oncology data resource at several geographic scales and between multiple collaborating cancer centers. It remains a complex and costly process to maintain a CR, and the constraints on its architecture is obvious, that is, it is difficult to be scaled up to manage big data in radiation oncology. A centralized data repository needs to coordinate uniform collection at many different data sources and manage exchanges between different types of storage formats. This requires every contributing party to first agree on, and then rigidly adhere to, the same data collection instrument and the same internal data structure. The types of statistical analysis are constrained by the limited data fields collected and further restricted by the operational objective(s) of the CR. Although some attempts have been made by certain CRs to make data available for research, the elements generally have to be extracted as specialized queries by a data manager at the CR.

A large amount of human resources (e.g., registrars and data managers) and infrastructure (e.g. servers, user interfaces) is required to build and maintain a CR, adding significantly to its financial cost. Furthermore, rigidly structured collaborations among many departments are needed to support a CR. For example, a population-based CR first requires that local hospitals, cancer centers, and other institutions extract and upload (automatically or entering by hand) very specific data collection forms pertaining to the condition of interest. Then the collected

data has to be processed and audited internally in the CR. Finally, with a significant time latency, results from a rigidly prescribed statistical analysis are reported to higher authorities and made available to the public.

Crucially, one of the most significant barriers to universal adoption of CRs pertains to how big data can be flexibly and securely shared amongst a large network of cancer institutes or research programs to answer a broad range of clinically relevant questions. First, the data is neither readily findable nor discoverable; no general query process available to physicians and researchers to ask if data pertaining to their clinical question resides in the CRs.

Second, data in CRs are generally inaccessible to physicians and researchers due to concerns over patient privacy, data sharing, or intellectual property rights. Where an entity outside the CR might be permitted to request some data, internal resources of the CR are required to program a data extraction query specific to each request received. For instance, researchers in a U.S. cancer center may be interested in the overall survival outcomes of lung cancer patients in a reasonably similar cohort in the Netherlands. Assuming the investigators in the United States even know what data fields to ask for, they would not be permitted to structure any external query that might address questions such as the following: (1) Is the case-mix of Dutch and U.S. populations approximately comparable? (2) What types of oncological and surgical interventions do Dutch lung cancer patients receive? (3) Do the follow-up protocols in both populations overlap to a sufficient degree to even contemplate a comparative study? In Dutch law, it is not allowed to share the data even for the purpose of clinical academic research. Patient privacy laws prevent effective data sharing, even in a highly geographically localized setting (for example, two neighboring cancer hospitals in the Netherlands).

Third, even if the administrative, legal and political barriers that prevent data sharing can be overcome, the technical challenge remains of reading data across incompatible computer systems that reside in many different formats. The exchange of "data dictionaries" between any two centers may permit a level of syntactic interoperability, but the parties also need to know how the exact structure of how the data elements are actually stored in memory, in order to construct a query. Exchange protocols such as Health Level-7 (HL7) and Fast Healthcare Interoperability Resources (FHIR) attempt to offer a degree of interoperability. However, each data transaction must be specifically customized to each collaboration site, rather than creating a single universal query that can work across all collaboration sites.

Therefore, one may not only consider a CR as a particular organization or entity (such as a hospital or a governance authority), but rather the CR must be considered as an architectural archetype for progressively more centralized data collection, storage, management, analysis and dissemination. At each layer of centralization, the variety and contextual richness of the data is incrementally lost, such that only the broadest and least-detailed summary statistics can be reported at the uppermost level (for example, a national cancer registry). The reduction in dimensionality and contextual depth is compensated by an increasing universality in population coverage.

Such a trade-off in CR architecture may be visualized as in **Figure 1**. At the first level, individual institutes (e.g., clinics and hospitals) are the point of generation of data that lies closest to the patient. Data collection is generally immediate and highly multidimensional (e.g., clinician notes, nursing observations, diagnostic imaging scans, treatment information, follow-up observations). The data is typically summarized and related to the parameters that may

address governance, quality and cost information. One level further, say at the level of a geographic region, each population-based CR may collect the data from every hospital in the region, but only for the conditions of specific interest to itself. The regional population-based CRs covers many more cancer cases and aggregates certain types of data elements, but also results in increasing fragmentation of clinical data as well as even further loss of dimensionality. Finally, population-based CRs typically summarize their data further to a global CR. Following the same process as the previous step, the global CR enables more universal coverage of cancer cases but with reduced data elements. The height of each layer represents the variety of the data (e.g., richness and dimensionalities). The bottom "layer" is thickest, indicating that it consists of the most varieties of data elements. In contrast, the upper "layer" is thinnest, indicating that it includes fewer dimensionalities of the data. Such data collection architecture can result in three serious issues:

1. The central CR cannot collect all the data, part of which are very important for modern cancer treatments and control (e.g., images data), from those data sources,
2. It usually spends much expense and takes a long time to report data to central CRs and collect data from data sources,
3. There is no complete network between different data sources, which seriously impedes linking, integrating and exchanging data.

The volume of data in oncology is large and increasing rapidly and it is already well beyond the processing capability of humans. Some companies (e.g., Google) are able to show that handling this volume of data is no longer a technical hurdle. A major problem is that most of the data in oncology is still unstructured, that means clinical knowledge implicit in the data cannot be explicitly mined by the use of machines. Without major disruption to the way data is stored and the manner in which data is used to generate clinical knowledge, essential insights for personalized medicine and truly participative decision support systems (that involve predictive modelling of treatment effects) will remain beyond our reach.

Big data in Radiation Oncology

With the advances in information technology, computing power, and digital medical equipment, the amount and types of data elements generated from just a single patient during routine cancer treatment are increasing, in terms of observational, biological, genetic, imaging and omics data. However, the data is usually collected at various points of care and then stored in varying formats

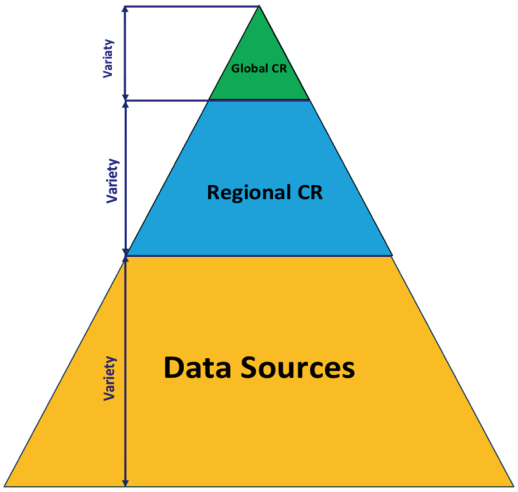


Figure 1: Illustration of Cancer Registry (CR) in terms of hierarchies of data collection. Data sources (clinics, hospitals) hold records of highest granularity (individuals) and greatest variety. At the level of a Regional CR, data from multiple sources are aggregated, but this typically reduces the granularity and variety of data elements. At the next level of aggregation, such as Global CRs, more universal coverage of the cancer cancers may be achieved across multiple regions and/or countries, but generally with greatly reduced dimensionality of data.

inside separate databases. As discussed in the preceding sections, this presents significant challenges when trying to aggregate, analyze and disseminate this data through traditional methods such as CRs. To employ these data to improve cancer treatment and health care of patients, an architecture specifically designed for big data is needed.

Big data research refers to the collection and analysis of a large volume of data elements and inter-relationships that are difficult to discover through traditional methods. Big data approaches have been used in many areas of medicine, including comparison of innovative techniques **[29]** and treatment modalities **[30]** in the field of radiation oncology.

A large volume of data has been generated within radiation oncology field, mainly due to the frequent use of innovative techniques (e.g., medical imaging) in diagnostic and therapeutic procedures, during the routine practice of modern radiotherapy treatment [31]. The data within radiation oncology displays all of the most significant hallmarks of big data, that is:

1. The use of data-intensive imaging modalities (**Volume**);
2. Imaging generates a large amount of data per time interval (**Velocity**);
3. There is an increasingly diverse spectrum of modalities available (**Variety**);
4. Objectivity and quality of collected data vary greatly, which can influence accurate analysis (**Veracity**).

As illustrated in **Figure 2**, a CR architecture is generally able to cover almost 100% of cancer cases (**Volume**), but only about 3% of potentially prognostic factors are collected, along with a moderate data loss rate of approximately 20%. A different data-generating paradigm in oncology (i.e., clinical trials), typically enrolls around 3% all eligible cancer cases [32-34]. However, due to stringent quality assurance and strict protocolization, nearly all of the factors

of interest are recorded (**Variety**) for only a relatively low (perhaps around, 5%) missing data rate.

What we conceive of as "Big Data" in oncology is represented by the entire graphic in **Figure 2**, where a vast volume and high throughput of data is continuously being generated by routine operations of modern cancer diagnosis and treatments. It is clear that neither CRs nor clinical studies adequately cover the full range of variety and volume of the clinical information available. However, as indicated in the figure, the consistency of data capture in the real world is generally low, thereby resulting in a high rate of missing data (perhaps around 80%). Furthermore, unlike the rigid protocols required by clinical trials and registry submissions, interobserver biases and divergent interpretations are additional data quality issues. With increasing use of automation and electronic data capture technology in oncology clinics, we expect that the potential bias and variation in the data (i.e., **Veracity**) will continue to improve.

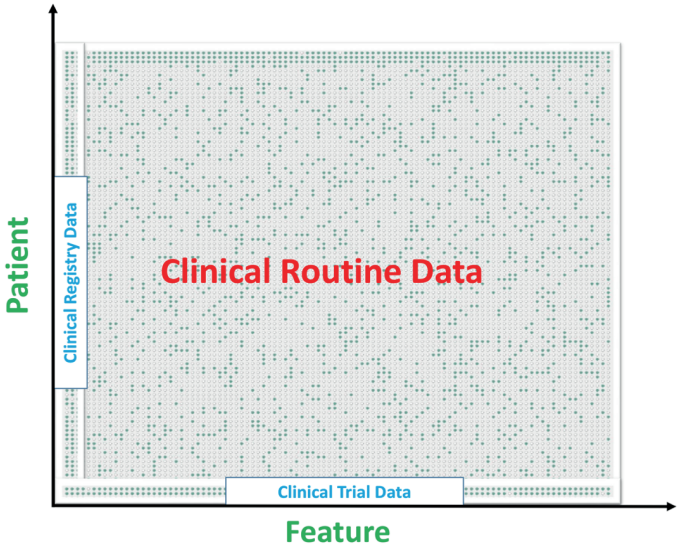


Figure 2: Schematic illustration of data fragmentation. Clinical trials generally record numerous variables per patient with high completeness, but only cover a small segment of the population. Conversely, a cancer registry tends to record a few key variables for a broad spectrum of the population. However, the schematic shows that these approaches miss an immense pool of data generated in routine clinical work, which however tends to be poorly structured and suffers from missing values.

There is strong motivation within the oncology field to exploit hitherto underutilized real-world data. Improvements in treatment outcomes guided by big data utilization, in combination with registry studies and clinical trials, are widely seen as the most effective avenue of delivering value-based healthcare [35-38]. Big data therefore lies at the core of a personalized approach to medicine that leads to increased value of treatment for a given financial outlay. The question of how to use real world big data to improve cancer
control and reduce treatment-related toxicity has gained increasing traction over time. Closely related questions of interest convers improving data collection coverage in real clinical settings,

imputation of missing data, and validation of outcome predictions in a wide variety of clinical settings.

As one of its primary objectives, big data research is expected to find the multiple clinical biological, and treatment variables that are related to treatment outcomes (e.g., overall survival toxicity) [39]. This benefits the creation of better predictive models that promote the advances of personalized therapies for each individual patient (e.g., delivering more aggressive therapies where needed and less aggressive treatments when appropriate). In order to develop reliable and robust prognostic outcome models, data sharing and exchange is required between multiple institutes. The reasons are threefold:

1. Obviously, each institute has a limited capability of data collection. The amount of new patients who are diagnosed or treated in a clinic may range from hundreds to thousands per year. Thus, it is reasonable to estimate that the amount of cancer cases stored in-house an individual clinic's past one or two decades within a clinic should be 20 to 200 thousand. Modern analytics methods (e.g., machine learning) are poised to satisfy the promise of identifying and guiding the response to variables influencing treatment outcomes of patients. However, the value of these analytics methods is highly dependent on data volume used for learning. As an illustrative example for size of training data, the hotly-debated artificial intelligence (AI) computer program in 2016, AlphaGo (designed by Alphabet Incorporated's Google DeepMind in London) was initially trained to mimic human gameplay from a historical games database containing approximately 30 million moves [40]. The data volume within an individual institute is often not sufficient to build such reliable and robust predictive outcome models through modern machine learning algorithms. Thus, data integration from multiple centers is necessary to develop realistic, understandable and robust predictive outcome models.
2. Data collected within a particular hospital usually refer to local patients with some specific types of cancer. For example, the incidences of esophageal cancer vary widely among countries, with approximately half of all cases occurring in China. For other countries, it is difficult to collect sufficient data elements of esophageal cancer data both for volume and dimensionalities due to the very small amount of patients with esophageal cancer. To generate prognostic outcome models, data exchange between two or more institutes seems a feasible and efficient approach.
3. As the predictive models are trained through local cohorts, it may be robust to predict the treatment outcome (e.g., overall survival) of unseen cancer cases within the local population. However, the predictive performance may be poor when applied to other populations, which impedes the usability and extension of the predictive model. Therefore, external validation is always necessary to measure the performance of a predictive model.

Because of the reasons explained above, data exchange between multiple clinics is a necessary procedure in radiation oncology field. Hence, there is a need to develop robust data exchange architectures instead of traditional methods to handle big data within radiation oncology field. The future data architectures must be able to: (1) scale to process ever-increasing amounts of data (**Volume)**; (2) have the throughput capacity to deal with high rates of data generation, in

particular from imaging modalities (**Velocity)**; (3) process many different types of data into a form that is amenable to machine-based analysis (**Variety)**; and (4) intercept issues of data quality (e.g., bias, non-reproducibility and abnormality) (**Veracity).**

Big Data Exchange in Radiation Oncology

Big Data Collection

Within radiation oncology, multiple types of data are routinely generated in the clinic from a variety of sources, which is the basis for big data research and provides an opportunity to improve cancer care.

One important source of big data is the patient demographics and clinical baseline factors obtained at the very beginning of the radiation oncology process. This information includes information about family history and personal health status. Crucially, this form of big data also consists of clinical observations and a baseline for treatment-related outcomes (especially comorbidities before treatment) by which the effectiveness of cancer interventions is evaluated.

Furthermore, an increasingly important source of big data, especially in regard to volume and variety, is the data stream produced by radiological and diagnostic imaging modalities. In radiation oncology, this routine data generation involves CT, PET, MRI. The volume of image-based data is increasing rapidly due to the use of daily verification imaging with the patient lying in the intended treated position (e.g., for cone beam CT). It is no surprise that the largest (by volume) repositories of data in radiation oncology is PACS.

Radiotherapy treatment planning is a highly sophisticated and computationally intensive process that generates big data in the form of organ delineations, beam geometry, radiation energy, collimation settings and spatial dose distribution. This data generally resides within the radiotherapy Treatment Planning System (TPS). With the growing trend towards adaptive radiotherapy responding to real-time imaging at the point of treatment delivery, this volume of data is also set to grow rapidly.

Lastly, a rapidly developing data source is the result of digital pathology and high-throughput specimen analysis from medical laboratories. This includes genomics, proteomics, metabolomics, histologics and hematologics. **Table 5** provides an overview of many of the possible radiotherapy research data types.

Table 5: Radiotherapy research data types [41].

DATA TYPE	DATA EXAMPLES
Baseline clinical data	Demographics information, TNM-stage, date of diagnosis, histopathology
Diagnostic imaging data	Diagnostic CT, MR and PET imaging
Radiotherapy treatment planning data	Delineation sets, planning-CT, dose matrix, beam set-up, prescribed dose and fractions

Radiotherapy treatment delivery data	Cone beam CTs, orthogonal EPID imaging, delivered fractions
Non-radiotherapy treatment data	Surgery, chemotherapy
Outcome data	Survival, local control, distant failure, toxicity, quality of life
Follow-up imaging data	Follow-up CT, MR and PET imaging
Biological data	Sample storage, shipping, tracing and lab results
Additional study conduct data	Study design, protocol, eligibility criteria

Source: Skripcak, T. et al., Radiother. Oncol., 113, 303–309, 2014.

Standard and Framework for Data Exchange

Efforts have been made to standardize data exchange between medical information archival systems with DICOM standard and HL7 interoperability standard.

DICOM [42] refers to a standard in medical imaging, which is supported by all imaging systems in medical field and used widely. DICOM has two identities: a type of file format and a network communication protocol. First, medical image systems generate DICOM files containing patient information (e.g., name, identifier, gender and date of birth) and then acquires information of medical image systems and corresponding settings. The images are stored in DICOM files. Second, the DICOM protocol can be employed to exchange data (e.g., image or patient information) between different systems that are connected to the network within the hospital. DICOM has shown its ability to improve data exchange in medicine field. Radiotherapy data are commonly exchanged using a subset of DICOM often referred to as DICOM-RT.

HL7 [43, 44] refers to a widely accepted standard-setting organization that provides standards to define the protocol, language and data type used for information communication among different systems. The most used version of HL7 is version 2 with which only a limited and not semantically rich data can be exchanged. HL7 version 3 had a much wider scope but is generally considered a failed standard due to its complexity and limited uptake. HL7 FHIR is the most recent standard and is receiving a lot of positive attention from the community and has resulted in real-world implementations by medical vendors.

Data Pooling Architectures

As described above, data integration is necessary procedure of modern studies in radiation oncology field. A data pooing architecture is used for data processing, storage, management and exchange within an individual institute or between multiple institutes. In this section, we will describe three data pooling architectures (i.e., centralized, decentralized and hybrid architecture), of which the infrastructures are shown in **Figure 3**.

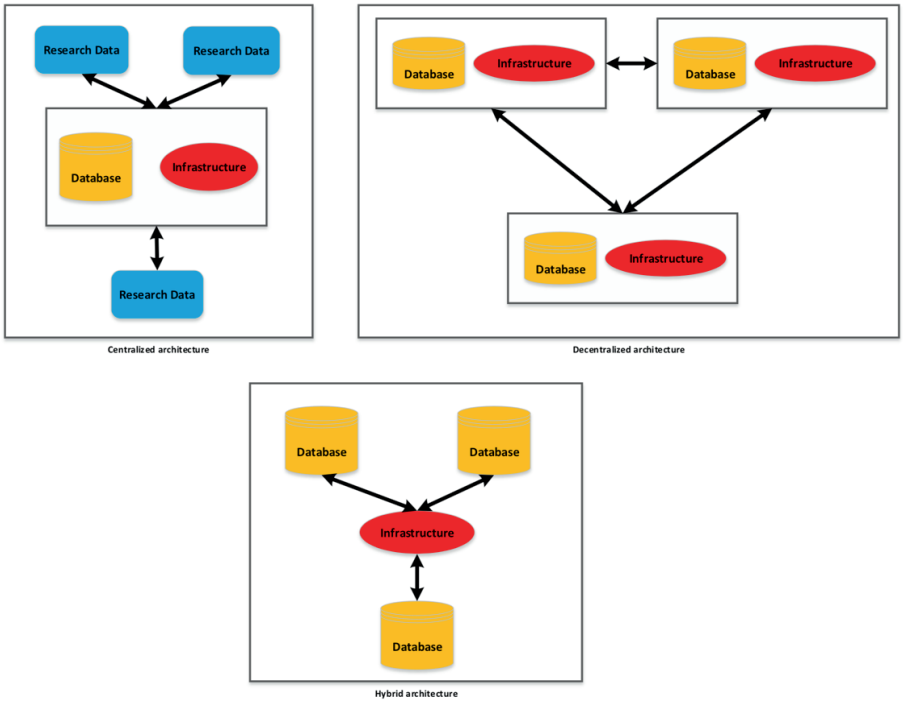


Figure 3: Illustration of three types of data pooling architectures. The centralized architecture stores data in a single centralized repository with one homogenized data structure and common infrastructure imposed on all data contributors. A decentralized architecture permits multiple data structures and more than one infrastructure, so collaboration has to be based on agreed standards for data exchange. A hybrid (or federated) architecture tries to combine elements from the previous two; heterogeneous data structures co-exist independently of each other, but a common infrastructure provides interoperability and connectivity between the databases.

Centralized architecture

A centralized architecture has complete control over the data pooled in a centralized repository. There is no direct real- or near real-time connection among participating institutions and operations (e.g., push/pull transaction and auditing occur in a central server). Although the architecture of centralized model is simple, it can raise several issues including privacy and anonymization, duplication of data, mapping the local data to the central data model (usually resulting in manual data entry/copying) and intellectual property (IP) rights **[41]**. The most significant advantage of centralized model is that the data is stored in a centralized repository, which makes data management and access easy.

The CR architecture described in **Section 2** can be considered as an instance of the centralized model. A centralized CR collects data of patients with cancer from different sub-data-sources (e.g., hospitals, cancer centers, and other institutes) through regular standards. Afterward, the centralized CR reports the analysis based on these data to the public and local government. However, an institute is usually not allowed to access the data within another institute due to data privacy of patients or local laws. It means there is usually no direct connection between these data sources.

Decentralized architecture

A decentralized architecture enables data exchange to occur among multiple institutions without any mediators, which is project-based via direct communication [41]. However, the infrastructure required to enable data exchange must first be established at each site and comply a standard exchange protocol. Shared data may be persistent (i.e., stored after exchange) or volatile (i.e., nothing is stored after exchange).

Hybrid architecture

A hybrid architecture attempts to combine the strengths of centralized and decentralized architectures. Data is transferred through direct communication across multiple sites. The difference between the hybrid and decentralized architectures is the information on infrastructure, data representation format, controlled terminologies, and other required metadata are stored in a central server, making the maintenance and modification of data exchange setting easier. Another advantage of a hybrid architecture is that big data generated in a local context is conceptually centralized but stored locally (i.e., inside the hospital or clinic). This hides the local complexities that differ for every clinic, below the level of the multicentric sharing. This can happen using the agreed-upon terminology. We prefer decentralized and hybrid architectures for data exchange across multiple centers based on the present IT systems within hospitals **[41]**.

Data Interoperability

Data interoperability has an important role in data exchange among multicenter: It is concerned with the capacity of a system to read and understand data transferred from another system. To implement complex and comprehensive data analyses, the data sources are required to be made fully interoperable across multiple IT systems. Data interoperability consists of two main sub-principles: (1) to enable data exchange between multiple institutes, all institutes need to have syntactic interoperability in reference to establishing uniform data formats and exchange protocols. In other words, data representation for writing and reading information should be identical among all institutes syntactic; and (2) syntactic and semantic interoperability should be in place, as described by [45]. The aim of data semantic interoperability is to make data consistently understandable by machines **[41]**.

In a real-world scenario, it is difficult to achieve data interoperability because of privacy of patient data, local policy, and even technical issues. To enable data interoperability among different IT platforms, certain technologies (e.g., Semantic Web and ontology) have been applied to big data exchange in radiation oncology field.

Semantic Web

The Semantic Web (also known as “Linked Data”) is an extension of the Web via many standards by the World Wide Web Consortium (W3C). The standards boost the development of data formats and communication protocols on the Web. Among the various data formats in the semantic web, Resource Description Framework (RDF) is the most fundamental format and is commonly used. The rationale behind the RDF data model is that any arbitrary statement about resources within the web can be represented by a simple triple (i.e., subject, predicate and object). Any levels of complexity in the descriptions of resources are possible using

multiple lines of triples. The subject and object here can considered as two resources. The predicate is the property of the subject and represents the relation between the subject and object. For example, a patient's survival age, biological sex and type of carcinoma can be described in the RDF format. **Figure 4** shows the virtual representation of this ontology.

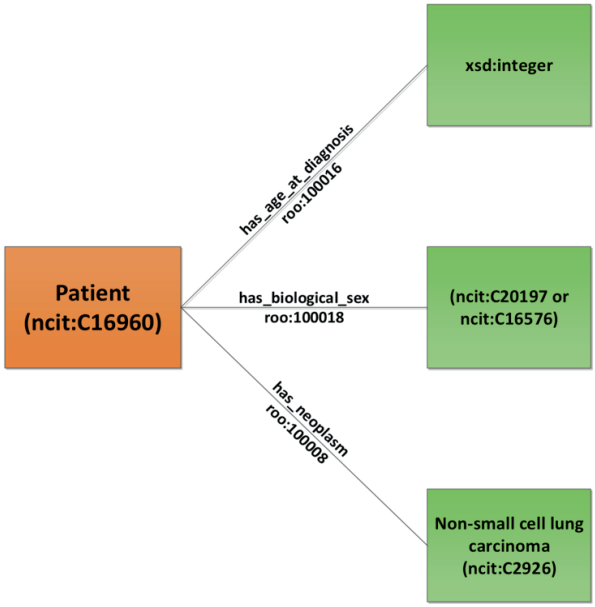


Figure 4: Graphical representation of the hypothetical statement "A 58-year old male patient with confirmed non-small cell lung cancer" in the Resource Description Framework (RDF) format. All resources (i.e., data elements) are represented in the form of "subject-predicate-object" triplets. Subjects and objects are typically members of a class of entities which are identified by unique codes, such as "Patient = ncit:C16960," "Male = ncit:C20197," and "non-small cell lung cancer = ncit:C2926." The relationships between class entities are defined using predicates such as "has age at diagnosis," "has biological sex," and "has neoplasm," each also having their own unique codes. Values or measurements such as "58" are literal quantities, but may themselves possess relationships defining the type of literal (integer) and its relevant units (years).

Storing data in the Semantic Web-based triple store is of great advantage for data access, usage and exchange when compared to a relational database (e.g., SQL). A situation that arises often is data exchange between two hospitals (A and B), which use different relational databases (i.e., the name of each column may be different as well as the internal linkages within the respective databases). Given that each database likely comprises many thousands of rows, the problem of integrating these two relational tables is a serious problem and requires both parties to have in-depth knowledge of the other's relational data structure. Querying a non-existent field in the other party's database may cause the query to crash. Therefore, one cannot add, delete or otherwise move records around without informing each other. The need for such knowledge necessarily precludes the ability to preserve privacy and data confidentiality. In contrast, it is trivially simple to integrate data that are stored in triple stores. Because everything is stored as statements with only three pieces of information per statement, a single additional line with a reference from one database into a unique resource identifier in the other is typically all that is needed. In triple stores, there is no constraint that a data structure must be known in

advance for a query to work (returns null instead of crashing), and therefore lines can be added, removed or ordered in any fashion without affecting the query.

Consider, as an example, that the data on a patient is stored in a relational table called "Patient_1" with several columns and that all the radiomic features of the same patient are stored in another separate table called "Radiomics_Patient_1". Linking the two tables is often not an easy task, which means that we have to add new columns of each radiomics feature in "Patient_1" table and enter the value. However, if the data of patient and his/her radiomic features are stored in separate RDF triple stores, data integration is trivial. To link the two triple stores, we just need one single sentence to define the triple: "Patient_1 (subject) has_RadiomicsFeature (predicate) Radiomics_Patient_1 (object)" via unique resource identifiers (URI) that represent patient_1 and his/her radiomics features. Now, the two databases are linked in the entirety of its data elements. For this triple, the class entities are defined in the Radiation Oncology Ontology (ROO) and the Radiomics Ontology (RO)

Based on the current development of technology, the Semantic Web is a more feasible and flexible choice for data representation in radiation oncology field. As the data is represented through ontologies, the Semantic Web enables seamless linkage of data that are stored in different data platforms. Another benefit of the Semantic Web is as that searching and accessing data can be done through web technologies that are known to be extremely scalable. Recently HL7 FHIR has also been specified in RDF format.

Ontology

An ontology refers to a terminology dictionary that defines the commonly used entities and relationships between entities in a particular domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. There are several ontologies available in radiation oncology field such as ROO (https://www.cancerdata.org/roo), the National Cancer Institute Thesaurus (NCIT) ontology [46] and International Classification of Diseases (ICD) ontology. When linking an entity to an ontology, the unique code in the corresponding ontology will be used to replace the value stored within the local database. The main purposes of using ontologies are as follows:

1. One of the primary purposes of using ontology is to share the common meaning of information among humans or machines **[47, 48].** For instance, the literal representation of a patient's biological sex may be "female" or "male". When linking to the NCI Thesaurus ontology, codes C16576 and C20197 are used to represent "female" and "male", respectively. One of the advantages is that a machine only needs to recognize the ontology codes that represent the common concept (e.g., biological sex) and ignore the literal representations stored within local databases (e.g., "f/m", "female/male" and "0/1").
2. Knowledge defined in one domain can be reused in another domain through using ontologies. For instance, different domains need to represent the concept of treatment of oncology including radiotherapy, chemotherapy, chemo radiotherapy, and surgery. If some group built an ontology on treatment of oncology, this knowledge can be reused simply in other domains. In addition, it is commonly to reuse a general ontology (e.g., NCIT ontology).

3. The differentiation of the domain knowledge from the operational knowledge is another important purpose. As an example for radiotherapy, we can define as domain knowledge that a planning target volume is based on margins around the clinical target volume while still allowing local operational knowledge form a trial or institutions to describe if and how the margins where applied.
4. If a declarative explanation of the term is available, it is possible to analyze domain knowledge. As an example, a radiation ontology specifying that radiation on the chest may result in nonbacterial radiation pneumonitis might be used in another context to prescribe steroids instead of antibiotics in these patients.

Data Exchange

As mentioned above, different types of data have been routinely generated in a clinic from a variety of sources, which can be benefit of cancer care. In order to ultimately reach FAIR data **[6]**, data exchange among different data sources is a necessary and important step. Therefore, architectures of data exchange should be built based on the rules such as efficiency, safety and veracity. For the purpose of data exchange, two manners (i.e., manual and automatic data exchange) are used for both internal and external data exchange. This operation can be considered as "send data out", which will be described in **Section 4.5.1**. However, "send data out" is not suitable and efficient for exchanging big data between multicenter in one or more countries because of the reasons such as IP rights, local policy, and patient data privacy. Instead of sending data out, it is better to keep data inside hospitals and "send questions in". This method is known as "distributed learning", of which the details will be described in **Section 4.5.2.** Finally, the comparison between centralized and distributed multicenter architectures for big data exchange will be described in **Section 4.5.3**.

Send data out

Internal exchange

In many situations, text data of patients with cancer are generated in a hospital, such as (1) personal information forms on demographic information and medical histories filled by patients, (2) diagnosis and treatment notes created by doctors, and (3) follow-up details. These data can be in free text or structured questionnaires, which are entered in departmental (e.g., Oncology Information System) and/or hospital systems (e.g., electronic health record [HER] system) for storage, management and analysis. The data are exchanged, usually through the EHR platform, from one department to another department within a hospital.

In addition, some types of data generated in the process radiotherapy can be transferred to computers automatically, including imaging scans (e.g., CT, PET or MRI), tumor delineations, and treatment plans. One of the common properties of these data is that they are generated by electronic equipment (e.g., scanners and computers). The medical imaging data is usually transferred using the DICOM protocol to a central imaging archive (e.g., PACS). Another type of data that is often digitally exchanged are laboratory results.

External exchange

For data exchange among multiple centers, one commonly used approach is based on multicenter clinical trials. Data are often transferred via mail, fax and email between two or more sites or to a central location. Although email is popular and convenient for receiving and sending information, it may result in three serious issues when involves medical data exchange: (1) missing data and data security (2) there is no standards for data exchange via email, and (3) transfer efficiency will be very low for large volume of data (e.g., image data).

Another approach of external data exchange is through web-based application. For example, Openclinica (**https://www.openclinica.com**) is developed for clinical data federation. Indeed, the use of Web-based products for data exchange have some benefits such as reliability and flexibility, although their use still can result in privacy-related issues.

Unlike in clinical research, data exchange for health care is not well developed and regular mail and faxes are still often the main manners of information exchange. Hence, there is a need for a more efficient and powerful approach for big data exchange among multiple centers.

Send Questions In

As described above, data sharing between multiple centers often raises privacy-related issues. A better manner is to keep data in hospitals and "send questions in" [49-51] rather than sharing data. The primary goal of data sharing is to mine knowledge from others' data. If there is an approach that can answer the research questions without allowing data outside of the hospitals, sending data out is unnecessary. The distributed solution allows advanced data analysis (e.g., knowledge sharing). Mathematical models are trained on local databases and shared to other hospitals. Because models contain only the "answer" of the question while the research data are kept within the hospital, using a mathematical model avoids privacy-related issues. Only some aggregate parameters are transferred between multicenter to reach the global convergence (consensus) of the mathematical model. This approach is known as "distributed learning" [5].

In addition, this distributed learning can be implemented on a Web-based learning environment (e.g., Varian Learning Portal). The learning platform can be considered as the master that merges knowledge models learned from different participating sites and continuously updates the model when more data are available [41].

Centralized vs Distributed Learning Architecture

Modern medical research has to process an increasing number of data generated from many fields such as medical imaging, genomic, and proteomics. However, the reality is that an individual hospital only has data on a limited number of patients, which may be not sufficient to medical research. From the experience of machine learning in other fields, we know one needs a sufficient number of events to build a reliable predictive model for cancer treatments. In general, the more data collected from different sources, the more robust a predictive model is. Thus, cooperation between two or more hospitals is needed to collect more data regarding patients with cancer. The architectures of centralized and distributed learning among multiple centers has been described in [52].

Figure 5 shows the general overview for the centralized multicenter architecture. This approach allows participating sites to build the institutional architectures based on local policies. In addition to the entry points of all institutions, there are two key components within

in a centralized learning architecture: (1) a central machine learning server is the place where learning occurs; and (2) a central collection point is responsible to perform the horizontal federation of data between all sites. As an example to explain the learning process, first participating Site A sends an algorithm to the central machine learning server. Second, the central server implements calculation of this algorithm on the centralized data repository. Finally, the results are sent back to Site A after the calculation is completed.

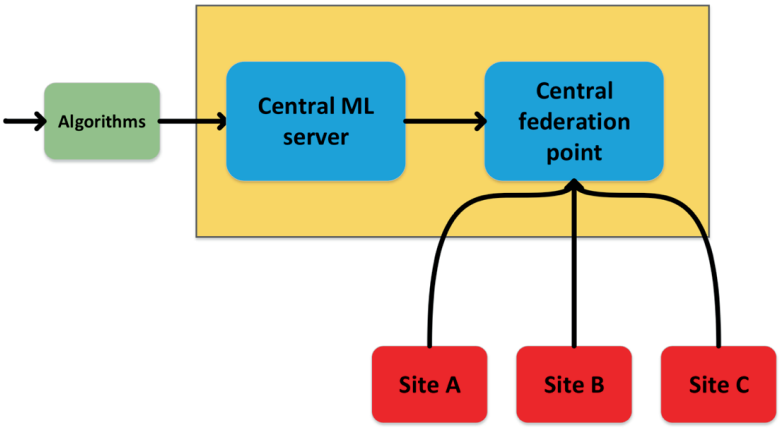


Figure 5: Schematic of a Centralized Multicenter Learning Infrastructure. The centralized learning approach forces each site to contribute structured data into a central accumulation point, whereupon algorithms can operate on the data via a machine-learning (ML) server co-located with the data to obtain the global result based on all the data.

On the other hand, the distributed learning architecture between multiple centers is different in terms of the places where computations happen [49, 52, 53]. **Figure 6** displays the general overview of distributed learning architecture. We can see that the local unit has been added and the central federation point has been removed. In this architecture, the responsibility of the central machine learning server is only coordination. First, a site submits an algorithm or query to the central machine learning server. The algorithm or query is split into several small sub-tasks. Second, the small sub-tasks are packed and sent to local machine learning units within each site. They will query the local data that is stored in RDF triple stores and the sub-algorithms are implemented in local sites. The local application learns a model from local data. Third, the central machine learning server will merge all the all results that have been computed on the local machine learning units of all sites [52]. Finally, if the preset criteria are met, a final model is generated. If not, the central machine learning server will send the models to all sites for re-learning until reaching the preset convergence criteria have been reached.

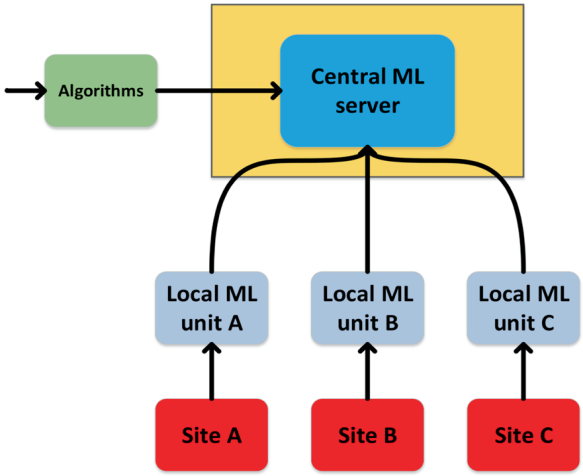


Figure 6: Schematic of a Distributed Multicenter Learning Infrastructure. In the distributed learning approach, a central machine-learning (ML) server splits the intended algorithm to each of the sites. Each copy of the algorithm operates only on the site-specific data via its own local ML unit. Results of computation from each site are returned independently and asynchronously to the Central ML server, which then has to process multiple site results into a single global result. Where necessary, the Central ML server may iterate through the site computation steps many times.

The significant difference between distributed and centralized machine learning architecture is the transfer of data versus the transfer of model weights. When performing centralized learning, data leaves the hospital and is send to the machine learning system. In contrast, data is kept within the hospital when we perform distributed learning. In this setting, the volume of data that is needed to be transferred is decreased compared to the centralized learning architecture; however, the transfer efficiency per task is increased. Boyd et al. **[54]** and Wu et al **[55]** have given a complete explanation of how distributed machine learning algorithms work in their publications.

As proof that the concepts covered in **Section 4** are indeed practical, a real-world machine learning project on distributed databases, known as Computer Assisted Theragnostics (CAT), has been proposed as illustrated schematically in **Figure 7**. First, data (e.g., image, genomics) are collected from a variety of data sources within each site and stored in local databases. Second, these data within each site are converted into the RDF data format and stored in the Semantic Web-based tripe stores. Third, as shown in **Figure 7**, researchers in (for example) Rome can send research questions to (for example) Oxford via a global learning server. Then, learning happens on the local learning server in Oxford. After finishing the local learning, the results are sent back to the global learning server. Finally, researchers in Rome can query the global server for the answers of the research questions they proposed. This pipeline enables the communication between two or more institutes that have participated in the CAT project. Data exchange via the CAT architecture leads to a few benefits as:

1. Because the information shared in the CAT pipeline is the results (i.e., answers of research questions), not the real data. Data exchange and knowledge mining occur without leaving data outside the local hospitals, avoiding data privacy-related issues;
2. All the data items stored in different databases within each hospital are standardized by the same ontology, leading to linked data;
3. In the example above, researchers in Rome only need to send the algorithms or query (known as a "research question"), which is a small package (around hundreds of kilobyte) comparing with sending the real data (large volume).

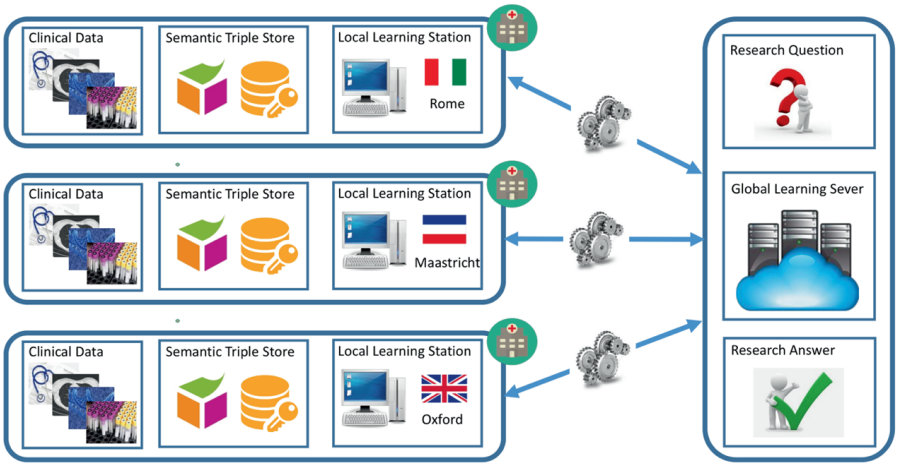


Figure 7: The diagram of a real-world machine learning project on decentralized databases, known as Computer Assisted Theragnostics (CAT). The schematic shows three hypothetical sites separated by language and geography, as well as site-specific data structures and local institutional infrastructures. The CAT workflow at each site always involves the same general steps of (i) removing personal identifiable information and mapping the local data into a tentative data structure, (ii) providing semantic interoperability using ontologies and describing the local data in Resource Description Framework (RDF), and finally (iii) connecting the site RDF data to a distributed multicenter learning infrastructure such as the CAT system.

A few projects (e.g., I2B2 (https://www.i2b2.org), EURegiOnal Computer Assisted Theragnostics [EuroCAT] [56] Validation of high Technology [VATE] [53], etc.) have demonstrated the feasibility of distributed multicenter machine learning. In addition, if we implement the procedures distributed learning correctly, the results can the same as the centralized learning approach [50]. Furthermore, it is possible to improve the robustness of prognostic models through validating on an external data sets [57], which has been demonstrated based on the results of the EuroCAT project [56].

Barriers of Big Data Exchange among Multicenters

This section will describe the barriers of big data exchange among multiple institutes in radiation oncology community. The exchange barriers involve four primary aspects: administrative, ethical, political and technical barriers. The details are as follows.

Administrator barrier

Two main barriers of administrative impede data exchange among multiple institutes: data completeness and bias**.** First, it is usually not possible to collect every data element of an individual patient and not all data elements are applicable, resulting in many missing data of the patient. It means there are no data to exchange. Second, system settings vary widely across sites. The bias in routine operation, protocols and equipment settings all can result in difference between data across different sites. Data bias highly impedes data integration among multiple sites.

Ethical barrier

The ethical issues refer to data privacy and reuse of research data. There may be a large difference in privacy explanation, application of confidentiality and legislation between counties and even ethical committees [41].

For both centralized and distributed multicenter infrastructures, privacy preservation is a major topic that must be considered. If correctly implemented as the applications described in preceding section, the distributed solution is generally more secure than centralized solution because only a few parameters are transferred among multiple centers rather than the real data, although we cannot draw the conclusion that the issues on privacy preservation have been overcome via the distributed solution. Actually, there are no standard methods to solve privacy-related issues currently. Various stakeholders will always have to find a balance between the value of information and anonymity of participating patients.

Political barrier

In some scenarios, people do not want to share their data to others because of the issues related to the culture and local policy. Thus, there is a need of more high-quality and published research articles that completely prove the benefits of data exchange (e.g., efficiency, robustness and security) to try to persuade data holders to participate in the collaborative research community and subsequently share their data.

Technical barrier

Even if these administrative, ethical and political issues are solved, the technical barriers such as interoperability between clinical departments and lack of a uniform standard of data collection may still impede data exchange among multiple institutes. First, obtaining internal clinical IT systems interoperability is important for the generation of local anonymized data sets, which enables a universal access to integrate research data. These data are managed by an institutional data warehouse and can be findable through the corresponding semantic models (ontologies). However, it is often difficult to reach this interoperability in real-world scenarios because of the differences in their support of internationally standardized protocols, formats and semantics. Although these can be solved, they often require an investment in resources that is not available in the operational setting.

Second, since the radiotherapy terminologies dictionary and ontologies are still under development, it is difficult to ensure that an element has a unique term and definition [41]. As an example, described above, different hospitals may use various representations to describe the biological sex of a patient such as “female and male”, “f and m” or “0 and 1”. While

performing computations on data from the two different origins, two incompatible representations will be encountered. The best way to overcome this issue is to link clinical variables to ontologies that can provide the standard definitions of these variables. The biological sex of a patient "female" and "male" are represented by NCI Thesaurus codes C16576 and C20197, respectively. Thus, the meaning of a clinical variable is only related to its ontological code rather than the literal representation.

Conclusion

In this chapter, we have seen that how different types of CR work as an organization for cancer data collection, management, storage, analysis and exchange. However, the architecture of a CR is not easily scalable to exploit some properties of big data in radiation oncology, most notably **Volume, Velocity, and Variety**). In addition, it remains a complex and costly process to maintain a CR, which needs a large amount of human resource (e.g., registrars and managers) and infrastructure (servers and user interfaces).

One of the most important challenges to the universal adoption of CRs is how big data can be flexible and securely shared among a large network of cancer institutes or research programs to answer a broad range of clinically relevant questions. Thus, there is a need to build robust data exchange architectures to handle big data within radiation oncology field instead of traditional methods (e.g., CR). "Send data out" is the commonly used approach of data exchange via mail, fax, email or Web-based applications, although it may result in the privacy-related date issues. The best manner to avoid privacy-related issues, while exploiting multicentric data, is to avoid sending data outside hospitals. This can be achieved through applying the method "send questions in", which is also known as distributed learning. The distributed approach only transfers a subprocess machine learning algorithm to a specific hospital and send the results back to the sender rather than transferring real data. It means that data/knowledge exchange occurs without allowing data outside of the hospital. Many collaborative projects (e.g., I2B2, EuroCAT, VATE etc.) have demonstrated the feasibility of this distributed learning architecture in handling big data in radiation oncology field. If the procedures are implemented correctly, it can produce the same results as a centralized learning architecture.

References

1. Deng, J., Big data in radiation oncology: challenges and opportunities. Cancer Sci Res Open Access, 2014. 1(2): p. 1-2.
2. Chen, M., S.W. Mao, and Y.H. Liu, Big Data: A Survey. Mobile Networks & Applications, 2014. 19(2): p. 171-209.
3. McNutt, T.R., K.L. Moore, and H. Quon, Needs and Challenges for Big Data in Radiation Oncology. Int J Radiat Oncol Biol Phys, 2016. 95(3): p. 909-915.
4. Roelofs, E., et al., Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol, 2013. 108(1): p. 174-9.
5. Lambin, P., et al., 'Rapid Learning health care in oncology' - an approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol, 2013. 109(1): p. 159-64.
6. Wilkinson, M.D., et al., The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 2016. 3: p. 160018.
7. Lambin, P., et al., Predicting outcomes in radiation oncology-multifactorial decision support systems. Nature Reviews Clinical Oncology, 2013. 10(1): p. 27-40.
8. Aerts, H.J.W.L., et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications, 2014. 5.
9. Dehing-Oberije, C., et al., Development and Validation of a Prognostic Model Using Blood Biomarker Information for Prediction of Survival of Non–Small-Cell Lung Cancer Patients Treated With Combined Chemotherapy and Radiation or Radiotherapy Alone (NCT00181519, NCT00573040, and NCT00572325). International Journal of Radiation Oncology* Biology* Physics, 2011. 81(2): p. 360-368.
10. Dehing-Oberije, C., et al., Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol, 2010. 97(3): p. 455-61.
11. Dehing-Oberije, C., et al., Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys, 2009. 74(2): p. 355-62.
12. Oberije, C., et al., A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: A step toward individualized care and shared decision making. Radiotherapy and Oncology, 2014. 112(1): p. 37-43.
13. Stacey, D., et al., Decision aids for people facing health treatment or screening decisions. Cochrane Database of Systematic Reviews, 2014(1).
14. Elwyn, G., et al., Shared decision making: a model for clinical practice. J Gen Intern Med, 2012. 27(10): p. 1361-7.
15. Santos, S.I., The role of cancer registries, in Cancer epidemiology, principles and methods, S.I. Santos, Editor. 1999. p. 385-403.
16. Young, J.L., The hospital-based cancer registry, in Cancer registration: principles and methods, O.M. Jensen, Editor. 1991, IARC. p. 177-184.

17. Coleman, M.P., et al., Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. Lancet, 2011. 377(9760): p. 127-38.
18. Sadjadi, A., et al., Cancer occurrence in Ardabil: results of a population-based cancer registry from Iran. Int J Cancer, 2003. 107(1): p. 113-8.
19. Parkin, D.M., The evolution of the population-based cancer registry. Nat Rev Cancer, 2006. 6(8): p. 603-12.
20. Van Leersum, N.J., et al., The Dutch surgical colorectal audit. Eur J Surg Oncol, 2013. 39(10): p. 1063-70.
21. Gliklich, R.E., N.A. Dreyer, and M.B. Leavy, Data Sources for Registries, in Registries for Evaluating Patient Outcomes: A User's Guide, R.E. Gliklich, N.A. Dreyer, and M.B. Leavy, Editors. 2014: Rockville (MD). p. 127-144.
22. MacLennan, R., Items of patient information which may be collected by registries, in Cancer registration: principles and methods
, O.M. Jensen, Editor. 1991, IARC. p. 43-63.
23. Powell, J., Data sources and reporting, in Cancer registration: principles and methods, O.M. Jensen, Editor. 1991, IARC. p. 29-42.
24. Armstrong, B.K., The role of the cancer registry in cancer control. Cancer Causes Control, 1992. 3(6): p. 569-79.
25. Parkin, D.M., The role of cancer registries in cancer control. Int J Clin Oncol, 2008. 13(2): p. 102-11.
26. Storm, H.H., Cancer registries in epidemiologic research. Cancer Causes Control, 1996. 7(3): p. 299-301.
27. Bray, F. and D.M. Parkin, Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer, 2009. 45(5): p. 747-55.
28. Navarro, C., et al., Population-based cancer registries in Spain and their role in cancer control. Ann Oncol, 2010. 21 Suppl 3: p. iii3-13.
29. Chen, A.B., Comparative effectiveness research in radiation oncology: assessing technology. Semin Radiat Oncol, 2014. 24(1): p. 25-34.
30. Aneja, S. and J.B. Yu, Comparative effectiveness research in radiation oncology: stereotactic radiosurgery, hypofractionation, and brachytherapy. Semin Radiat Oncol, 2014. 24(1): p. 35-42.
31. Lustberg, T., et al., Big Data in radiation therapy: challenges and opportunities. Br J Radiol, 2017. 90(1069): p. 20160689.
32. Movsas, B., et al., Who enrolls onto clinical oncology trials? A radiation patterns of care study analysis. International Journal of Radiation Oncology Biology Physics, 2007. 68(4): p. 1145-1150.
33. Grand, M.M. and P.C. O'Brien, Obstacles to participation in randomised cancer clinical trials: a systematic review of the literature. J Med Imaging Radiat Oncol, 2012. 56(1): p. 31-9.
34. Murthy, V.H., H.M. Krumholz, and C.P. Gross, Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA, 2004. 291(22): p. 2720-6.

35. Murdoch, T.B. and A.S. Detsky, The inevitable application of big data to health care. JAMA, 2013. 309(13): p. 1351-2.
36. Khoury, M.J. and J.P. Ioannidis, Medicine. Big data meets public health. Science, 2014. 346(6213): p. 1054-5.
37. Larson, E.B., Building trust in the power of "big data" research to serve the public good. JAMA, 2013. 309(23): p. 2443-2444.
38. Schneeweiss, S., Learning from big health care data. N Engl J Med, 2014. 370(23): p. 2161-3.
39. Rosenstein, B.S., et al., How Will Big Data Improve Clinical and Basic Research in Radiation Therapy? International Journal of Radiation Oncology Biology Physics, 2016. 95(3): p. 895-904.
40. Metz, C., In major AI breakthrough, Google system secretly beats top player at the ancient game of go. 2016.
41. Skripcak, T., et al., Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiother Oncol, 2014. 113(3): p. 303-9.
42. Mildenberger, P., M. Eichelberg, and E. Martin, Introduction to the DICOM standard. Eur Radiol, 2002. 12(4): p. 920-7.
43. Dolin, R.H., et al., The HL7 Clinical Document Architecture. J Am Med Inform Assoc, 2001. 8(6): p. 552-69.
44. Dolin, R.H., et al., HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc, 2006. 13(1): p. 30-9.
45. Valentini, V., H.-J. Schmoll, and C.J. van de Velde, Multidisciplinary Management of Rectal Cancer: Questions and Answers. 2012: Springer Science & Business Media.
46. Sioutos, N., et al., NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform, 2007. 40(1): p. 30-43.
47. Musen, M.A., Dimensions of knowledge sharing and reuse. Comput Biomed Res, 1992. 25(5): p. 435-67.
48. Gruber, T.R., A translation approach to portable ontology specifications. Knowledge acquisition, 1993. 5(2): p. 199-220.
49. Damiani, A., et al. Distributed learning to protect privacy in multi-centric clinical studies. in Conference on Artificial Intelligence in Medicine in Europe. 2015. Springer.
50. Wiessler, W., PO-0886: Privacy-preserving, multi-centric machine learning across hospitals and countries: does it work? Radiotherapy and Oncology, 2013. 106: p. 343.
51. Lindell, Y. and B. Pinkas. Privacy preserving data mining. in Annual International Cryptology Conference. 2000. Springer.
52. van Soest, J.P., et al., Application of machine learning for multicenter learning, in Machine Learning in Radiation Oncology. 2015, Springer. p. 71-97.
53. Meldolesi, E., et al., VATE: VAlidation of high TEchnology based on large database analysis by learning machine. 2014.
54. Boyd, S., et al., Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning, 2011. 3(1): p. 1-122.

55. Wu, Y., et al., Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. Journal of the American Medical Informatics Association, 2012. 19(5): p. 758-764.
56. 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, 2017. 4: p. 24-31.
57. Dekker, A., et al., PD-0496: Multi-centric learning with a federated IT infrastructure: application to 2-year lung-cancer survival prediction. Radiotherapy and Oncology, 2013. 106: p. S193-S194.

Chapter 3

Data-Sharing and Toxicity Modelling - A Vision of the Near Future

Zhenwei Shi, Rianne Fijten, Zhou Zhen, Andre Dekker and Leonard Wee

Adapted from
Shi, Zhenwei, et al. "Data Sharing and Toxicity Modelling: A Vision of the Near Future." Modelling Radiotherapy Side Effects. CRC Press, 2019. 365-399.
DOI: http://dx.doi.org/10.1201/b21956-15

Introduction

Optimal cancer treatment requires maximizing the chance of tumor eradication while simultaneously minimizing the risk of adverse treatment-induced side-effects. The success of modern oncology implies that more and more patients (generally) live longer with the adverse outcomes of their treatment. Therefore, the ability to predict the likely trajectory of treatment-related toxicity is indispensable in value-based intervention against cancer. This can be achieved with predictive modelling of specific toxic events and their presumed severity. Estimation of probable toxicity touches multiple aspects of oncology decision-making, from modality selection down to individually-adapted treatment plan optimization.

In this chapter, a self-contained and compact overview of *data-driven toxicity modelling and prediction* is offered. The chapter opens with an examination of the data landscape and general data requirements to develop quantitative toxicity models. Next, data architectures that support data aggregation and data sharing for toxicity modelling are reviewed, along with advantages and disadvantages of each respective architecture. Lastly, we consider the current condition of toxicity models and prediction model development, followed by an exposition of some of the current developments in toxicity modelling that may soon be realized in routine clinical practice.

General Data Requirements

Available data

One of the major actions in the development of multi-variate prediction models of treatment-related toxicity is locating relevant data that exists (i) in highly structured format, (ii) as a fully digital archive and (iii) with a sufficiently high probability of data completeness. One needs to understand the *data landscape* – where we locate the data needed to answer the prediction question – before considering how to consolidate disparate elements together into a data set on which a prediction model can developed, i.e. the data corpus.

Structured data versus unstructured data

Structured data refers to information that is organized into *key-value* pairs, such that the keys (or labels of *variables*) are uniformly used throughout the data corpus, and the values consist of a known range of results with a uniformly applied format through the corpus.

Typical keys include data fields such as a patient record number, their date of birth, clinical tumor staging and one or more outcomes of interest. Keys are generally defined in such a manner as to enable efficient processing by software – no spaces, no mixtures of upper and lower cases and only alphanumeric characters – but should nonetheless bear close resemblance to its human-readable label. Formats of values must be consistent and uniformly applied – for example, date formats are "*dd-mm-yyyy*" throughout the data corpus. Other permissible formats could be alphanumeric strings (e.g. "MRN012345"), Boolean labels (e.g. "y/n", "0/1" or "true/false"), categorical labels (e.g. "Grade 1", "Grade 2", "Grade 3", etc.) or floating point

numbers (e.g. “50.4”). A very common form of structured data is the table format, such that the keys are arranged as columns and the values are arranged in rows (example in **Table 1**). This is by no means the only manner of arranging the key value pairs; a different but equally valid format of structured data is shown in **Figure 1**.

Unstructured data consists of information that has not been organized into key-value pairs, or where the keys do not exist or are not uniformly applied, or where the values are neither of consistent range nor of uniform format. The two most common forms of unstructured data are free natural-language text (e.g. digitized speech, consultation notes, radiology reports, pathology findings) and images (e.g. medical imaging scans, facsimiled laboratory reports, photos of hand-written notes). Examples of unstructured data are shown in **Figure 2**. In comparison to structured data, it is very much evident that unstructured data requires sophisticated data pre-processing, either by human experts or by machine algorithms, in order to be appropriate for building mathematical toxicity models.

Table 1. Example presentation of keys (column names) and corresponding values (row entries) as a rectangular table (adapted from https://datatables.net/examples/ajax/custom_data_flat.html).

id	name	position	location	start_date	salary
1	Airi Satou	Accountant	Tokyo	2008/11/28	162700
2	Angelica Ramos	Chief Executive Officer	London	2009/10/09	1200000
3	Ashton Cox	Junior Technical Author	San Francisco	2009/01/12	86000

More recently, there has been increased effort to record clinical findings in semi-structured format by defining consistent and well-defined sections within a textual report. However, the information within each of the sections remains free natural text, out of which semantically concrete values must be derived by pre-processing. Conversely, some efforts have been made to enforce labelled statements to be included in text reports such as “TUMOUR VOLUME: 3.5 CC” or “ESOPHAGITIS: GRADE 2”. While this makes extraction of quantitative value from free text easier, such conventions are generally not feasible to uniformly enforce in the clinical setting, nor would not be any guarantee that any one clinic’s labelling convention would be acceptable or used the same way by a different clinic.

Efforts towards structured data collection are more advanced in certain domains, such as *structured reporting standards* for radiology notes, e.g. BI-RADS [1] and PI-RADS [2]. However, such structured reporting standards are not yet universally adopted, and even in places where they are used, continuous quality assurance efforts are required to ensure compliance to the requirements.


```
[
  {
  "id": 1,
  "name": "Airi Satou",
  "position": "Accountant",
  "salary": 162700,
  "start_date": "2008/11/28",
  "location": "Tokyo"
   },
```

```
  {
  "id": 2,
  "name": "Angelica Ramos",
  "position": "Chief Executive Officer",
  "salary": 1200000,
  "start_date": "2009/10/09",
  "location": "London"
  },
  {
  "id": 3,
  "name": "Ashton Cox",
  "position": "Junior Technical Author",
  "salary": 86000,
  "start_date": "2009/01/12",
  "location": "San Francisco"
  }
]
```


Figure 1. Example Ajax data, containing exactly the same information as in **Table 1**, as a flattened list of key-value pairs (adapted from https://datatables.net/examples/ajax/custom_data_flat.html).

Gross Description:
The specimen is received in two parts. They are labeled #1, "biopsy bladder tumor", and #2, "scalene node, left". Part #1 consists of multiple fragments of gray-brown tissue which appear slightly hemorrhagic. They are submitted in their entirety for processing. Part #2 consists of multiple fragments of fatty yellow tissue which range in size from 0.2 to 1.0 cm in diameter. They are submitted in their entirety for processing.
Microscopic:
Section of bladder contains areas of transitional cell carcinoma. No area of invasion can be identified. A marked acute and chronic inflammatory reaction with eosinophils is noted together with some necrosis. Sections are examined at six levels. Section of lymph node contains normal node with reactive germinal centers.
Diagnosis:
1. Papillary transitional cell carcinoma, grade II, bladder, biopsy
2. Acute and chronic inflammation, most consistent with recent biopsy procedure
3. Scalene lymph node, left, no pathologic diagnosis

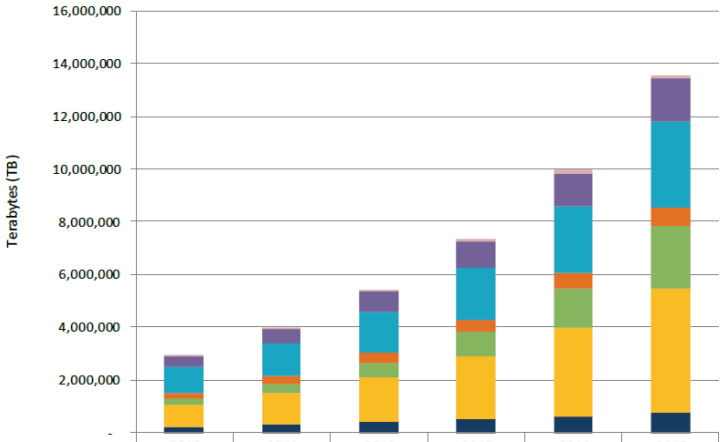
Figure 2. Example fragment of a pathology report from the National Cancer Institute training materials (https://training.seer.cancer.gov/abstracting/procedures/pathological/histologic/operative/example/ex1.html).

Therefore, a major challenge before engaging in toxicity modelling is to impose a well-defined structure onto unstructured data before it can be used in quantitative analyses. Recent surveys of the types of data in medical care centers suggest that approximately 80% of healthcare data by volume is unstructured (see **Figure 3**). Advances in natural language processing (NLP) [3-5] may help to transform free text into structured data, but comprehensive clinical validation of such approaches are still lacking. The difficulty of obtaining clinician-expert labels on medical text, in order to provide machine algorithms with suitable data on which to "learn", remains the critical bottleneck in NLP development for medical text analysis. Better progress has been made in the area of deriving quantitative metrics from medical images [6-9], but once again the availability of expert clinician-annotated image sets is the limiting factor. An area that has seen significant development in recent years are digitization of medical records. Specifically, optical-character-recognition (OCR) tool are available to convert images of written or typed notes into digital text. Increasingly, speech-to-text may be used to transcribe dictations and spoken conversations directly into digital text. However, the extant problems associated with extracting meaningful values from natural text, as described above, persist.

Re-use of clinical trials data

In light of difficulties in obtaining structured data for use in quantitative toxicity modelling, many researchers turn to secondary analysis, i.e. "re-use", of previously collected clinical trials data. In this respect, clinical trials generally excel, since vast amounts of trial financial resources would have been devoted to centralized data storage, rigorous data quality checks and systematic record keeping. Whenever this kind of data can be obtained, it is highly structured and has a high degree of data completeness.

However, as can be seen in **Figure 3**, structured research data (much of which can be attributed to the conduct of clinical trials) comprises only a small fraction of the entire healthcare data stack. In a rapidly evolving discipline such as radiotherapy, a randomized trial may be unethical [10] or unsuited to demonstrating clinical effectiveness [11]. Furthermore, clinical trials have been criticized for being unrepresentative of the wider patient population having the target disease [12,13] and for being excessively focused on tumor control endpoints rather than treatment-induced toxicity [14,15]. Clinician under-reporting of toxicity in trials has been well-documented [16]. Sample size in most trials have been determined using clinical endpoints *other than toxicity*, therefore it is highly unlikely that reliable toxicity models might be obtained from secondary analysis of clinical trial data. A possible solution would be to perform meta-analysis for toxicity using individual patient-level data derived from multiple institutions and/or multiple trials. However, such efforts are generally impeded by concerns about patient privacy and legalities of material (data) transfer between institutions.



	2010	2011	2012	2013	2014	2015
Research Data	45,701	57,429	73,471	91,330	112,749	139,393
Non-Clinical Imaging	440,070	576,422	752,877	973,819	1,252,292	1,606,082
General Unstructured Data/File Services	950,980	1,219,132	1,562,994	1,998,615	2,551,631	3,262,842
E-Mail	242,711	300,632	372,166	457,259	559,598	684,522
Electronic Health Records	219,719	356,398	574,443	922,804	1,500,001	2,380,643
Clinical Imaging	824,493	1,179,802	1,695,289	2,397,323	3,371,689	4,730,891
Administrative Applications	272,337	340,466	423,563	523,403	643,292	789,160

Figure 3. Estimated volume of medical data in all North American hospitals from 2010 – 2015, showing proportions of different data types. The majority of the data is unstructured, and of these, medical imaging contributes the largest share. Reproduced with permission from Enterprise Strategy Group Research Publication "North American Health Care Provider Information Market Size & Forecast", January 2011.

Utilization of real-world clinical data

Real-world data refers to information about patients, medical interventions and clinical findings that have been derived from routine procedures in the standard-of-care care setting. The sheer *volume* and *variety* of real-world data available makes it attractive for developing multi-dimensional prediction models that can correctly stratify patients towards optimal interventions, according to their individual characteristics [17] i.e. personalized medicine. Real-world data also accumulates at a significantly higher *velocity* than clinical trial data, thereby opening up the prospect of rapidly learning healthcare systems [18].

However, the two major barriers in the way of real-world data utilization are (i) incompleteness of data collection in the routine clinical setting and (ii) high degree of fragmentation within the individual patient data landscape.

The first issue, data incompleteness or "missing data", refers to potentially useful explanatory variables that are either not measured during routine clinical practice or are not uniformly reported/recorded within the normal care setting. In the oncology setting (i.e. the empty circles in **Figure 4**), we suggest that data completeness rates could be routinely around 95% within clinical trial datasets and around 80% for specialized cancer registries, but one may reasonably expect upwards of 80% of real-world clinical data to be missing. An omnipresent risk of bias in real-world modelling studies is that the missing values may (or may not) be "missing purely at random", yet it would be difficult to prove that there is no systematic pattern in the data loss. The second issue, data fragmentation, refers to the well-known fact that data elements of oncology patients are generally widely distributed across multiple hospital record systems (e.g., electronic patient journal, radiotherapy information system, treatment planning system and radiology image archive) and across multiple disciplinary divisions (e.g., surgery, medical oncology, radiology and radiotherapy). The problem of *horizontal partitioning* exists when one database contains all the variables, but only for a subset of patients. A clinical trial dataset is an idealized example of *horizontal* partitioning (see **Figure 4a**), where one records 100% of the variables of interest, but only for 3% of the target population. On the other hand, the problem of *vertical partitioning* exists when a database contains information about all of the patients, but only for subset of variables of interest. The idealized example of *vertical* partitioning is a regional cancer registry, such that 100% of the patients are included in the register, but only for 3% of the variables of interest. In practice, one would find that real-world data is subject to both horizontal and vertical partitioning; each of the departmental databases – surgery, oncology and radiology – only records some of the variables for some of the patients, but with incomplete overlap (see **Figure 4b**).

In general, a researcher needs to be able to: (i) use data derived from different settings (e.g. clinical trial, cancer registry and clinical routine), and (ii) connect variables and outcomes from

multiple sources of data, that is also likely to involve data integration across multiple database architectures and non-uniform ways of encoding the data values.

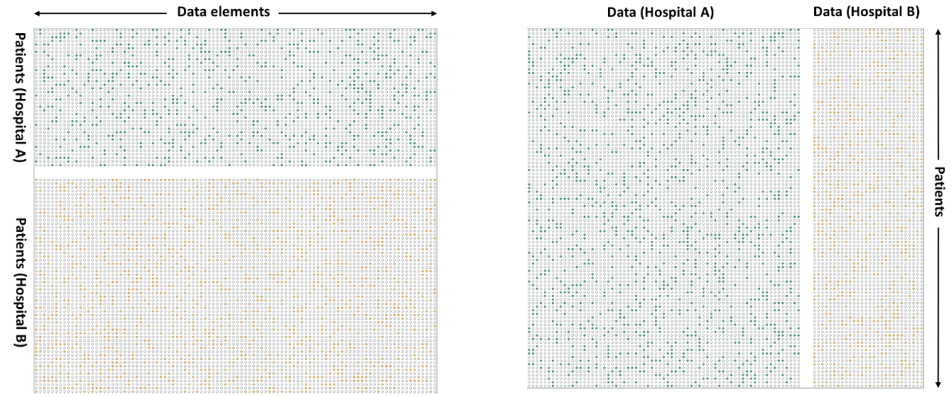


Figure 4. Illustration of two types of patient data partitioning; on the left (a) horizontal partitioning and, on the right (b) vertical partitioning. The unfilled circles illustrate the approximate prevalence of "missing values".

Generic data schema

Having obtained a clear understanding of the possible complexities in the data landscape, we turn our attention to compiling a *data schema* comprising of all of the relevant clinical outcomes of interest, along with a wide range of potential explanatory variables that could be used to predict those outcomes. A comprehensive schema should not only identify the names of the data fields, but also the likely location of potential data, the possible range of values in the data and the existing (or desired) format of the values.

For this purpose, "standard sets" of fields, where published by international multi-disciplinary efforts across will be particularly useful. The prime example of such interdisciplinary sets are the recommended outcomes tables produced by the International Consortium for Health Outcomes Measurement (ICHOM) [19]. These standard sets are particularly strong in the domain of relevant outcomes per cancer type, and include those case-mix details, clinical baselines and diagnostic work-up elements that would be necessary for multivariate modelling studies.

In respect to intervention details, however, the ICHOM consensus sets are not comprehensive enough for modelling requirements in oncology. Cancer type-specific and intervention modality-specific expertise is required to understand the interplay between the clinical outcomes and the treatment variables. This process must be informed by an understanding of the data landscape, as the some of the variables of interest is not feasible to access within a given data landscape.

Patient population characteristics

Modelling activities must always include certain general patient characteristics such as demographics, baseline clinical features, diagnostic findings (including any biopsy sampling performed as part of the diagnostic pathway) and any comorbidities.

Personal information is of importance in retrospective and/or secondary data analyses involving multiple cohorts, where basic personal characteristics modulate the outcome(s) of interest. This especially applies when incorporating quality of life or other patient-reported symptom where, for example, a person's marital status and living arrangements could affect the perception of outcome. Levels of educational attainment and ethnicity are known to correlate with quality of follow up, ability to respond to outcomes questionnaires (if used) and likely socio-economic disadvantages that may affect long-term general health outcomes [20]. In some data fields, such as those pertaining to matters of cultural, political or social sensitivity, it is often preferable to allow "undisclosed" as a valid value, rather than to leave missing values in the data.

Baseline clinical features and comorbidities help to clarify the particular clinical case mix that the modelling study addresses and, where strong statistical heterogeneity may be encountered, multivariate correction for differences in case-mix may be attempted. Comorbidities should be recorded as potentially multiple-value categories using a universal co-morbidity instrument, such as the Charlson [21] or the patient self-reported variant of Charlson (PRO-CCI) [22], but without collapsing all of information into a single score.

Table 2 illustrates some of the general population variables that should be reasonably easily obtained for toxicity modelling studies, since these typically reside in the patient electronic medical record as a matter of routine registration for treatment.

The above fields must be supplemented by other baseline and diagnostic details deemed clinically relevant for the target disease to be addressed by modelling. For instance, menopausal status is likely to be highly relevant if studying breast cancer, and Gleason score and Prostate-Specific Antigen levels are important for prostate cancer models. Additional lifestyle information, such as the estimated number of pack-years of nicotine use, may be relevant for lung cancer. For toxicity modelling, baseline 'toxicity' (i.e. complaints before treatment) is often crucial information and significantly predictive of post-therapy toxicity. Such information are likely to be located within the systematic medical departments e.g. urology, gynaecology or thoracic medicine, etc.

Since oncology becomes ever more reliant on multi-modality medical imaging, essential diagnostic information will also reside within the radiology department. Gene sequence analyses (i.e. of tumor specimen and of circulating tumor DNA) and proteomics analyses are also increasingly prominent in the oncology diagnosis pathway. If biopsies or tissue specimens were taken, it may be relevant to obtain the results of pathology findings such as the degree of tumor cell differentiation and surgical margin status. Wherever available and appropriate, diagnostic information from multiple imaging and pathologic modalities should be included.

Intervention details

In many prediction studies, the clinical question of interest relates to how specific elements of the cancer treatment are likely to modulate the outcome(s). For instance, escalating a radiotherapy dose to the tumor may seem like an attractive option for a case with poor initial prognosis, but would this impact on the likelihood of severe adverse effects on this individual patient? Conversely, an oncologist may propose to skip chemotherapy altogether for a patient with poor overall health, but is the reduced risk for toxicity sufficiently appealing to the patient to outweigh their increased chance of earlier death?

Table 2. Representative subset of patient population characteristics that may be required for toxicity modelling, including possible data formats and examples of valid values.

Variable	Explanation	Data type	Example valid values
	Personal and demographic information		
Unique identifier	Denotes patients without using their name	Alphanumeric	MRN012345
Last name, first name	Patient's preferred name	Free text	Doe, John
Biological sex	Biological sex at birth, as opposed to gender which is a social/cultural construct	Categorical, single value	0 = female 1 = male 99 = undisclosed
Date of birth	Date of birth as defined on official documents such as passport or birth certificate	Date format	15-01-1970
Ethnicity	Generally defined in the context of the country where the study was performed	Categorical, single value	0 = Asian 1 = African 2 = Caucasian . . . 98 = Other 99 = undisclosed
Educational level	Highest educational level attained at time of enrolment in database	Categorical, single value	0 = No schooling 1 = Primary 2 = Secondary 3 = Tertiary 99 = undisclosed
Marital status	Relationship status at time of enrolment in database	Categorical, single value	0 = Never married or partnered 1 = Married or partnered 2 = Divorced or separated 3 = Widowed 99 = undisclosed
Living arrangement	Household living arrangements at time of enrolment in database	Categorical, single value	0 = Sole occupant 1 = Co-occupant, adult(s) 2 = Co-occupant - adult(s) and minor(s) 3 = Co-occupant, minor(s) 99 = undisclosed
	Clinical baseline and diagnosis information		
Comorbidities	Existing medical conditions, either clinically reported or patient self-reported, using an internationally recognized and validated schema such as the 19-item Charlson.	Categorical, one or more values	0 = no other conditions 1 = myocardial infarction . . . 19 = AIDS 99 = unknown
Height (in unit)	Height measured in clinic during enrolment, with pre-defined units e.g.cm or ft	Numeric	165
Weight (in unit)	Weight measured in clinic during enrolment, with pre-defined units e.g.kg or lbs	Numeric	78
Body mass index	Calculated from recorded height and weight	Numeric	28.7

Performance status	The ECOG/WHO performance status score	Categorical, single value	0 = WHO PS 0 . . . 4 = WHO PS 4 99 = unknown
Date of diagnosis	Date of diagnosis of the target condition	Date format	15-10-2015
Diagnosis code	International Classification of Disease (ICD)	Alphanumeric	C50.112
Diagnosis code version	ICD codebook version number	Numeric	10
Clinical T stage	Clinical tumor stage at time of diagnosis	Alphanumeric	IIa (or X if unknown)
Clinical N stage	Clinical nodal stage at time of diagnosis	Alphanumeric	0 (or X if unknown)
Clinical M stage	Clinical metastatic stage time of diagnosis	Alphanumeric	0 (or X if unknown)

Table 3. An example subset of interventional data that may be required for toxicity modelling.

Surgical notes *(probably free text information archived in the local patient journal system in the treating hospital)*	Pre-operative surgery Main cancer surgery date Main cancer surgical method Exploration during main surgery (and any relevant findings) Complications during any surgery Complications within 30 days of any surgery Wound complications (if any) - Wound complication type - Wound complication date
Chemotherapy notes *(possibly structured fields in medical oncology information systems or otherwise free text; text mining of drug codes and billing codes may be feasible)*	Chemotherapy agent(s) Dosage of chemotherapy agent(s) Number of chemotherapy cycles Timing of chemotherapy relative to other interventions Interval between chemotherapy cycles First cycle start date Last cycle end date
Radiotherapy (RT) details *(most likely to be found as structured fields in RT information, treatment planning and verification systems; requires significant IT support when interacting with multiple computerized systems)*	Prescribed dose Prescribed fraction size Number of fractions per day Elapsed time between fractions Treatment modality (x-rays, brachytherapy, protons, etc.) Timing of radiotherapy to surgery (pre-operative, post-operative, intra-operative, etc.) Timing of radiotherapy to chemotherapy (induction, concurrent, etc.) Start date of radiotherapy End date of radiotherapy *Details of immobilization (e.g. stereotactic frame, abdominal compression, etc.) *Details of localization imaging (e.g. cone-beam CT, port films, etc.) *Organs at risk delineations archived in universal data interchange format such as DICOM-RT *Organs at risk doses archived in universal data interchange format such as DICOM-RT *Planning/simulation images (e.g. PET and CT) archived in universal data format such as DICOM

The relevant treatment parameters needed to address these kinds of prediction questions are naturally highly specific to the target disease and must be significantly more detailed in the particular intervention that is supposed to be tailored towards an individual patient's outcome. However, over time, we may reasonably expect that a wider palette of surgical, pharmaceutical and radiotherapeutic options will become available to treating physicians and their patients;

this will lead to a high demand for the development of predictive models that handle variations in surgery, chemotherapy and radiotherapy treatment parameters. General data schema elements pertaining to each treatment modality are suggested in **Table 3**, along with caveats relating to data structure that may affect the extraction of such information.

Outcomes data

The most crucial aspect defining the potential success of a predictive toxicity model is the availability and quality of outcomes data. In practice, the accessibility and reliability of outcomes data is the limiting factor on sample size. In comparison to treatment parameters and patient characteristics, the quantity of real-world outcomes data is often the single aspect in data-driven modelling studies that is most commonly over-estimated by practicing clinicians. Simultaneously, the quality of real-world outcomes data is likely to be the least understood by practicing clinicians. The quality aspect arises from a widespread assumption that the presence of a note or dictation in the patient record suffices for quantitative analysis, whereas such notes are often impractical to use as modelling data due to ambiguity and poorly structured reporting. It may be sometimes advantageous to consider pairing baseline observations and treatment parameters from real-world data sets to secondary analysis of clinical trials outcomes data, rather than to rely solely on systematic collection of outcomes data in routine clinical practice.

Acute versus late toxicity

Short-term follow-up for treatment outcomes and induced toxicity are generally available for modelling studies. Surgical teams keep detailed notes of surgical complications, up to approximately 30 days following surgery. Likewise, one may expect observations during chemotherapy and radiotherapy to be reasonably systematic during the course of treatment, up to approximately 90 days after the end of treatment. Audits of follow-up in the routine oncology setting show that completeness of clinical observation data falls sharply after the initial period. This implies that toxicity models are more likely to address short-term adverse events (acute toxicity) following oncological treatment, rather than long-term (late) toxicity. This naturally creates the potential for a conflict of interest between clinicians and patients in decision-making, since patients tend to care more about long-term toxicities that are perceived as debilitating or life-changing [23]. It has been recognized that increasing numbers of patients are living with the long-term side-effects of cancer treatment [24], requiring oncology physicians to pay more attention to the dual impact of toxicity on quality *and* length of life after care.

Clinician graded toxicities

The preferred method for clinician-assessed treatment-related toxicity is a structured data field, whereby the key can be unambiguously related to a specific toxicity in an internationally recognized toxicity schema such as the Common Terminology Criteria for Adverse Events, CTCAE (https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm), and the toxicity criteria of the Radiation Therapy Oncology Group [25]. The outcome should be objectively reported as a numerical grade according to one of the abovementioned toxicity criteria. Free text descriptions of either the type of toxicity or its severity should be avoided,

because this will require NLP pre-processing and/or human-expert re-interpretation before the data can be used for modelling – both of these are potentially time-consuming and/or error-prone.

Patient-reported outcomes

In light of systematic and logistic problems collecting long-term follow-up data in the routine clinical setting, some researchers have proposed either prospective [26,27] or retrospective collection of patient-reported outcomes to supplement clinician assessments, though the latter is known to be particularly subject to recall bias [28]. Some proponents argue that patient reported outcomes (PROs) are potentially more useful for toxicity models informing patient-sensitive choices, since the side-effects of treatment are couched in terms of the adverse experiences of the patient [29,30]. However, opponents of PROs counter that a multitude of non-clinical factors may modulate the perception of toxicity, and therefore might not be accurate as an academic research instrument [32]. However, clinically-validated PRO collection instruments now exist in multiple languages: such as the EORTC general quality of life [32] and its allied cancer-specific questionnaires, as well as the PRO variant of the CTCAE (PRO-CTCAE) [33].

Timing of toxicity measurements

Prediction models may address the probability of events within a given time interval (e.g., radiation pneumonitis within 6 months of completing chemo-radiotherapy for lung cancer), or most probable time-to-event. Repeated observations (or measurements) of toxicity over a long period of times may be instructive for the former class of model, since treatment-induced effects in normal tissue are known to grow and ebb away over time, but are mandatory for the latter class of model. For the reasons of data accessibility and quality stated previously, the required data to develop the latter class of model is significantly more challenging to obtain. Hence, the majority of toxicity prediction studies have focused on the risk of events within a finite time interval.

Linking data from multiple sources

The problem of populating a schema with variables and outcomes for individual patient data is one of *extraction, transformation and loading* to link the data elements from multiple different sources. This requires a detailed understanding of the data landscape, since it is highly unlikely (unless re-using previously curated datasets) that all of the required information exists in a structured form within a single data corpus.

The complexity of the linking challenge depends on the size of the data set and the multiplicity of raw data sources. Smaller scale projects of several hundred cases, provided the data is already highly structured as tables, can be managed in office-desk applications such as STATA, SPSS and Excel. Less structured data may require sophisticated data reshaping functions available in R, Matlab and Mathematica. In R, the "*dplyr*" library includes functions for efficiently handling search, sub-setting and other manipulation of data shapes (including

melting, casting and aggregation). The programming language python also supports an advanced data manipulation library (*pandas*). Data integration tools (such as Pentaho) permits linking of multiple data sources to compile a single database (PostGreSQL) without the need for database programming knowledge. However, large datasets (i.e. greater than the memory capacity of a single device) may be handled in a distributed computational framework such as Hadoop or Spark.

Extracting real-world data from live operational systems (such as electronic patient journals, hospital billing systems, etc.) requires transformation from the raw source to structured data. Some systems may have pre-defined APIs that allow users to dump structured data in plain text formats (e.g. character-separated tables). Other systems may publish their database architecture, therefore (assuming some programming skill) special queries (e.g. in SQL or Crystal Reports) may be written to extract data values directly from such systems. Patient demographics, procedural codes, diagnosis codes and billing numbers may be accessible in this manner.

However, other relevant information (such as radiology reports, pathology findings and surgical notes) will frequently be encountered as only free text. In some cases, a form of semi-structured reporting may have been used (see section 2.1.1). In any situation where values are embedded within free text, some form of natural language processing [34] and semantic extraction approaches [35] might be used to processed these into structured data.

The combination of these approaches leads to the feasibility of data lakes, where large amounts of archived data are labelled with descriptive metadata (i.e. semantics) such that structured data can be efficiently retrieved using a semantic query language. A problem that quickly arises is semantically-labelled data in one such "lake" may not be inter-operable with other "lakes" unless there has been some uniformity in how the same concept shall be described. One has a choice of enforcing the same data structure over everyone, or undertake a major challenge to re-map the metadata to a common scheme; but neither of these approaches are scalable in a practical manner.

The indefinitely scalable approach defined by Tim Berners-Lee in 2001 is the *Semantic Web*, based around formal, publicly open and infinitely extensible semantic ontologies. Semantic Web defines the protocol by which any entity and any arbitrary relationship between entities can be assigned a persistently unique web address that is resolvable by widely-used web communication norms such as Hypertext Transfer Protocol (http). The indivisible data element is hence a triple, i.e. a statement in the format *subject-predicate-object*. Data that is stored as semantic triples can be easily rendered globally *Findable, Accessible, Interoperable and Re-useable* (FAIR) [36].

Data Sharing

Advantages of data sharing

Until recently, the majority of evidence-based healthcare practice has focussed on population-based treatment evaluation and randomized clinical trials. The problem with this is three-fold. First, population-based findings only consider the mean effect size, and not the fact that some patients within the cohort perform much better than other patients within the same cohort; that is, there is always a range of probable effect sizes when looking at the outcome of any given patient. Personalized medicine asks – why would this be so, what are the other covariates that a population-based study glosses over that might help doctor identify which patient will respond better (or which one will respond poorly)?

Secondly, addressing modern oncology treatment options by controlled trials is inefficient[11], particularly in two instances:

i. for technological innovations that either requires large up-front capital investment or where the technological process evolves rapidly with respect to the time scale of clinical trials (as is usually the case in surgery and radiation oncology), and
ii. when multiple factors are sought to stratify the variation in treatment outcome (thus leading to many study arms or multi-factorial randomization, both of which significantly increase the numbers needed to recruit).

Furthermore, as previously stated, most clinical trials have been designed to address clinical disease control or survival, and are hence underpowered to answer questions about toxicity.
A third problem that becomes evident in any systematic review of toxicity prediction models is that the input variables are divergent depending on the dataset used to develop the model. For example, there are multiple models of radiation-induced esophagitis [37-40], but some models required volume-based dose metrics of the esophagus, while others require the surface area of irradiated esophagus as the input.

Into this arena, the combination of real-world big data and data sharing approaches can be used to address the above concerns. In the Netherlands, several initiatives exist that encourage data sharing among Dutch radiotherapy clinics. For instance, the Translational Research IT (TraIT, http://www.ctmm-trait.nl) program allows for data sharing for translational research projects within one clinic or among different (Dutch) clinics. Dutch national infrastructure initiatives that will exploit toxicity models based on this shared data include *PRODECIS* [41] and *proTRAIT*.

PRODECIS is a prediction platform that evaluates the benefit a patient might experience from proton therapy compared to conventional x-ray therapy. This benefit is calculated on three levels: (i) the dosimetric level, where it is evaluated whether a radiotherapy plan meets a pre-defined dosimetric threshold for the organs at risk; (ii) the toxicity level, where any differences in toxicity of normal tissue are expected; and (iii) the cost-effectiveness level, where it is evaluated whether the benefits for the patient outweigh the extra costs for proton therapy. *ProTRAIT* (http://www.rug.nl/news/2017/08/1_5-miljoen-euro-voor-onderzoek-protonentherapie?lang=en) is a new infrastructure project, funded by the Dutch Cancer Society, to enable data exchange for proton therapy between institutes serving both clinical and research purposes.

Data sharing approaches combining multiple datasets from different clinics have the advantage of reducing the risk of cohort-dependent biases when developing a prediction model, resulting in a more globally-applicable consensus model across multiple centers. Data sharing enables either a larger combined data corpus for training and cross-validation, or the option for a model developed on one corpus to be independently validated against the other corpus.

An additional advantage of data sharing is the ability to link data from multiple sources and/or overcome the data partitioning problems discussed in the earlier section of this chapter. Data from cancer registries, population databases, clinical trials and electronic medical records could be integrated, particularly using a semantic data lakes approach with a universal domain ontology. A much richer (i.e. more potential explanatory variables) and possibly larger sample size may be feasible with data sharing versus an institutionalized approach. This is particularly advantageous for rare cancers (e.g., anal cancers and sarcomas) where it is exceedingly unlikely that a single center receives enough cases to develop a broadly externally valid model using solely its own data.

One example of this approach is the study described by Dekker et al. [42] who built a model for overall survival in Stage IV lung cancer patients at the MAASTRO clinic in The Netherlands, then validated this model in a similar set of patients treated at two centers in Australia. The model successfully differentiated between a good prognosis group and a medium/poor prognosis group, suggesting that inter-operability between different clinics is possible. Several initiatives apply this approach to combine data from different clinical centers for outcome and toxicity modelling: euroCAT [43], ozCAT [44] and meerCAT [45]. This approach will likely also be successful in the case of toxicity modelling, especially in rare cancer types. Data sharing further allows researchers to exploit heterogeneity in the data in order to learn something of immense clinical importance, such as a dose-dependent response to treatment.

Barriers to data sharing

Despite the advantages of data sharing, many barriers exist that hinder data sharing. These can be divided in administrative, ethical and technical barriers.

Two main administrative barriers hinder data sharing among multiple institutes are data completeness and coding inconsistency. Data completeness is difficult, if not impossible, to accomplish. It is generally not practical to collect every available data element of an individual patient, which severely hampers analysis. A pragmatic compromise always results in an incomplete dataset. The second barrier, coding inconsistency, occurs when system-dependent settings (such as standard operating procedure, treatment protocols, equipment settings and the way different observations are recorded) vary widely among sites. These variations impeding data integration and, if not correctly reconciled under a consistent data schema, may lead to incompatibility or clinically biased results between datasets across different sites.

Ethical barriers refer to data privacy regulations and other restrictions pertaining to re-use of patient data for some other purpose than originally collected. First, data privacy is very different in its definition, legislation and implementation across countries [46]. This privacy barrier needs to be overcome for multi-center infrastructure in order to achieve data sharing between institutions, implemented in either a centralized or distributed infrastructure.

A distributed solution may be more secure than a centralized solution, since very few (rather than all) of the potentially sensitive data parameters are communicated between multiple centers, thus overcoming the privacy barrier partially even if not perfectly entirely. A security breach in the distributed sharing network compromises hardly any individual patient-level data, whereas the same failure in a centralized system necessarily compromises *all* of the data. However, the specific methods to overcome this privacy barrier needs to be tailored to the project at hand, as there is no global solution that works in all cases.

Furthermore, re-use of research data may be a potential barrier since data sharing is not yet fully embedded in the scientific culture, although this is improving in recent times. Nevertheless, if data is shared among scientists, more data will be thus become available for modelling, resulting in better models for toxicity predictions. To promote data sharing in the scientific community, more scientific evidence of the benefits of data exchange needs to be published, thus persuading data holders to participate in the collaborative research community and subsequently share their data.

Even if these administrative and ethical barriers can be overcome, technical barriers such as interoperability and standards remain, which may still impede open data sharing between institutes. First, achieving interoperability between IT systems within a clinic is important for the generation of anonymized data, which should ideally be managed by the institutional data warehouse and be accessible through semantic web technologies. However, it is often difficult to achieve this ideal situation due to (oftentimes) incomplete support of internationally standardized protocols, formats and semantic web technologies. In order to improve the support for these technologies, investments are needed to implement these resources if not yet present at each contributing site.

Second, standardization of radiotherapy-specific terminologies and ontologies is still under development [46], resulting in a situation where various hospitals may use different representations to describe patient-specific information, such as the biological sex of a patient, which is denoted as “female” or “male” in one clinic and “f” or “m” in another. When combining data from two clinics, this difference results in the presence of both representations in this combined dataset. This problem can be solved by linking each clinical representation to a semantic ontology that can provide the standard definition of the variable. For instance, the biological sex of a patient is represented by NCI Thesaurus codes *C16576* for female and *C20197* for male. Thus, the meaning of a clinical variable is only related to its ontological label rather than through its literal representation as a text string.

Data sharing architectures

Centralized data sharing

When data is shared within a centralized architecture, the infrastructure itself has complete control of all of the data. This data is not stored in each of the individual clinics, but must be pooled in a centralized repository. In this situation, all operations occur at a central location and no real-time communication occurs between participating institutions. Even though this architecture type is conceptually simple, the ethical barriers mentioned above have to be managed at the central meta-institutional level. Furthermore, as new data fields are added, or the existing data elements are periodically re-organized in some fashion, the effort of centralized data management increases much more rapidly than either size or complexity. Additionally, there is redundancy related to duplication of data, transformation of the local data to the central data model (usually resulting in manual data entry and/or manual copying) and negotiation of intellectual property (IP) rights in the form of data ownership agreements. National cancer registries and health service quality monitoring databases commonly adopt a centralized data architecture.

Decentralized data sharing

In a decentralized sharing architecture, each local source of data retains full ownership (in terms of IP) over its own data. Unlike a centralized model, peer-to-peer data sharing is now possible without the need for a centralized governance structure to pool data from all of the sources. A decentralized data sharing architecture allows for multiple levels of granularity for data access control (e.g. project-by-project based or network-wide protocols) between multiple institutes [46]. The downside of the arrange is, every data station need to be established and exposed to the other partners in an inter-operable manner at each site to enable effective data exchange, and each clinic needs to comply with a standard data communication protocol, such as *http*. Governance of the network is almost non-existent, involving only self-imposed adoption of communication protocols and inter-operability standards. This architecture scales readily with size of network and dimensional complexity of data, since there is no overarching central coordination whatsoever. The most commonly quoted example of a decentralized architecture is the World Wide Web.

Hybrid architectures

The final option is to use a hybrid data sharing architecture, in which elements of both centralized and decentralized architectures are combined. In this situation, direct peer-to-peer communication and data sharing occurs between sites, but the information about the infrastructure, data representation format, controlled terminologies and other required metadata are maintained at a central location, which facilitates the maintenance and modification of data exchange. One obvious advantage of a hybrid architecture is that the data is stored locally at each site, but it is only conceptually centralized by means of the controlled protocols, thus addressing some of the technical barriers such as site-specific terminology at the level of multi-centric sharing[46].

Distributed machine learning

An increasing amount of data is being generated in different fields of modern medicine including medical imaging, transcriptomics, metabolomics and proteomics. This highly dimensional data is potentially very valuable for machine learning to build a reliable predictive model for diagnosis and treatment outcomes. Generally speaking, the more data that is used from diverse sources, the more robust and externally valid a predictive model becomes. Each institution generally possesses data for a limited number of patients, that may be insufficient for reliable predictive modelling. Therefore, multi-centric collaboration is key to ensuring that a sufficient amount of data is collected for predictive modelling. This can be practically implemented in 2 archetypal ways, as described in Section 3.3.

A general overview for the centralized multi-center architecture for machine learning purposes is shown in **Figure 5a**. In this configuration, a central collection point is responsible for storing all of the data transmitted by each contributing institution and then processing the site-specific data into a uniform format for analysis. Each institute therefore provides one communication point, which communicates in one direction only with the central data server. Statistical analysis and model building thus occurs only at the center (which typically hosts high-performance computing hardware), which results in a global model that is internally valid across all of the contributing centers.

An alternative method for machine learning across multiple institutions is the distributed learning architecture (see **Figure 5b**). It differs from the centralized architecture in regard to the location at which the numerically intensive computations are performed [47,48]. According to this paradigm, an institution with a research question (which may also be any one of the centers intending to contribute data) splits the problem into multiple sub-computations which will be deployed to each participating site. The data-intensive computations are performed within each site on its own respective database, and the coefficients of the local model are shared with the researcher. Crucially, none of the individual patient-level data needs to leave the contributor's database. The multiple local models must then be compared at the original source of the research question[48]. Depending on the algorithm used, such as Boyd's Alternating Direction Method of Multipliers [49], modified constraints on the coefficients may be sent back out to the contributing sites for re-calculation, and the above cycle repeated many times until a globally convergent model emerges. For specific machine-learning algorithms (such as logistic regression), it can be mathematically proven that the final convergent model obtained by decentralized learning must be the same as a centrally-learned model [49]. In-depth explanations of distributed machine-learning algorithms may be found in literature reviews [49,50].

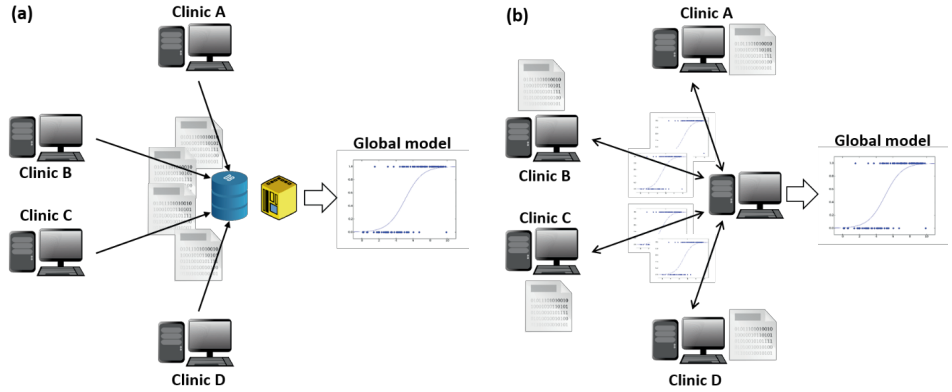


Figure 5. Centralized and decentralized approaches to sharing data for model-building. (a) Centralized architecture where data is sent to a central repository, where all of the processing and analysis takes place to build a global multi-institutional model. (b) Decentralized architecture where local data is retained within each institution and only fitting coefficients are transmitted to a central processor, which combines the local models (possibly over multiple iterations) to produce a global model.

Semantic Web technologies

For the abovementioned distributed learning methodology to work, the local data needs to be parsed in a format that is fully machine-readable and machine-understandable (i.e. objectively semanticized). This is achieved by defining domain ontologies conforming to standard of the Semantic Web. The semantic web is an extension of the internet in which data is intentionally designed to be interpreted by machines rather than humans. In order to achieve machine-readability of the data, structured data is labelled with a publicly accessible semantic ontology. Machine-learning algorithms can thus access this via a universal Resource Description Framework (RDF) and the "SPARQL Protocol and RDF Query Language" (SPARQL). Both sources use Uniform Resource Identifier (URIs) as links between expert semantic meaning and actual physical resources, so that data published through a web "endpoint" is amenable to queries by both machines and people.

Computer Assisted Theragnostics (CAT)

A real-world decentralized machine-learning network where the concepts described above have been realized in practice is known as "Computer Assisted Theragnostics" (CAT). Within CAT and its related projects, data is curated from a variety of sources at each collaborating institute and stored locally in standard (PostgreSQL) data warehouses. The "warehoused" data is then converted into a semantic ontology-labelled RDF store and published to other collaboration partners as a SPARQL-compatible endpoint. A set of site-based connector applications and webpage access portals are then used to distribute machine-learning algorithms to address prediction modelling questions. The use of CAT methodology allows research networks to address the barriers to data sharing and model-building as discussed earlier in Section 3.2, such as patient privacy, data ownership and lack of inter-operability between data warehouses.

A number of real-world projects have demonstrated the practical feasibility of distributed multi-centric machine learning including: I2B2 (https://www.i2b2.org), EuroCAT [43], VATE [51], meerCAT [45]. Additionally, improving the robustness of the models by external validation will benefit the reproducibility and validity of prediction models [42]. This has also been demonstrated within the EuroCAT project [43].

Toxicity Modelling

The efficacy of radiotherapy against malignant neoplasms is based on the deterministic effect of ionising radiation-induced damage on living cells. Increasing the absorbed dose of radiation always leads to greater detrimental effect on cellular function. The relationship between the physics behind the interaction between ionising radiation and matter and the macroscopic effects observable at the clinical level is a complex process. Energy transfer from radiation to matter via ionization and collision events induces reactive oxygen species (ROS; free radicals) that attack deoxyribonucleic acid (DNA), resulting in sub-lethal damage (eventually repaired by the cell) and lethal damage (triggers apoptosis, i.e. cell death). Chemical reactions also trigger a cascade of stress-related responses, immune system responses, release of tumor growth and tumor necrotic factors, and release of DNA repair molecules. Understanding radiation toxicity through mechanistic arguments remains a large and active area of research. Furthermore, certain genetic mutations may predispose or protect against some form of radiation-induced toxicity [52].

What is lacking from mechanistic explanations of the interaction between radiation and cells is an understanding of how highly localized, compartmentalized functional damage accumulates over spatial and temporal scales to give rise to clinically-observable toxicity (see **Table 4**). Radiation induced toxicity is known to have dependencies on dose magnitude, spatial dose distribution and dose-time (fractionation) patterns.

The outstanding technical success of modern-era radiotherapy, specifically ultra-conformal planning techniques [53-66] and use of image-guidance during treatment [57,58] have been based on one relatively simple paradigm; increasing the absorbed dose to the tumor while simultaneously reducing, as much as possible, the incidental dose to the healthy tissues in the vicinity of the tumor. Constraints imposed by the physics of radiation transport in matter makes it impossible to completely eliminate radiotherapy-related toxicity. Therefore, the ability to predict toxicities associated with radiotherapy is essential.

One approach to reduce toxicity to healthy tissues is phenomenologically-based toxicity modelling, where a deeply mechanistic description of the underlying process is temporarily set aside in favor of using only a few explanatory variables to predict clinically measurable toxicity.

The most commonly quoted Normal Tissue Complication Probability (NTCP) models such as QUANTEC [59,60] are based on the mathematical formalism developed by Lyman [61], Kutcher [62] and Burman [63] to equate a non-uniform dose distribution on a healthy organ to a generalized “uniform” dose. Given a sparse collection of toxicity data, such as Emami et al.

[64], the coefficients of the Lyman-Kutcher-Burman (LKB) model allows the probability of a given toxicity to be estimated for any arbitrary dose distribution in the healthy organ of concern. Of paramount importance in the phenomenological approach is the availability of large volumes of individual level patient data on treatment parameters as well as systematic longitudinal (i.e. repeated over time) observations of treatment-related side-effects. Richly multidimensional data is required to characterize an individual patient's phenotype as completely as possible, along the lines discussed in the preceding sections of this chapter, so as to accurately predict the expected toxicity.

Table 4. Some examples of commonly observed toxicities listed by radiotherapy treatment site.

Radiotherapy Site	Acute Toxicities	Late Toxicities
Any / unspecific location	Skin erythema Skin desquamation Dry and/or painful skin Tiredness Nausea Anorexia Oedema (swelling) Paresthesia	Telangiectasia Fibrosis Ulceration Obstruction Stenosis Radiation osteonecrosis
Head and Thoracic	Oral mucositis Esophagitis Odynophagia Dysphagia Radiation pneumonitis Dyspnea	Lymphoedema Breast atrophy Xerostomia Dyspnea Myelopathy Cognitive decline Hearing loss Hypopituitarism
Abdominal	Gastritis Stomach pain Vomiting Diarrhea Malabsorption Mucus colitis Bleeding (melena or haematochezia)	Strictures Adhesions Fistulae
Pelvic	Dysuria Frequent micturition Bladder cystitis Radiation nephritis Mucosal oedema Leukopenia Thrombocytopenia	Proteinuria Incontinence Nocturia Fistulae Proctitis Bleeding
Reproductive organs	Vaginal mucositis Cessation of menstruation Dyspareunia	Infertility or sterility Induced menopause Erectile dysfunction Vaginal stenosis Vaginal obstruction Vaginal dryness Vaginal fistulae

A data-sharing approach to toxicity modelling [65] is important due to the relative sparsity of toxicity observations in post-treatment follow-up (compared to clinical outcomes such as survival or local control) as well as strong heterogeneity in individual radio-sensitivity within populations. The data sharing architectures previously described are of significance here, as suitable data sets for toxicity modelling are more likely if individual patient-level observations from multiple institutions have been pooled together. Pooled individual-level toxicity data from multiple institutions has the added advantage that the distribution of radiation doses to organs-at-risk will naturally extend over a wider range, leading to a more robust statistical model. Some heterogeneity in the model covariates is an absolute requirement to identify predictive features and hence estimate the change in toxicity outcome with respect to changing one or more of these features.

Current state-of-the-art

Toxicity modelling process

The procedure for modelling treatment-related toxicity is generally similar and broadly analogous to the procedure for developing diagnostic tests, genetic markers, blood-borne biomarkers and predictive models of tumor control (e.g. overall survival or time to recurrence). The general procedure begins with a clear formulation of the clinical question at hand, and proceeds onwards to identifying the relevant covariates and assembling the data into a structured format. Thereafter, technicalities of the methods used may diverge somewhat, depending on (i) the clinical question, (ii) the data landscape and (iii) the data sharing architecture. To conclude the process, there must be some evaluation of the predictive performance of the model. The model should be published according to a standard quality of reporting checklist such as TRIPOD [66] and, ideally, an anonymized data repository is made openly accessible to other researchers. A methodological framework [67] is shown in **Figure 6**.

As previously mentioned, re-use of clinical trial data and observations collected from routine clinical practice is common practice for toxicity model development; one must remain cognizant of potential problems that could impact the model. For example, clinical trial data tends to be relatively complete for a selective sample of patients with a given condition. In contrast, real-world data might be unselectively sampled and available in huge quantity, but is likely to be missing many of the covariates that would be relevant for modelling.

Handling missing values

The question of imputation of missing data (either covariates that are entirely missing or only missing values) is crucially important and is described in detail elsewhere [68,69]. Generally, a small proportion of values that are missing purely at random may be imputed. However, it is not advisable to impute values that are missing in some systematic (but potentially unexplained) way. In particular regard to toxicity observations, one must be wary that moderate toxicities ($\leq$ Grade 2) might be either significantly under-reported by clinicians or not recorded at all; the potential bias introduced in the model would be to assume such missing values as “no toxicity”.

The difficulty that arises in routine treatment data, where a large proportion of values (e.g., greater than 80%) may be missing, is that a systematic pattern of missing values may be impossible to detect. Missing covariates may be imputed, but hinges on strong correlation with other (non-missing) covariates. If so, the benefit of imputing the missing covariate is somewhat questionable, since it is now some functional combination (and therefore no longer statistically independent) of the other surrogate variables.

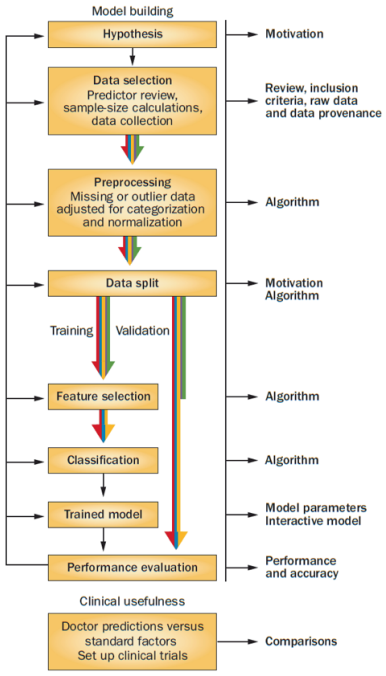


Figure 6. Schematic overview of methodological processes in clinical decision-support system development, describing model development, assessment of clinical usefulness and what ideally to publish. The coloured, parallel lines represent heterogeneous data, which have been split early for independent validation (but without internal cross-validation). Reproduced with permission from Lambin et al. [67].

Model fitting

In the next phase of the process, a carefully curated dataset is used to determine the coefficients of the model such that it describes the observed outcomes with the least bias in residual error. Suitable models for categorical or dichotomous (i.e. binary) outcomes include logistic regression, support vector machines [70], Bayesian network classifiers [71] and classification trees [72]. Treatment-related toxicities are almost always recorded as binary outcomes (patient either does or does not have a given severity of toxicity) with asymptotic upper and lower bounds on probability, therefore one of the above models would be appropriate for toxicity modelling.

Toxicity outcomes may sometimes be framed as a time-to-event problem, such as an expected time interval when a certain severity of late toxicity is observed. For these, Cox proportional hazards or competing risks models are commonly used.

Unless there is a prospective plan for an independent dataset to validate the model, the available data should first be partitioned into a "training" set and a "testing" set. Splitting the overall sample 80/20 between training and testing, respectively, is a commonly applied rule-of-thumb. The splitting may be determined entirely at random, or one may retain the same relative proportion of outcome events on each side of the split (i.e. a stratified split). The latter may be advisable if toxicity events are rare, since a purely random split might result in no events falling within the training subset. Splitting based on the outcome has an attendant risk of unknowingly contaminating the testing subset with features that are surrogates of the outcome, and would therefore not be truly independent from the training subset.

Bias-variance trade-off

An over-fitted model is one that so specifically and so exclusively describes the training subset (i.e. low variance in the model's residual errors) that it performs poorly when presented with a previously unseen testing subset (i.e. highly biased). Conversely, an under-fitted model is poor at describing the observed variation of outcomes in the training subset (high variance in the residual) but has the same performance in a testing subset (low bias). Neither of the above is desirable in prediction modelling; an optimally fitted model needs to be a trade-off between in-training set variance versus in-validation set bias.

Various strategies can be employed, either singly or in combination, to reduce the risk of over-fitting. First, the number of covariates relative to the number of events should be as low as possible. Suitable candidates for elimination are covariates that are either strongly correlated to others or where its values hardly change.

Secondly, data-driven covariate selection methods, such as stepwise-forward or stepwise-backward regression, may be used. Maximum likelihood estimation or information criteria metrics, such as Akaike's and Bayes' Information Criteria [73,74], can be used to iteratively build up an optimal set of explanatory covariates (or drop them from further consideration). However, the implicit risk is that potentially predictive variables may be eliminated because they do not appear to be sufficiently predictive in the training cohort. Another recommended method is ranked univariate selection, where each candidate covariate is ordered according to the strength of its correlation with the outcome. A maximally parsimonious model can then be designed by progressively dropping covariates from the bottom of the ranking, until the best predictive performance is obtained using the smallest number of covariates. In a similar way, regularization approaches [75] may be used to progressively and automatically filter out covariates.

Thirdly, the training subset may be further split randomly into multiple internal subgroups (i.e. "folds"). Each fold is tested against a model trained using all of the other folds combined, and the process iterates over all folds. This is known as k-fold cross-validation [76,77], where k

represents the number of folds. The value of the model coefficient is averaged over all folds. If there is a concern about selective sampling among the folds, the above cycle may be repeated several times with a new pseudo-random number of generator seed in each cycle.

Evaluating prediction performance

Minimally, the predictive performance of the model needs to be evaluated in the testing cohort. The typical measure of discrimination/classification performance is the area under the receiver-operator characteristic curve (AUC). Overall accuracy, sensitivity and specificity, and calibration index are also commonly reported for categorical and/or binary toxicity outcomes. However, for time-to-adverse event predictions, the calibration index and the hazard ratio between different patient sub-populations are commonly reported.

The number of published multifactorial prediction models has grown rapidly in recent years. Attention to methodological transparency and adherence to reporting guidelines supports the evaluation of models, and helps identify potential sources of bias that are to be avoided. Objectiveness in model performance evaluations ought to be based on repeatability and reproducibility of the published results. Hence, it is important that journals encourage open access to datasets and the software or computer code used to generate the prediction models.

Clinical impact

At present, toxicity prediction models can only distinguish between broad categories of interventions (such as radical surgery versus active surveillance, or sequential versus concurrent chemo-radiotherapy for lung cancer). In the radiotherapy domain, physicians and patients are likely more interested in tailoring each specific radiotherapy intervention, at an individual patient level, towards avoidance of radiation-induced injury. The vast majority of present day toxicity models are neither sufficiently detailed nor independently validated to be used to select interventions and included in the radiotherapy treatment plan optimization (i.e. inverse planning). The present generation of toxicity models are largely based on dose-volume metrics alone.

Ultimately, the utility of all multifactorial toxicity prediction models must be established within a clinical context [67] The predictive model needs to demonstrate added clinical value when used in the treatment decision-making process, which is over and above the utility of existing prognostic indicators. A positive clinical impact may be framed in terms of reduction in frequency (and /or severity) of adverse events due to treatment, or better quality of life of patients in after-treatment care. After a prediction model has been shown to have significant clinical impact and be broadly valid over a range of clinical settings, one may then expect rapid adoption and wide usage of the model in routine clinical decision- making.

Toxicity modelling in the near future
Incorporating diverse data types

Though present-day NTCP models in radiotherapy are still largely based on only on dose and/or dose-volume metrics, efforts are under way to include a wider diversity of clinical data types to improve the predictive performance of such models. In the domain of "big data" relevant to oncology, there are major research efforts under way to develop more holistic models in which covariates from many different disciplines (such as surgery, radiology, chemo-radiotherapy, genetic testing, immunotherapy and blood-borne biomarkers) can be combined. The first among these can be seen in efforts to incorporate clinical risk factors into the QUANTEC dose-based NTCP models [78].

Another deep vein of richly multi-dimensional data that can be mined for potential predictors is routine medical imaging. Medical images constitute a large proportion of real-world healthcare data on individual patients. Extraction of data from images requires high-throughput and (ideally) fully automated image processing pipelines to reduce the qualitative aspects of the images into a finite number of structured features. Such "radiomic" features have been shown in diagnostic and prognostic studies to be potentially correlated with tumor histology and overall survival [6-8,79,80]. Forthcoming developments in radiomics-boosted clinical toxicity prediction models will address radiation responses in tumor and healthy tissues over time (i.e. "delta radiomics"), including the individual patient's susceptibility for treatment-induced toxicity.

A further active area of research involves the study of genetic susceptibilities, including innate radio-sensitivity based on single-nucleotide polymorphism [81], to inform toxicity prediction. Drug-gene interaction studies are being performed in the chemotherapeutic domain. In radiotherapy, the rapidly advancing field of "radiogenomics" utilizes genome-wide association studies (GWAS) to correlate genetic markers with individual predisposition to radiation-induced toxicity [82-85]. Therefore, one may also expect to see trials and validation studies of gene-boosted toxicity prediction models in the near future.

It has been hypothesized that inherent differences in radio-sensitivity at the cellular level is the dominant contributing factor to clinically-observable normal-tissue reactions [67]. This suggests that the effect estimate in a cohort study of toxicity may be unduly biased by the potential inclusion of a small number of highly radiosensitive patients. Rapid-turnaround blood assays, potentially exploiting lab-on-a-chip technology, may be used to measure inflammatory responses and other molecular expressions of cellular damage (such as tumor growth factors). This could be used to stratify patient sub-groups by innate cellular radiosensitivity, and the potentially predictive molecular markers may themselves be incorporated into future predictive models of toxicity.

Exploiting differences for personalizing treatment

Tailoring interventions to the specific characteristics of the individual patient has become an important focus of modern cancer research. This necessarily requires that individual variability, that is, the comprehensive "human phenotype" must be taken into account when preparing an intervention [86]. However, the observable phenotype extends beyond purely genetic and

biological aspects – the patient must be also be understood as a holistic combination of cognitive abilities, experiences and preferences, as well as the accumulated effects of lifestyle and environmental exposures [87].

Pre-existing illnesses and co-morbidities are known to be correlated with radiation-induced injuries. A study by Nalbantov et al. [88] proposed that cardiac co-morbidity and baseline dyspnea were significantly correlated with dyspnea at 6 months after radiotherapy. However, few toxicity models presently take account of pre-existing medical conditions and co-morbidities as potential predictive features of post-treatment toxicity. Chen et al. [89] suggests that a subset of elderly cancer patients might tolerate and benefit from more aggressive radiotherapy, but more precise stratification of such patients based on age and presence of comorbid illnesses would be required. This hypothesis was supported by Dekker et al. [42] who reported that a subset of palliatively-treated lung cancer patients with promising treatment response (according to a multifactorial prediction model) may have had an additional 18-month benefit in median survival time, if they had been treated aggressively.

The existence of two or more treatment options within a given range of clinical equipoise satisfies a necessary condition for preference-sensitive decision-making. Consensus guidelines now recommend that post-mastectomy radiotherapy for women with T1-2 breast cancer (with 1 to 3 positive axillary nodes) should be discussed with patients in connection to their individual situation and personal preferences [90]. Similarly, paucity of toxicity data from large international trials in anal cancer chemo-radiotherapy has led to equivocation about the optimal radiotherapy prescription dose [91-94], even though radiation-related side-effects occur frequently. Although there is no evidence pointing to any single dose prescription being clinically superior, the risks of normal tissue complication are thought to be greater at the highest dose levels. Rønde et al. [95] examined the feasibility of including physician and patient preferences for treatment outcome into treatment plan optimization, but were unable to pose the optimization problem in regards to probabilities of genito-urinary and gastro-intestinal toxicity due to lack of data.

Accumulation and sharing of multi-dimensional toxicity data, along the lines discussed in this chapter, would enable the development of detailed toxicity models that support physician decision-making, patient counselling and shared decision making. Shared decision making is a structured collaborative consultation between a patient and their treating physician, such that each brings their own perspective into a conversation about the trade-off between risks and benefits among a set of preference-sensitive treatment options [96,97], and in so doing to arrive at a mutually agreed choice of treatment.

Use of machine learning and artificial intelligence

Up until recently, the vast majority of the data used for radiotherapy modelling has been processed by human hands. That is, the data is collected, cleaned, linked and structured by human operators. Thereafter, the data is fed to computers to compute statistical models and machine classifiers. While this approach may be suited to small amounts of information, it is

hopelessly inappropriate for scaling to the volume, variety and velocity of data required to address the clinical challenges addressed in this chapter.

The only possible option for the future is to generate data handling workflows where machines perform almost the entirety of the work of data pre-processing and data linkage from a large number of multiple sources. Such workflows can be adapted to any one of the data architectures previously discussed, and thus efficiently manage the modelling workflow for human experts to afterwards validate and use.

With increasing size and dimensionality of data, a form of machine-based intelligence is required to rapidly derive clinical insight from the raw data with minimal need for human guidance [98]. Artificial neural networks (i.e. neural "nets" or ANNs) are increasingly being set this task in oncology, including prediction of treatment toxicity.

Such neural nets consist of multiple layers of individual "neurons", that are each univariate logistic regression classifiers. The input to every neuron (other than the lowest input layer) is a linear combination of the output of every neuron in the layer below. Therefore, a "stack" of such layers are capable of encoding astoundingly complex and nuanced "responses" to a given input. The drawback of such an approach is, a vast volume of data (in the magnitude of thousands or possibly millions of independent cases) are required in the training set.

The earliest examples of application already show that ANNs can feasibly predict nocturia and rectal bleeding following prostate radiotherapy [99,100], as well as radiation pneumonitis in lung radiotherapy [101,102]. The current performance of ANNs remains sub-optimal, and this may be a combined effect of poor data quality, paucity of individual-patient level data and low "learning" efficiency of ANNs (for the typical number of training cases that can be feasibly mustered in radiation oncology).

"Deep learning", as a variant of simpler ANNs, consists of multiple stacks of neurons in the intermediate zone between the input layer and the output layer. In non-medical applications, deep learning networks have been shown to be capable of remarkable feats of object recognition, target object segmentation, natural language processing, competitive game-playing and spontaneous generation of simple narratives (i.e. auto-generating captions when given a previously unseen photograph). The anticipation is that in the near future, deep learning systems will also have immense impact in the field of radiotherapy in the form of diagnosis support systems, region-of-interest autosegmentation, conversion of natural text (e.g. pathology reports and follow-up notes) into structured data and sophisticated prediction models utilizing a wide range of "-omics" inputs.

Towards a data sharing "culture" in oncology

The future development of reliable and consistent toxicity model for clinical use will be critically dependent on the evolution of a data-sharing culture across multiple disciplines that intersect in oncology. As Deasy et al. [65] have pointed out, NTCP models based only on

summary reports of dose-volume effects can be inconsistent. These authors make a compelling case for data pooling (rather than the so-called "data-to-trash can" approach). Given the natural heterogeneity and variability inherent in these once-off cohort studies, secondary analysis of pooled of individual patient-level data for toxicity modelling is a valuable and under-exploited opportunity for clinical learning, given that data architectures now exist to support large-scale learning and modelling efforts that do not divulge personal patient information (see Section 3 above). Fortunately, a number of data linkage efforts are under way to render institutional data stores FAIR and discoverable for the purpose of collaborative learning, including the development of multi-institutional toxicity models (e.g. OncoSpace and CORAL).

Conclusion

Development of reliable, accurate and clinically-validated prediction models of toxicity requires a concerted effort to address systemic healthcare challenges pertaining to (i) structured data collection and data linkage, (ii) multiple barriers against sharing data for cooperative clinical learning of toxicity models and (iii) methodological approaches for clinical learning on vast volumes of richly-dimensional healthcare data. We conclude on a note of optimism, that present-day research and development projects are attempting to solve these systemic challenges, and hence to usher in an era where improved toxicity prediction models will support precision personalized medicine and shared decision making in the very near future.

References

1. Spak, D. A., Plaxco, J. S., Santiago, L., Dryden, M. J. & Dogan, B. E. BI-RADS(®) fifth edition: A summary of changes. Diagn. Interv. Imaging 98, 179–190 (2017).
2. Weinreb, J. C. et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur. Urol. 69, 16–40 (2016).
3. Kreimeyer, K. et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J. Biomed. Inform. 73, 14–29 (2017).
4. Pons, E., Braun, L. M. M., Hunink, M. G. M. & Kors, J. A. Natural Language Processing in Radiology: A Systematic Review. Radiology 279, 329–343 (2016).
5. Yim, W.-W., Yetisgen, M., Harris, W. P. & Kwan, S. W. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2, 797–804 (2016).
6. Kumar, V. et al. Radiomics: the process and the challenges. Magn. Reson. Imaging 30, 1234–1248 (2012).
7. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer Oxf. Engl. 1990 48, 441–446 (2012).
8. Larue, R. T. H. M., Defraene, G., De Ruysscher, D., Lambin, P. & van Elmpt, W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br. J. Radiol. 90, 20160665 (2017).
9. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. Phys. Med. Biol. 61, R150 (2016).
10. Sullivan, R. et al. Delivering affordable cancer care in high-income countries. Lancet Oncol. 12, 933–980 (2011).
11. van Loon, J., Grutters, J. & Macbeth, F. Evaluation of novel radiotherapy technologies: what evidence is needed to assess their clinical and cost effectiveness, and how should we get it? Lancet Oncol. 13, e169-177 (2012).
12. Geifman, N. & Butte, A. J. DO CANCER CLINICAL TRIAL POPULATIONS TRULY REPRESENT CANCER PATIENTS? A COMPARISON OF OPEN CLINICAL TRIALS TO THE CANCER GENOME ATLAS. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 21, 309–320 (2016).
13. Jin, S., Pazdur, R. & Sridhara, R. Re-Evaluating Eligibility Criteria for Oncology Clinical Trials: Analysis of Investigational New Drug Applications in 2015. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 35, 3745–3752 (2017).
14. Bentzen, S. M. & Trotti, A. Evaluation of early and late toxicities in chemoradiation trials. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 25, 4096–4103 (2007).
15. Secord, A. A. et al. Patient-reported outcomes as end points and outcome indicators in solid tumours. Nat. Rev. Clin. Oncol. 12, nrclinonc.2015.29 (2015).
16. Gilbert, A. et al. Systematic Review of Radiation Therapy Toxicity Reporting in Randomized Controlled Trials of Rectal Cancer: A Comparison of Patient-Reported Outcomes and Clinician Toxicity Reporting. Int. J. Radiat. Oncol. Biol. Phys. 92, 555–567 (2015).

17. Lambin, P. et al. ‘Rapid Learning health care in oncology’ - an approach towards decision support systems enabling customised radiotherapy’. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 109, 159–164 (2013).
18. Abernethy, A. P. et al. Rapid-learning system for cancer care. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 28, 4268–4274 (2010).
19. Kelley, T. A. International Consortium for Health Outcomes Measurement (ICHOM). Trials 16, O4 (2015).
20. von dem Knesebeck, O. et al. Social inequalities in patient-reported outcomes among older multimorbid patients – results of the MultiCare cohort study. Int. J. Equity Health 14, 17 (2015).
21. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J. Chronic Dis. 40, 373–383 (1987).
22. Habbous, S. et al. Validation of a one-page patient-reported Charlson comorbidity index questionnaire for upper aerodigestive tract cancer patients. Oral Oncol. 49, 407–412 (2013).
23. Kunneman, M., Pieterse, A. H., Stiggelbout, A. M. & Marijnen, C. A. M. Which benefits and harms of preoperative radiotherapy should be addressed? A Delphi consensus study among rectal cancer patients and radiation oncologists. Radiother. Oncol. 114, 212–217 (2015).
24. Rowland, J. H., Hewitt, M. & Ganz, P. A. Cancer survivorship: a new challenge in delivering quality cancer care. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 24, 5101–5104 (2006).
25. Cox, J. D., Stetz, J. & Pajak, T. F. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int. J. Radiat. Oncol. Biol. Phys. 31, 1341–1346 (1995).
26. Basch, E. et al. Recommendations for Incorporating Patient-Reported Outcomes Into Clinical Comparative Effectiveness Research in Adult Oncology. J. Clin. Oncol. 30, 4249–4255 (2012).
27. Riesen, I. N., Boersma, L., Brouns, M., Dekker, A. & Smits, K. PO-0755: Implementation of structural patient reported outcome registration in clinical practice. Radiother. Oncol. 123, S398 (2017).
28. Lee, C., Sunu, C. & Pignone, M. Patient-reported Outcomes of Breast Reconstruction after Mastectomy: a Systematic Review. J. Am. Coll. Surg. 209, 123–133 (2009).
29. Dueck, A. C. et al. Validity and Reliability of the US National Cancer Institute’s Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). JAMA Oncol. 1, 1051–1059 (2015).
30. Kluetz, P. G., Chingos, D. T., Basch, E. M. & Mitchell, S. A. Patient-Reported Outcomes in Cancer Clinical Trials: Measuring Symptomatic Adverse Events With the National Cancer Institute’s Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Am. Soc. Clin. Oncol. Educ. Book Am. Soc. Clin. Oncol. Meet. 35, 67–73 (2016).
31. Atkinson, T. M. et al. The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO): a

systematic review. Support. Care Cancer Off. J. Multinatl. Assoc. Support. Care Cancer 24, 3669–3676 (2016).
32. Fayers, P., Bottomley, A., EORTC Quality of Life Group & Quality of Life Unit. Quality of life research within the EORTC-the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer. Eur. J. Cancer Oxf. Engl. 1990 38 Suppl 4, S125-133 (2002).
33. Basch, E. et al. Development of the National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). JNCI J. Natl. Cancer Inst. 106, (2014).
34. Jovanović, J. & Bagheri, E. Semantic annotation in biomedicine: the current landscape. J. Biomed. Semant. 8, 44 (2017).
35. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 128–144 (2008).
36. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, sdata201618 (2016).
37. Belderbos, J. et al. Acute esophageal toxicity in non-small cell lung cancer patients after high dose conformal radiotherapy. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 75, 157–164 (2005).
38. Bradley, J., Deasy, J. O., Bentzen, S. & El-Naqa, I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. Int. J. Radiat. Oncol. Biol. Phys. 58, 1106–1113 (2004).
39. De Ruyck, K. et al. Development of a multicomponent prediction model for acute esophagitis in lung cancer patients receiving chemoradiotherapy. Int. J. Radiat. Oncol. Biol. Phys. 81, 537–544 (2011).
40. Palma, D. A. et al. Predicting esophagitis after chemoradiation therapy for non-small cell lung cancer: an individual patient data meta-analysis. Int. J. Radiat. Oncol. Biol. Phys. 87, 690–696 (2013).
41. Cheng, Q. et al. Development and evaluation of an online three-level proton vs photon decision support prototype for head and neck cancer – Comparison of dose, toxicity and cost-effectiveness. Radiother. Oncol. 118, 281–285 (2016).
42. Dekker, A. et al. Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. Radiother. Oncol. 113, 47–53 (2014).
43. Deist, T. M. et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin. Transl. Radiat. Oncol. 4, 24–31 (2017).
44. Lustberg, T. et al. Implementation of a rapid learning platform: Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting. Oncotarget 7, 37288–37296 (2016).
45. Jochems, A. et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int. J. Radiat. Oncol. • Biol. • Phys. 99, 344–352 (2017).

46. Skripcak, T. et al. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 113, 303–309 (2014).
47. Damiani, A. et al. Distributed Learning to Protect Privacy in Multi-centric Clinical Studies. in Artificial Intelligence in Medicine 65–75 (Springer, Cham, 2015). doi:10.1007/978-3-319-19551-3_8
48. Soest, J. P. A. van, Dekker, A. L. A. J., Roelofs, E. & Nalbantov, G. Application of Machine Learning for Multicenter Learning. in Machine Learning in Radiation Oncology 71–97 (Springer, Cham, 2015). doi:10.1007/978-3-319-18305-3_6
49. Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found Trends Mach Learn 3, 1–122 (2011).
50. Wu, W.-H. et al. Review of trends from mobile learning studies: A meta-analysis. Comput. Educ. 59, 817–827 (2012).
51. Meldolesi, E. et al. Standardized data collection to build prediction models in oncology: a prototype for rectal cancer. Future Oncol. Lond. Engl. 12, 119–136 (2016).
52. West, C. M. & Barnett, G. C. Genetics and genomics of radiotherapy toxicity: towards prediction. Genome Med. 3, 52 (2011).
53. Bauman, G. et al. Intensity-modulated radiotherapy in the treatment of prostate cancer. Clin. Oncol. R. Coll. Radiol. G. B. 24, 461–473 (2012).
54. Co, J., Mejia, M. B. & Dizon, J. M. Evidence on effectiveness of intensity-modulated radiotherapy versus 2-dimensional radiotherapy in the treatment of nasopharyngeal carcinoma: Meta-analysis and a systematic review of the literature. Head Neck 38 Suppl 1, E2130-2142 (2016).
55. Marta, G. N. et al. Intensity-modulated radiation therapy for head and neck cancer: systematic review and meta-analysis. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 110, 9–15 (2014).
56. Zhang, B. et al. Intensity-modulated radiation therapy versus 2D-RT or 3D-CRT for the treatment of nasopharyngeal carcinoma: A systematic review and meta-analysis. Oral Oncol. 51, 1041–1046 (2015).
57. Gwynne, S. et al. Image-guided radiotherapy for rectal cancer: a systematic review. Clin. Oncol. R. Coll. Radiol. G. B. 24, 250–260 (2012).
58. Jadon, R. et al. A systematic review of organ motion and image-guided strategies in external beam radiotherapy for cervical cancer. Clin. Oncol. R. Coll. Radiol. G. B. 26, 185–196 (2014).
59. Bentzen, S. M. et al. Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): An Introduction to the Scientific Issues. Int. J. Radiat. Oncol. Biol. Phys. 76, S3–S9 (2010).
60. Marks, L. B. et al. Use of normal tissue complication probability models in the clinic. Int. J. Radiat. Oncol. Biol. Phys. 76, S10-19 (2010).
61. Lyman, J. T. Complication probability as assessed from dose-volume histograms. Radiat. Res. Suppl. 8, S13-19 (1985).

62. Kutcher, G. J., Burman, C., Brewster, L., Goitein, M. & Mohan, R. Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. Int. J. Radiat. Oncol. Biol. Phys. 21, 137–146 (1991).
63. Burman, C., Kutcher, G. J., Emami, B. & Goitein, M. Fitting of normal tissue tolerance data to an analytic function. Int. J. Radiat. Oncol. Biol. Phys. 21, 123–135 (1991).
64. Emami, B. et al. Tolerance of normal tissue to therapeutic irradiation. Int. J. Radiat. Oncol. Biol. Phys. 21, 109–122 (1991).
65. Deasy, J. O. et al. IMPROVING NORMAL TISSUE COMPLICATION PROBABILITY MODELS: THE NEED TO ADOPT A "DATA-POOLING" CULTURE. Int. J. Radiat. Oncol. Biol. Phys. 76, S151–S154 (2010).
66. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350, g7594 (2015).
67. Lambin, P. et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. Nat. Rev. Clin. Oncol. 10, nrclinonc.2012.196 (2012).
68. Schafer, J. L. & Olsen, M. K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivar. Behav. Res. 33, 545–571 (1998).
69. Sterne, J. A. C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338, b2393 (2009).
70. Noble, W. S. What is a support vector machine? Nat. Biotechnol. 24, nbt1206-1565–1565 (2006).
71. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. Mach. Learn. 29, 131–163 (1997).
72. Loh, W.-Y. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1, 14–23 (2011).
73. Akaike, H. Akaike's Information Criterion. in International Encyclopedia of Statistical Science (ed. Lovric, M.) 25–25 (Springer Berlin Heidelberg, 2011). doi:10.1007/978-3-642-04898-2_110
74. Posada, D., Buckley, T. R. & Thorne, J. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Syst. Biol. 53, 793–808 (2004).
75. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–22 (2010).
76. Stone, M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. J. R. Stat. Soc. Ser. B Methodol. 39, 44–47 (1977).
77. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Ser. B Methodol. 36, 111–147 (1974).
78. Appelt, A. L., Vogelius, I. R., Farr, K. P., Khalil, A. A. & Bentzen, S. M. Towards individualized dose constraints: Adjusting the QUANTEC radiation pneumonitis model for clinical risk factors. Acta Oncol. Stockh. Swed. 53, 605–612 (2014).
79. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5, ncomms5006 (2014).

80. Miles, K. A., Ganeshan, B. & Hayball, M. P. CT texture analysis using the filtration-histogram method: what do the measurements mean? Cancer Imaging Off. Publ. Int. Cancer Imaging Soc. 13, 400–406 (2013).
81. Kerns, S. L. et al. The Prediction of Radiotherapy Toxicity Using Single Nucleotide Polymorphism−Based Models: A Step Toward Prevention. Semin. Radiat. Oncol. 25, 281–291 (2015).
82. Andreassen, C. N., Schack, L. M. H., Laursen, L. V. & Alsner, J. Radiogenomics - current status, challenges and future directions. Cancer Lett. 382, 127–136 (2016).
83. Herskind, C. et al. Radiogenomics: A systems biology approach to understanding genetic risk factors for radiotherapy toxicity? Cancer Lett. 382, 95–109 (2016).
84. Kerns, S. L. et al. Meta-analysis of Genome Wide Association Studies Identifies Genetic Markers of Late Toxicity Following Radiotherapy for Prostate Cancer. EBioMedicine 10, 150–163 (2016).
85. Kerns, S. L., Ostrer, H. & Rosenstein, B. S. Radiogenomics: Using Genetics to Identify Cancer Patients at Risk for Development of Adverse Effects Following Radiotherapy. Cancer Discov. 4, 155–165 (2014).
86. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. N. Engl. J. Med. 372, 793–795 (2015).
87. Wild, C. P. Complementing the genome with an ‘exposome’: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. 14, 1847–1850 (2005).
88. Nalbantov, G. et al. Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients. Radiother. Oncol. 109, 100–106 (2013).
89. Chen, R. C., Royce, T. J., Extermann, M. & Reeve, B. B. Impact of Age and Comorbidity on Treatment and Outcomes in Elderly Cancer Patients. Semin. Radiat. Oncol. 22, 265–271 (2012).
90. Recht, A. et al. Postmastectomy Radiotherapy: An American Society of Clinical Oncology, American Society for Radiation Oncology, and Society of Surgical Oncology Focused Guideline Update. Pract. Radiat. Oncol. 6, e219–e234 (2016).
91. Fakhrian, K. et al. Chronic adverse events and quality of life after radiochemotherapy in anal cancer patients. A single institution experience and review of the literature. Strahlenther. Onkol. Organ Dtsch. Rontgengesellschaft Al 189, 486–494 (2013).
92. Han, K. et al. Prospective evaluation of acute toxicity and quality of life after IMRT and concurrent chemotherapy for anal canal and perianal cancer. Int. J. Radiat. Oncol. Biol. Phys. 90, 587–594 (2014).
93. Kachnic, L. A. et al. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. Int. J. Radiat. Oncol. Biol. Phys. 86, 27–33 (2013).
94. Northover, J. et al. Chemoradiation for the treatment of epidermoid anal cancer: 13-year follow-up of the first randomised UKCCCR Anal Cancer Trial (ACT I). Br. J. Cancer 102, 1123–1128 (2010).

95. Rønde, H. S., Wee, L., Pløen, J. & Appelt, A. L. Feasibility of preference-driven radiotherapy dose treatment planning to support shared decision making in anal cancer. Acta Oncol. 56, 1277–1285 (2017).
96. Barry, M. J. & Edgman-Levitan, S. Shared decision making--pinnacle of patient-centered care. N. Engl. J. Med. 366, 780–781 (2012).
97. Stiggelbout, A. M., Pieterse, A. H. & De Haes, J. C. J. M. Shared decision making: Concepts, evidence, and practice. Patient Educ. Couns. 98, 1172–1179 (2015).
98. Bibault, J.-E., Giraud, P. & Burgun, A. Big Data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett. 382, 110–117 (2016).
99. Tomatis, S. et al. Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model. Phys. Med. Biol. 57, 1399–1412 (2012).
100. Gulliford, S. L., Webb, S., Rowbottom, C. G., Corne, D. W. & Dearnaley, D. P. Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate. Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 71, 3–12 (2004).
101. Su, M. et al. An artificial neural network for predicting the incidence of radiation pneumonitis. Med. Phys. 32, 318–325 (2005).
102. Chen, S. et al. A neural network model to predict lung radiation-induced pneumonitis. Med. Phys. 34, 3420–3427 (2007).

Chapter 4

Ontology-guided radiomics analysis workflow (O-RAW)

Zhenwei Shi, Alberto Traverso, Johan van Soest, Andre Dekker and Leonard Wee

Adapted from
Shi, Zhenwei, et al. "Ontology-guided radiomics analysis workflow (O-RAW)." Medical Physics 46.12 (2019): 5677-5684.
DOI: https://doi.org/10.1002/mp.13844

Abstract

Purpose: Radiomics is the process to automate tumor feature extraction from medical images. This has shown potential for quantifying the tumor phenotype and predicting treatment response. The three major challenges of radiomics research and clinical adoption are: (a) lack of standardized methodology for radiomics analyses, (b) lack of a universal lexicon to denote features that are semantically equivalent, and (c) lists of feature values alone do not sufficiently capture the details of feature extraction that might nonetheless strongly affect feature values (e.g. image normalization or interpolation parameters). These barriers hamper multi-center validation studies applying subtly different imaging protocols, pre-processing steps and radiomics software. We propose an open-source Ontology-guided Radiomics Analysis Workflow (O-RAW) to address the above challenges in the following manner: (a) distributing a free and open-source software package for radiomics analysis, (b) deploying a standard lexicon to uniquely describe features in common usage and (c) provide methods to publish radiomic features as a semantically-interoperable data graph object complying to FAIR (Findable Accessible Interoperable Reusable) data principles.

Methods: O-RAW was developed in Python, and has three major modules using open-source component libraries (PyRadiomics Extension and PyRadiomics). First, PyRadiomics Extension takes standard DICOM-RT (Radiotherapy) input objects (i.e. a DICOM series and an RTSTRUCT file) and parses them as arrays of voxel intensities and a binary mask corresponding to a volume of interest (VOI). Next, these arrays are passed into PyRadiomics, which performs the feature extraction procedure and returns a Python dictionary object. Lastly, PyRadiomics Extension parses this dictionary as a W3C-compliant Semantic Web "triple store" (i.e., list of subject-predicate-object statements) with relevant semantic meta-labels drawn from the Radiation Oncology Ontology and Radiomics Ontology. The output can be published on an SPARQL endpoint, and can be remotely examined via SPARQL queries or to a comma separated file for further analysis.

Results: We showed that O-RAW executed efficiently on three datasets with differing modalities, MMD (CT), CROSS (PET) and THUNDER (MR). The test was performed on an HP laptop running Windows 7 operating system and 8GB RAM on which we noted execution time including DICOM images and associated RTSTRUCT matching, binary mask conversion of a single VOI, batch-processing of feature extraction (105 basic features in PyRadiomics), and the conversion to an RDF object. The results were (RIDER) 407.3, (MMD) 123.5, (CROSS) 513.2 and (THUNDER) 128.9 seconds for a single VOI. In addition, we demonstrated a use case, taking images from a public repository and publishing the radiomics results as FAIR data in this study on www.radiomics.org. Finally, we provided a practical instance to show how a user could query radiomic features and track the calculation details based on the RDF graph object created by O-RAW via a simple SPARQL query.

Conclusions: We implemented O-RAW for FAIR radiomics analysis, and successfully published radiomic features from DICOM-RT objects as semantic web triples. Its practicability and flexibility can greatly increase the development of radiomics research and ease transfer to clinical practice.

Introduction

Imaging has developed rapidly in the healthcare field and is commonly used in clinical practice. When integrated into clinical decision support systems (CDSS) [1], medical imaging could play a key role in precision medicine [2] that could lead to better customized healthcare at an individual patient level. Multimodality medical imaging is routinely used in clinical practice, and plays a critical role in how doctors diagnose and treat cancer.

With rising utilization of imaging technology, there has been an increasing interest in the use of quantitative tumor markers derived from imaging data. Radiomics [3-5] is an important development in quantitative image analysis, where digitally encoded medical images containing information related to tumor pathophysiology are converted into high-dimensional mineable features [5, 6]. Radiomics requires a high-throughput computerized tumor feature extraction process that can operate on vast quantities of digital imaging data. The features extracted from imaging data have been associated with key clinical outcomes (e.g., overall survival). Previous studies [7-11] have shown the value of radiomics on quantifying the tumor phenotype and predicting treatment response in clinical settings. By developing diagnostic and prognostic signatures, radiomics is expected to provide additional and complementary information to clinical factors for decision support.

However, three major challenges impede the pace of radiomics research and its clinical adoption: (i) lack of standardized methodology for radiomics analyses; (ii) insufficient information in the feature lexicon to fully characterize the pre-processing steps leading up to feature extraction; and (iii) insufficient information in the extracted feature values for an independent investigator to reproduce the same values (such as image normalization or interpolation parameters). These issues above hamper multi-center studies because of subtly different imaging protocols, pre-processing steps and extraction software. As a result, the development of radiomics research has been impeded. There is a need for an open-source package to make radiomic features more readily comparable for researchers and clinical users. We hypothesize that comparative research will be supported if we not only share radiomic features values, but also information about the pre-processing and computational steps that led to that specific feature value.

An option to address the sharing problem is to progressively lengthen a human-readable label as the feature name, for example *log.sigma.3.0.mm.3D_firstorder_Kurtosis*, but this can become unwieldy if the complexity of metadata increases. An alternative is to provide comprehensive dictionaries so that feature definitions can be cross-referenced, however this also becomes cumbersome when multiple software packages, imaging settings and processing steps come into play. The Semantic Web approach has added value here, since each calculated value of a feature can be defined with a unique identifier independently of its human-readable feature name and additional unique identifiers can be attached which acts as metadata describing that feature. For the purpose of easy comparison, sharing and validation of radiomic results, this is a feasible approach is to build FAIR (Findable, Accessible, Interoperable,

Reusable) [12] radiomic data via an open and extensible semantic ontology for annotating radiomic feature values with metadata and unique identifiers.

In this article, we propose an open-source Ontology-guided Radiomics Analysis Workflow (O-RAW) to address the above challenges in the following manner: (i) distributing a free and open-source software package for radiomics analysis, (ii) using a domain-specific semantic web ontology to uniquely describe features in common usage, and (iii) providing methods to publish radiomic features as a semantically-interoperable data graph object complying to FAIR data principles. With this resource, we aim to support further standardization radiomics analysis with the use of ontologies, promote multi-center collaboration via a novel learning approach using Semantic web (i.e., Resource Description Framework (RDF)) ([13-15] and hence increase the potential for wide external validation and validity of radiomics-assisted clinical prediction models.

Materials and Methods

a. Datasets

Imaging data from different modalities (i.e., CT, PET and MRI) were used in this study:

1) RIDER test-retest dataset [16]: 31 sets of lung tumor CT scans with associated RTSTRUCT;
2) Multi-delineation (MMD) dataset [17]: 21 sets of lung tumor CT scans and corresponding RTSTRUCT with manual delineations from 5 different of oncologists;
3) CROSS trial dataset [18]: 79 sets of esophageal tumor PET scans and corresponding RTSTRUCT.
4) THUNDER trial dataset: 23 Apparent diffusion coefficient (ADC) maps in locally advanced rectal cancer patients. Corresponding RTSTRUCT was delineated manually by three different observers.

Of the above, the RIDER and MMD datasets are publicly available via an image repository (http://xnat.bmia.nl).

b. O-RAW architecture

The O-RAW (version 2.0) workflow package (https://github.com/zhenweishi/O-RAW) was developed using the Python programming language, which encapsulates the workflow in three major steps and uses two open-source component packages (PyRadiomics as radiomics feature extractor [19] and PyRadiomics Extension) [20]). PyRadiomics is an open-source package for radiomics extraction, which can be applied on both two and three-dimensional medical imaging. The primary goal of PyRadiomics is to build an open-source platform that could provide standardized methods for easy and reproducible radiomics extraction and analysis. To achieve this goal, four steps are applied in PyRadiomics: (i) loading and pre-processing of scans and associated segmentation; (ii) employment of wavelet and filters (e.g., Laplacian of Gaussian filter); (iii) radiomic features extraction from first-order statistics, shape, and texture feature classes; and (iv) returning a python dictionary object containing, configuration information, feature names and values. There are several available open-source radiomics software such as MITK [21], Mazda [22], PyRadiomics [19], IBEX [23] and CERR [24]. Apte et al.[24],

described the main characteristics including limitations of some software (supplementary Table-S1). The current limitations of PyRadiomics are partially solved by O-RAW, such as (i) directly take original DICOM images and RTSTRUCT files as input; (ii) describing the process of feature extraction by a universal lexicon (i.e., an ontology) rather than literal expressions. The PyRadiomics Extension package aims to extend the functionality of PyRadiomics on both the input and output sides and allows users to employ native DICOM series and RTSTRUCT directly for radiomics extraction, and convert the radiomic features (python dictionary object) to RDF using the relevant semantic ontology (i.e., radiomics ontology [25]). The conversion is done by mapping individual radiomic feature of PyRadiomics to unique identifiers defined by the Image Biomarker Standardisation Initiative (IBSI). If features do not exist or do not match with the IBSI identifiers, these are defined and labelled using the domain-specific Radiomics Ontology.

Figure 1 shows the workflow of radiomics analysis proposed in O-RAW. First, imaging data from a local repository or web data repository (i.e., XNAT) are retrieved (pyxnat library [26]). Second, PyRadiomics Extension takes standard DICOM-RT inputs (DICOM images and the associated RTSTRUCT file) and parses them as arrays of voxel intensities and a binary mask for each volume of interest (VOI). Next, the arrays are passed into PyRadiomics that performs the abovementioned feature extraction and returns a python dictionary object to PyRadiomics Extension. Then, PyRadiomics Extension parses the dictionary as a W3C-compliant semantic web "triple store" (i.e., RDF) with metalabels attached from the Radiation Oncology Ontology [27] and the IBSI compliant Radiomics Ontology [25]. The RDF result can be published to an http-accessible endpoint, and examined via SPARQL Protocol and RDF Query Language (SPARQL) queries. The information stored in RDF format include the radiomics feature unique identifier, its name and value, pre-processing approaches, VOI(s), the patient identifier, and the radiomics software used. Finally, an external application can perform machine learning algorithms on the RDF triple store and return results back to learning application. Briefly, PyRadiomics is the radiomics feature extractor, and PyRadiomics Extension is the input and output extension of PyRadiomics to handle DICOM images and RDF object. O-RAW is the workflow incorporating these tools to make radiomics study easily and connect to external application. In principle this modular set-up should allow for other modules e.g. other binary conversion methods (e.g. Plastimatch or CERR) or other radiomics feature extractor software (e.g. IBEX, CERR), if in- and output are known.

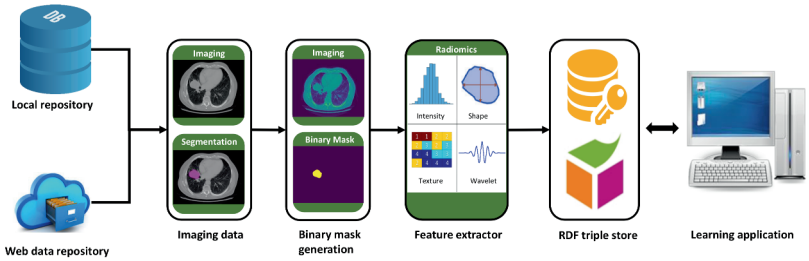


Figure 1: The generalized workflow of O-RAW. First, DICOM imaging data are captured from a local or web repository. Second, DICOM images and the associated RTSTRUCT files are converted into binary masks according the feature extractor input requirements. The radiomics extractor then calculates features and exports

these in a (usually) custom output format. The features are then mapped to RDF to achieve semantic interoperability and published in an http-accessible endpoint, from which they can be queried with SPARQL and used in a learning application. O-RAW, Ontology-guided Radiomics Analysis Workflow; RDF, resource description framework.

Results

In order to assess O-RAW, three tests were performed in this study. In the first test, the ability of O-RAW to handle multiple modalities was verified on four datasets, RIDER (CT), MMD (CT), CROSS (PET) and THUNDER (MR). The results show that O-RAW can perform radiomics analysis on different imaging modalities. O-RAW executed efficiently on these datasets on a laptop running Windows 7 operating system and 8GB RAM, on which the execution time was noted of a common radiomics analysis. This included DICOM images and associated RTSTRUCT matching, binary mask conversion, feature extraction (105 basic features), and conversion of RDF object. The results were (RIDER) 407.3, (MMD) 123.5, (CROSS) 513.2 and (THUNDER) 128.9 seconds for a single VOI.

Second, to evaluate the method of binary mask conversion in O-RAW, we used Plastimatch (version 1.7.3), PyRadiomics Extension and CERR (MATLAB) to convert binary masks for 100 randomly selected patients from MMD and CROSS datasets with CT and PET scans. This test led to a mask in NRRD format per VOI per application. The flowchart is shown in **Figure 2** (pipeline 1, 2 and 3). The Dice similarity coefficient [28] of binary masks converted by three approaches were all exactly unity, which indicated that conversion by either Plastimatch, PyRadiomics Extension and CERR result in the same binary mask.

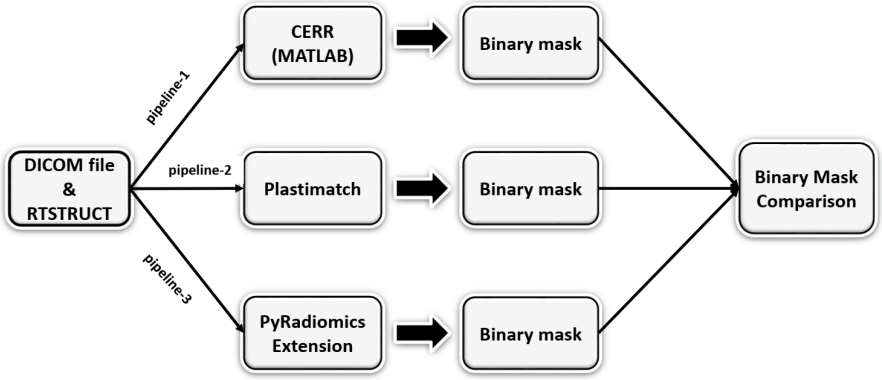


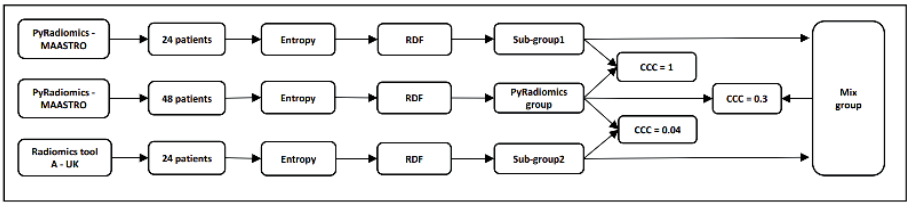
Figure 2: Three approaches (CERR, Plastimatch, and PyRadiomics Extension) shown in pipeline 1, 2, 3 were used for binary mask conversion. All three methods converted binary mask in NRRD format. One hundred randomly selected patients from MMD and CROSS datasets with CT and PET scans. The Dice similarity coefficient was used as the comparison measurement.

O-RAW provides two user-configurable options of output format, CSV (Comma delimited) and RDF (default). On the one hand, the simpler csv file output can be saved in the export directory given in the configuration file of O-RAW, which contains the information of patient ID, VOI(s), and values of radiomic features. To track more details of radiomic features

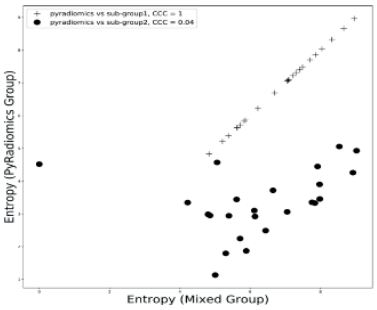
extraction (e.g., pre-processing methods), information should be saved in other flat tables, which is the limitation of using relational tables to present data using a rigid and pre-defined structure (known as a schema). Another and preferred option it to save radiomic features in the RDF format, which allows full expressivity of features and their details. The information can be retrieved by using SPARQL queries from a SPARQL endpoint (e.g., Blazegraph), which not only includes patient identifiers, VOI(s), and values of radiomic features, but also feature units, pre-processing and radiomics software details.

In **Figure 3**, we show a real-world example, which demonstrates the importance of tracking computation details and the feasibility of describing radiomics via an RDF graph object. As shown in **Figure 3a**, a dataset with 48 patients is first randomly split into two sub-groups without overlap, 24 patients for each. One sub-group data was sent to our partner in the UK. The radiomics of the sub-group1 were computed by PyRadiomics in MAASTRO and converted into RDF format. The radiomics of the sub-group2 were computed by Radiomics tool A and converted into RDF format as well. We combined radiomics of the two sub-groups into a mixed group. Second, all 48 patients were computed by PyRadiomics only as PyRadiomics group. As a demonstration, we only select the feature entropy, which is defined identically, in terms of definition, formula and IBSI [29] code, in both two software (PyRadiomics and Radiomics tool A) implementations. Concordance correlation coefficients (CCC) was calculated between the PyRadiomics group and mixed group, which yielded a CCC score of 0.3. The feature entropy values of two groups are shown in **Figure 3-b**, where some cases in the mixed group were approximately identical to feature values in the PyRadiomics group, but some are not. Then, a simple SPARQL query shown in **Figure 3-c** was used to track the computation details. According to the returned result (**Figure 3-d**), the reason of the low CCC score is caused by using different software (PyRadiomics vs MATLAB radiomics toolbox). By analyzing more computational details such as image pre-processing or filter methods, one could further investigate why these two entropy values are different between software implementation.

a



b



c

```
PREFIX ncit:<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX roo: <http://www.cancerdata.org/roo/>
PREFIX ro: <http://www.radiomics.org/RO/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?patient ?radiomicfeatures ?programming_language ?software_version ?featureValue
WHERE {
                # patient defined using the ncit ontology
                ?patient a ncit:C16960 .
                # patient -> has_radiomicfeature- > feature name (entropy)
                ?patient ro:P00088 ?radiomicfeatures.
                #radiomics -> computed_usIng -> calculationrun
                ?radiomicfeatures ro:P00002 ?calculationrun .
                # calculationrun -> performed_by -> software
                ?calculationrun ro:0283 ?software .
                # software -> has_version -> version
                ?software ro:0010192 ?software_version .
                # software -> has_programminglanguage -> programming_language
                ?software ro:0010195 ?programming_language .
                #radiomicfeature -> has_value -> feature value
                ?radiomicfeatures ro:010191 ?featureValue .
}
```

d

Patient	Radiomicsfeatures	Programming_language	Software_version	Feature Value
Patient-001	Entropy	MATLAB	V9.3, R2017b	3.3315
Patient-002	Entropy	MATLAB	V9.3, R2017b	3.0625
Patient-003	Entropy	MATLAB	V9.3, R2017b	3.4564
Patient-004	Entropy	MATLAB	V9.3, R2017b	2.441
. . .	. . .	. . .	. . .	. . .
Patient-025	Entropy	Python	3.5	8.288166
Patient-026	Entropy	Python	3.5	9.026876
Patient-027	Entropy	Python	3.5	8.041159
Patient-028	Entropy	Python	3.5	5.395222
. . .	. . .	. . .	. . .	. . .

Figure 3: A real world example showing the use of an ontology-supported description of a radiomics feature rather than just using the feature name. a) Flowchart indicating the use of two different radiomics feature extraction software implementation (PyRadiomics and MATLAB). b) Comparison of entropy values calculated by a mix of MATLAB and PyRadiomics (x-axis) to PyRadiomics alone (y-axis). The black dots indicate entropy calculated by MATLAB which has a concordance correlation coefficient of only 0.04 with PyRadiomics calculated entropy. c) SPARQL query to retrieve patient ID, radiomic feature name, programming language, software version, and feature value. d) Returned results of SPARQL query, where one can see the feature entropy was computed by two software implementations (MATLAB and Python/PyRadiomics).

The visualization of RDF graph data generated by O- RAW is presented in **Figure 4**. This RDF graph object complies with the radiomics output structure of IBSI that as described in the Radiomics Ontology. Moreover, we showed an example in the **Fig. S1**, which describes how radiomic features are queried and computation details are tracked by a simple SPARQL query.

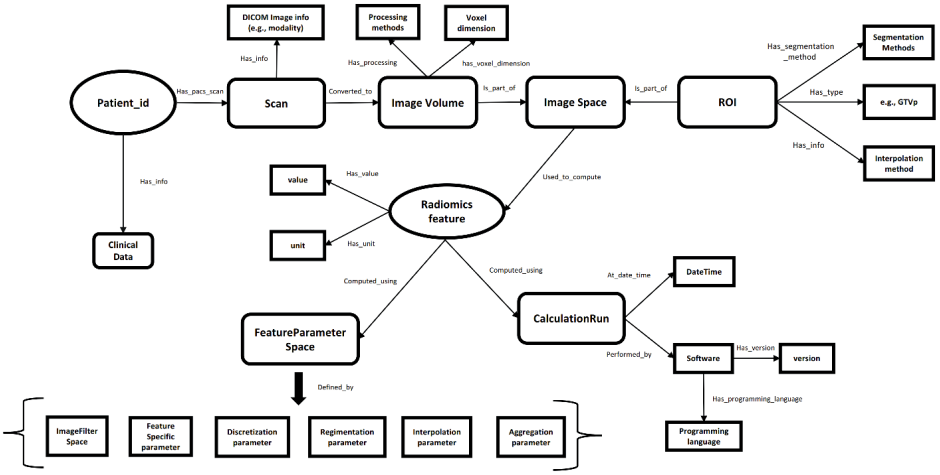


Figure 4: Visualization of nodes and relations as an RDF graph object generated by O-RAW. The structure follows the IBSI compliant Radiomics Ontology. More details, see [30]. The radiomics graph is able to link to a clinical data graph via Patient ID.

Discussion

Our work (O-RAW) aims to address the lack of a standardized methodology for radiomics analysis. Most radiomics toolboxes are developed in-house without public and standardized documentation on the details of the radiomics calculation and analysis, which makes it difficult to reproduce and validate radiomics studies. Besides in-house developments, there are radiomics toolboxes which are publicly available, such as MITK [21], Mazda [22], PyRadiomics [19], IBEX [23], CERR [24] and LifeX [31], and also commercial implementations exist. Each of these have different capabilities and limitations [24], showcasing the need for standardization efforts such as presented in this study.

The O-RAW workflow package was developed using two open-source component packages (PyRadiomics, and PyRadiomics Extension), of which functionalities are clarified here. PyRadiomics plays the role of a radiomics feature extractor using native file formats as input (e.g., NRRD format) and output (e.g., CSV). PyRadiomics Extension allows the use of standardized file formats as input (DICOM images and RTSTRUCT) and output (RDF) based on an ontology (i.e., Radiomics Ontology and Semantic DICOM ontology). The O-RAW package integrate PyRadiomics and PyRadiomics Extension to implement batch processing including DICOM handling, ROI selection and exclusion, conversion from RDF to additional standardized file formats (CSV), and so on. The main innovation of O-RAW is thus in orchestrating the workflow of a radiomics study. It can work with any radiomics feature extraction software, provided that they accept standard formats for input (i.e., file formats that can be read by ITK) and export data according to the Radiomics Ontology.

We selected PyRadiomics as the feature extractor in O-RAW, as it best fits the concept of O-RAW currently, in terms of well standardized documentation, universal programming language (python), fully open-sourced code, rapid maturation and an active community of user-

developers. With O-RAW also fully open source, the process of making radiomic features FAIR using associated ontologies can be reproduced easily by others. Although we used PyRadiomics, it is important to note that the modular nature of O-RAW allows easy integration of other radiomics toolkits as long as its input and output are known. When including a different radiomics toolkit, syntactic and semantic interoperability has to be created. With regard to syntactic interoperability, O-RAW uses DICOM as its input syntax and RDF as its output syntax. In current radiomics feature extraction tools, either DICOM is already accepted (e.g., in CERR) requiring no change. If another standardized input format (e.g., NRRD) is accepted, then tools exist to convert DICOM into these formats (e.g., ITK). Output formats of radiomics toolkits vary (Python dictionary, CSV, etc.) but many tools exist to convert such application data into RDF (https://www.w3.org/wiki/ConverterToRdf) – requiring simple configuration of those tools. Achieving syntactic interoperability is therefore not very difficult. With regard to semantic interoperability, recent standardization efforts for radiomics features, including the Radiomics Ontology, which has also implemented the IBSI standard, have emerged but have not yet been implemented widely. Custom code is thus necessary to map the native nomenclature to semantically standardized radiomics features. Also, details on how the radiomic features were calculated (e.g. filtering, software version etc.) need to be configured in O-RAW. As an example, the current study used an in-house MATLAB (Radiomics tool A) implementation from our UK partner and mapped its MATLAB output to RDF. In practice, this is a two-step process: (i) Filling in the toolkit details in the configuration file of O-RAW including computational details as described in the Table 1 of the literature [32]. This configuration table is used to create a base RDF graph containing information on the radiomics toolkit and its settings. (ii) Mapping native names of features to Radiomics Ontology codes, by filling in mapping table with two columns: one is the feature names output from the users' radiomics toolkit and the other one is the radiomics ontology codes. This mapping table is then used to create the individual radiomic features in RDF. The two extra steps can be avoided when using PyRadiomics, as the process of mapping is implemented automatically via the PyRadiomics Extension in O-RAW.

In all cases, O-RAW allowed users to track the details (e.g., feature calculation approaches or parameters) of each step within a typical radiomics analysis workflow. We feel the primary aim and benefit of O-RAW was thus demonstrated which is to support reproducible and interoperable radiomics research with the use of ontologies, promote multi-center collaboration and hence increase the potential for wider external validation studies of radiomics-assisted clinical prediction models.

The use of a standard and publicly accessible lexicon, the IBSI compliant Radiomics Ontology [25], explicitly documents the definitions and mathematical formulas of radiomic features. Using an ontology is a major improvement over using a human-readable label alone, which is not sufficient to guarantee semantic equivalence and interoperability. We feel creating semantic interoperability through the use of ontologies is essential for the comparison and validation of radiomic studies, given the diverse software implementations, pre-processing approaches and feature labels which are in active use.

For example, using an ontology first forces one to choose if a feature called "entropy" is the Intensity Histogram Entropy (IBSI ID = TLU2) [25] or textural feature Joint Entropy (IBSI ID = TU9B) [25]. Second, using an ontology, two users who compute Joint Entropy can also note what pixel spacing and software implementation was used to compute their respective Joint Entropies. We have shown in this study that O-RAW can offer such detailed expressiveness by using the Radiomics Ontology to describe features and attach metadata for those features, resulting in semantic interoperability and ultimately FAIR data.

Finally, flat tables for radiomics output do not sufficiently capture the methodological steps that affect feature values. For instance, image resampling prior to features extraction might affect the result [33, 34]. However, a two dimensional table that only describes the radiomic feature names and their values is not sufficient to determine which methodologies of pre- and post-processing are used for radiomics calculation. It is impossible to know if radiomic features from two flat tables use the same pre-processing method(s) without any further information, though their value might be equal.

We identified the challenges for radiomics research as: (i) lack of standardized methodology for radiomics analyses, (ii) lack of a universal lexicon to denote features that are semantically equivalent, and (iii) lists of feature values alone do not sufficiently capture the details of feature extraction that might nonetheless strongly affect feature values (e.g. image normalization or interpolation parameters). We have demonstrated that O-RAW is capable of handling these challenges. First, the radiomics extractor used in O-RAW is PyRadiomics, which is the largest open-source radiomics package and has attracted more and more attention of researchers in the radiomics community. Second, the IBSI compliant radiomics ontology is applied in O-RAW to guide radiomics analysis, which offers unambiguous metadata to note if two radiomic features are equivalent semantically or not. Third, the default output format is RDF within O-RAW, which can link radiomic features and values to related meta-data, such as patient ID, VOI, unit, pre-processing, and software version. The uniqueness of a radiomic feature is not the name of the feature, but the details describing how the feature is computed. As shown in **Figure 4**, when two features have an identical name, they may not be identical. Their computation settings, such as image processing and filter methods, may be different, which can be tracked by using ontologies such as proposed in O-RAW.

One of the benefits of using RDF graphs to store radiomics data which we will study in future work, that is it allows the linkage to other RDF graphs e.g. containing clinical data such as histology and other biological characteristics of the tumor and outcomes which are important in many Radiomics studies to derive relations between the imaging phenotype and tumor genotype and to make clinically relevant prediction models. Similarly, RDF allows as to leverage work done in the Semantic DICOM ontology (https://www.ncbi.nlm.nih.gov/pubmed/25160167) and store the complete DICOM header in RDF and link it to the Radiomics RDF. This is part of our future works as it would make it unnecessary to derive and store DICOM header information (e.g. slice thickness and pixel spacing) in O-RAW as is done now.

Finally, O-RAW is able to generate FAIR data: (i) radiomics data and extraction details could be published with a Findable(F) and unique identifier; (ii) radiomics data and metadata are described with the radiomics ontology, which make them accessible(A) and understandable by machines and humans; (iii) data uses a formal, standardized and applicable ontology for knowledge representation, which makes interoperability(I) among multi-centers possible; (iv) data offers explicit information on provenance and licenses for reuse (R).

A current limitation of the O-RAW package is that multi-center studies based on querying the feature RDFs must be based on PyRadiomics or converted to RDF triples. Future work will involve three aspects. First, we will to extend our method to convert features generated in native format and nomenclature into IBSI compliant, ontology-based to other radiomics software, and will cooperate with other willing developers to address the challenge to create syntactic and semantic interoperability between radiomics studies. With such interoperability O-RAW would allow identification of differences in feature calculation between different packages or vendors. Second, a distributed learning study among multi- centers will be performed to link clinical outcome, DICOM header and radiomic features via O-RAW. Finally, accurate and robust automatic segmentation of tumor tool will be integrated into our workflow. It means that O-RAW will extract radiomics from original DICOM images without a requirement for any other (manual) annotation information (i.e., RTSTRUCT).

Conclusion

In this study, we successfully implemented O-RAW for radiomics analysis from radiotherapy-based images to ontology guided FAIR data. Its practical use and flexibility can greatly promote the advance of radiomics research and may help the associated achievements transfer to clinical practice. The development goal of O-RAW is to help radiomics users on both input and output sides. First, it allows to import original DICOM images and RTSTRUCT files that are commonly used in the radiation oncology field. Second, the output is machine-readable data with related ontologies, which promotes the standardization in terms of radiomic features, pre- and post-processing.

References

1. Lambin, P., et al., *Predicting outcomes in radiation oncology—multifactorial decision support systems.* Nature reviews Clinical oncology, 2013. **10**(1): p. 27.
2. Hood, L. and S.H. Friend, *Predictive, personalized, preventive, participatory (P4) cancer medicine.* Nature reviews Clinical oncology, 2011. **8**(3): p. 184.
3. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic resonance imaging, 2012. **30**(9): p. 1234-1248.
4. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* European journal of cancer, 2012. **48**(4): p. 441-446.
5. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nature communications, 2014. **5**: p. 4006.
6. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nature Reviews Clinical Oncology, 2017. **14**(12): p. 749.
7. Coroller, T.P., et al., *CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.* Radiotherapy and Oncology, 2015. **114**(3): p. 345-350.
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, 2016. **34**(18): p. 2157-2164.
9. Coroller, T.P., et al., *Radiomic phenotype features predict pathological response in non-small cell lung cancer.* Radiotherapy and Oncology, 2016. **119**(3): p. 480-486.
10. Leijenaar, R.T., et al., *External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma.* Acta Oncologica, 2015. **54**(9): p. 1423-1429.
11. Parmar, C., et al., *Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer.* Scientific reports, 2015. **5**: p. 11044.
12. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.
13. 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, 2017. **4**: p. 24-31.
14. 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, 2017. **99**(2): p. 344-352.
15. 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, 2016. **121**(3): p. 459-467.
16. Zhao, B., et al., *Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer.* Radiology, 2009. **252**(1): p. 263-272.
17. Van Baardwijk, A., et al., *Pet-ct–based auto-contouring in non–small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes.* International Journal of Radiation Oncology• Biology• Physics, 2007. **68**(3): p. 771-778.

18. Shapiro, J., et al., *Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial.* The lancet oncology, 2015. **16**(9): p. 1090-1098.
19. van Griethuysen, J.J., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer research, 2017. **77**(21): p. e104-e107.
20. Shi, Z. *PyRadiomics Extension (Py-rex)*. 2017 [cited 2017; Available from: https://github.com/zhenweishi/Py-rex.
21. Götz, M., et al., *MITK Phenotyping: An open-source toolchain for image-based personalized medicine with radiomics.* Radiotherapy and Oncology, 2019. **131**: p. 108-111.
22. Szczypiński, P.M., et al., *MaZda—a software package for image texture analysis.* Computer methods and programs in biomedicine, 2009. **94**(1): p. 66-76.
23. Zhang, L., et al., *IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics.* Medical physics, 2015. **42**(3): p. 1341-1353.
24. Apte, A.P., et al., *Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research.* Medical physics, 2018. **45(8), 3713-3720.**
25. Traverso, A. *Radiomics Ontology*. 2017; Available from: https://bioportal.bioontology.org/ontologies/RO.
26. Schwartz, Y., et al., *PyXNAT: XNAT in python.* Frontiers in neuroinformatics, 2012. **6**: p. 12.
27. Traverso, A., et al., *The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques.* Medical physics, 2018. **45.10**: p. e854-e862.
28. Dice, L.R., *Measures of the amount of ecologic association between species.* Ecology, 1945. **26**(3): p. 297-302.
29. Zwanenburg, A., et al., *Image biomarker standardisation initiative-feature definitions.* Preprint at arXiv: https://arxiv.org/abs/1612.07003, 2016.
30. 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, 2018. **78**(16): p. 4786-4789.
31. Traverso, A. *RadiomicsOntologyIBSI*. 2018 [cited 2018; Available from: https://github.com/albytrav/RadiomicsOntologyIBSI.
32. Ibrahim, A., et al., *Radiomics Analysis for Clinical Decision Support in Nuclear Medicine.* Seminars in nuclear medicine, 2019. **49**(5): p. 438-449.
33. Fave, X., et al., *Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer.* Translational Cancer Research, 2016. **5**(4): p. 349-363.
34. Lu, L., et al., *Robustness of Radiomic Features in [11 C] Choline and [18 F] FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization.* Molecular Imaging and Biology, 2016. **18**(6): p. 935-945.

Supplementary

Table S1: The primary characteristics of publicly available open-source radiomics extraction tools [56].

	Programming language	IBSI feature definitions	Full OS compatibility	DICOM-RT Import	Integrated visualization	Radiomics metadata storage	Built-in segmentation	Radiomics maps
ITK	C++	No	Yes	Yes	No	No	No	Yes
MaZda	C++/Delphi	No	No	No	Yes	No	Yes	Yes
PyRadiomics	Python	Yes	Yes	No	No	No	No	No
IBEX	Matlab/C++	No	No	Yes	Yes	Yes	Yes	No
CERR	Matlab	Yes	Yes	Yes	Yes	Yes	Yes	Yes

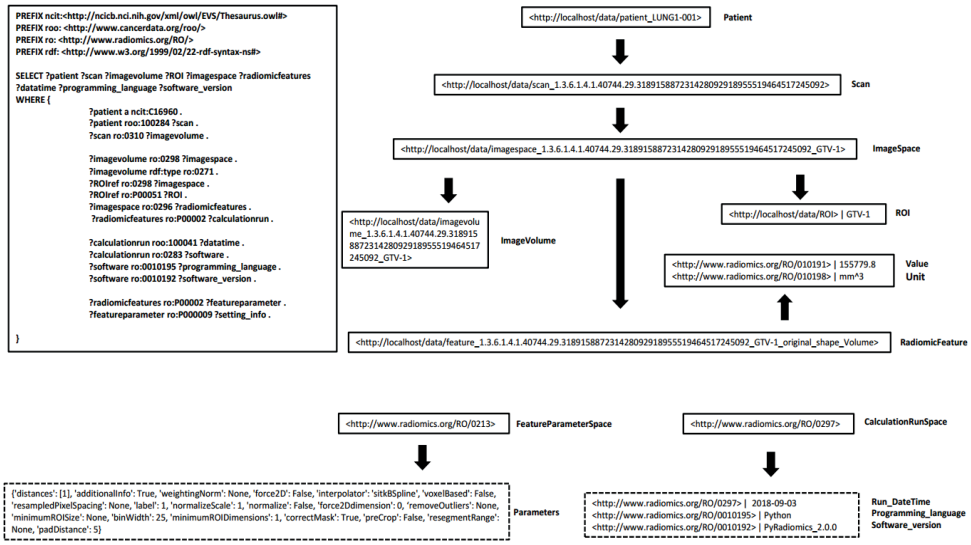


Figure S1: shows an example, which describes how radiomic features are queried and computation details are tracked by a simple SPARQL query. The radiomic features could be linked to the clinical data of the patient by patient ID.

Chapter 5

Prediction of Lymph Node Metastases Using Pre-Treatment PET Radiomics of the Primary Tumor in Esophageal Adenocarcinoma: an External Validation Study

Zhenwei Shi, Chong Zhang, Petros Kalendralis, Philip Whybra, Craig Parkinson, Maaike Berbee, Emiliano Spezi, Ashley Roberts, Adam Christian, Tom Crosby, Andre Dekker, Leonard Wee and Kieran G Foley

In preparation

Abstract

Purpose: To improve clinical lymph node staging (cN-stage) in esophageal adenocarcinoma by developing and externally validating three prediction models with 1) clinical variables 2) positron emission tomography (PET) radiomics and 3) a combined clinical and radiomics model.

Method: Consecutive patients with fluorodeoxyglucose (FDG) avid tumors treated with neo-adjuvant therapy between 2010 and 2016 in two international centers (n=130 and n=60, respectively) were included. Four clinical variables (age, gender, clinical T-stage and tumor regression grade) and PET radiomics derived from the primary tumor were used to construct the three models. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, area under curve (AUC), discrimination and calibration were calculated for each model. The prognostic significance was also assessed.

Results: The incidence of lymph node metastasis (LNMs) was 58% in both cohorts. The AUCs of the clinical, radiomics and combined models were 0.79, 0.69 and 0.82 in the developmental cohort, and 0.65, 0.63 and 0.69 in the external validation cohort, with good calibration demonstrated. In comparison, the AUC of current cN-stage in development and validation cohorts was 0.60 and 0.66, respectively. For overall survival, the combined clinical and radiomics model achieved the best performance to discriminate the external validation cohort (X^2 6.08, CI = 0.60, df 1, p = 0.01).

Conclusion: Accurate diagnosis of LNMs is crucial for predicting prognosis and guiding treatment decisions. Despite obtaining signal for improved prediction in the development cohort, the models using PET radiomics derived from the primary tumor were not fully replicated in an external validation cohort.

Introduction

Esophageal carcinoma (EC) is the eleventh most common cancer and the sixth leading cause of cancer-associated death worldwide [1, 2] with adenocarcinoma being the most common histological cell type in many Western countries. The overall five-year survival rate of EC patients is 15%, with less than 40% of patients suitable for potentially curative therapy at presentation [3]. Importantly, lymph node metastases (LNMs) are a significant prognostic indicator of survival in EC [4]. Accurate knowledge of LNMs influences patient stratification, selection for radical therapy, treatment decision-making and planning.

Lymph nodes are assessed using computed tomography (CT), endoscopic ultrasound (EUS) and positron emission tomography with CT (PET/CT) as part of clinical Tumor Node Metastasis (TNM) staging [5]. Recent data suggests that the accuracy of lymph node staging with CT, EUS and PET/CT is poor (54.5%, 55.4% and 57.1%, respectively) [6]. The poor accuracy has been attributed to a high incidence of micro-metastases within morphologically normal-sized lymph nodes. These data are supported by a similar study which also demonstrated suboptimal N-staging accuracy (75.6%, 77.2% and 74.5%, respectively [7]). Thus, existing EC staging techniques are unlikely to detect small LNMs, so alternative biomarkers that improve diagnostic accuracy should be sought. The current difficulty in identifying LNMs is likely to be a contributor of poor treatment outcomes.

Advances in quantitative medical image data-mining techniques, broadly known as radiomics, enable the non-invasive decoding of tumor heterogeneity by translating medical images into abstract numerical features for analysis. CT-derived radiomics have enabled superior prediction of LNMs in colorectal cancer [8], bladder cancer [9], and esophageal cancer [10, 11]. Previous pre-operative CT studies achieved satisfactory detection of LNMs for esophageal squamous cell carcinoma, reporting area under curve (AUC) statistics of 0.806 and 0.758 in development cohorts, and 0.771 and 0.773 in the validation cohorts, respectively [11, 12]. Similar results have been reported using magnetic resonance imaging (MRI) radiomics, although MRI is often not routinely performed in clinical workup [13].

PET radiomics have been significantly associated with overall survival [14], response to neo-adjuvant therapy [15] and metastases [16], but the performance of PET radiomics to predict LNMs for esophageal adenocarcinoma has not been validated. Increasing scientific evidence demonstrates that metastatic spread from the primary tumor is driven by biological changes in the underlying microenvironment of the primary tumor [17]. Generally, the additional value of radiomics extracted from small regions of interest, such as lymph nodes, over that of simple metrics such as volume is felt to be limited [18]. Accurate delineation is difficult and time-consuming which hinders the clinical utility of lymph node radiomics, unlike larger primary tumors which are more reliably outlined with less error [19]. Therefore, our study attempted to improve currently poor lymph node staging accuracy by extracting pre-treatment PET radiomics from the primary tumor to quantify its metastatic potential and predict the presence of LNMs following surgery.

In this study, we investigated the predictive value of PET radiomic features for LNMs by comparing three models: (1) a model based on clinical variables alone; (2) a model based on PET radiomics alone and (3) a combined model developed by clinical variables and PET radiomic features. The prognostic significance of developed LNM models was also assessed.

Materials & Methods

This study was a review board-approved Transparent Reporting of a multi-variable prediction model for Individual Prognosis or Diagnosis (TRIPOD) type 3 study (model development and external independent validation) [20]. Research ethics committee approval was granted (reference 19/WA/0119).

Patients

To minimize selection bias, consecutive patients (n=190) with biopsy proven FDG-avid esophageal adenocarcinomas treated with neo-adjuvant therapy and surgery between 2010 and 2016 were included in this retrospective study. The development cohort (hereafter called "STAGE") comprised 130 patients receiving either surgery alone, neo-adjuvant chemo (NCT) or neo-adjuvant chemoradiotherapy (NCRT) followed by surgery in the *blinded* [21]. The external validation cohort (hereafter called "CROSS") comprised 60 patients who underwent NCRT at *blinded*. In both cohorts, only patients without stents were included in the study. The PET/CT was performed prior to any treatment, but not repeated after neo-adjuvant therapy. This is common practice in many countries because the examination is expensive [22] and evidence for its cost-effectiveness in clinical practice is currently lacking. A proportion of these patients have previously been reported; 138 of 403 STAGE patients were reported in [14] and 46 of 60 CROSS patients in [23]. These prior articles developed a prognostic model for overall survival and validated the results in an external cohort. In the present study, we use standardized features to predict LNMs using radiomics from the primary tumor. The CROSS cohort were treated with the CROSS regimen [24-26] followed by resection of the esophagus after NCRT. **Figure 1** details the number of patients and exclusion criteria in each cohort.

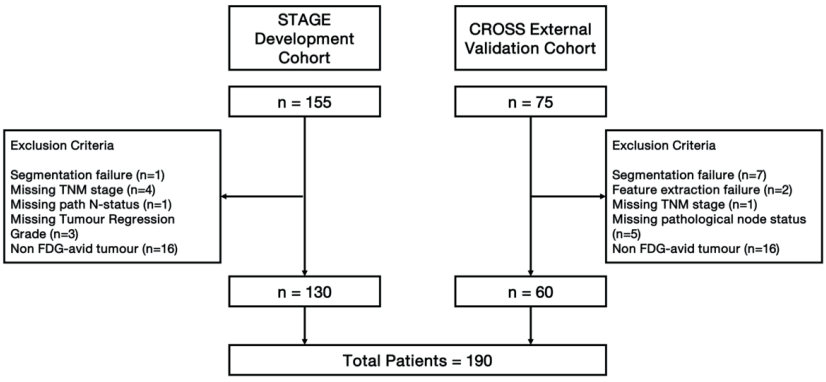


Figure 1: Study flowchart describing the numbers of patients in each cohort and reasons for exclusions from the CROSS cohort.

Clinical Parameters

Routine clinical demographics were collected. Age was recorded at the time of diagnosis. Tumor location was recorded from a combination of endoscopic and radiological examinations. Radiological staging was assigned according to TNM 7th edition, which was used during the study period [27]. Tumor regression grade (TRG) was defined according to a Mandard Score [28]. The primary outcome was LNM status, determined by histopathological examination. Overall survival was defined in months from the date of diagnosis until date of death or last follow-up.

Radiomics Feature Extraction and Tumor Segmentation

To reduce interobserver variability, esophageal primary gross tumor volumes (GTVs) were systematically delineated on PET images using "Automatic Decision Tree Learning Algorithm for Advanced Segmentation" (ATLAAS) [29]. ATLAAS was implemented in MATLAB (The Mathworks, Natick, USA) as a plug-in to the Computational Environment for Radiological Research (CERR) [30]. PET images were re-sampled into 0.5 standardized uptake value (SUV) equally-sized bins, which has been recommended in [31]. In total, 154 radiomic features were extracted from the GTV using the Spaarc Pipeline for Automated Analysis and Radiomics Computing (SPAARC) at Cardiff University [32]. SPAARC radiomic features comply with the Image Biomarker Standardization Initiative (IBSI) [33]. Different scanners and imaging protocols were used across the two centers. Radiomic features could be changed significantly as a function of scanner, image acquisition or reconstruction settings, hence we performed the post-reconstruction Combat harmonization method [34] to harmonize features extracted from images acquired across different centers.

Feature Selection and Prediction Model Development

A flowchart describing feature selection and model development in the STAGE cohort is shown in **Figure 2**. Recursive Feature Elimination (RFE) and a Least Absolute Shrinkage and Selection Operator (LASSO) method were used to select the optimal clinical feature combination (selected from age, gender, tumor location, histological cell type, clinical T-stage, type of neo-adjuvant treatment and tumor regression grade (Mandard score)), using the AUC measurement from the receiver operator curve (ROC). Clinical N-stage (cN-stage) was collected for comparative analysis and to evaluate the baseline staging accuracy in each cohort, but was not included in the multivariable models to avoid multi-collinearity and prevent influencing the model by using potentially inaccurate data.

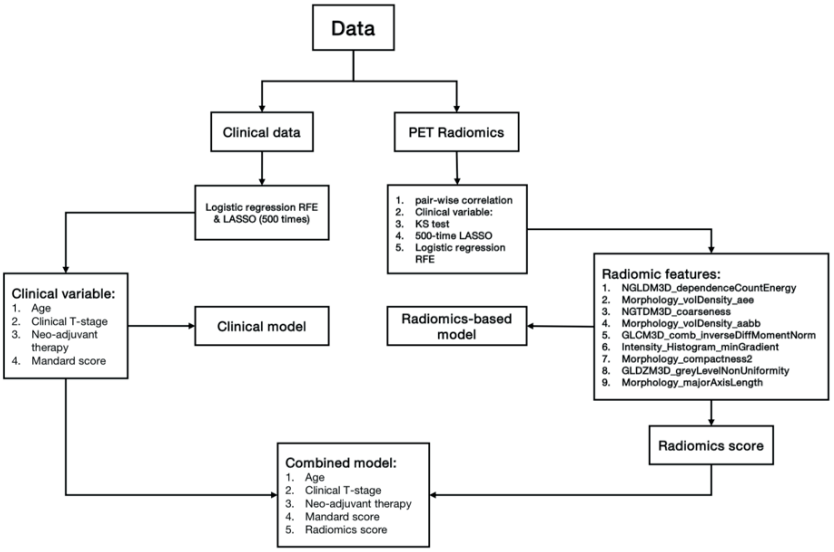


Figure 2: The flowchart of the used feature selection and model development approaches.

For radiomic features, pair-wise Pearson correlation of radiomic features was calculated and the threshold was set to 0.85. The cut-off logic assessed the mean absolute correlation of each radiomic feature and removed the feature with the largest mean absolute correlation. A non-parametric Kolmogorov-Smirnov test statistic was calculated for each feature between resected node positive (pN+) and negative (pN0) outcomes and only features with a p-value < 0.05 were retained to ensure the two classes were significantly distinguished. A LASSO method was used to tabulate the most frequently-selected radiomic features over 500 repetitions of internally separating STAGE into training (70%) and validation (30%) subsets. The best lambda of LASSO was automatically selected in each repetition based on AUC. Finally, we used RFE with 5-fold cross validation to search for combinations of features with non-zero frequency, to find an optimal combination by its AUC.

Feature value normalization after RFE was performed using the mean and standard deviation of selected features in STAGE. Prediction models were developed using multivariable logistic regression in the STAGE cohort and a radiomics score (Rad-score) for each patient was then computed using the coefficients weighted by regression model. A combined model was developed using the selected clinical variables and the Rad-score. External validation in CROSS was performed using the same data transformations that were applied in STAGE.

Statistical Analysis

Statistical analyses were performed in R (v3.30). Clinical demographic differences between STAGE and CROSS were examined by two-sided t-test (continuous variables) or chi-square/fisher test (categorical variables). Estimates of 95% confidence intervals were derived from 2000 stratified bootstrap replicates. Appropriate calibration of the models was assessed using calibration plots and Hosmer-Lemeshow test statistics.

Prognostic significance was explored by entering the selected clinical variables and radiomic features to a Cox proportional hazards model with censoring. We computed the Harrell concordance index and performed log-rank tests of significance for the survival models. To ensure the higher-order radiomics variables were not just surrogates for simple tumor characteristics, we also compared the concordance indices and log-rank tests against primary metabolic tumor volume (MTV) and total lesion glycolysis (TLG).

Results

The patient characteristics of the STAGE and CROSS cohorts are detailed in **Table 1**. The incidences of pathological lymph node metastasis (pLNMs) were 58% (75/130) and 58% (35/60) in STAGE and CROSS, respectively. In STAGE, the majority (62.3%) had NCT whereas all CROSS patients had NCRT. The cohorts differed significantly for cN staging and tumor location. mean follow-up times were 25.6 months (95% confidence interval (CI): 22.7-28.4) in the STAGE and 28.5 months (95% CI: 23.6-33.4) in the CROSS cohorts, respectively.

Table 1: Patient Characteristics in STAGE and CROSS Cohort.

Characteristic; Frequency (%)	STAGE Development Cohort (n=130)	CROSS Validation Cohort (n= 60)	p-value
Tumor type			$P = 1.00
Adenocarcinoma	130 (100%)	60 (100%)	#P = 0.63
Age mean ± SD, years	64.33 ± 9.54	63.15 ± 8.68	
Gender			$P = 0.38
Male	111 (85.4%)	54 (90.0%)	
Female	19 (14.6%)	6 (10.0%)	
Tumor location			+P =0.0059
Distal third esophagus	45 (34.6%)	34 (56.6%)	
Mid third esophagus	7 (5.4%)	1 (1.7%)	
Esophagitis junction	78 (60%)	24 (41.7%)	
Clinical T stage			+P = 0.12
T1	5 (3.8%)	0 (0.0)	
T2	14 (10.7%)	12 (20%)	
T3	101 (77.7%)	46 (76.7%)	
T4a	10 (7.8%)	2 (3.3%)	#P < 0.001
Clinical N stage			
N0	60 (46.2%)	15 (25.0%)	
N1	50 (38.5%)	17 (28.3%)	
N2	13 (10.0%)	15 (25.0%)	
N3	7 (5.3%)	13 (21.7%)	
Stage Groups			+P = 0.16
Stage 1	17 (13.1%)	4 (6.7%)	
Stage 2	43 (33.1%)	15 (25.0%)	
Stage 3	70 (53.8%)	41 (68.3%)	
TRG Score			+P < 0.08
1	12 (9.2%)	11 (18.3%)	
2	12 (9.2%)	11 (18.3%)	
3	13 (10.0%)	14 (23.3%)	
4	37 (28.5%)	16 (26.7%)	
5	26 (20.0%)	8 (13.4%)	
Not applicable	30 (23.1%)	0 (0.0)	
Neo-adjuvant therapy			+P < 0.001
NCRT	19 (14.6%)	60 (100%)	
NCT	81 (62.3%)	0 (0.0)	
Surgery alone	30 (23.1%)	0 (0.0)	
Overall survival			$P = 0.11
Alive	77 (59.2%)	28 (46.7%)	

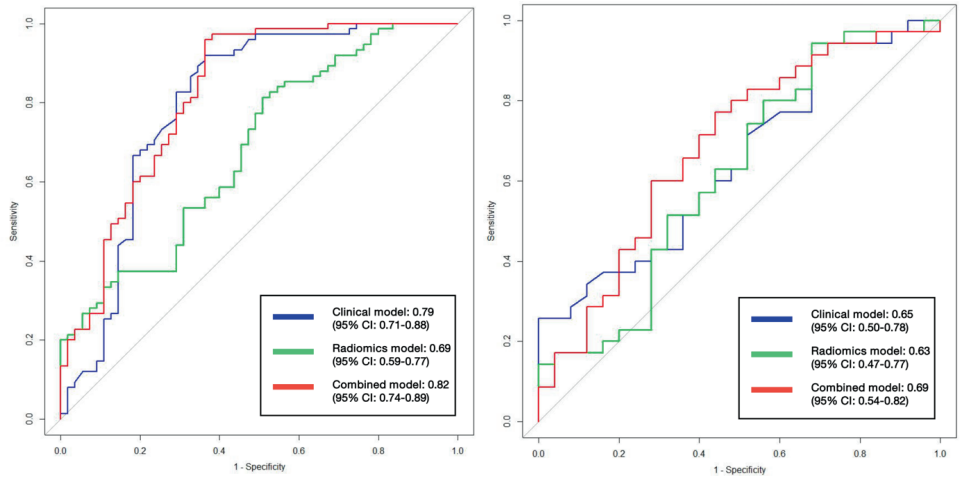
Dead	53 (40.8%)	32 (53.3%)	
Radiomics score, mean ± SD	0.49 ± 1.80	0.72 ± 2.78	#P =0.56

NCT neo-adjuvant chemotherapy; NCRT neo-adjuvant chemoradiotherapy; $ chi-square test; # t-test; + fisher test

Four clinical features; age, clinical T-stage, neo-adjuvant therapy and Mandard score, were included in the multivariable model after applying RFE method optimized for AUC. These features were within the top four most-frequently selected directly through LASSO during 500 random splits of STAGE. This multivariable clinical model yielded mean AUCs of 0.79 (95% CI: 0.71-0.88)) in STAGE and 0.65 (95% CI: 0.50-0.78) in CROSS. In the same cohorts, a cN-based model gave mean AUCs 0.60 (95% CI: 0.52-0.69) and 0.66 (95% CI: 0.55-0.78), respectively.

Nine radiomic features were selected for a radiomics-based model, but resulted in lower mean AUCs of 0.69 (95% CI: 0.59-0.77) in STAGE and 0.63 (95% CI: 0.47-0.77) in CROSS. A combined clinical and radiomics-based model yielded mean AUCs of 0.82 (95% CI: 0.74-0.89) and 0.69 (95% CI: 0.54-0.82) in these cohorts, respectively. In validation, there was no statistically significant difference in AUC performance across the three models. **Figure 3** gives the ROC plots of the above models with their respective mean AUCs. Results of AUC, accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are reported in **Table 2**. cN-stage results for each cohort were calculated and are included in **Table 2** for comparison with each of the three models.

The calibration plots of the models in both cohorts are shown in **Supplementary Figures 1**. The Hosmer-Lemeshow test indicated that the combined model was well calibrated in for development (p=0.11) and validation (p=0.47), although calibration was suboptimal for clinical only (p=0.02) and radiomics only models (p=0.01) in the development cohort.



(a) (b)

Figure 3: ROC plots of clinical model (blue line), radiomics-based model (green line), and combined model (red line) in the (a) development and (b) external validation cohorts. The results show the combined model achieved the best performance with AUCs 0.82 (95% CI: 0.74-0.90) and 0.71 (95% CI: 0.59-0.83) in the development and external validation cohorts. 95% CI was computed with 2000 times bootstrapping. CI: confidence interval.

Table 2: The statistic comparison of clinical N-staging, clinical model, radiomics-based model, and combined model in the development and external validation cohorts.

	Development Cohort				External Validation Cohort			
	Clinical N-stage	**Clinical model**	**Radiomics model**	**Combined model**	**Clinical N-stage**	**Clinical model**	**Radiomics model**	**Combined model**
Incidence	58%	58%	58%	58%	58%	58%	58%	58%
AUC	0.60 (95% CI: 0.52-0.69)	0.79 (95% CI: 0.71-0.88)	0.69 (95% CI: 0.59-0.77)	0.82 (95% CI: 0.74-0.89)	0.66 (95% CI: 0.55-0.78)	0.65 (95% CI: 0.50-0.78)	0.63 (95% CI: 0.47-0.77)	0.69 (95% CI: 0.54-0.82)
Accuracy	0.61 (95% CI: 0.52-0.69)	0.79 (95% CI: 0.70-0.85)	0.66 (95% CI: 0.57-0.74)	0.76 (95% CI: 0.71- 0.79)	0.70 (95% CI: 0.57-0.81)	0.57 (95% CI: 0.43-0.69)	0.65 (95% CI: 0.52-0.75)	0.65 (95% CI: 0.51-0.76)
Sensitivity	0.63 (95% CI: 0.60-0.65)	0.88 (95% CI: 0.83-0.90)	0.77 (95% CI: 0.83-0.90)	0.81 (95% CI: 0.79-0.92)	0.89 (95% CI: 0.86-0.90)	0.52 (95% CI: 0.44-0.59)	0.74 (95% CI: 0.65-0.77)	0.63 (95% CI: 0.56-0.67)
Specificity	0.58 (95% CI: 0.57-0.60)	0.65 (95% CI: 0.62-0.69)	0.49 (95% CI: 0.47-0.50)	0.66 (95% CI: 0.62-0.69)	0.44 (95% CI: 0.41-0.46)	0.64 (95% CI: 0.44-0.69)	0.48 (95% CI: 0.46-0.54)	0.64 (95% CI: 0.59-0.70)
PPV	0.67 (95% CI: 0.64-0.69)	0.78 (95% CI: 0.74-0.79)	0.68 (95% CI: 0.64-0.71)	0.76 (95% CI: 0.72-0.78)	0.69 (95% CI: 0.67-0.73)	0.667 (95% CI: 0.56-0.69)	0.67 (95% CI: 0.63-0.70)	0.71 (95% CI: 0.69-0.74)
NPV	0.53 (95% CI: 0.51-0.57)	0.80 (95% CI: 0.76-0.82)	0.61 (95% CI: 0.56-0.70)	0.72 (95% CI: 0.70-0.90)	0.73 (95% CI: 0.70-0.756)	0.49 (95% CI: 0.43-0.59)	0.57 (95% CI: 0.52-0.62)	0.55 (95% CI: 0.53-0.66)

AUC area under curve; CI confidence interval; PPV positive predictive value; NPV negative predictive value.

In univariate analysis, the AUC for MTV to predict LNMs was 0.58 (95% CI: 0.48-0.66) and 0.60 (95% CI: 0.47-0.72) in STAGE and CROSS cohorts, respectively. For TLG, the AUC was 0.58 (95% CI: 0.48-0.66) and 0.58 (95% CI: 0.45-0.71), respectively.

Results of survival analysis are tabulated in **Supplementary Table 1** and **2 (ST-1, 2)** and Kaplan-Meier curves are given in **Supplementary Figure 2, 3, 4 (SF-2, 3, 4)**. In both STAGE and CROSS cohorts, the true pathological lymph node status (X^2 13.76, df 1, $p < 0.001$, and X^2 4.36, df 1, $p = 0.04$, respectively, **ST-1&2**) was significantly associated with overall survival. In addition, the combined clinical and radiomics model was also significantly associated with overall survival in the external CROSS cohort (X^2 6.08, df 1, $p = 0.01$, **Supplementary Figure 4**) and performed better than the other developed models. Finally, **Supplementary Figure 5, 6, 7 (SF-5, 6, 7)** show the performance of the three types of models in the development cohort according to the subgroups divided by treatment types.

Discussion

In this study, we developed and externally validated three prediction models; a model using clinical variables only, PET radiomics only and a combined clinical-radiomics model. A combined clinical and radiomics model developed in the STAGE cohort showed potentially

improved diagnostic accuracy compared with current cN-stage results but this was not replicated in the external validation CROSS cohort.

In terms of prognostic significance, a combined clinical-radiomics model demonstrated good discrimination between patient groups in the external cohort but this was not the case in the development cohort. The external cohort included patients recruited into the CROSS trial [26], in which the pathological stage following NACRT (ypTN+ stage) was significantly associated with overall survival. There were significant differences in cN-stage status between cohorts, with a higher proportion of patients having cN+ disease in the external CROSS cohort. This was reflected in the sensitivity and specificity results obtained. The sensitivity was reduced in the STAGE compared to CROSS cohorts with the opposite true for specificity, indicating that radiologists were more likely to 'under-stage' disease in STAGE compared to radiologists in CROSS. [7] The variability in staging classification maybe explained by reporting practice differences between the two countries.

Failure to replicate models is a common finding in external validation studies. The lack of full validation may be attributable to the relatively small sample size of the external validation cohort, inter-scanner differences such as varying slice thickness, voxel size and acquisition parameters, and inter-patient differences such as time from injection to imaging. However, PET radiomics have been shown to have potential clinical value in EC when incorporated into a prognostic model [14].

Initially, the results showed that PET radiomics derived from the primary tumor volume may have added predictive value for LNM detection, but this effect was not replicated. Common concerns are that firstly, clinical PET images have a spatial resolution too large for radiomic analysis and secondly, higher order radiomic variables are surrogates of simple MTV [35]. However, simpler PET metrics such as MTV and TLG were excluded as potential confounders, through a detailed process of radiomic feature selection. Furthermore, MTV and TLG had no predictive value for LNMs or overall survival. Despite comprehensive clinical and radiological data, we failed to show that a radiomics signature was significantly superior to either cN-stage or the clinical multivariable model for predicting LNMs in esophageal adenocarcinoma.

Our results add evidence that current cN-staging accuracy remains poor [6, 7]. The Union for International Cancer Control (UICC) Tumor Node Metastasis (TNM) classification is largely reliant on anatomical definitions [27]. CT has sensitivity for LNMs as low as 18% [36] because it relies on morphology, and cannot differentiate between occult malignant metastases and normal-sized lymph nodes. It is now thought that exosomes are excreted by aggressive primary tumors into the bloodstream. This circulating tumor DNA (ctDNA) seeds in lymph nodes, giving rise to synchronous micro-metastases [37]. Patients with esophageal adenocarcinomas commonly present with LNMs due to lack of esophageal serosa [38]. EC staging must become more accurate and better at risk-stratifying patients [36].

Commonly, treatment decisions often hinge on the accurate diagnosis of a lymph node. (**Figure 4**) For example, equivocal lymph nodes located away from the primary tumor may be considered un-resectable or be outside of the maximum radiotherapy field possible [39]. Often,

tissue confirmation is attempted with EUS fine needle aspiration (EUS-FNA) but on occasions where the aspirate is normal or insufficient, concerns over under-sampling exist. Additional predictive information would add confidence to this important treatment decision. Similarly, improved diagnostic accuracy of non-regional lymph nodes in the abdomen that are inaccessible by FNA or core biopsy may prevent a harmful major resection that is unlikely to yield long-term survival gain. Finally, in the case of T1-T2 N0 tumors, the decision to proceed directly to surgery is standard practice. However, the risk of pLNMs is 45–75% for T2 tumors and 80–85% for >= T3 tumors [40]. Administration of neo-adjuvant therapy would be preferable in these scenarios. Non-invasive imaging biomarkers that suggest that the primary tumor has high metastatic potential would guide the clinical decision towards neo-adjuvant treatment prior to surgery. Further research that focusses on this disease stage group is warranted but would require a large sample size to adequately power such a study.

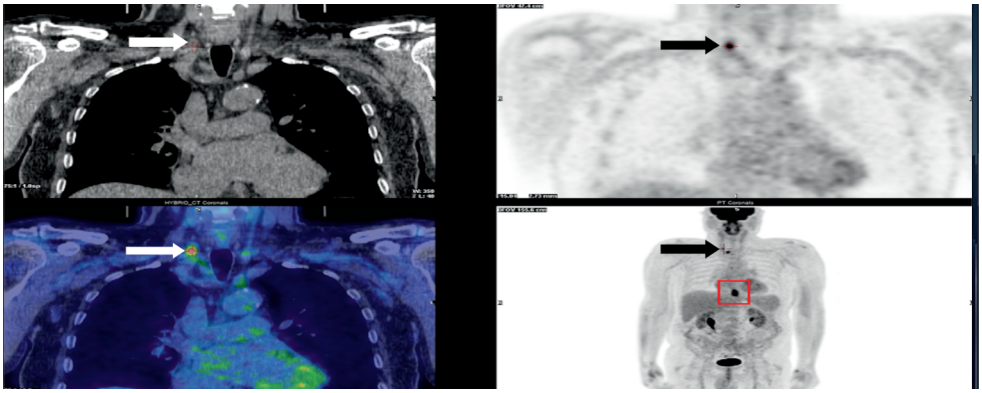


Figure 4: Clockwise from top left; a non-contrast CT, a magnified maximum intensity projection (MIP), a whole body MIP and fused PET/CT image of a patient with a distal esophageal tumor (Red box) and two LNMs; one supra-clavicular and one high right para-esophageal (black and white arrows).

Strengths of study

External validation studies are rarely performed, particularly in the field of radiomics. To eliminate the inter-observer variability of delineation, we used a standardized auto-segmentation approach (ATLAAS) to outline the tumor on PET images. Standardized biomarkers (IBSI) were also used. These reproducible methods allow further validation in different centers. Furthermore, a reproducible radiomic feature selection method was employed. In principle, as fewer important features are used in the model, the chance of model overfitting reduces. A common criticism of radiomics studies are that large numbers of predictor variables are used for a relatively small cohort size [41]. Therefore, we performed a relatively strict feature reduction approach to maximize the event per variable ratio.

Limitations

The main limitation was the lack of PET radiomics following neo-adjuvant therapy. As discussed, PET/CT is not repeated prior to surgery in many countries due to limited evidence about its cost-effectiveness, therefore quantifying the change in radiomics over time is not

possible. As a result, lymph node response to neo-adjuvant treatment on PET cannot be assessed in this study and subsequently there is indirect comparison between baseline radiological features and the final pathological lymph node evaluation following surgery. However, 63% and 40% of patients had a TRG of 4 or 5 in the STAGE and CROSS cohorts, respectively, indicating that a substantial proportion of patient had little or no response to neo-adjuvant therapy. This provides some reassurance that an indirect comparison provides some meaningful data. The differences in TRG rates can be explained by the differences in treatment between the two cohorts, with CROSS patients receiving NCRT but the majority of STAGE receiving NCT. Secondly, only primary tumors were analyzed. Integrated radiomics analysis of the primary tumor and individual lymph nodes may potentially provide more prediction information [42] but this process is more time-consuming and unlikely to be adopted into busy clinical practice. In addition, only FDG-avid adenocarcinomas were included in the study. Analyses of squamous cell carcinoma and non FDG-avid tumors would be equally valuable.

Conclusions

Accurate diagnosis of LNMs is crucial for predicting prognosis and guiding treatment decisions in esophageal adenocarcinoma, but radiological cN-staging is currently suboptimal. Despite obtaining signal for improved prediction in a development cohort, this study showed that models using clinical variables and PET radiomics derived from the primary tumor were not fully replicated in an external validation cohort from an international center. New techniques for improving the diagnostic accuracy of LNMs are required. We plan to further validate and confirm these findings in larger external cohorts.

References

1. Fitzmaurice, C., et al., *Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study.* JAMA oncology, 2017. **3**(4): p. 524-548.
2. Torre, L.A., et al., *Global cancer statistics, 2012.* CA: a cancer journal for clinicians, 2015. **65**(2): p. 87-108.
3. *Cancer Research UK: Oesophageal cancer statistics*. [cited 2019 08, May]; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer#heading-Two.
4. Kayani, B., et al., *Lymph node metastases and prognosis in oesophageal carcinoma–a systematic review.* European Journal of Surgical Oncology (EJSO), 2011. **37**(9): p. 747-753.
5. Allum, W., et al., *Guidelines for the management of oesophageal and gastric cancer.* Gut, 2002. **50**(suppl 5): p. v1-v23.
6. Foley, K., et al., *Accuracy of contemporary oesophageal cancer lymph node staging with radiological-pathological correlation.* Clinical radiology, 2017. **72**(8): p. 693. e1-693. e7.
7. Bunting, D., et al., *Loco-regional staging accuracy in oesophageal cancer—How good are we in the modern era?* European journal of radiology, 2017. **97**: p. 71-75.
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, 2016. **34**(18): p. 2157-2164.
9. Wu, S., et al., *A radiomics nomogram for the preoperative prediction of lymph node metastasis in bladder cancer.* Clinical Cancer Research, 2017. **23**(22): p. 6904-6911.
10. Shen, C., et al., *Building CT radiomics based nomogram for preoperative esophageal cancer patients lymph node metastasis prediction.* 2018. **11**(3): p. 815-824.
11. Tan, X., et al., *Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma.* European Radiology, 2019. **29**(1): p. 392-400.
12. Shen, C., et al., *Building CT radiomics based nomogram for preoperative esophageal cancer patients lymph node metastasis prediction.* Translational oncology, 2018. **11**(3): p. 815-824.
13. Qu, J., et al., *The MR radiomic signature can predict preoperative lymph node metastasis in patients with esophageal cancer.* European radiology, 2019. **29**(2): p. 906-914.
14. 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, 2018. **28**(1): p. 428-436.
15. van Rossum, P.S., et al., *The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer.* Journal of Nuclear Medicine, 2016. **57**(5): p. 691-700.

16. Dong, X., et al., *Three-dimensional positron emission tomography image texture analysis of esophageal squamous cell carcinoma: relationship between tumor 18F-fluorodeoxyglucose uptake heterogeneity, maximum standardized uptake value, and tumor stage.* Nuclear medicine communications, 2013. **34**(1): p. 40-46.
17. Walker, R.C. and T.J. Underwood, *Molecular pathways in the development and treatment of oesophageal cancer.* Best Practice & Research Clinical Gastroenterology, 2018.
18. Wu, W., et al., *Exploratory study to identify radiomics classifiers for lung cancer histology.* Frontiers in oncology, 2016. **6**: p. 71.
19. Hatt, M., et al., *Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211.* Medical physics, 2017. **44**(6): p. e1-e42.
20. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* BMC medicine, 2015. **13**(1): p. 1.
21. Foley, K.G., et al., *Prognostic significance of novel 18F-FDG PET/CT defined tumour variables in patients with oesophageal cancer.* European journal of radiology, 2014. **83**(7): p. 1069-1073.
22. Schreurs, L.M., et al., *Value of EUS in determining curative resectability in reference to CT and FDG-PET: the optimal sequence in preoperative staging of esophageal cancer?* Annals of surgical oncology, 2016. **23**(5): p. 1021-1028.
23. Foley, K.G., et al., *External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer.* Radiotherapy and Oncology, 2019. **133**: p. 205-212.
24. Noordman, B.J., et al., *Effect of neoadjuvant chemoradiotherapy on health-related quality of life in esophageal or junctional cancer: results from the randomized CROSS trial.* Journal of Clinical Oncology, 2018. **36**(3): p. 268-275.
25. Shapiro, J., et al., *Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial.* The lancet oncology, 2015. **16**(9): p. 1090-1098.
26. van Hagen, P., et al., *Preoperative chemoradiotherapy for esophageal or junctional cancer.* New England Journal of Medicine, 2012. **366**(22): p. 2074-2084.
27. Sobin, L.H., M.K. Gospodarowicz, and C. Wittekind, *TNM classification of malignant tumours*. 2011: John Wiley & Sons.
28. Mandard, A.M., et al., *Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations.* Cancer, 1994. **73**(11): p. 2680-2686.
29. Berthon, B., et al., *ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography.* Physics in Medicine & Biology, 2016. **61**(13): p. 4855.
30. Apte, A.P., et al., *Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research.* Medical physics, 2018. **45(8), 3713-3720.**

31. Leijenaar, R.T., et al., *The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis.* Scientific reports, 2015. **5**: p. 11075.
32. Whybra, P., et al., *Assessing radiomic feature robustness to interpolation in 18 F-FDG PET imaging.* Scientific reports, 2019. **9**(1): p. 9649.
33. Zwanenburg, A., et al., *Image biomarker standardisation initiative-feature definitions.* Preprint at arXiv: https://arxiv.org/abs/1612.07003, 2016.
34. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
35. Cook, G.J., et al., *Radiomics in PET: principles and applications.* Clinical and Translational Imaging, 2014. **2**(3): p. 269-276.
36. Choi, J.Y., et al., *Improved detection of individual nodal involvement in squamous cell carcinoma of the esophagus by FDG PET.* Journal of Nuclear Medicine, 2000. **41**(5): p. 808-815.
37. Becker, A., et al., *Extracellular vesicles in cancer: cell-to-cell mediators of metastasis.* Cancer cell, 2016. **30**(6): p. 836-848.
38. Rustgi, A.K., *Gastrointestinal cancers: a companion to Sleisenger and Fordtran's Gastrointestinal and Liver disease*. 2003: WB Saunders.
39. Foley, K., et al., *Impact of positron emission tomography and endoscopic ultrasound length of disease difference on treatment planning in patients with oesophageal cancer.* Clinical Oncology, 2017. **29**(11): p. 760-766.
40. Hölscher, A.H., et al., *Prognostic factors of resected adenocarcinoma of the esophagus.* Surgery, 1995. **118**(5): p. 845-855.
41. Chalkidou, A., M.J. O'Doherty, and P.K. Marsden, *False discovery rates in PET and CT studies with texture features: a systematic review.* PloS one, 2015. **10**(5): p. e0124165.
42. Coroller, T.P., et al., *Radiomic-based pathological response prediction from primary tumors and lymph nodes in NSCLC.* Journal of Thoracic Oncology, 2017. **12**(3): p. 467-476.

Supplementary: External Validation of a Prediction Model Using PET Radiomics to Improve the Diagnostic Accuracy of Lymph Node Metastases in Esophageal Adenocarcinoma

Image acquisition (STAGE cohort)

The STAGE imaging protocol has previously been published in Foley et al. [85]. Patients were fasted for at least 6 hours prior to tracer administration. Serum glucose levels were routinely checked and confirmed as less than 7.0 mmol/L prior to imaging. Patients received an ^{18}F-FDG dose of 4 MBq/kg. PET/CT imaging was performed after 90 minutes uptake time with a GE 690 PET/CT scanner (GE Healthcare, Buckinghamshire, UK) [85]. PET images were acquired at 3 minutes per field of view. The length of the axial field of view was 15.7 cm. Images were reconstructed with the ordered subset expectation maximization algorithm, with 24 subsets and 2 iterations. Matrix size was 256 x 256 pixels, using the VUE Point ™ time of flight algorithm. Computed tomography (CT) images were acquired in a helical acquisition with a pitch of 0.98 and a tube rotation speed of 0.5 seconds. Tube output was 120 kVp with output modulation between 20 and 200 mA. Matrix size for the CT acquisition was 512 x 512 pixels with a 50cm field of view. No oral or intravenous contrast was administered.

Image acquisition (CROSS cohort)

For NCRT radiotherapy dosimetry, CT scans were performed on either Sensation Open or Biograph 40 (Siemens Healthcare GmbH, Erlangen, Germany) clinical scanners. The imaging protocol parameters were: 120kVp, 500 mAs, 500 mm field of view (FOV), 512x512 grid, reconstructed slice thickness and axial pixel spacing of 3mm and 0.98 mm, respectively. In 3 subjects, a wider FOV (685 – 820 mm) was used to encompass the body outline required for radiotherapy dosimetry; leading to larger axial pixel spacing (1.3 – 1.6 mm). For 3 other subjects, slices had been reconstructed at 6, 9 and 12 mm. Hounsfield values had been calibrated in the scanner, such that pure water was zero.

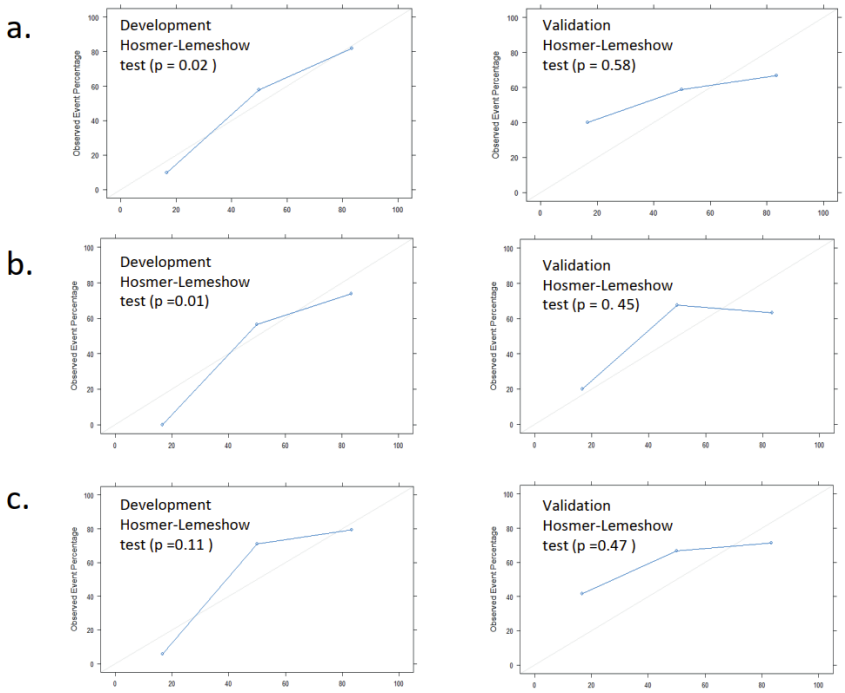
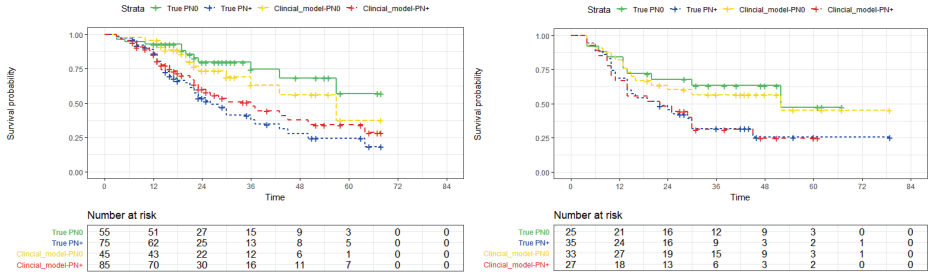
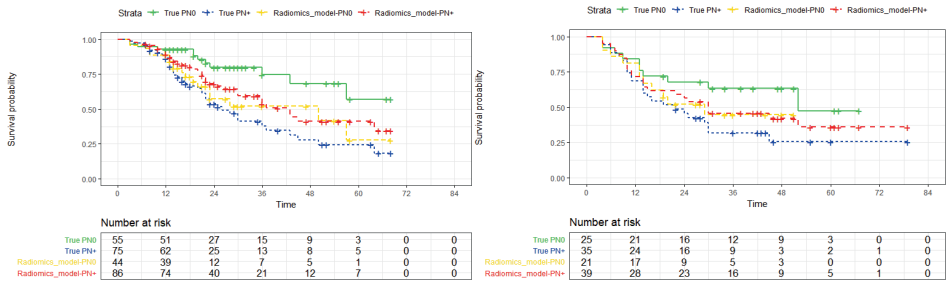


Figure 1: The calibration curves of the three models in the development and validation cohorts. (a) clinical variables model only; (b) radiomics model only; c) combined clinical and radiomics model



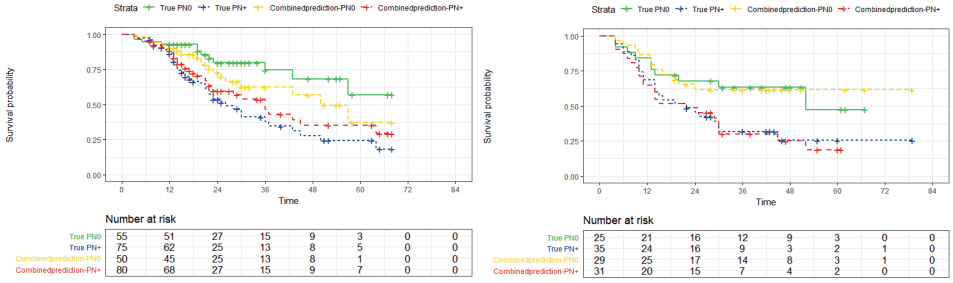
(a) (b)

Figure 2: Kaplan–Meier curves and number of patients at risk for (a) lymph node status predictions by clinical model, (X^2 3.35, df 1, p = 0.07) with a c-index of 0.58 (SE: 0.033) vs pathological lymph node status (X^2 13.41, df 1, p < 0.001) with a c-index of 0.63 (SE: 0.036) in the development cohort; (b) lymph node status predictions by clinical model (X^2 3.41, df 1, p = 0.06) with a c-index of 0.57 (SE: 0.045) vs pathological lymph node status (X^2 4.36, df 1, p = 0.04) with a c-index of 0.58 (SE: 0.046) in the external validation cohort. Survival times are in months. c-index: concordance index; SE: standard error; df: degree of freedom.



(a) (b)

Figure 3: Kaplan–Meier curves and number of patients at risk for (a) lymph node status predictions by radiomics-based model (X^2 0.52, df 1, p = 0.50) with a c-index of 0.52 (SE: 0.037) vs pathological lymph node status (X^2 13.41, df 1, p < 0.001) with a c-index of 0.63 (SE: 0.036) in the development cohort; (b) lymph node status predictions by clinical model, (X^2 0.01, df 1, p = 0.90) with a c-index of 0.51 (SE: 0.044) vs pathological lymph node status (X^2 4.36, df 1, p = 0.04) with a c-index of 0.58 (SE: 0.046) in the external validation cohort. Survival times are in months. c-index: concordance index; SE: standard error; df: degree of freedom.



(a) (b)

Figure 4: Kaplan–Meier curves and number of patients at risk for (a) lymph node status predictions by combined model (X^2 1.69, df 1, p = 0.20) with a c-index of 0.55 (SE: 0.037) vs pathological lymph node status (X^2 13.41, df 1, p < 0.001) with a c-index of 0.63 (SE: 0.036) in the development cohort; (b) lymph node status predictions by clinical model, (X^2 6.08, df 1, p = 0.01) with a c-index of 0.60 (SE: 0.045) vs pathological lymph node status (X^2 4.36, df 1, p = 0.04) with a c-index of 0.58 (SE: 0.046) in the external validation cohort. Survival times are in months. c-index: concordance index; SE: standard error; df: degree of freedom.

Table 2: Univariable proportional hazards regression analysis using pLNMs status, and the prediction of pLNMs by the three models in the development cohort.

	Coefficient	C-index (SE)	Log-rank test	Hazard Ratio (HR)
True pLNMs status	1.156	0.63 (0.036)	X^2 13.76, df 1, p < 0.001	3.18
Clinical model	0.564	0.58 (0.033)	X^2 3.35, df 1, p = 0.07	1.76
Radiomics-based model	-0.140	0.52 (0.037)	X^2 0.52, df 1, p = 0.50	0.81
Combined model	0.38	0.55 (0.037)	X^2 1.69, df 1, p = 0.20	1.46

pLNMs pathological lymph node metastasis; SE standard error; df degree of freedom

Table 2: Univariable proportional hazards regression analysis using pLNMs status, and the prediction of pLNMs by the three models in the external validation cohort.

	Coefficient	C-index (SE)	Log-rank test	Hazard Ratio (HR)

True pLNMs status	0.773	0.58 (0.046)	X^2 4.36, df 1, p = 0.04	2.2
Clinical model	0.630	0.57 (0.045)	X^2 3.41, df 1, p = 0.06	1.9
Radiomics-based model	-0.029	0.51 (0.044)	X^2 0.01, df 1, p = 0.90	1.0
Combined model	0.878	0.60 (0.045)	X^2 6.08, df 1, p = 0.01	2.4

pLNMs pathological lymph node metastasis; SE standard error; df degree of freedom

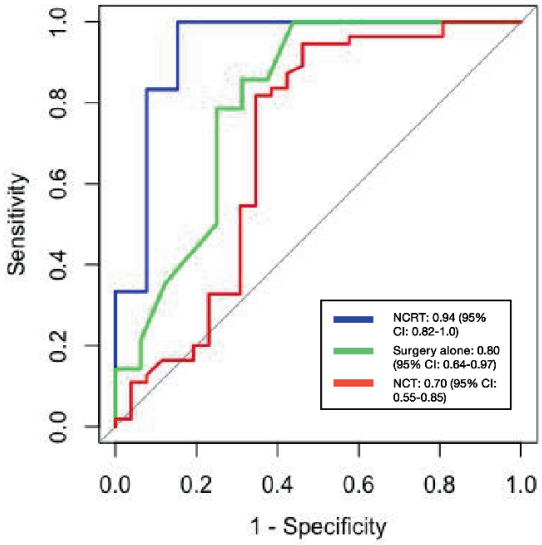


Figure 5: ROC plots of clinical model in three sub-treatment groups in the development cohort. Neo-adjuvant chemo-radiotherapy (NCRT, blue), surgery alone (green), and neo-adjuvant chemo-therapy (NCT).

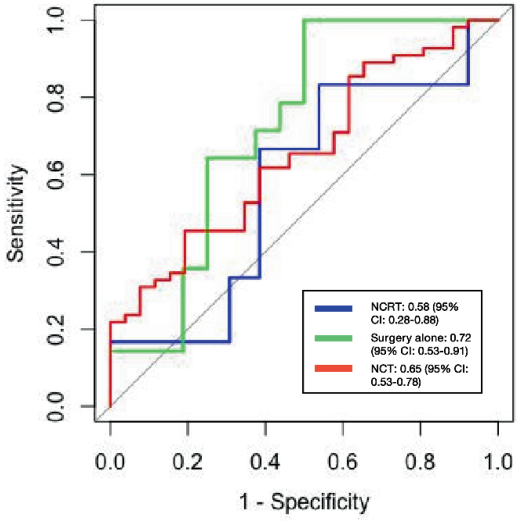


Figure 6: ROC plots of radiomics model in three sub-treatment groups in the development cohort. Neo-adjuvant chemo-radiotherapy (NCRT, blue), surgery alone (green), and neo-adjuvant chemo-therapy (NCT).

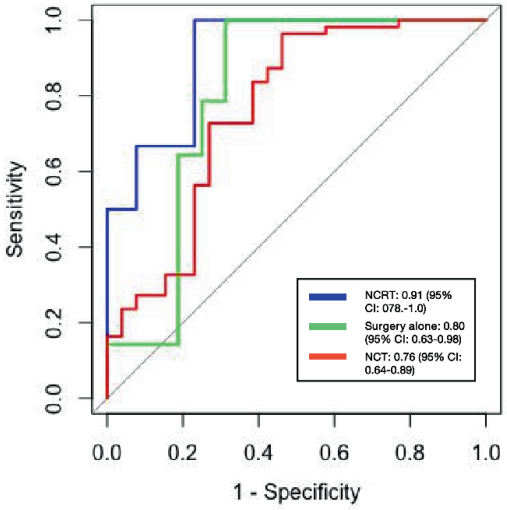


Figure 7: ROC plots of the combined model in three sub-treatment groups in the development cohort. Neo-adjuvant chemo-radiotherapy (NCRT, blue), surgery alone (green), and neo-adjuvant chemo-therapy (NCT).

Reference

1. Foley, K.G., et al., *Prognostic significance of novel 18F-FDG PET/CT defined tumour variables in patients with oesophageal cancer.* European journal of radiology, 2014. **83**(7): p. 1069-1073.

Chapter 6

External Validation of a Prognostic Model Incorporating Quantitative PET Image Features in Esophageal Cancer

Kieran G. Foley, Zhenwei Shi, Philip Whybra, Petros Kalendralis, Ruben Larue, Maaike Berbee, Meindert N. Sosef, Craig Parkinson, John Staffurth, Tom D.L. Crosby, Stuart Ashley Roberts, Andre Dekker, Leonard Wee, Emiliano Spezi

Underscore indicates equal contribution

Adapted from:
Foley, K. G., Shi, Z., Whybra, P., Kalendralis, P., Larue, R., Berbee, M., ... & Roberts, S. A. (2019). External validation of a prognostic model incorporating quantitative PET image features in esophageal cancer. Radiotherapy and Oncology, 133, 205-212.
DOI: https://doi.org/10.1016/j.radonc.2018.10.033

Abstract

Aim: Enhanced prognostic models are required to improve risk stratification of patients with esophageal cancer so treatment decisions can be optimized. The primary aim was to externally validate a published prognostic model incorporating PET image features. Transferability of the model was compared using only clinical variables.

Methods: This was a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 study. The model was validated against patients treated with neoadjuvant chemoradiotherapy according to the Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for esophageal or junctional cancer (CROSS) trial regimen using pre- and post-harmonized image features. The Kaplan-Meier method with log-rank significance tests assessed risk strata discrimination. A Cox proportional hazards model assessed model calibration. Primary outcome was overall survival (OS).

Results: Between 2010 and 2015, 449 patients were included in the development (n=302), internal validation (n=101) and external validation (n=46) cohorts. No statistically significant difference in OS between patient quartiles was demonstrated in prognostic models incorporating PET image features (X^2=1.42, df=3, p=0.70) or exclusively clinical variables (age, disease stage and treatment; X^2=1.19, df=3, p=0.75). The calibration slope β of both models was not significantly different from unity (p=0.29 and 0.29, respectively). Risk groups defined using only clinical variables suggested differences in OS, although these were not statistically significant (X^2=0.71, df=2, p=0.70).

Conclusion: The prognostic model did not enable significant discrimination between the validation risk groups, but a second model with exclusively clinical variables suggested some transferable prognostic ability. PET harmonization did not significantly change the results of model validation.

Highlights

- PET image features have shown additional prognostic value in esophageal cancer
- Harmonization of PET images to standardize slice thickness is possible
- The prognostic model did not enable discrimination between the external risk groups
- A second model suggested transferable prognostic ability between cohorts

Introduction

The prognosis of patients with esophageal cancer is poor with overall 5-year survival approximately 15% [1]. Esophageal cancer is the eighth most common malignancy worldwide, accounting for around 400,000 deaths each year [2].

Treatment strategies of patient with esophageal cancer are currently informed by radiological staging. Accurate staging is vital to inform clinicians of the likely prognosis of each patient and to appropriately risk stratify patients, ensuring the best individual management plan is decided upon. However, the failure of survival rates to increase significantly in recent decades suggests that staging accuracy, treatment selection and prognosis could be improved further. For example, lymph node metastases (LNMs) are one of the major prognostic indicators in esophageal cancer, but there is evidence that regional lymph node staging (N-stage) is presently suboptimal [3, 4]. Therefore, enhanced staging methods are required to improve prognostication and subsequent risk stratification of patients.

Esophageal cancer is typically confirmed by a small-sample biopsy taken during endoscopic examination. Despite advances in genomics, no molecular prognostic markers are currently in routine clinical use [5]. It has been proposed that additional tumor phenotype information may be derived by quantitative analysis of Positron Emission Tomography (PET) scans. [6] "Radiomics" broadly refers to automated, computerized and high-throughput extraction of quantitative image markers (features) from a large corpus of radiological images. [7] Radiomic features typically include histogram metrics (e.g. mean and maximum), shape descriptors (e.g. longest axis length and compactness) and textures (e.g. continuous length of voxels with similar intensities) [8]. These features can be sensitive to differences in image parameters such as slice thickness [9]. Post-reconstruction harmonization methods have been proposed to adjust for these differences, thus promoting standardized research between centers [10].

The primary aim of this study was to externally validate the results of a UK esophageal cancer prognostic model incorporating radiomic features [11] firstly pre-harmonization, then post-harmonization, in a cohort of esophageal cancer patients treated with neo-adjuvant chemoradiotherapy (NACRT) according to the Dutch Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for esophageal or junctional cancer (CROSS) trial regimen [12]. An enhanced prognostic model incorporating radiomic features of primary tumors may provide clinicians with complimentary data to traditional prognostic factors that will assist treatment decision making and risk stratification [11, 13]. The secondary aim was to compare prognostic models with and without PET image features between cohorts to provide further validation.

Materials & Methods

This study was designed as a Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD) type 3 external independent validation study [14]. A previously published prognostic model had been developed and internally validated in patients with esophageal cancer. Details of model development have been provided in Foley et

al [11]. Briefly, the prognostic model had only been evaluated by same-center internal validation in patients managed by the South-East Wales Regional Upper Gastrointestinal (GI) Cancer Multi-Disciplinary Team (MDT), United Kingdom. A suitable independent cohort was not accessible at the time of publication. Institutional board review (IRB) approval was granted for the development of the prognostic model (REF 14/WA/1208). The prognostic model was developed as part of a larger study investigating the prognostic significance of image texture analysis in gastro-esophageal cancer (STAGE), and from here-on will be known as the STAGE cohort. The external validation cohort comprised patients treated with the CROSS regimen in The Netherlands. IRB permission was obtained for the external validation cohort.

Patient cohorts

In total, 449 patients were included in the development and validation of this prognostic model. **Figure 1** details the number of patients in each cohort and the reasons for exclusion of patients from the CROSS validation cohort. The largest number of patient exclusions (n=23) from the CROSS cohort were because of the pre-defined metabolic tumor volumes (MTV) adopted in Foley et al [11] and used in this current study for consistency. Other main reasons for patient exclusion were different calibration units (n=11) and ATLAAS segmentation failure (n=7).

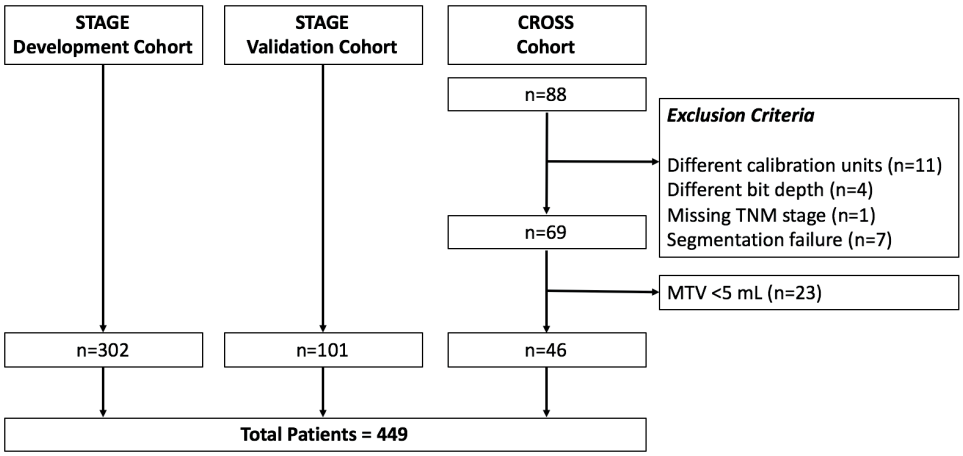


Figure 1: Study flowchart describing the numbers of patients in each cohort and reasons for exclusions from the CROSS cohort.

Primary Outcome

The primary endpoint of the published prognostic model is overall survival, defined as the number of months survived after the date of diagnosis until death or last day of follow-up. Dates of death were obtained from the Cancer National Information System Cymru (CaNISC) database (Velindre NHS Trust, Wales), reported by the Office for National Statistics. Dates of death of patients in the CROSS cohort were obtained from the national registry. In both cohorts, local researchers were not blinded to the dates of death. A uniform and standardized procedure

for autosegmentation and radiomics computation was implemented at each center to ensure consistent methodology.

Tumor Segmentation

Primary tumors were segmented on PET images using an automatic tree-based learning algorithm for advanced segmentation (ATLAAS) [15]. The benefit of ATLAAS is that inter-observer variability in contouring is eliminated. Full details regarding the use of ATLAAS in this study are provided in Foley et al. and Berthon et al. [11, 15].

The following model equation (Eq. 1) was used to calculate a prognostic score for each patient. This equation was derived using published methods [16].

*Prognostic score = Stage Group*0.397 - Treatment*1.094 + Age*0.024 - log(Histogram Energy)*1.320 + log(TLG)*1.748 + Histogram Kurtosis*0.198* *Eq. 1*

External Validation

The ATLAAS code and equations to calculate each of the PET image features were shared between institutions. The primary tumors on the PET scans of the CROSS patients were then segmented using ATLAAS and the MTVs produced were visually assessed for adequacy for quality control. Validation was firstly performed with pre-harmonization metrics and then repeated with post-harmonization PET features to adjust for potential differences between scanners. Fully anonymized data was then shared between institutions.

Different PET/CT scanners and protocols were used across the cohorts (Appendix A). Radiomic features are known to change significantly as a function of scanner model, image acquisition or reconstruction settings, therefore we explored using the post- reconstruction Combat harmonization method to harmonize features extracted from images acquired across different scanners. Slice thickness was chosen for harmonization because images from one scanner had different thickness values, which resulted in 5 categories (Appendix A, Table A.1). Further details of the cohorts, treatments received, PET/CT protocols, metric equations, variation in image features and the post-reconstruction PET harmonization Combat method, used to adjust for batch effects across different datasets, have been provided in Appendix A.

Statistical analysis

Categorical data are described as frequency (percent) and continuous variables as median (range) and differences assessed with appropriate non-parametric tests. There was no missing data in the development cohort and cases with missing data were excluded from the validation CROSS cohort. Patient characteristics at staging were compared for each cohort. Boxplots were generated locally on each cohort to compare the distributions of the model variables. Firstly, the published model was applied to 46 suitable patients in the CROSS cohort prior to PET harmonization. A second model validation was then performed using image features calculated post-harmonization. Model discrimination was evaluated using the log-rank test; a p-value of

<0.05 was defined as statistically significant. Model calibration followed a standard test procedure detailed in [17], and which has been previously implemented in [18]. In this study, we define model discrimination as preserved if the p-value of the calibration slope $\beta = 1$ is >0.05. Thirdly, we performed the same validation steps for a prognostic model developed on the same STAGE cohort, but exclusively using clinical variables (age at diagnosis, stage and treatment) and no imaging-based variables. Statistical analysis was performed with SPSS version 23.0 (IBM, Chicago, USA) and MATLAB version 9.0 (MathWorks, Natick, MA).

Results

The baseline characteristics of the STAGE development, validation and CROSS cohorts are detailed in **Table 1**. The median overall survival of the CROSS cohort was 25 months (95% confidence interval (CI) 23.0 to 31.4). The median overall survival of the STAGE development and validation cohorts was 16.0 months (95% CI 13.8-18.2) and 14.0 months (95% CI 10.4-17.6), respectively.

Table 1. Baseline Characteristics of Patients in Development, Validation and CROSS Cohorts

Frequency (%)	**STAGE Development Cohort (n=302)**	**STAGE Validation Cohort (n=101)**	**CROSS cohort (n= 46)**	**p-value***
Median Age	67.0 years (Range 39-83)	69.0 years (Range 39-84)	64.5 years (Range 47-77.8)	0.114
Gender (M: F)	227 (75.2): 75 (24.8)	78 (77.2): 23 (22.8)	38 (82.6): 8 (17.4)	0.528
Histology				0.602
Adeno	237 (78.5)	79 (78.2)	39 (84.8)	
SCC	65 (21.5)	22 (21.8)	7 (15.2)	
Tumor Location				0.010
Esophagus	192 (63.6)	47 (46.5)	28 (60.9)	
Gastro-esophageal junction	110 (36.4)	54 (53.5)	18 (39.1)	
Stage Groups				0.018
Stage 1	17 (5.6)	2 (2.0)	2 (4.4)	
Stage 2	56 (18.5)	24 (23.8)	10 (21.7)	
Stage 3	160 (53.1)	57 (56.4)	33 (71.7)	
Stage 4	69 (22.8)	18 (17.8)	1 (2.2)	
Treatment				<0.001
Curative	158 (52.3)	50 (49.5)	46 (100)	
SA	24 (15.2)	4 (8.0)	0 (0.0)	
NACT	67 (42.4)	23 (46.0)	0 (0.0)	
NACRT	13 (8.2)	7 (14.0)	46 (100)	
dCRT	54 (34.2)	16 (32.0)	0 (0.0)	
Palliative	144 (47.7)	51 (50.5)	0 (0.0)	
Overall Survival				<0.001
Alive	70 (23.2)	43 (42.6)	20 (43.5%)	
Dead	232 (76.8)	58 (57.4)	26 (51.5%)	

SCC squamous cell carcinoma; SA surgery alone; NACT neo-adjuvant chemotherapy; NACRT neo-adjuvant chemoradiotherapy; dCRT definitive chemoradiotherapy; *chi-square test

Boxplots were constructed to compare the values of log(TLG), log(Histogram Energy) and Histogram Kurtosis in between the STAGE and CROSS cohorts. (**Figure 2**) Additional boxplots and descriptive statistics of PET feature values pre- and post-harmonization are included in Appendix B. There were similar mean values and distributions of the 3 variables between STAGE and CROSS cohorts, although a greater number of outliers were observed for Histogram Kurtosis in the STAGE cohort. This is probably due to a larger number of patients and greater range in MTV of the primary tumors included in the STAGE cohort. (Table B.1)

A prognostic model containing clinical variables only was calculated from the STAGE development cohort using identical data from the original study. Age at diagnosis (HR 1.025, 95% CI 1.011-1.040, p<0.001), stage (0.337, 0.243-0.468, p<0.001) and treatment (1.462, 1.187-1.802, p<0.001) were all independently and significantly associated with overall survival.

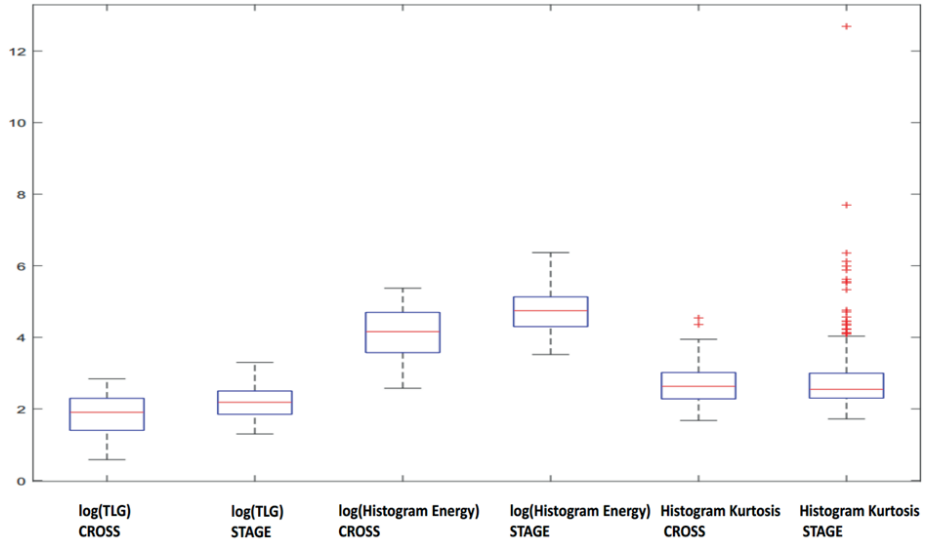


Figure 2: Boxplots displaying pre-harmonization mean values and interquartile ranges of log(TLG), log(Histogram Energy) and Histogram Kurtosis in STAGE and CROSS cohorts.

Prognostic model developed by clinical and radiomic features

Pre-harmonization

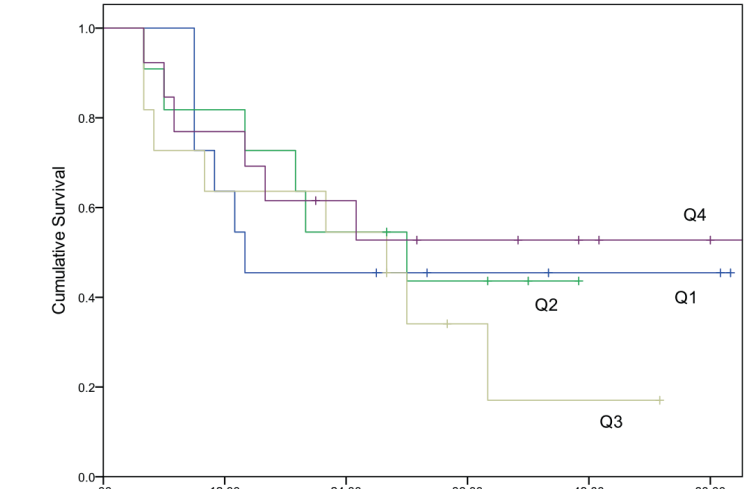
Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.27, df=3, p=0.74). (**Figure 3**) The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.89 (95% CI 0.29-2.75), 1.36 (95% CI 0.47-3.92) and 0.78 (95% CI 0.25-2.41), respectively. The calibration slope β of the prognostic

score in the CROSS cohort was 1.09 (standard error (SE) 0.41). β is not significantly different from 1 (p=0.84), which indicates that model discrimination is preserved.

The mean overall survival for patient quartiles 1-4 were 34.0 months (95% CI 19.0-49.2), 29.5 months (95% CI 19.5-39.5), 25.9 months (95% CI 14.8-37.0) and 41.2 months (95% CI 25.9-56.4), respectively. Median overall survival could not be calculated for all quartiles. The median prognostic score for quartiles 1-4 was -0.51 (n=11, range -1.14 to -0.37), -0.15 (n=11, range -0.36 to 0.01), 0.20 (n=11, range 0.04 to 0.30) and 0.48 (n=13, range 0.30 to 1.16), respectively.

Post-harmonization

Following post-reconstruction PET harmonization, repeated Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.42, df=3, p=0.70). (**Figure 3**) The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.78 (95% CI 0.24-2.55), 1.47 (95% CI 0.50-4.25) and 1.15 (95% CI 0.39-3.40), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 1.26 (standard error (SE) 0.22). β is not significantly different from 1 (p=0.29), which indicates that model discrimination is preserved. The adjusted survival data for the patient quartiles is available in Appendix B.



Number at risk						
Q1	11	7	5	3	2	2
Q2	11	9	6	4	0	0
Q3	11	7	6	2	1	0
Q4	13	10	7	5	3	1

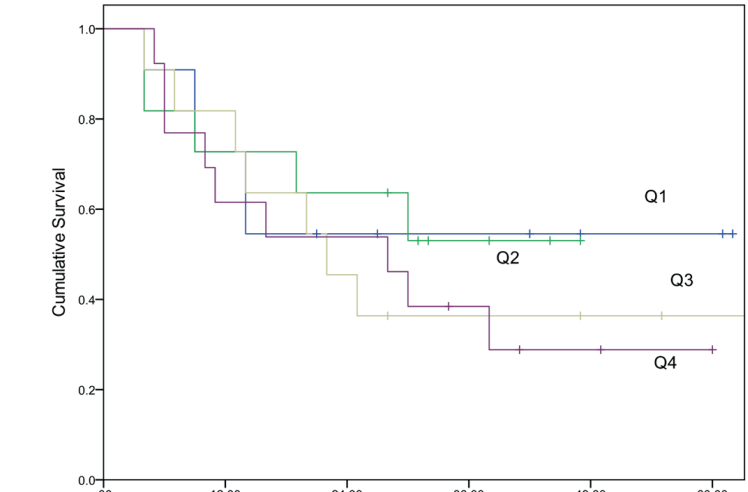
Figure 3: Cumulative survival curves of patient quartiles (Q1-4) in CROSS cohort using model developed with clinical and radiomic features (X^2=1.27, df=3, p=0.74).

These results indicate that PET harmonization did not have a substantial effect on model validation, with similar results obtained using both methods.

Prognostic model developed with clinical features only

The median prognostic score of the model developed with clinical variables only was -2.68 (range -4.89 to -0.17). As shown in **Figure 4**, Kaplan-Meier analysis did not demonstrate a significant difference in overall survival between patient quartiles in the CROSS cohort (X^2=1.19, df=3, p=0.75). The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 0.93 (95% CI 0.27-3.23), 1.41 (95% CI 0.45-4.43) and 1.53 (95% CI 0.51-4.57), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 2.15 (SE 0.72). β is not significantly different from 1 (p=0.29), which indicates that model discrimination is preserved.

In the prognostic model with clinical variables only, patients in quartiles 2 & 3 were combined to create an intermediate risk group, following a previously published method [19] (**Figure 5**) Applying Bonferroni correction, there was no statistically significance difference between the low, intermediate and high risk groups (X^2 0.712, df 2, p=0.701) but a separation in overall survival curves was observed (intermediate risk vs low risk HR 1.16 (95% CI 0.41-3.30 and high risk vs low risk HR 1.53 (95% CI 0.51-4.58)). The calibration slope β= 2.15 (SE .72, p-value 0.29) indicating model discrimination was preserved.



Number at risk						
Q1	11	8	6	6	6	6
Q2	11	8	7	6	6	6
Q3	11	9	5	4	4	4
Q4	13	8	7	5	4	4

Figure 4: Cumulative survival curves of patient quartiles (Q1-4) in CROSS cohort using model developed with clinical features only (X2=1.19, df=3, p=0.75).

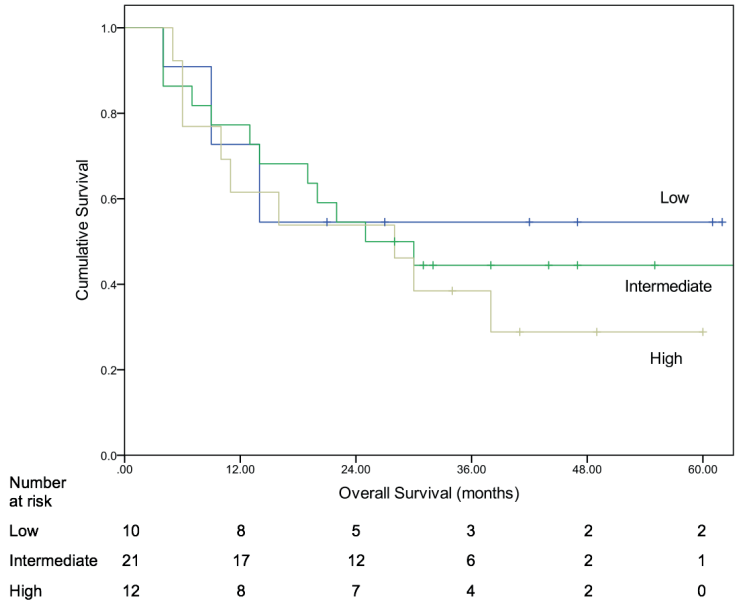


Figure 5: Cumulative survival curves of combined risk groups in CROSS cohort using model developed with clinical features only. The original quartile 1 corresponds to the low-risk group, quartiles 2 & 3 were combined to create an intermediate risk group and quartile 4 corresponds to the high-risk group.

Discussion

Patients with esophageal cancer have a poor prognosis and the incidence of the disease is increasing [20]. Despite advances in modern healthcare, survival rates remain low. Enhanced staging algorithms are required to improve the accuracy of staging, which informs clinicians of the likely prognosis and provides subsequent patient risk stratification. Prognostic models incorporating radiomic features are one strategy being investigated for this purpose.

This external validation study has shown that results of a developed prognostic model combining clinical risk factors and PET radiomic features was not replicated in a cohort of patients treated with the CROSS trial regimen. However, when a prognostic model including only clinical variables from the STAGE development cohort was tested, some aspects of the model were indicative of transferability to the CROSS cohort. Our data shows that clinical features of esophageal cancer remain prognostic across different countries and studies.

Despite not being able to replicate the validation results of the published prognostic model, this study remains clinically important because more accurate staging of esophageal cancer is essential to improve survival rates. Validated prognostic and predictive radiomics models are one strategy to improve radiological staging of esophageal cancer [21]. Greater staging accuracy will improve patient risk stratification, which is critically important for optimizing personalized treatment decision-making. Once validated, staging algorithms incorporating

radiomics may enable clinicians to decide upon the best management plan from the outset of diagnosis, therefore providing the greatest chance of survival for each patient.

A number of important methodological reasons in the modelling process may have contributed to the lack of external validity of the prognostic model when transported to the CROSS observations. First, the PET image acquisition protocols in the CROSS regimen cohort may not have been as strictly policed as in the STAGE study, leading to divergence in PET acquisition parameters. (Table A.1) All patients in STAGE (n=403) were staged using the same PET/CT scanner and protocol. However, different PET/CT scanners and protocols were used in both the STAGE and CROSS cohorts (Appendix A). Radiomic features are known to change significantly as a function of image acquisition settings such as slice thickness [9] therefore we explored using the post-reconstruction Combat harmonization method described in [10] to standardize slice thickness across different scanners. Harmonizing PET image features demonstrated little improvement in the model validity between cohorts, which supports this post-reconstruction method in external validation radiomics studies and suggests that harmonization had little influence in these cohorts. A consensus on uniformly standardized PET imaging protocols is required for multi-institutional validation of prognostic/predictive models incorporating radiomics [22].

Second, the prognostic model excluded patients with small MTV < 5 mL, thus further reducing the number of CROSS patients that were eligible for validation. The small patient numbers in the external validation cohort limits the ability to replicate the results of the STAGE prognostic model. This study is likely to be under-powered and improved validation could be achieved by increasing the cohort size. Patients with a smaller MTV were more likely to be suitable for radical therapy and therefore eligible for recruitment into the CROSS trial. In addition, evidence at the time of prognostic model development suggested possible unstable segmentation at smaller MTVs and an increase in redundant radiomic data that can be extracted [23]. There is no clear consensus on minimum MTV in PET radiomics studies. One study recommends excluding MTVs of < 45 mL, although only local entropy was evaluated in this study [24]. Other studies have previously recommend excluding patients with a primary MTV of < 10 mL [25, 26]. However, prognostic models including image features extracted from small tumor volumes can still be developed. [8] The possibility for including redundant data exists but providing the study is appropriately powered, the model can still be compared to those containing only clinical variables.

Third, the development of the previous prognostic model did not include an exhaustive radiomic feature selection steps to identify features that would be robustly reproducible within the STAGE cohort and hence more likely to be transferable to the CROSS cohort [8]. Details of the PET variables implemented in the developed prognostic model can be found in Foley et al. [11] These variables were shown to have prognostic significance in the early radiomics literature [27-29] and were implemented identically.

More studies are required to test the reliability, robustness and additional value of PET image features across a range of MTVs and between different PET/CT scanners [9, 25]. Previous

studies have found significant associations between higher order features and overall survival [28] and that the amount of complementary radiomic information gained increases with larger MTVs [25]. Despite this, the original development study did not demonstrate prognostic significance of any higher order features, although only 3 such features were investigated.

Advanced correction algorithms are being developed to harmonize features extracted from scans with different acquisition parameters, which could greatly benefit multi-center radiomic studies and reduce variation in metrics [30].

Standardization efforts such as the Internal Biomarker Standardisation Initiative (IBSI) [31] are an important methodological step towards reducing sensitivity of radiomic features to computation (image extraction) software. Deployment of the same autosegmentation tool (ATLAAS [15]) reduced inter-observer variability in contouring and the same feature extraction software that was executed locally was used in both participating centers. These techniques are examples of standardized processes that improve the robustness of radiomic features.

Lastly, a relatively small proportion of the STAGE cohort received NACRT or surgery alone (Table 1). These differences may not have been adjusted for completely by the original model multivariate regression. The STAGE cohort is relatively heterogeneous cohort of patients compared to the CROSS cohort, because it was collected during an observational cohort study recruiting all patients with esophageal cancer. Patients in the CROSS cohort were all treated with NACRT, so they share more similar characteristics. Differences between validation cohorts are important in external validation studies because the generalization of the model can be tested.

All prognostic models must be validated in an independent external cohort before being considered for use in clinical practice because many models present optimistic and over-fitted results from development cohorts [32]. However, external validation studies are rarely performed. A review of the performance of prognostic models showed that 11% are externally validated [33]. This may explain why few developed prognostic models are adopted into clinical practice [34]. Our collaborative research group is planning to update this prognostic model and perform a further external validation study with more robust feature selection and standardized feature extraction algorithms using all tumor volumes.

In conclusion, this initial TRIPOD type 3 external validation study evaluated a prognostic model developed in esophageal cancer patients staged with PET/CT. The prognostic model did not enable significant discrimination between patient risk groups in the CROSS cohort, but a second model including clinical variables only (age, disease stage and treatment) demonstrated transferable prognostic factors between international cohorts.

Acknowledgements

The authors wish to acknowledge the contributions of Professor Robert K Hills who developed the original prognostic model, Professor Wyn G Lewis who helped with data collection in the STAGE cohort, Professor Christopher Marshall (Director of the Positron-Emission Tomography Imaging Centre (PETIC) in Cardiff and members of the South-East Wales Upper GI Cancer MDT committee.

Ethical Statement
Institutional review board approval was obtained.

Data Availability
The data that has been used in this study is confidential and cannot be shared

Funding
The study was partially funding by a UK Tenovus Cancer Care Grant (TIG2016/04).

Competing interests
The authors declare that they have no competing interests.

Author contributions
KF, AR, LW and ES conceived and designed the study. RL, MB, MS, PK and TC collected the data. ZS, PW, CP and PK preformed the data analysis. KF, LW, JS, TC and AD drafted the manuscript. All authors read and approved the final manuscript.

References

1. Cancer Research UK. *Oesophageal Cancer Statistics*. 2016 November 22nd 2016]; Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer.
2. Ferlay, J., et al., *Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-386.
3. Kayani, B., et al., *Lymph node metastases and prognosis in oesophageal carcinoma-a systematic review.* Eur J Surg Oncol, 2011. **37**(9): p. 747-53.
4. Foley, K.G., et al., *Accuracy of contemporary oesophageal cancer lymph node staging with radiological-pathological correlation.* Clin Radiol, 2017. **72**(8): p. e691-e697.
5. McCormick Matthews, L.H., et al., *Systematic review and meta-analysis of immunohistochemical prognostic biomarkers in resected oesophageal adenocarcinoma.* Br J Cancer, 2015. **113**(1): p. 107-18.
6. Cook, G.J.R., et al., *Radiomics in PET: principles and applications.* Clin Transl Imaging, 2014. **2**: p. 269-276.
7. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* Eur J Cancer, 2012. **48**(4): p. 441-6.
8. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. **5**: p. 4006.
9. Desseroit, M.C., et al., *Reliability of PET/CT Shape and Heterogeneity Features in Functional and Morphologic Components of Non-Small Cell Lung Cancer Tumors: A Repeatability Analysis in a Prospective Multicenter Cohort.* J Nucl Med, 2017. **58**(3): p. 406-411.
10. Orlhac, F., et al., *A post-reconstruction harmonization method for multicenter radiomic studies in PET.* J Nucl Med, 2018.
11. 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.* Eur Radiol, 2018. **28**(1): p. 428-436.
12. van Hagen, P., et al., *Preoperative chemoradiotherapy for esophageal or junctional cancer.* N Engl J Med, 2012. **366**(22): p. 2074-2084.
13. Tan, X., et al., *Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma.* Eur Radiol, 2018.
14. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* BMJ, 2015. **350**: p. g7594.
15. Berthon, B., et al., *ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography.* Phys Med Biol, 2016. **61**(13): p. 4855-4869.
16. Moons, K.G., et al., *Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker.* Heart, 2012. **98**(9): p. 683-90.

17. Royston, P. and D.G. Altman, *External validation of a Cox prognostic model: principles and methods.* BMC Medical Research Methodology, 2013. **13**: p. 33.
18. Leijenaar, R.T., et al., *External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma.* Acta Oncol, 2015. **54**(9): p. 1423-9.
19. Dekker, A., et al., *Rapid learning in practice: a lung cancer survival decision support system in routine patient care data.* Radiother Oncol, 2014. **113**(1): p. 47-53.
20. Cancer Research UK. *Oesophageal Cancer Incidence Statistics*. 2016 December 20th 2016]; Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer/incidence.
21. van Rossum, P.S., et al., *The emerging field of radiomics in esophageal cancer: current evidence and future potential.* Transl Cancer Res, 2016. **5**(4): p. 410-423.
22. Hatt, M., et al., *Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211.* Med Phys, 2017. **44**(6): p. e1-e42.
23. Wu, W., et al., *Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology.* Front Oncol, 2016. **6**: p. 71.
24. Brooks, F.J. and P.W. Grigsby, *The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake.* J Nucl Med, 2014. **55**(1): p. 37-42.
25. Hatt, 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.* J Nucl Med, 2015. **56**(1): p. 38-44.
26. Orlhac, F., et al., *Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis.* J Nucl Med, 2014. **55**(3): p. 414-22.
27. Hatt, M., et al., *Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology.* Eur J Nucl Med Mol Imaging, 2011. **38**(7): p. 1191-202.
28. Tixier, F., et al., *Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer.* J Nucl Med, 2011. **52**(3): p. 369-78.
29. Yip, C., et al., *Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy.* Radiology, 2014. **270**(1): p. 141-8.
30. Mackin, D., et al., *Harmonizing the pixel size in retrospective computed tomography radiomics studies.* PLoS One, 2017. **12**(9): p. e0178524.
31. Zwanenburg, A., et al. *Image biomarker standardisation initiative - feature definitions*. 2016 March 20th 2017]; Available from: https://arxiv.org/abs/1612.07003v3.
32. Altman, D.G. and P. Royston, *What do we mean by validating a prognostic model?* Stat Med, 2000. **19**(4): p. 453-73.

33. Mallett, S., et al., *Reporting performance of prognostic models in cancer: a review.* BMC Med, 2010. **8**: p. 21.
34. Reilly, B.M. and A.T. Evans, *Translating clinical research into clinical practice: impact of using prediction rules to make decisions.* Ann Intern Med, 2006. **144**(3): p. 201-9.

Appendix A
Additional Materials and Methods

Patient Cohorts

STAGE cohort

Patients from the STAGE cohort were followed up until July 2016, for a minimum of 12 months. The development and internal validation cohorts were chronologically separated; between September 2010 and September 2014, and between September 2014 and July 2015, respectively. All patients were deemed to have potentially curable esophageal cancer following contrast-enhanced computed tomography (CECT) staging investigation. [1] Stage was defined classified according to the Union for International Cancer Control (UICC) Tumor Node Metastasis (TNM) 7th edition [2].

All patients were identified at the Regional Upper GI Cancer multi-disciplinary team (MDT) meeting. Inclusion criteria were: biopsy proven esophageal cancer, histological cell type of adenocarcinoma or squamous cell carcinoma (SCC), a fluorodeoxyglucose (FDG) avid tumor ($SUV_{max} \geq 3.0$) and a metabolic tumor volume (MTV) ≥ 5 mL. Exclusion criteria were: patients aged < 18 years, a non- or poorly FDG-avid primary tumor ($SUV_{max} < 3$), a histological cell type other than adenocarcinoma or SCC, a MTV < 5 mL, a synchronous primary malignancy or an esophageal stent in situ at the time of PET/CT. [1]

CROSS cohort

The external validation cohort consisted of esophageal cancer patients treated with neo-adjuvant chemoradiotherapy (NACRT) according to the CROSS trial regimen. [3] The inclusion and exclusion criteria matched those of the STAGE cohort. The CROSS cohort patients were recruited from 3 centers in The Netherlands.

Treatment Protocols

As patients were recruited in the STAGE cohort based on intention-to-treat prior to PET/CT, both radical and palliative treatments were included in the prognostic model. Full details of the treatment protocols can be found in Foley et al. [1] In summary, patients had surgery alone (SA), neo-adjuvant chemotherapy (NACT) or NACRT prior to surgery, definitive chemo-radiotherapy (dCRT) or palliative therapy. The optimum treatment strategy was decided by consensus at the MDT. In general, fit patients with tumors preoperatively staged as T3/T4a, N0/N1 were pre-operatively treated with NACT or NACRT (45 Gy delivered over 25 fractions). Less fit patients, or those with T1/2 N0 disease, had surgery alone. Patients deemed unsuitable for surgery because of co-morbidity and/or performance status, extensive loco-regional disease or personal choice received dCRT (50 Gy delivered over 25 fractions).

Full details of treatment in the CROSS trial regimen have previously been published. [3, 4] In summary, patients that were randomized to NACRT received a total radiation dose of 41.4 Gy given in 23 fractions of 1.8 Gy each, with 5 fractions administered per week, starting on the first day of the first chemotherapy cycle. For chemotherapy, carboplatin targeted at an area under the curve of 2 mg/mL/minute and paclitaxel at a dose of 50 mg/m^2 of body-surface area were administered intravenously. Esophagectomy was performed within 4-6 weeks of the staging investigations.

PET/CT protocols

STAGE cohort

The PET/CT protocol used in all patients included in the STAGE cohort has previously been published in Foley et al [1] and is included here. Patients were fasted for at least 6 hours prior to tracer administration. Serum glucose levels were routinely checked and confirmed as less than 7.0mmol/L prior to imaging. Patients received a dose of 4MBq of ^{18}F-FDG/kg. Uptake time was 90 minutes, standard practice at our institution. A GE 690 scanner (GE Healthcare, Buckinghamshire, UK) was used.

CT images were acquired in a helical acquisition with a pitch of 0.98 and tube rotation speed of 0.5 seconds. Tube output was 120 kVp with output modulation between 20 and 200 mA. Matrix size for the CT acquisition was 512 x 512 pixels with a 50 cm field of view. No oral or intravenous contrast was administered.

PET images were acquired at 3 minutes per field of view. The length of the axial field of view was 15.7 cm (skull base to mid-thigh). Images were reconstructed with the ordered subset expectation maximization algorithm, with 24 subsets and 2 iterations. Matrix size was 256 x 256 pixels, using the VUE Point™ time of flight algorithm. Slice thickness was 3.27 mm and pixel size was 2.73 x 2.73. All PET scans in the STAGE cohort were reconstructed with BitsAllocated: 16, BitsStored: 16 and HighBit: 15.

CROSS cohort

Four different PET/CT scanners were used in the CROSS cohort (n=88). Each scanner used different protocols and had varying acquisition parameters. These are summarized in the table below. Scans in units of BQML were consistently rescaled to SUV units prior to feature extraction. The dominant heterogeneity in the CROSS cohort were the different slice thicknesses used for PET acquisition.

Table A.1. Variety in Parameters of the Different PET/CT Scanners Used in the CROSS Cohort

Scanner	Total	Units	Bits Allocated	Bits Stored	HighBit	Pixel size (mm)	Matrix Size	Slice Thickness (mm)

1	15	BQML	16	16	15	2.04	400x400	5
2	21	BQML (n=14) CNTS (n=6) CPS (n=1)	16	16 (n=19) 12 (n=2)	15 (n=19) 11 (n=2)	4.00	144x144	4
3	50	BQML	16	16 (n=47) 12 (n=3)	15 (n=47) 11 (n=3)	4.06	168x168	5 (n=26) 2 (n=24)
4	2	BQML	16	16	15	5.31	128x128	3.38

Post-reconstruction Combat Harmonization Method

The Combat harmonization method was used to adjust for potential differences in parameters between the scanners described above. This method originally described by Johnson et al [5] is described in further detail in Orlhac et al [6]. In our study, we harmonized our acquired multi-center data based upon slice thickness using the Combat methodology freely available in the R statistical computing environment. [7] The image features were calculated after post-reconstruction harmonization and model validation was repeated using these feature values.

Metric Equations for Tumor Lesion Glycolysis (TLG), Histogram Energy and Histogram Kurtosis

TLG is calculated as the product of SUV_{mean} and MTV.

Histogram Energy, implemented as in Eq. A.1 [8], was calculated by:

$$Histogram\ Energy = \sum_i \big(P(i)\big)^2$$

Eq. A.1

where $P(i) = \frac{N_i}{N}$, with N_i the number of voxels of intensity I, and N, the total number of voxels.

Histogram Kurtosis, Eq. A.2 [8] was calculated by:

$$Histogram\ Kurtosis = \frac{\frac{1}{N}\sum_i (I(i)-\mu)^4}{\left(\frac{1}{N}\sum_i (I(i)-\mu)^2\right)^2}$$

Eq. A.2

where N is the number of voxels in the image, I(i) is the positive intensity value in the 3D matrix and μ is the mean intensity value.

Developed Prognostic Model

Full details of the prognostic model can be found in Foley et al. [1] The prognostic model was developed by entering age at diagnosis (years), radiological TNM stage of disease (stages 1-4) and treatment (curative or palliative) plus 16 candidate PET image features into a Cox proportional hazards regression model. The 16 PET image features were calculated from the primary tumor following segmentation with ATLAAS, an automatic decision tree-based learning algorithm for advanced image segmentation in PET. [9] Six variables were found to be significantly and independently associated with overall survival in multivariate analysis, including 3 PET image features; age [HR =1.02 (95% CI 1.01-1.04), $p < 0.001$], radiological stage [1.49 (1.20-1.84), $p < 0.001$], treatment [0.34 (0.24–0.47), $p < 0.001$], log(TLG) [5.74 (1.44–22.83), $p = 0.013$], log(Histogram Energy) [0.27 (0.10–0.74), $p = 0.011$] and Histogram Kurtosis [1.22 (1.04–1.44), $p = 0.017$]. The median overall survival of the development and validation STAGE cohorts was 16.0 months [95% confidence interval (95% CI) 13.8-18.2] and 14.0 months (95% CI 10.4-17.6), respectively. The incremental value of adding PET image features was evaluated and confirmed with the Akaike Information Criterion (AIC). [10] The AICs of prognostic models with and without PET image features were 2238.007 and 2247.693, respectively. [1] Patients were ranked according to the prognostic score (with a low score favoring better prognosis) split in quartiles to represent risk groups. The prognostic score demonstrated significant differences in overall survival between quartiles in both the development (X_2 143.14, df 3, $p < 0.001$) and validation cohorts (X^2 0.621, df 3, $p < 0.001$). [1]

References

[1] Foley KG, Hills RK, Berthon B, Marshall C, Parkinson C, Lewis WG, et al. Development and validation of a prognostic model incorporating texture analysis derived from standardised segmentation of PET in patients with oesophageal cancer. Eur Radiol. 2018;28:428-36.
[2] Sobin LH, Gospodarowicz MK, Wittekind CH. UICC TNM Classification of Malignant Tumours. 7th ed. New York: Wiley; 2009.
[3] van Hagen P, Hulshof MCCM, van Lanschot JJB, Steyerberg EW, van Berge Henegouwen MI, Wijnhoven BPL, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. N Engl J Med. 2012;366:2074-84.
[4] Shapiro J, van Lanschot JJ, Hulshof MC, van Hagen P, van Berge Henegouwen MI, Wijnhoven BP, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. Lancet Oncol. 2015;16:1090-8.
[5] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118-27.
[6] Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018.
[7] R Core Team. R: A language and environment for statistical computing. R Foundation for Stastical Computing. Vienna, Austria. 2013 [Accessed 23th April 2018]; Available from: http://www.R-project.org/.

[8] Orlhac F, Soussan M, Maisonobe JA, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med. 2014;55:414-22.
[9] 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;61:4855-69.
[10] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19:716-23.

Appendix B
Additional Results

Table B.1 describes the descriptive statistics of the 6 PET image features calculated in both STAGE development and CROSS cohorts. Similar values of the image features are observed between the 2 cohorts, although the mean MTV is greater in the STAGE cohort which is reflective of the fact that a large range of primary tumors were included in this patient group.

Table B.1. Descriptive Statistics of Selected PET Image Features in the STAGE development (n=302) and CROSS cohorts (n=46)

		PET Image Feature					
		SUV_{max}	SUV_{mean}	MTV	log(TLG)	log(Histogram Energy)	Histogram Kurtosis
Pre-harmonization							
STAGE	Mean	16.55	9.14	25.8	2.2	4.74	2.81
	Min	3.56	2.07	5.04	1.3	3.52	1.72
	Max	59.97	35.06	132.47	3.3	6.37	12.69
CROSS	Mean	15.14	9.28	23.35	2.16	4.43	2.76
	Min	3.08	1.62	5.6	1.27	3.1	1.87
	Max	30.63	19.81	99.49	2.84	5.37	4.54
Post-harmonisation							
	Mean	16.45	9.60	22.33	2.08	4.56	2.79
	Min	1.68	0.30	0.29	0.52	1.71	0.81
	Max	69.12	54.04	129.39	3.42	6.47	12.26

MTV metabolic tumor volume; TLG tumor lesion glycolysis

Table B.2 below shows the p-values of Kruskal-Wallis tests obtained before and after Combat harmonization. Figures B.1-B.6 show the distribution of extracted radiomic features before and after Combat harmonization. Kruskal-Wallis tests (p=0.05) were used to compare the distributions of the extracted radiomic features from different slice thicknesses before and after harmonization with the Combat algorithm available in R.

Table B.2. Results of Kruskal-Wallis Analysis of Radiomic Features Extracted from Different Slice Thicknesses Pre- and Post-Harmonization with Combat Method

	p-value	
PET Image Feature	**Pre-Combat Harmonization**	**Post-Combat Harmonization**
SUV_{max}	0.17	0.99
SUV_{mean}	0.91	0.99
MTV	0.0004	0.99
TLG	0.01	0.82
Histogram Energy	<0.001	0.14
Histogram Kurtosis	0.02	0.89

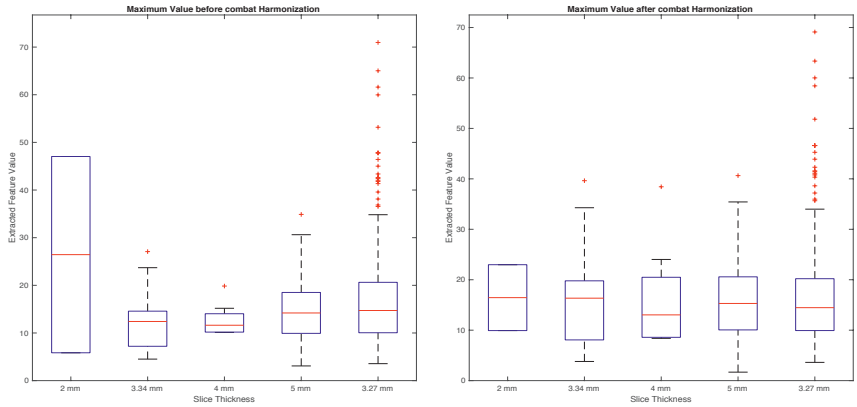


Figure B.1. SUVmax distribution before and after Combat Harmonization

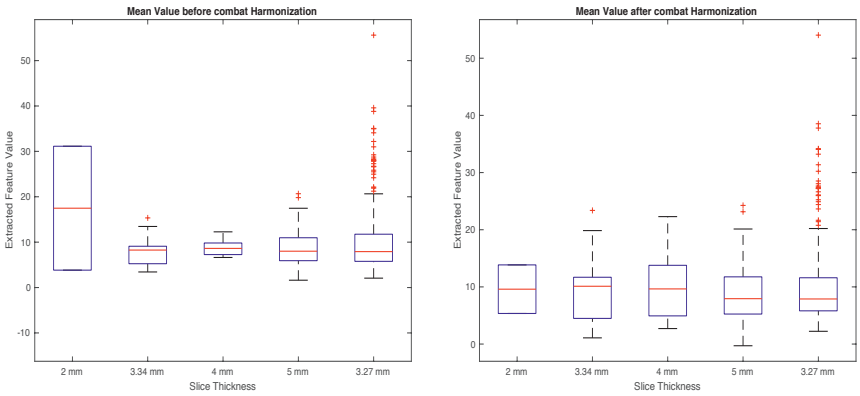


Figure B.2. SUVmean distribution before and after Combat Harmonization

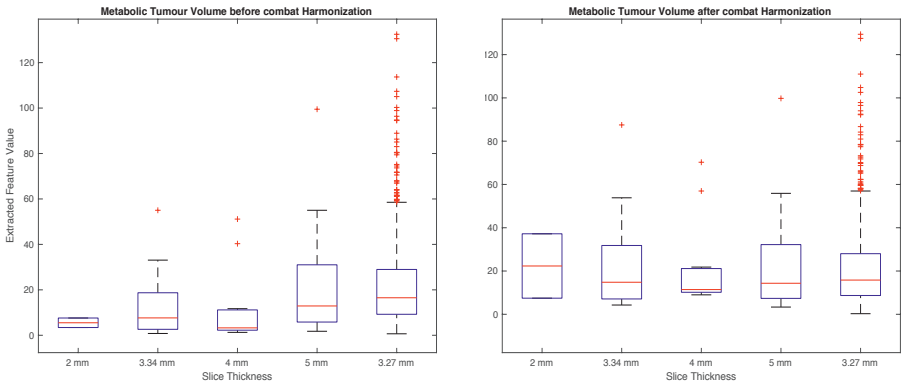


Figure B.3. Metabolic Tumor Volume before and after Combat Harmonization

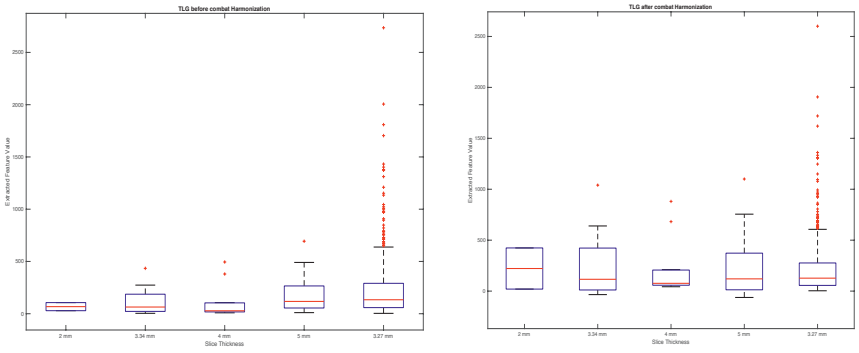


Figure B.4. Tumor Lesion Glycolysis before and after Combat Harmonization

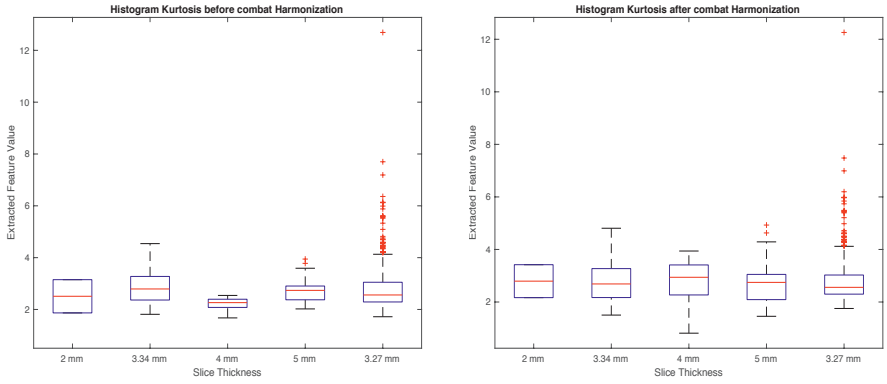


Figure B.5. Distribution of Histogram Kurtosis before and after Combat harmonization

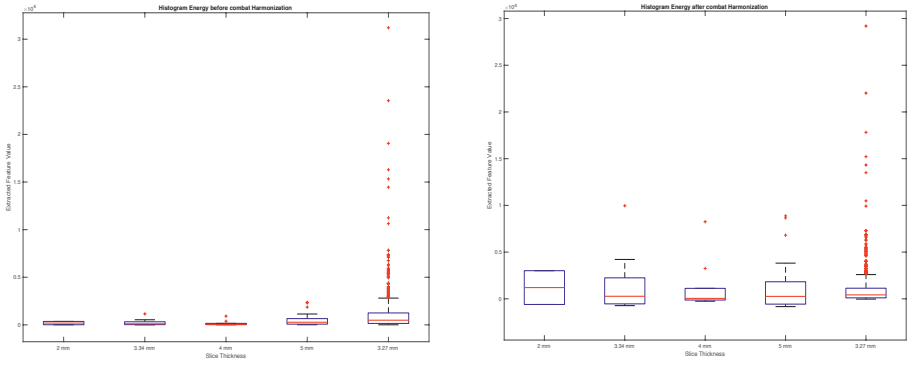


Figure B.6. Distribution of Histogram Energy before and after Combat harmonization

Adjusted survival data for patient quartiles following model validation with PET harmonization

The mean overall survival for patient quartiles 1-4 were 29.5 months (95% CI 15.7-43.2), 30.3 months (95% CI 18.0-42.6), 25.1 months (95% CI 12.7-37.4) and 23.0 months (95% CI 12.4-33.6), respectively. Median overall survival could not be calculated for all quartiles. The median prognostic score for quartiles 1-4 was -1.84 (n=11, range -2.40 to -1.09), -0.76 (n=11, range -1.08 to -0.60), -0.29 (n=11, range -0.57 to -0.09) and 0.29 (n=13, range -0.02 to 2.54), respectively.

Sensitivity Analysis Using all MTVs

This sensitivity analysis includes the 23 cases that were initially excluded because of an MTV < 5 mL. A non-significant association remained between patient quartiles and overall survival in the CROSS cohort (X^2=3.85, df=3, p=0.28) suggesting that inclusion of smaller MTVs had little effect on model validation. (Fig B.7) The HRs of quartiles 2, 3 and 4 compared to quartile 1 was 2.00 (95% CI 0.77-5.18), 2.43 (95% CI 0.95-6.18) and 1.85 (95% CI 0.74-4.89), respectively. The calibration slope β of the prognostic score in the CROSS cohort was 1.24 (standard error (SE) 0.15). β is not significantly different from 1 (p=0.16), which indicates that model discrimination is preserved.

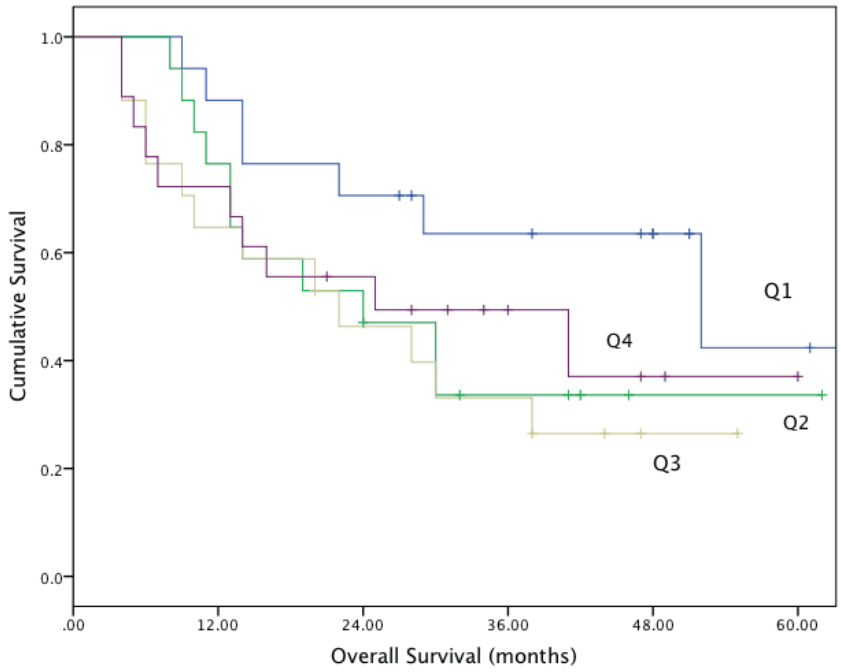


Figure B.7. Cumulative survival curves of patient quartiles (Q1-4) in CROSS cohort using model developed with clinical and radiomic features (X^2=3.85, df=3, p=0.28).

Chapter 7

External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients

Zhenwei Shi, Kieran Foley, Juan Pablo de Mey, Emiliano Spezi, Philip Whybra, Tom Crosby, Johan van Soest, Andre Dekker and Leonard Wee

Adapted from:
Shi, Z., Foley, K., Pablo De Mey, J., Spezi, E., Whybra, P., Crosby, T., ... & Wee, L. (2019). External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients. Frontiers in Oncology, 9, 1411.
DOI: https://doi.org/10.3389/fonc.2019.01411

Abstract

Purpose: Radiation-induced lung disease (RILD), defined as dyspnea in this study, is a risk for patients receiving high-dose thoracic irradiation. This study is a TRIPOD (Transparent Reporting of A Multivariable Prediction Model for Individual Prognosis or Diagnosis) Type 4 validation of previously-published dyspnea models via secondary analysis of esophageal cancer SCOPE1 trial data. We quantify the predictive performance of these two models for predicting the maximal dyspnea grade ≥ 2 within 6 months after the end of high-dose chemo-radiotherapy for primary esophageal cancer.

Material and methods: We tested the performance of two previously published dyspnea risk models using baseline, treatment and follow-up data on 258 esophageal cancer patients in the UK enrolled into the SCOPE1 multi-center trial. The tested models were developed from lung cancer patients treated at MAASTRO Clinic (The Netherlands) from the period 2002 to 2011. The adverse event of interest was dyspnea ≥ Grade 2 (CTCAE v3) within 6 months after the end of radiotherapy. As some variables were missing randomly and cannot be imputed, 212 patients in the SCOPE1 were used for validation of model 1 and 255 patients were used for validation of model 2. The model parameter Forced Expiratory Volume in 1s (FEV_1), as a predictor to both validated models, was imputed using the WHO performance status. External validation was performed using an automated, decentralized approach, without exchange of individual patient data.

Results: Out of 258 patients with esophageal cancer in SCOPE1 trial data, 38 patients (14.7%) developed radiation-induced dyspnea (≥ Grade 2) within 6 months after chemo-radiotherapy. The discrimination performance of the models in esophageal cancer patients treated with high-dose external beam radiotherapy was moderate, area under curve (AUC) of 0.68 (95% CI 0.55 – 0.76) and 0.70 (95% CI 0.58 - 0.77), respectively. The curves and AUCs derived by distributed learning were identical to the results from validation on a local host.

Conclusion: We have externally validated previously published dyspnea models using an esophageal cancer dataset. FEV_1 that is not routinely measured for esophageal cancer was imputed using WHO performance status. Prediction performance was not statistically different from previous training and validation sets. Risk estimates were dominated by WHO score in Model 1 and baseline dyspnea in Model 2. The distributed learning approach gave the same answer as local processing, and could be performed without accessing a validation site's individual patients-level data.

Introduction

In radiation therapy, radical radiation doses are expected to provide better local control than lower palliative doses, however the risk of radiation-induced adverse events is increased. Clinical symptoms of radiation-induced lung disease (RILD) include dyspnea, cough, and fever, which can have a serious effect on the patient's quality of life. Approximately 10-20% of patients with lung cancer who receive (chemo)-radiotherapy developing moderate to severe symptomatic RILD [1].

Radiation-induced dyspnea (RILD in this study) is a side-effect for patients treated with high-dose thoracic irradiation. Studies have reported the predictors for radiation-induced dyspnea for lung cancer patients treated with (chemo)radiotherapy [2, 3]. The risk factors for RILD include dosimetric factors, clinical factors, pathological factors and blood biomarkers [2-16]. In our knowledge, there is no published study reporting the risk factors of radiation-induced dyspnea for patients with primary esophageal cancer, which might be explained by the fact that dyspnea is not routinely assessed during follow-up of esophageal cancer treatment.

The current study conducted a TRIPOD (Transparent Reporting of A Multivariable Prediction Model for Individual Prognosis or Diagnosis) Type 4 validation [17] of previously-published dyspnea models M1 [2] and M2 [3] via secondary analysis of the SCOPE1 [18, 19] dataset. SCOPE1 was a randomized controlled trial investigating the effects of chemo-radiotherapy with and without additional cetuximab in patients with esophageal cancer, including follow-up assessments of dyspnea. We quantify the predictive performance of these two models for predicting the maximal dyspnea grade ≥ 2 within 6 months after the end of high-dose chemo-radiotherapy for primary esophageal cancer. The goal of this study is to verify two hypotheses: (I) that a common thoracic RILD model may be feasible for a different index tumor and (II) that it is feasible to perform an external validation of a toxicity model between two sites via a distributed learning approach without any exchange of patient-specific records.

Methods and Materials

Model development cohorts

Patient characteristics in the development and validation cohorts are detailed in **Table 1.** The first radiation-induced dyspnea model (M1) [2] was developed from 438 patients with either non-small cell lung cancer (NSCLC) Stage I-IIIB or limited disease small cell lung cancer, treated with curatively-intended (chemo)radiotherapy between January 2002 till January 2007. Patients in this cohort were predominantly male (328/438, 74.8%) with confirmed NSCLC histology (292/438, 66.7%) and a spread of chemotherapy regimens (concurrent 70/438, 16%; sequential 203/438, 46%; no chemotherapy 159/438, 36%, unspecified 6/438, 1%). RILD, including dyspnea, was scored according to CTCAE (v3.0) [20] during radiotherapy (RT) and up to a maximum of 6 months after RT. A range of radiotherapy prescribed doses from 46.9 Gy to 79.2 Gy were used, with fraction doses not exceeding 2 Gy.

Table 1: Patient characteristics.

Variable	D1 Maastro clinic (N=438)	D2 Maastro clinic (N=259)	V1 SCOPE1 (N = 212)	V2 SCOPE1 (N = 255)
Gender Male Female	 328 (74.9%) 110 (25.1%)	 163 (62.9%) 96 (37.1%)	 120 (56.6%) 92 (43.4%)	 145 (56.2%) 113 (43.8%)
Age (years)	Mean 68 (SD 9)	Mean 67.5 (SD 10.1)	Mean 72.8 (SD 8.95)	Mean 72.9 (SD 9.02)
Smoking status Current smoker	 77 (29.7%)	 NA	 NA	 NA
WHO-PS 0 1 ≥2	 119 (27.9%) 223 (52.3%) 84 (19.7%)	 63 (24.3%) 153 (59.1%) 43 (16.6%)	 110 (51.9%) 102 (48.1%) 0	 130 (50.9%) 125 (49.1%) 0
CCI 0 1 2 ≥3 Missing	 132(30.9%) 128 (30.0%) 95 (22.2%) 72 (16.8%) 0	 No: 184 (71.0%) Yes: 75 (29%)	 NA	 NA
Cardiac comorbidity No Yes Missing	 132(30.9%) 295 (69.0%) 1 (0.1%)	 No: 184 (71.0%) Yes: 75 (29.0%)	 208 (98.1%) 2 (1.0%) 2 (1.0%)	 252 (98.8%) 3 (1.2%) None
Baseline dyspnea score 0 1 ≥2 Missing	 NA NA NA NA	 78 (30.1%) 140 (54.1%) 38 (14.7%) 3 (1.1%)	 197 (92.9%) 10 (4.7%) 3 (1.4%) 2 (1.0%)	 238 (93.3%) 14 (5.5%) 3 (1.2%) None
dyspnea score after RT 0 1 ≥2 Missing	 NA NA NA NA	 NA NA NA NA	 135 (63.7%） 46 (21.7%) 31 (14.3%)	 164 (64.3%） 53 (20.8%) 38 (14.9%)
FEV_1 (%)	Mean 70.0 (SD 23)	Mean 76.0 (SD 21.86)	NA	NA
Chemotherapy No Yes Missing	 159 (36.8%) 273 (63.2%) 0	 44 (17.0%) 197 (76.1%) 18 (6.9%)	 0 212 (100%) 0	 0 255 (100%) 0
Tumor location Lower/Middle lobe Upper lobe	 245 (56.3%) 190 (43.7%)	 76 (29.3%) 83 (32.1%)	 NA NA	 NA NA
Mean lung dose (Gray) Min Max Median Missing	13.5 (SD 4.5)	15.7 (SD 4.44)	9.8 (SD 2.8) 0.01 17.9 10.0 None	9.83 (SD 2.8) 0.01 17.9 9.9 45 (9.80%)
V_{20} (%)	Mean 21.0 (SD 7.3)	Mean 25.5 (SD 9.9)	NA	NA

***Abbreviations*:** WHO-PS, World Health Organization performance scale; CCI, Charlson comorbidity index; FEV_1, forced expiratory volume (1s); V_{20}, volume of the lung receiving ≥ 20 Gy, SD, standard deviation. D1and D2 are development cohorts for the validated model 1[2] and model 2 [2]. V1 and V2 are validation cohorts.

A second radiation-induced dyspnea model was developed from 259 lung cancer patients treated with curatively intended chemo(radiotherapy) between 2008 and 2011, Stage I-IIIB and fractional dose ≤ 3 Gy were used to develop a second radiation-induced dyspnea model (M2) [3]. These patients were treated in two hospitals, underwent PET/CT for radiotherapy treatment planning and had lung volumes delineated in the planning system. This cohort was drawn from an earlier iso-toxicity dose escalation radiotherapy trial (*clinicaltrials.gov* identifier NCT00572325 and NCT00573040) with maximum tumor dose not exceeding 69 Gy. This cohort was predominantly male (163/259, 62.9%) with confirmed NSCLC histology (198/259, 75.6%), received concurrent chemotherapy (148/259, 57.1%) and had no surgery prior to radiotherapy (236/259, 91.1%). Carboplatin and gemcitabine were given for sequential chemotherapy, and cisplatin and etoposide for concurrent chemotherapy. RILD, including dyspnea, was scored according to CTCAE (v3.0), by either thoracic physicians or radiation oncologists, at baseline and every 3 months following RT.

External validation cohort

Two hundred and fifty-eight esophageal cancer patients were enrolled in the SCOPE1 [18, 19] trial from 36 UK centers between Feb 7, 2008 and Feb 22, 2012. The inclusion criteria were: non-metastatic, histologically confirmed carcinoma of the esophagus (adenocarcinoma, squamous-cell, or undifferentiated carcinoma) or gastro-esophageal junction (Siewert type 1 or 2 with <2 cm extension into the stomach); selected for definitive chemo-radiotherapy by a designated multidisciplinary team; aged 18 years or older; WHO performance status 0 or 1; stage I-III disease (TNM stage 6); and esophageal tumor length < 10 cm as measured by endoscopic ultrasound. The study protocol has been published [19] and the trial was coordinated by the Wales Cancer Trials Unit (WCTU). Recruitment in SCOPE1 was halted due to futility, but follow-up of at least 24 weeks on all recruited patients was available for secondary analysis.

All patients received 4 cycles of cisplatin and capecitabine; 2 cycles were given prior to commencement of RT, and 2 cycles were given concurrently with RT. This chemotherapy regimen was the most commonly used for esophageal cancer treatment in the UK. Chemotherapy dose was modulated for potential hematological toxicity (based on neutrophil and platelet counts) and kidney function (based on glomerular filtrate rate). Chemotherapy cycles were also withheld for serious non-haematological adverse events until resolution to grade 0 or 1. Half of these patients were randomized to additional cetuximab for their chemotherapy.

All 3D conformal RT plans were based on contrast CT 3 mm slices, for a prescribed dose of 50 Gy in 25 once-daily fractions. The esophageal clinical target volume (CTV) was manually delineated as a 2 cm distal and 2 cm proximal expansion along the esophagus from the gross primary tumor, and a 1 cm radial expansion. The planning target volume was an additional 1

cm proximal-distal expansion from the CTV and an extra 0.5 cm radially. Lung volume receiving 20 Gy or higher was constrained to be less than 25% of the total lung volume.

None of the SCOPE1 patients in the validation cohort received post-RT surgery. The majority of patients were male (145/258, 56%) with either mid- or lower-esophageal tumors (226/258, 87.6%) and mean endoscopy-defined tumor length of 5.6 cm. Toxicity scoring according to CTCAE (v3.0) was carried out at baseline, during each chemotherapy cycle, at 24 weeks and then every 3 months thereafter.

Previously published dyspnea model parameters

The model M1 [2] consisted of the following predictors: age, WHO performance status (WHO-PS) before start of RT, nicotine use (non-/ex-smoker versus current smoker), FEV_1 at baseline and mean lung dose in Gy. The predictors used in model M2 [3] were dyspnea score before start of RT, cardiac comorbidity, FEV_1 at baseline, tumor location (upper versus middle/lower lobes of lung) and sequential chemotherapy. Multivariate logistic regression analysis was performed to build M1 and M2. The coefficients used in the models are summarized in **Table 2**. Both models defined adverse outcomes as dyspnea grade 2 or higher within 6 months of the end of (chemo)-radiotherapy.

Table 2: Coefficients obtained from the multivariate logistic regression in the first (M1) [2] and second (M2) [3] dyspnea models.

Variable		Model coefficients (M1)	Model coefficients (M2)
Intercept		-2.2767	-1.512
Performance status			
	WHO-PS = 1	0.28	-
	WHO-PS ≥ 2	0.57	-
Current smoker		-0.45	-
Age		0.02	-
Mean Lung Dose		0.05	-
Baseline dyspnea		-	0.990
Cardiac comorbidity		-	0.826
Sequential chemotherapy		-	0.610
Tumor in middle/lower lung lobe		-	-0.290
Baseline FEV_1		-0.02	-0.007

Model assumptions and missing-values imputation

The previous M1 and M2 had been developed on, and validated in, primary lung cancer patients. However, Forced Expiratory Volume (i.e. FEV_1), smoking status and lung tumor location (lobe) were uniformly absent from the esophageal SCOPE1 dataset. We assumed (based on the trial protocol) that all SCOPE1 patients received chemotherapy and we simulated different population scenarios for smoking status. For the model M2, we further assumed that unintended radiation dose for esophageal cancers were most analogous to RT for lung tumors in lower and/or middle lung lobes.

Since FEV_1 was a predictor in both M1 and M2, we imputed the missing FEV_1 measurements of the SCOPE1 patients from available data in the model M1 development cohort while blinded to the dyspnea outcome. The imputation was based on categorical regression for WHO-PS = 0, WHO-PS = 1 and WHO-PS ≥2. A statistically significant fit for FEV1 (in % of total expired volume) was found using the model:

$$FEV1\ (in\ \%) = 82.0\ if\ WHO - PS = 0$$

$$FEV1\ (in\ \%) = 74.7\ if\ WHO - PS = 1$$

$$FEV1\ (in\ \%) = 67.3\ if\ WHO - PS \geq 2$$

Distributed learning

External validation was performed using the same distributed methodology as published by Deist et al., Jochems et al. and Shi et al. [21-23] using the Varian Learning Portal (VLP, Varian Medical Systems, Palo Alto, CA) v1.0. A validation algorithm containing model coefficients of M1 and M2 were remotely distributed from the investigator site to the validation site via a secured http channel. The SCOPE1 data was parsed using a radiation oncology-specific semantic ontology into the Web 3.0-standard resource descriptor format (RDF). The distributed validation algorithm executes as a purely site-specific local computation by querying the local RDF repository. Only the summary classification results of validation on the SCOPE1 cohort was returned to the investigator site. Security and privacy settings within VLP blocked transfer and exposure of patient-level records from the validation site to the investigator. Previous studies [21-23] have proven that the algorithm converges to the same result as if all of the patient data was locally processed on site by an investigator. The workflow of the distributed learning approach is shown in **Figure 1**.

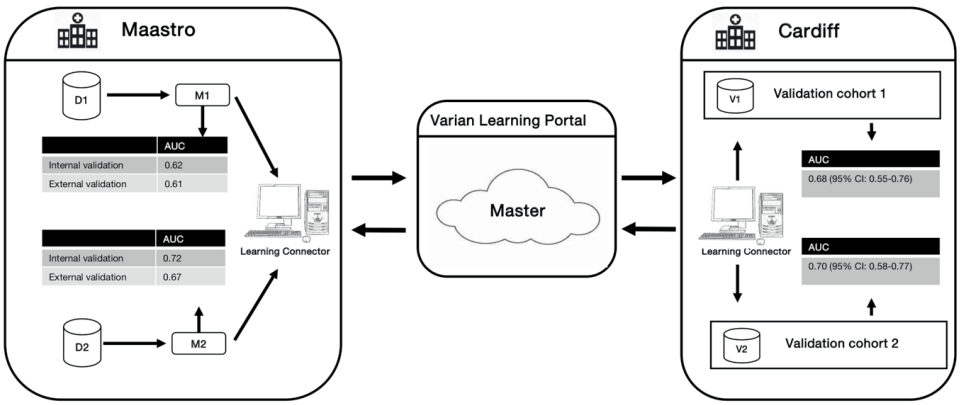


Figure 2: Generalized workflow of the distributed learning approach used in this study. D1 and D2 indicate the development cohorts used to develop the original RILD models M1 and M2. V1 and V2 indicate the validation cohorts for M1 and M2, respectively. CI indicates confidence interval.

Statistical analysis

The validation algorithm was deployed in MATLAB, version 9.0 (MathWorks, Natick, MA). Discrimination of predictive model was evaluated using the area under the receiver-operator curve (AUC) metric [24]. The AUC metric was estimated by bootstrapping (1000 resamples). Calibration of the predictive model was assessed using calibration plots. The logistic recalibration was performed through fitting a logistic regression model by the linear predictor as the only covariate, which leads to an updated model without changing discrimination performance [25, 26].

Results

Out of 258 available validation cases in the SCOPE1 dataset, 46 and 3 patients, respectively, were excluded from the validation due to missing values of mean lung dose for validation of model M1 and baseline scores of cardiac comorbidity and dyspnea for validation of model M2. A total of 212 patients and 255 patients were available to externally validate model M1 and M2. In the validation cohort for M1 (V1), there were 31 patients (14.3%) manifesting dyspnea grade 2 or higher within 6 months of RT. In the validation cohort for M2 (V2), 38 patients (14.9%) manifested dyspnea at the equivalent time point.

To investigate the effect of smoking status on the performance of M1 in the external validation cohort, smoking status was assigned to (i) all smokers, (ii) non-smokers, and (iii) randomly and repeat 1000 iterations. The test yielded the AUC of 0.68 ± 0.053, 0.68 ± 0.054, and 0.65 ± 0.04 respectively by bootstrap sampling. Although the smoking status a missing predictor for esophageal validation cohort, there was no statistically significant difference in performance observed based on a bootstrapped Wilcoxon test between the three scenarios ($p = 0.34$, $p = 0.17$, $p = 0.11$). Therefore, we set it randomly in the validation cohort.

The receiver operator curves (ROCs) of the models on external validation sets V1 and V2 are shown in **Figure 2**. The AUC of both models measured in the previous studies were 0.62 and 0.72 in internal validation and 0.61 and 0.67 in external validation. Compared to the previous studies, the AUC of the two models on V1 and V2 were 0.68 (95% CI: 0.55-0.76) and 0.70 (95% CI: 0.58-0.77), respectively. No statistically significant difference in performance was observed between M1 and M2 in the previous training cohorts and current external validation cohorts (AUC of M1 0.62 vs 0.68, $p = 0.17$; AUC of M2 0.72 vs 0.70, $p = 0.45$, Wilcoxon test). The detailed assessment of accuracy, sensitivity, specificity, positive predictive value and negative predictive value are shown in the Supplementary Table-S1. Both prognostic models (M1 and M2) showed poor calibration performance and tended towards underestimation of dyspnea in the test population, which is shown in the calibration plots (**Figure 3a).** Recalibration was performed to update the prognostic models (**Figure 3b**). As expected, the recalibration resulted in higher predicted risks without changing the AUCs. The calibration line of the recalibrated M1 was shifted be closer to the ideal line, whereas the calibration line of M2 was not improved overall by the recalibration.

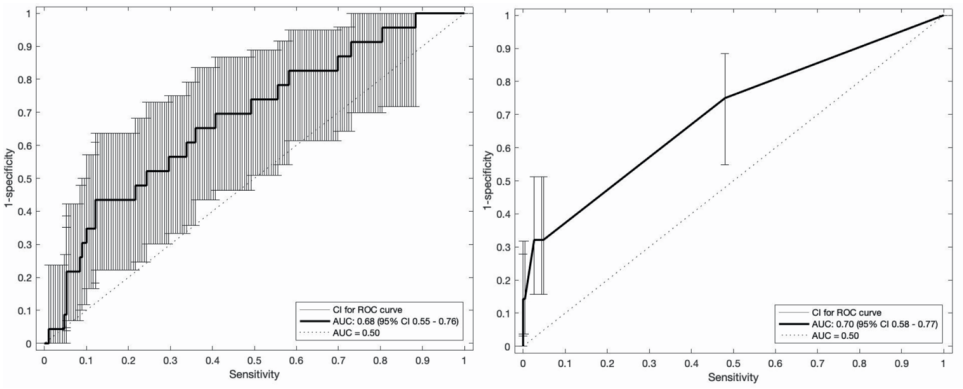


Figure 3: Receiver operating characteristic curves of the prognostic models a: M1 and b: M2 with 95% CI of area under the receiver-operator curve (AUC). CI: confidence interval.

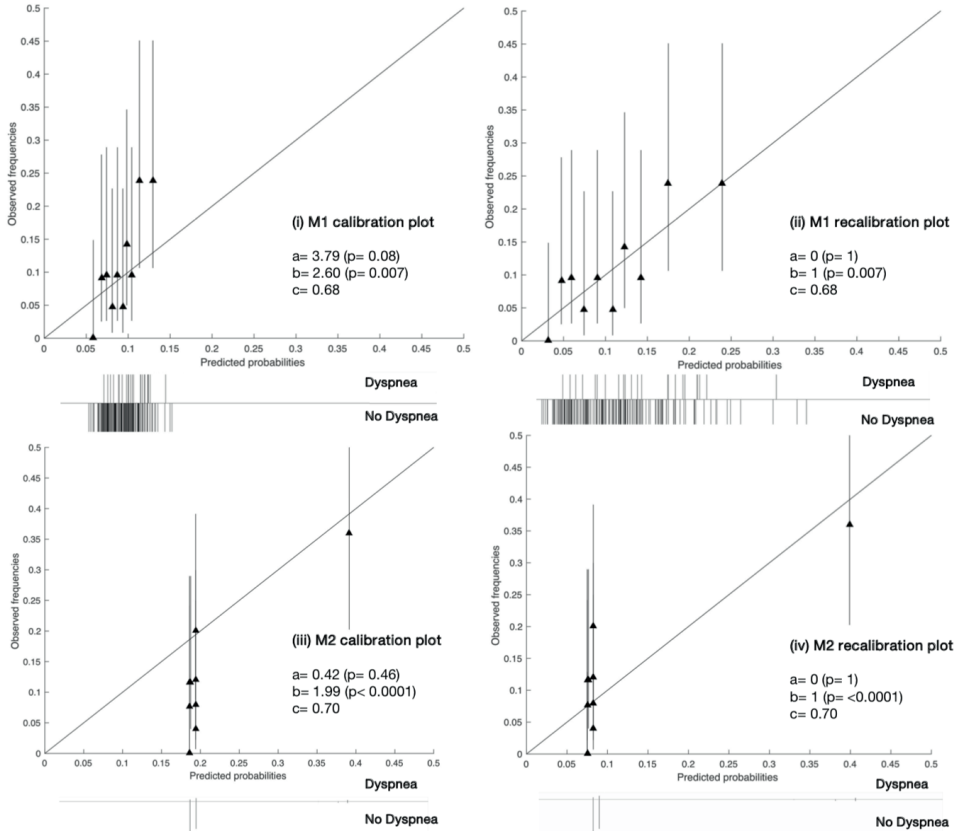


Figure 4: Calibration and recalibration plots of M1 and M2 on the V1 and V2 cohorts, respectively. Perfect calibration is represented by the solid line through the origin with slope =1. Ten quantile groups were used to compare the predicted probability and the corresponding observed frequencies with a triangle. Histogram of outcomes (i.e., dyspnea or no dyspnea) is shown below each plot. a: calibration-in-the large; b: calibration slop; c: area under the receiver-operator curve (AUC).

Discussion

The current study has tested two previously-published RILD models M1 and M2 [2, 3] on the independent validation sets V1 and V2 of the SCOPE1 trial data [18, 19], which comprises esophageal cancer patients treated with chemo-radiotherapy. Moreover, external validation was successfully implemented using an automated and decentralized approach without exchange of individual patient data.

As is well known, high-dose of thoracic radiation can often provide better local tumor control and survival for patient with cancer. Previous studies have shown that additional radiation in an appropriate range can improve locoregional tumor control and increase survival of patients with lung cancer [27-29]. However, the irradiation dose in the radiotherapy treatment of esophageal cancer can have an adverse effect on lung tissue resulting in RILD, such that it leads to disutility of care and have a serious negative impact on patients' quality of life. RILD usually manifests itself in the acute (<6 months) phase as radiation pneumonitis (RP) and in the later (> 6 months) phase as chronic pulmonary fibrosis [30, 31]. RP is the most common dose-limiting complication of thoracic radiation with clinical symptoms such as dyspnea, cough, and sometimes fever [32]. Therefore, it is a trade-off between better tumor control (i.e., better survival or lower death rate) and RILD.

The prognostic models are regarded as the basis of clinical decision support systems (CDSS) [33] that can relieve clinicians from the pressure of analyzing the large volume of publications and data by applying discoveries from research into a data-analytics architecture [34, 35]. However, it is difficult to apply the results of research in clinical practice to predict which patients with esophageal cancer will likely suffer from RILD. The first reason is that many studies have investigated the risk predictors of RILD including dosimetric, clinical, pathological factors or blood biomarkers [2-16], but results between studies are highly variable or even contradictory [32, 36]. In the meantime, there is no standardized lung toxicity grading system and no standard data models (so-called umbrella protocols) to guide prospective collection on routine cases. On the other hand, few publications report the risk predictors of RILD, (e.g., severe dyspnea) for patients with esophageal cancer. This difficulty might be explained by the fact that dyspnea is not routinely assessed during diagnosis and prognosis of esophageal cancer.

At present, it is widely acknowledged that a prognostic model cannot be applied in clinical practice before its feasibility and practicability have been certified via validation on different levels [17, 37]. External validation of a prognostic model should be performed on an/some independent cohort(s), because most models present optimistic results in the development cohorts. Validation of prognostic models involves two aspects [38]. First, generalizability of a prognostic model can be described by validation on similar (reproducibility) or different (transferability) cohorts. The similarity or difference between cohorts refer to temporal, geography, methodology or investigator, which aims to distinguish from the development cohort of the original model [17, 39, 40]. One primary goal of the current study to investigate

the transferability of two previously-published lung toxicity models M1 and M2 under these "different" situations.

Second, accuracy performance of a prognostic model shows the statistical validity [41]. Discrimination and calibration, in general, measure the accuracy performance. (i) Discrimination describes whether an individual with higher predictive probability is indeed experience RILD more often. Area under the receiver-operator curve (AUC) [24] was used to assess the discrimination performance, which is shown in **Figure 2**. The model M1 achieved a better discrimination performance (i.e., AUC) on V1 compared to the internal and external validation performed in the original study. The M2 obtained a better AUC on V2 than the AUC of the external validation but was consistently degraded in AUC from the internal validation of the original study. (ii) Calibration reflects the agreement between observed event and predicted risk. The calibration performance was assessed by calibration plots, which are shown in **Figure 3**. A perfectly calibrated model means that the predicted probabilities of RILD are identical to the observed frequencies of RILD for all patient groups. The calibration-in-the-large (i.e., intercept) of M1 and M2 were 3.79 ($p = 0.08$) and 0.42 ($p = 0.46$), and calibration slope were 2.60 ($p = 0.007$) and 1.99 ($p < 0.0001$), which indicates that predicted risks of M1 and M2 in SCOPE1 were systematically under-estimated and there was insufficient variation of covariates in V1 and V2 sets. A possible explanation may involve systematic under-reporting of clinical toxicity in the retrospectively-collected training sets. By testing different assumptions about smoking status in the test cohorts, there is no evidence to support an effect of smoking in either aggravating or protecting against dyspnea. It is also possible that the original models in lung cancer were improperly calibrated, but there was no additional information in the published articles to confirm this. However, a systematic underestimation of the dyspnea rate would be consistent with an offset error in the linear fit of FEV1 using the WHO performance score. This potential source of error could only be circumvented by measuring the FEV1 for the SCOPE1 test cases, which was not done. To correct poor calibration performance, recalibration can be performed through fitting a logistic regression model by the linear predictor as the only covariate, which leads to an updated model without changing discrimination performance [25, 26, 42]. The calibration performance of M1 was moderate after conducting recalibration. The M2 model still had poor calibration performance even after recalibration, which means care should be taken applied in real clinical practice.

Strengths of the analysis

The SCOPE1 trial data, as an independent validation cohort, satisfied the conditions of separation in terms of temporal (different treatment time of patients in SCOPE1 and previous training cohorts), geographic (different regions, Cardiff vs Netherlands) and investigator (different people from different institutes) from the development cohort of lung cancer. It means that the SCOPE1 was a sufficiently challenging dataset to externally validate the transferability of a prediction model between different index cancers [39, 41]. Second, we have shown the RILD models (e.g., M1) can be robustly transferred to other diseased sites (e.g., esophagus) that only having the incidentally irradiated normal tissues in common without losing accuracy performance. Thirdly, this study was implemented using an automated and

distributed approach without exchanging any patient data. Due to the confidentiality of patient data, local laws and technical issues, it can be prohibitively difficult to exchange patient data among hospitals. Compared to the centralized learning approach, the distributed learning approach can avoid privacy-related issues by sending research questions among institutes. The distributed learning can be achieved by transferring a machine learning algorithm to a target site and returning the results back to the sender rather than transferring real data. This process means knowledge exchange occurs without important clinical data leaving hospitals and there is no loss of validation integrity when performed distributed learning.

Weakness of the analysis

The current study has some limitations worthy of mention. First, some outcome data and predictor variables were missing in the validation cohorts, and data was not missing completely at random. If the missing data were compulsory predictors for the prognostic models (M1 and M2) and cannot be imputed, the corresponding patients had to be removed from the validation cohort. In addition to this, there are non-random missing data, which might be explained by the fact that the information about lung cancer were not be registered for patients with esophageal cancer in the SCOPE1 trial, such as tumor location, smoking status, and FEV_1. For tumor location, we assumed that all of these esophageal cancer patients treated with radiation were similar to lung patients with a tumor in the lower lung lobe. For the missing FEV_1, WHO-PS was used to impute as mentioned above. Second, there are some differences between the development (D1 and D2) and validation cohorts (V1 and V2), of which the effect on the model performance are the subject of future work. (i) SCOPE1 randomized half of the patients between cetuximab or not, whereas patients in D1 and D2 were not treated with cetuximab. (ii) All patients received chemo-radiotherapy in V1 and V2, while only 273 (63.2%) and 197 (76.1%) patients received chemotherapy in D1 and D2. (iii) The numbers of patients in D2 with baseline score 0, 1, ≥2 are 78 (30.1%), 140 (54.1%) and 48 (14.7%), whereas these numbers in V2 are 238 (93.33%), 14 (5.49%) and 3 (1.18%). It indicates that more patients had low-grade or no dyspnea overall in V2 compared with patients in D2. The effects of these uncertainties on the performance of prognostic models M1 and M2 remain unclear and are the subject of future studies.

Finally, another potential limitation is about the validated models' selection, that is the performance of M1 is moderate in terms of AUC and M2 does not include lung dose volume parameters. Although the discrimination performance of M1 is moderate, we found it achieved a similar and even better discrimination performance in the external validation cohort, which demonstrated that M1 has a good generalization. M2 was developed using multivariable regression approach. The original study [3] did evaluate mean lung dose and V20Gy as potential risk factors, but then dropped it from the final regression model because their contributions were small and/or could not be shown to be statistically significant.

Future work

Future work would involve two aspects. First, the M1 could be tested on a similar dataset to validate the reproducibility. Second, we would like to re-train the lung toxicity model on D1 and D2 via combining different types of features, such as image, pathological or generic features.

Conclusion

In this study, we have externally validated previously published dyspnea models using an esophageal cancer dataset. First, the discrimination performance of the models in esophageal cancer patients treated with high-dose external beam radiotherapy are moderate, AUC of 0.68 (95% CI 0.55 – 0.76.) and 0.70 (95% CI 0.58 -0.77), respectively. Second, risk estimates were strongly determined by WHO score in Model 1 and baseline dyspnea in Model 2. Third, the distributed learning approach gave the same answer as local validation but is feasible without accessing a validation site's patient-level data. Finally, the clinical contribution of the dyspnea prognostic model is that it would help doctors to identify patients who will likely suffer from severe dyspnea and who could therefore benefit from dose de-escalation in (chemo)-radiotherapy. Although we cannot conclude that a common thoracic RILD model is feasible for a different primary tumor, it can be deemed as a "benchmark" for further investigation of RILD prognostic models of thoracic tumor.

Data availability statement

The datasets generated for this study will not be made publicly available. The data used in this study was generated in the external validate center. The corresponding author cannot see the data, which was the reason why we performed the distributed learning to avoid data sharing in this study.

Ethics statement

The studies involving human participants were reviewed and approved by Velindre Cancer Centre, Cardiff, UK. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

ZS implemented the distributed MATLAB code via VLP, converted clinical data of SCOPE1 to RDF format, performed analysis on the results using MATLAB, and made a major contribution to the writing of the manuscript as the first author. KF and TC were responsible for data preparation and quality check of SCOPE1 dataset. JP implemented the imputation analysis to deal with the missing data in SCOPE1 dataset. ES and PW were responsible for VLP setup in Cardiff for distributed learning. JS provided technical support for external validation analysis through VLP. AD and LW acted in the capacity of joint senior authors who motivated the study, set the general methodology and had overall scientific responsibility for this investigation. All co-authors contributed to proof-reading of the manuscript.

Funding

This work has been supported by a Dutch STW-Perspectief grant: Radiomics STRaTegy (file number 14930) and NWO grant: BIONIC (629.002. 205).

References

1. Mehta, V., *Radiation pneumonitis and pulmonary fibrosis in non–small-cell lung cancer: Pulmonary function, prediction, and prevention.* International journal of radiation oncology* biology* physics, 2005. **63**(1): p. 5-24.
2. Dehing-Oberije, C., et al., *The importance of patient characteristics for the prediction of radiation-induced lung toxicity.* Radiother Oncol, 2009. **91**(3): p. 421-6.
3. Nalbantov, G., et al., *Cardiac comorbidity is an independent risk factor for radiation-induced lung toxicity in lung cancer patients.* Radiother Oncol, 2013. **109**(1): p. 100-6.
4. Hope, A.J., et al., *Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters.* Int J Radiat Oncol Biol Phys, 2006. **65**(1): p. 112-24.
5. Jenkins, P. and J. Watts, *An improved model for predicting radiation pneumonitis incorporating clinical and dosimetric variables.* Int J Radiat Oncol Biol Phys, 2011. **80**(4): p. 1023-9.
6. Kim, M., et al., *Factors predicting radiation pneumonitis in locally advanced non-small cell lung cancer.* Radiat Oncol J, 2011. **29**(3): p. 181-90.
7. Kwa, S.L., et al., *Evaluation of two dose-volume histogram reduction models for the prediction of radiation pneumonitis.* Radiother Oncol, 1998. **48**(1): p. 61-9.
8. Madani, I., et al., *Predicting risk of radiation-induced lung injury.* J Thorac Oncol, 2007. **2**(9): p. 864-74.
9. Marks, L.B., et al., *Radiation dose-volume effects in the lung.* Int J Radiat Oncol Biol Phys, 2010. **76**(3 Suppl): p. S70-6.
10. Palma, D.A., et al., *Predicting radiation pneumonitis after chemoradiation therapy for lung cancer: an international individual patient data meta-analysis.* Int J Radiat Oncol Biol Phys, 2013. **85**(2): p. 444-50.
11. Rancati, T., et al., *Factors predicting radiation pneumonitis in lung cancer patients: a retrospective study.* Radiother Oncol, 2003. **67**(3): p. 275-83.
12. Stenmark, M.H., et al., *Combining physical and biologic parameters to predict radiation-induced lung toxicity in patients with non-small-cell lung cancer treated with definitive radiation therapy.* International Journal of Radiation Oncology* Biology* Physics, 2012. **84**(2): p. e217-e222.
13. Vinogradskiy, Y., et al., *A novel method to incorporate the spatial location of the lung dose distribution into predictive radiation pneumonitis modeling.* International Journal of Radiation Oncology* Biology* Physics, 2012. **82**(4): p. 1549-1555.
14. Iwata, H., et al., *Correlation between the serum KL-6 level and the grade of radiation pneumonitis after stereotactic body radiotherapy for stage I lung cancer or small lung metastasis.* Radiotherapy and Oncology, 2011. **101**(2): p. 267-270.
15. Voets, A.M., et al., *No association between TGF-β1 polymorphisms and radiation-induced lung toxicity in a European cohort of lung cancer patients.* Radiotherapy and Oncology, 2012. **105**(3): p. 296-298.
16. Novakova-Jiresova, A., et al., *Transforming growth factor-β plasma dynamics and post-irradiation lung injury in lung cancer patients.* Radiotherapy and oncology, 2004. **71**(2): p. 183-189.

17. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement.* BMC Med, 2015. **13**: p. 1.
18. Crosby, T., et al., *Chemoradiotherapy with or without cetuximab in patients with oesophageal cancer (SCOPE1): a multicentre, phase 2/3 randomised trial.* Lancet Oncol, 2013. **14**(7): p. 627-37.
19. Hurt, C.N., et al., *SCOPE1: a randomised phase II/III multicentre clinical trial of definitive chemoradiation, with or without cetuximab, in carcinoma of the oesophagus.* BMC Cancer, 2011. **11**: p. 466.
20. Trotti, A., et al. *CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment*. in *Seminars in radiation oncology*. 2003. Elsevier.
21. Shi, Z., et al., *Distributed radiomics as a signature validation study using the Personal Health Train infrastructure.* Scientific data, 2019. **6**(1): p. 1-8.
22. Jochems, A., et al., *Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept.* Radiotherapy and Oncology, 2016. **121**(3): p. 459-467.
23. Deist, T.M., et al., *Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT.* Clinical and translational radiation oncology, 2017. **4**: p. 24-31.
24. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.
25. Janssen, K.J., et al., *A simple method to adjust clinical prediction models to local circumstances.* Can J Anaesth, 2009. **56**(3): p. 194-201.
26. Steyerberg, E.W., et al., *Validation and updating of predictive logistic regression models: a study on sample size and shrinkage.* Stat Med, 2004. **23**(16): p. 2567-86.
27. Kong, F.-M., et al., *High-dose radiation improved local tumor control and overall survival in patients with inoperable/unresectable non–small-cell lung cancer: Long-term results of a radiation dose escalation study.* International Journal of Radiation Oncology• Biology• Physics, 2005. **63**(2): p. 324-333.
28. Pignon, J.-P., et al., *A meta-analysis of thoracic radiotherapy for small-cell lung cancer.* New England Journal of Medicine, 1992. **327**(23): p. 1618-1624.
29. Warde, P. and D. Payne, *Does thoracic irradiation improve survival and local control in limited-stage small-cell carcinoma of the lung? A meta-analysis.* Journal of Clinical Oncology, 1992. **10**(6): p. 890-895.
30. Bernchou, U., et al., *Time evolution of regional CT density changes in normal lung after IMRT for NSCLC.* Radiotherapy and Oncology, 2013. **109**(1): p. 89-94.
31. Jiang, Z.-Q., et al., *Long-term clinical outcome of intensity-modulated radiotherapy for inoperable non-small cell lung cancer: the MD Anderson experience.* International Journal of Radiation Oncology• Biology• Physics, 2012. **83**(1): p. 332-339.
32. Rodrigues, G., et al., *Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review.* Radiotherapy and oncology, 2004. **71**(2): p. 127-138.

33. Lambin, P., et al., *Decision support systems for personalized and participative radiation oncology.* Advanced drug delivery reviews, 2017. **109**: p. 131-153.
34. Abernethy, A.P., et al., *Rapid-learning system for cancer care.* Journal of Clinical Oncology, 2010. **28**(27): p. 4268-4274.
35. Lambin, P., et al., *Rapid Learning health care in oncology'–an approach towards decision support systems enabling customised radiotherapy.* Radiotherapy and Oncology, 2013. **109**(1): p. 159-164.
36. Mehta, V., *Radiation pneumonitis and pulmonary fibrosis in non–small-cell lung cancer: Pulmonary function, prediction, and prevention.* International Journal of Radiation Oncology• Biology• Physics, 2005. **63**(1): p. 5-24.
37. Steyerberg, E.W., et al., *Prognosis Research Strategy (PROGRESS) 3: prognostic model research.* PLoS medicine, 2013. **10**(2): p. e1001381.
38. Van Soest, J., et al., *Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility.* Medical physics, 2017. **44**(9): p. 4961-4967.
39. Justice, A.C., K.E. Covinsky, and J.A. Berlin, *Assessing the generalizability of prognostic information.* Annals of internal medicine, 1999. **130**(6): p. 515-524.
40. Moons, K.G., et al., *Risk prediction models: II. External validation, model updating, and impact assessment.* Heart, 2012. **98**(9): p. 691-698.
41. Altman, D.G. and P. Royston, *What do we mean by validating a prognostic model?* Statistics in medicine, 2000. **19**(4): p. 453-473.
42. Lamain-de Ruiter, M., et al., *External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study.* bmj, 2016. **354**: p. i4338.

Supplementary

Table S1: The performance assessment of the validated dyspnea model 1 and model 2 on the external validation cohorts V1 and V2.

	External validation cohort (V1)	External validation cohort (V2)
Incidence	11%	11%
AUC	0.68 (95% CI: 0.55-0.76)	0.70 (95% CI: 0.58-0.77)
Accuracy	0.54 (95% CI: 0.47-0.61)	0.88 (95% CI: 0.84-0.92)
Sensitivity	0.70	0.32
Specificity	0.52	0.95
PPV	0.15	0.45
NPV	0.93	0.92

AUC area under curve; CI confidence interval; PPV positive predictive value; NPV negative predictive value.

Chapter 8

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 and Leonard Wee

Underscore indicates equal contribution

Adapted from:
Shi, Z., Zhovannik, I., Traverso, A., Dankers, F. J., Deist, T. M., Kalendralis, P., ... & Dekker, A. (2019). Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. Scientific data, 6(1), 1-8.
DOI: https://doi.org/10.1038/s41597-019-0241-0

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 tumor 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-center 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.

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

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

The specific use case was to learn a radiomic signature "ZS2019" for non-small cell lung cancer (NSCLC) overall survival at one institution and validate it at a remote institution in a distributed fashion. We included two of the NSCLC subject cohorts used by Aerts et al. [2], however, with independently reviewed annotations (tumor 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") [26] was generated at Maastricht University, which was used exclusively for model training, thus obtaining coefficients for a four-feature signature in ZS2019. The second of these datasets (hereafter "Lung2") was generated at Radboud University remains in a private hospital collection that could not be shared publicly for privacy reasons; Lung2 was used exclusively for model validation.

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

Results

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

median follow-up time, the median survival time and the overall 2-year survival rate were similar in both cohorts.

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

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

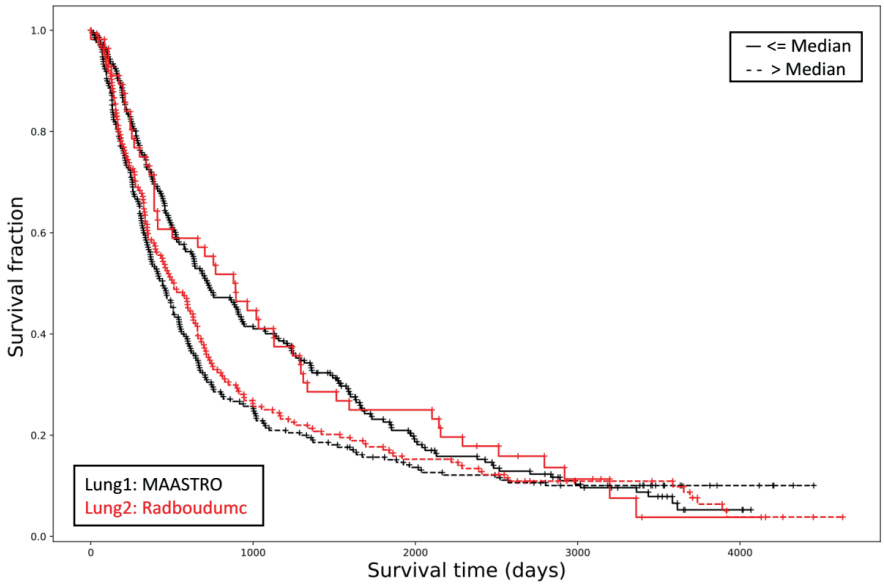


Figure 1: The performance of radiomic signature ZS2019 according to Kaplan-Meier survival analysis. The signature was developed in Lung1 (MAASTRO; black line) and then distributedly validated in Lung2 (Radboudumc; red line). The upper and lower survival curves were split according to the median of the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

We confirmed that the same ZS2019 result was obtained when trained centrally on Lung1 and validated in Lung2. The analysis is given in a Python v3.6 Jupyter notebook that is made

publicly available (https://gitlab.com/UM-CDS/distributedradiomics). 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.

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-center 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[27]. Development of an optimal radiomic signature for NSCLC overall survival would require a detailed re-examination of features and feature selection in the updated datasets, which is not the primary objective of the present study.

The PHT approach utilizes existing data to answer key questions in personalized healthcare, preventive medicine and value-based healthcare. PHT is one of a number of innovative approaches (DataSHIELD[28] and WebDISCO[29]) where the research question is coded as machine-learning algorithms sent to wherever data may reside, instead of centralizing 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) [28], WebDisco (https://omictools.com/webdisco-tool) [29] and ppDLI (https://distributedlearning.ai/blog) may also be used for distributed radiomics. The differences between the present study and the

original study may be traced to : (i) the original Matlab code is commercial confidential and not available to the authors, so we used PyRadiomics developed by Aerts et al. [2] as a practical alternative and (ii) we tried our best to replicate the original method using the documented steps in the original manuscript, but we also improved the survival follow-up such that many right-censored events were now confirmed deaths.

Conclusion

This study demonstrates the proof of concept for multi-center 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.

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.

Tumor delineations

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

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

Outcomes

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

Data processing

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

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

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

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

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

- energy from the intensity histogram feature class, which estimates the overall density of the region of interest,
- compactness from the morphological feature class, which describes the volume of the object relative to that of a perfect sphere,

- grey level run-length matrix (GLRLM) non-uniformity from the textural feature class, which is a measure of intensity heterogeneity averaged over 13 different directions in a 3D matrix of values, and
- wavelet-filtered (HLH) GLRLM non-uniformity, which was the same as (iii) after applying a wavelet decomposition filter over the original image.

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

Semantic web ontologies

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

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

Distributed approach

The VLP distributed learning architecture has been described in deep detail elsewhere [21-23] . In brief, VLP consists of (i) a global web-based clinical learning environment that spans across any number of participating institutes for a given learning project, and (ii) a local connector application that runs exclusively inside the IT firewall of each institute. The former coordinates access permission, asynchronous messaging, web security and site privacy protocols across the learning network, while the latter hosts a local FAIR data repository. Radiomic feature values were hosted in the respective VLP local connector application (v2.0.1) as RDF.

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

Model training

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

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

Cox model evaluation

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

2-year overall survival

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

Data Availability

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

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

Code availability

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

a. D2RQ folder: contains the raw feature value to RDF mapping (D2RQ) script and the SPARQL query used to retrieve the local data into the local VLP connector application.
b. VLP folder: contains the MATLAB codes submitted by the user into VLP, which then transmits it to the participating site for model validation and analysis.
c. Analysis Centralized Learning folder: contains the Jupyter notebook from Radboudumc for model development and evaluation on the aggregated datasets.

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

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

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

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

References

1. McKnight, J., B. Babineau, and J. Gahm, *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, 2014. **5**: p. 4006.
3. Gillies, R.J., P.E. Kinahan, and H. Hricak, *Radiomics: images are more than pictures, they are data.* Radiology, 2015. **278**(2): p. 563-577.
4. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic resonance imaging, 2012. **30**(9): p. 1234-1248.
5. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nature Reviews Clinical Oncology, 2017. **14**(12): p. 749.
6. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* European journal of cancer, 2012. **48**(4): p. 441-446.
7. Coroller, T.P., et al., *CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.* Radiotherapy and Oncology, 2015. **114**(3): p. 345-350.
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, 2016. **34**(18): p. 2157-2164.
9. Parmar, C., et al., *Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer.* Scientific reports, 2015. **5**: p. 11044.
10. Nie, K., et al., *Rectal cancer: assessment of neoadjuvant chemo-radiation outcome based on radiomics of multi-parametric MRI.* Clinical cancer research, 2016. **22.21**: p. 5256-5264.
11. Zhang, B., et al., *Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma.* Clinical Cancer Research, 2017. **23.15**: p. 4259-4269.
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, 2018. **28**(1): p. 428-436.
13. Leijenaar, R.T., et al., *Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability.* Acta oncologica, 2013. **52**(7): p. 1391-1397.
14. Apte, A.P., et al., *Extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research.* Medical physics, 2018. **45.8** p. 3713-3720.
15. van Griethuysen, J.J., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer research, 2017. **77**(21): p. e104-e107.
16. Zhang, L., et al., *IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics.* Medical physics, 2015. **42**(3): p. 1341-1353.
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, 2018. **78**(16): p. 4786-4789.

18. Clark, K., et al., *The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.* Journal of digital imaging, 2013. **26**(6): p. 1045-1057.
19. Berners-Lee, T., J. Hendler, and O. Lassila, *The semantic web.* Scientific american, 2001. **284**(5): p. 28-37.
20. van Soest, J., et al., *Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data.* Studies in health technology and informatics, 2018. **247**: p. 581-585.
21. Jochems, A., et al., *Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries.* International Journal of Radiation Oncology• Biology• Physics, 2017. **99**(2): p. 344-352.
22. Jochems, A., et al., *Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital–A real life proof of concept.* Radiotherapy and Oncology, 2016. **121**(3): p. 459-467.
23. Deist, T.M., et al., *Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT.* Clinical and translational radiation oncology, 2017. **4**: p. 24-31.
24. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.
25. Zwanenburg, A., et al., *Image biomarker standardisation initiative-feature definitions.* Preprint at arXiv: https://arxiv.org/abs/1612.07003, 2016.
26. Aerts, H.J.W.L., Rios Velazquez, Emmanuel, Leijenaar, Ralph T. H., Parmar, Chintan, Grossmann, Patrick, Carvalho, Sara, … Lambin, Philippe. *(2015). Data From NSCLC-Radiomics. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI*.
27. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. (Supplementary).* Nature communications, 2014. **5**: p. 4006.
28. Wolfson, M., et al., *DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data.* International journal of epidemiology, 2010. **39**(5): p. 1372-1382.
29. Lu, C.-L., et al., *WebDISCO: a web service for distributed cox model learning without patient-level data sharing.* Journal of the American Medical Informatics Association, 2015. **22**(6): p. 1212-1219.
30. Lowekamp, B.C., et al., *The design of SimpleITK.* Frontiers in neuroinformatics, 2013. **7**: p. 45.
31. Traverso, A., et al., *The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques.* Medical physics, 2018. **45.10**: p. e854-e862.
32. Kaplan, E.L. and P. Meier, *Nonparametric estimation from incomplete observations.* Journal of the American statistical association, 1958. **53**(282): p. 457-481.
33. Harrell, F.E., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in medicine, 1996. **15**(4): p. 361-387.

34. Peto, R. and J. Peto, *Asymptotically efficient rank invariant test procedures.* Journal of the Royal Statistical Society: Series A (General), 1972. **135**(2): p. 185-198.

Supplementary

The contents of this file are listed in the table below with page number and brief description.

	Page number	Brief description
Table A	3	Radiomic feature names and formulas in the original article and PyRadiomics.
Figure A	5	The distribution of four features in training and testing datasets after Z-score scaling.
Figure B	5	The distribution of four features in training and testing datasets after Log10 data transformation and Z-score scaling.
Table B	5	Radiomic features distributions and split medians.
Figure C	6	Kaplan Meier survival curves of single radiomic feature.

Radiomic features

The following four radiomic features were used in this study. Table A lists the names in the original paper and PyRadiomics. The definition and formula are described as follows.

Table A: **Radiomic feature names in the original article and PyRadiomics**.

	Hugo's feature name	Pyradiomics name
Feature 1	Statistics Energy	original_firstorder_Energy
Feature 2	Shape Compactness	original_shape_Sphericity ^3
Feature 3	Grey Level Nonuniformity	original_glrlm_GrayLevelNonUniformity
Feature 4	wavelet Grey Level Nonuniformity HLH	wavelet-HLH_glrlm_GrayLevelNonUniformity

Feature 1: original_firstorder_Energy

$$energy = \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$$

Here, c is optional value, defined by `voxelArrayShift`, which shifts the intensities to prevent negative values in $\mathbf{X}$. This ensures that voxels with the lowest gray values contribute the least to Energy, instead of voxels with gray level intensity closest to 0.

Energy is a measure of the magnitude of voxel values in an image. A larger values implies a greater sum of the squares of these values.

Original formula [Aerts et al]:

$$energy = \sum_{i}^{N} \mathbf{X}(i)^2$$

Feature 2: original_shape_Sphericity

Predictors	raw mean	raw std	scaled mean	scaled std	split median
original_firstorder_Energy	8.426e+08	1.431e+09	-4.487e-16	1.001	-4.915e-03
original_shape_Compactness	0.2569	0.1088	3.296e-17	1.001	7.861e-02
original_glrlm_GrayLevelNonUniformity	2542	4261	-6.137e-17	1.001	1.49e-01
wavelet-HLH_glrlm_GrayLevelNonUniformity	4674	7103	-6.219e-17	1.001	1.363e-01

$$sphericity = \frac{\sqrt[3]{36\pi V^2}}{A}$$

Sphericity is a measure of the roundness of the shape of the tumor region relative to a sphere. It is a dimensionless measure, independent of scale and orientation. The value range is $0 < sphericity \leq 1$, where a value of 1 indicates a perfect sphere (a sphere has the smallest possible surface area for a given volume, compared to other solids).

The **Compactness2** feature is highly correlated to sphericity, so the compactness2 was computed via the third power of sphericity. Original formulae [Aerts et al]:

$$compactness\ 2 = 36\pi \frac{V^2}{A^3}$$

$$sphericity = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}$$

Feature 3: original_glrlm_GrayLevelNonUniformity

$$GLN = \frac{\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r} \mathbf{P}(i,j|\theta)\right)^2}{N_z(\theta)}$$

GLN measures the similarity of gray-level intensity values in the image, where a lower GLN value correlates with a greater similarity in intensity values.

Original formula [Aerts et al]:

$$GLN = \frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} p(i,j|\theta)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j|\theta)}$$

Feature 4: wavelet-HLH_glrlm_GrayLevelNonUniformity

The wavelet analysis was conducted using the python library PyWavelets (version 0.5.2) and 'Coif1' wavelet was applied as the filter on the original CT images. The formula of feature 4 is the identical to the feature 3, but on the filtered images.

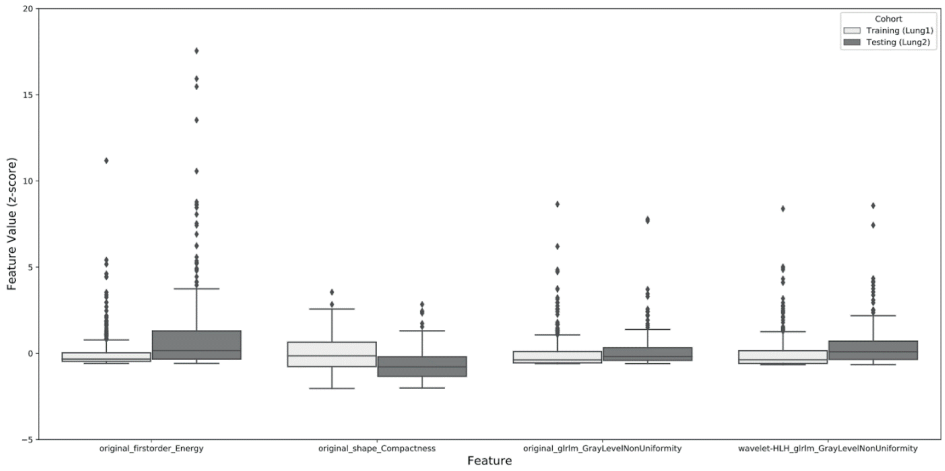


Figure A: The distribution of four features in training and testing datasets after Z-score scaling.

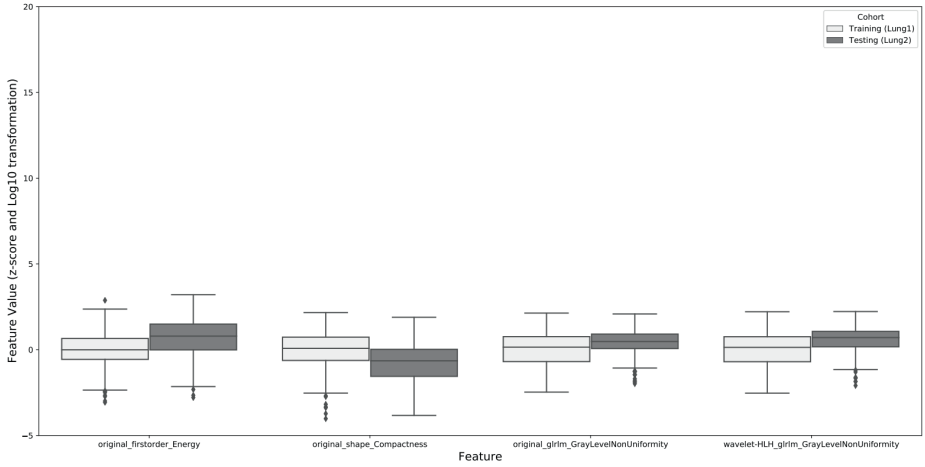


Figure B: The distribution of four features in training and testing datasets after Log10 data transformation and Z-score scaling.

Table B: features distributions and split medians.

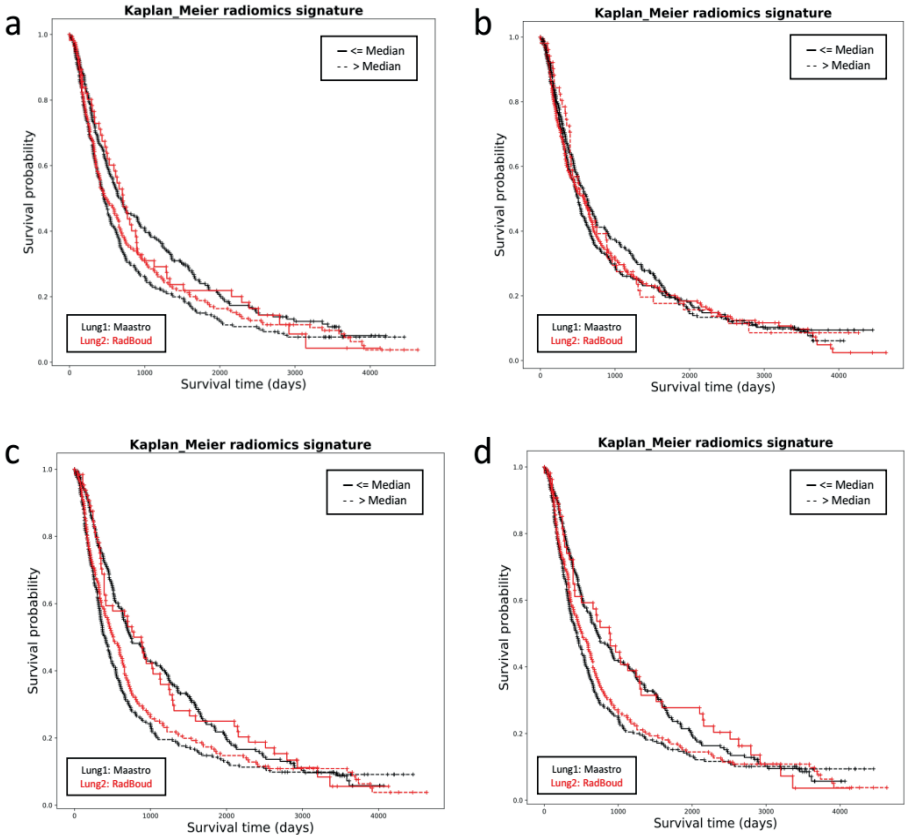


Figure C: Kaplan Meier survival curves of single radiomic features. a original_firstorder_Energy, b original_shape_Compactness, c original_glrlm_RunLengthNonUniformity, d wavelet-HLH_glrlm_RunLengthNonUniformity.

The statistics and split median for each of the features of the cox regression model are shown in Table B. The performance of the radiomic signature was validated in the dataset Lung2 using the Harrell concordance index (HCI), which is a generalization of the area under the ROC curve. The radiomic signature has a reasonable performance to split in the validation set (HCI = 0.58, $P = 3.9 \times 10^{-18}$, Wilcoxon test, n = 221). The Kaplan Meier curves of each single feature are shown in Figure C.

Chapter 9

Discussion and future perspectives

Zhenwei Shi

In this thesis, we have been discussed the current status of big data and treatment toxicity modelling in radiation oncology (**Chapters 2-3**). Of specific interest has been the feasibility of using quantitative imaging features derived from medical images to predict radiotherapy treatment outcomes and side effects (**Chapters 5-6**). As a necessary adjunct to the way that big data is partitioned, we have tackled the challenges of applying artificial intelligence (AI) techniques, especially machine learning on big data, in a decentralized fashion. In order to solve this challenge, a novel approach called federated learning was proposed in [1-3] as a potential solution to allow for training machine learning algorithms on multiple data sources located in different institutes without patient-level data leaving the institutes, so that patient privacy is preserved (**Chapters 7-8**). To make quantitative imaging analysis (i.e., radiomics) feasible especially federated learning across multiple centers, we introduced an ontology-guided radiomics analysis workflow (**Chapter 4**) which helps to present radiomics data in a more Findable, Accessible, Interoperable, and Reusable (FAIR) [4] form as the pre-emptive step for federated learning.

In this chapter, we will discuss (1) challenges and possible solutions of using quantitative imaging analysis to predict cancer treatment outcomes and side effects; (2) feasibility of federated learning in radiation oncology field and (3) some future perspectives.

Prediction of treatment outcomes and side effects by radiomics

Due to the rapid advance of imaging technology, there has been a rising attention on the utilization of quantitative imaging biomarkers derived from medical images. Quantitative imaging analysis (i.e., radiomics [5-7]) has become a promising field for translational clinical research and aims to convert a large volume of routine clinical imaging data into a mineable big data resource to promote research into better cancer treatments and clinical decision-making. Such quantitative assessment of non-invasive biomarkers has been demonstrated to hold predictive/prognostic information in numerous cancer types in addition to molecular and clinical characteristics [5, 8, 9].

The radiomic features extracted from images can be associated with key clinical outcomes (e.g., treatment response and overall survival). Previous studies [10-14] have shown the value of radiomics for quantifying the tumor phenotype and predicting treatment outcomes in clinical settings. By developing diagnostic and prognostic models, radiomics is expected to provide additional and complementary information to clinical factors for decision support. In this thesis, we have investigated the ability of radiomic features extracted from PET (**Chapters 5-6**) and CT (**Chapter 8**) imaging modalities to treatment outcomes (e.g., survival) and side effects (e.g., dyspnea) after radiotherapy treatment.

When integrated into clinical decision support systems (CDSS) [15], these predictive or prognostic models could play a key role in precision medicine [16] that could lead to better customized healthcare at an individual patient level. Furthermore, the automatic process would be able to provide a stable and time-efficient prediction performance. These characteristics could be key elements for treatment shared decision-making systems.

Handcrafted and deep learning-based radiomics

Radiomics analyses usually rely heavily on supervised machine learning. Basically, there are two major types of radiomics commonly used in radiation oncology field [17], that is, handcrafted radiomics and deep learning-based radiomics.

For handcrafted radiomics, the process can be divided into five fundamental parts: (1) image acquisition, (2) identification of the region of interest (ROI), (3) extraction of thousands of human-defined and curated quantitative features from the ROI, (4) selection of potentially prognostic features from a vast set of computed features that are associated with a given clinical endpoint, and finally (5) model development and validation. However, the injection of human expertise characteristic of handcrafted radiomics methods has been criticized recently, as it may introduce subjectivity bias into the process [18], which has given rise to the concerns such as reproducibility [19] and variation. The reproducibility issue might be caused by the intra-reader and inter-reader variability that results from the reliance on manual segmentation of the tumor. The variation issue might also be caused by varying imaging methods and processing techniques for feature extraction. Moreover, the value of handcrafted radiomics has been challenged by the rapid advance of deep learning techniques and the subsequent proof of concept studies [20, 21].

More recently, the deep learning-based radiomics workflow [22-24] has emerged which is different from the former handcrafted radiomics workflow. It does not necessarily need a segmented ROI, and the feature extraction and analysis processes are partially of fully automated. Hence, here are general two steps: (1) image data acquisition (e.g., the entire CT slice) and (2) development of deep neural networks and evaluation. For deep learning-based radiomics studies, ROI identification is either based on a single point placed within the tumor volume or on the entire slice containing the tumor, essentially replacing full tumor segmentations with approximate localization and minimizing the need for human interventions. Furthermore, deep learning approaches allow automated learning of relevant features from images without the need for any pre-definition by humans. The abstract representations used by deep learning have shown a larger learning capacity, boosting generalizability and accuracy while reducing potential bias [25].

Challenges and possible solutions

Even though quantitative imaging analysis is able to decode information about a tumor and its phenotype, there are still some challenges impeding the pace of quantitative imaging analysis research, in particular for studies across multiple centers.

Standardization of radiomics

Although there is an increasing clinical interest in quantitative imaging, radiomics studies often face difficulties of reproducibility and external validation [26-28]. Even using the identical imaging data, radiomic feature values may be different because of different software

implementations, different feature definitions, and different processing schemes. Apart from a requirement of reporting sufficient details of radiomics processing as more as possible, it is necessary to use standardized radiomics. Recently, the Image Biomarker Standardization Initiative (IBSI) [29] has been proposed with the aims to establish standardized definitions of radiomic features, a general processing scheme and a set of report guidelines.

Semantic Web ontology is one kind of technique to describe radiomic features, not only for feature definitions but also for the processes used to calculate these features. To this end, the Radiomics Ontology (www.bioportal.bioontology.org/ontologies/RO) has been developed to offer a standardized manner of publishing radiomic features and approaches, so that one can more concisely report the implementation details of a given radiomics processing workflow.

In this thesis, we showed two radiomics studies (**Chapter 5 and 8**), where the features complied with IBSI, that is, they were described by using the Radiomics Ontology. By doing this, the reproducibility of these two studies are increased.

Radiomics research across multi-centers

For multi-center handcrafted radiomics research, there are three major challenges: (1) lack of standardized methodology for image capture and radiomics analyses; (2) insufficient information in the feature lexicon to fully characterize the preprocessing steps leading up to feature extraction; and (3) insufficient information in the extracted feature values for an independent investigator to reproduce the same values (such as image normalization or interpolation parameters). These issues above hamper multi-center studies because of subtly different imaging protocols, preprocessing steps and extraction software. It is obvious that comparative research will be supported if we not only share the values of radiomic features, but also information about the imaging, preprocessing and computational steps that led to that specific feature value.

The intuitive solution describing radiomic features is to progressively lengthen a human-readable label, for instance *log.sigma.3.0.mm.3D_firstorder_Kurtosis* (a feature name printed by PyRadiomics [30]). However, it will become unpractical if the complexity increases. As an alternative, the Semantic Web approach has added value for this issue, because each calculated value of a feature can be defined with a unique identifier independently of its human-readable feature name and additional unique identifiers can be attached which acts as metadata describing that feature. Also, the details about how we calculate these features can be described easily in graph data by using the Semantic Web approach.

In **Chapter 4**, an open-source ontology-guided radiomics analysis workflow (O-RAW) was proposed to address the aforementioned challenges in the following manner: (1) distributing a free and open-source software package for radiomics analysis, (2) using a domain-specific Radiomics Ontology to uniquely describe features in common usage, and (3) providing methods to publish radiomic features as a semantically interoperable data graph object complying to FAIR data principles. The O-RAW package is expected to support further

standardization of radiomics analyses with the use of ontologies. We have shown multi-center collaboration via a novel learning approach using Semantic Web in the previous studies [31-35].

Feature selection in radiomics

For handcrafted radiomic features, a robust and comprehensive feature selection procedure plays a major role to reduce the risk of over-fitting and in developing a compact, parsimonious final model that only contains the essential predictive variables [36]. Feature selection in machine learning generally comes in three distinct flavors of method, each with their own advantages and disadvantages: Filter methods (e.g. chi-squared test), wrapper methods (e.g. stepwise feature inclusion/elimination) and embedded methods (e.g. regularization) [37]. Previous studies [9, 38-41] have investigated the consequences of different feature selection methods on radiomics model performance.

It is clear that inappropriate feature selection method can adversely affect the performance of a radiomics signature [42]. However, there is no known a priori best method for selecting the most prognostic radiomic features from the beginning. Specifically, we need feature selection to deal with the common issue in many clinical transnational radiomics studies, that is, a limited patient sample size (i.e., a small number of outcome events) and a vast feature dimensional space (which could be one or more orders of magnitude larger than the patient sample size). To increase the clinical applicability and wider generalizability of radiomics, appropriate methods for feature selection and signature compilation are much needed in the field.

Validation of radiomics models

It is necessary to validate clinical prediction models on different levels before applying them into a CDSS. There are two major aspects that should be assessed for a radiomics model as follows [43].

The first aspect is performance of the radiomics model itself, i.e., the statistical validity of the predictive model. It is generally determined by discrimination and calibration, where the discrimination ability describes how well a model can classify a sample into right category correctly and calibration describes the agreement of the frequency between observed and predicted events [44].

The second aspect is the generalizability of a radiomics model, which indicates how it performs in training and validation cohorts. For binary classification models, it can be tested by the method of cohort membership [44], as a test of reproducibility and transferability. The generalization of a radiomics model is often a difficulty for developers, because it is influenced by many factors from the development and validation datasets, such as patient cohorts, imaging protocols, assessment methodologies, and radiomic features (definition, pre-processing approach, and software). Among these factors, we suggest using standardized approaches to

describe radiomic features using rich metadata that could help with the assessment of a radiomics model. We have used Radiomics Ontology in this thesis.

The TRIPOD statement [45] has provided report guidelines suggesting that the report should be full and clear for a prediction model. It provides a checklist with 22 items considering essential information of each section should be reported in prediction model studies. Note that, the TRIPOD statement is not a quality evaluation tool of a prediction model. For radiomics studies specifically, Lambin et al. [6] proposed a radiomics quality score (RQS) to assess the quality of radiomics studies. The RQS overlaps with TRIPOD on matters of general statistical methodology, but differs in that it focusses only on the radiomics domain. Specifically, the RQS awards points for describing the imaging protocol, feature space reduction (or multiple testing correction) and making the data openly available. As described by Sanduleanu et al. [46] the RQS aims to guide users in their appraisal of the radiomics analysis workflow, rather than assess the impact of a radiomics study or significance of a radiomics model.

In many cases, a radiomics model will achieve a promising performance in training and even in internal validation sets, but the external validation performance drops dramatically which is also called over-fitting. This issue reflects the weakness of radiomics in terms of robustness. Basically, there are three common reasons: (1) data has noise (e.g., outliers or errors), (2) too complicated model which is more common in deep learning, and (3) data for training may be not sufficient in terms of sample size and diversity. Fortunately, there are some approaches to avoid over-fitting. Obviously, the first and most direct approach is to use more data. The main reason of over-fitting is that we only have a small dataset, but we try to learn from it. The algorithm will learn a rule to satisfy all the data points exactly in this limited dataset. If we can increase data volume with ample data diversity, the algorithm will be forced to generalize and come up with a good model that suits most data points in this larger dataset. However, increasing data volume is always possible. In these situations, cross validation is the technique to provide a fairly good estimate of model, because the model is tested on different partitions of data to generalize it as much as possible.

Furthermore, although we discover specific findings of the study in the training process and these findings are demonstrated in the internal validation dataset, the findings are still limited due to the selection bias of population. When transferring these findings to general population, it might fail because of the dissimilarity between the data used in training and general population in terms of patient data or even radiomic features. Hence, it requires external validity of applying the findings of a scientific study outside the context of that study to assess the robustness of a prediction model based on radiomics. However, we often lack external validation datasets, mainly because of the concerns of political, ethical, legal, privacy and technical natures, which are so-called data sharing issues between centers.

Deep learning-based radiomics

Although issues, such as intra-reader and inter-reader variability that results from the reliance on manual segmentation of the tumor, and feature selection of human defined features, do not

have a dominant role in deep learning-based radiomics, it presents its own set of challenges. For instance, one of the biggest challenges in deep learning is that it requires a large number of samples for training the deep neural networks. As described in the **Introduction Section**, the medical domain does contain big data that is able to feed the advance of deep learning. Unfortunately, these data are not allowed to be shared among institutes even for research purposes because of the privacy of patient data. Therefore, there is an urgent need for big data or artificial intelligence techniques that can perform deep learning on multiple federated data sources across institutes. Regarding to the technical issue, federated learning is a potential solution which allow for machine learning algorithms on multiple data sources located in different institutes without patient-level data leaving the institutes, so that patient privacy is preserved. We will discuss federated learning in the next section.

Federated learning

While machine learning has developed rapidly in the medical field, its performance, especially deep learning, is highly dependent on the number and diversity of data [10]. In the context of medical imaging, this has become a big challenge because the required input data for machine learning are not available/accessible in a single institute, due to the low incidence rate and limited number of patients. Furthermore, it is not feasible to share patient-level data outside of hospitals because of privacy regulations of medical data such as HIPAA [47, 48] and GDPR [49, 50].

As we described in the **Introduction Section**, the healthcare domain actually owns "big data" [51]. The problem is that most healthcare data are locked in local data repositories inside hospitals. Due to aforementioned concerns of political, ethical, legal, privacy and technical nature, these data are often not allowed to be shared across centers, especially among international centers. These privacy concerns are important and worth to consider cautiously, but the negative aspect is that they have limited the healthcare domain from fully maximizing the benefits of AI.

The medical field is not the only field having this issue of privacy of data. Fortunately, there is a cutting-edge technique that has the potential to have a huge effect on the future of machine learning in medical domain, that is, federated learning [1-3]. Federated learning allows collaborative and distributed training of machine learning algorithms on local nodes (i.e., individual hospital) without exchanging patient-level data. It should be stressed that the patient-level data remains private to each node and is never exchanged in the process of training. Only the model's trainable weights or updates are shared, hence preserving the privacy of patient-level data. In this way, federated learning succinctly avoids many of the data security challenges by leaving the data where they are and enables multi-center collaboration. In **Chapters 7** and **8**, we have shown that federated learning, also called distributed learning, is feasible and allows machine learning on federated data stored inside hospitals.

To achieve the goal of federated learning, data stored in multiple sources must be understood by machine learning algorithms. One possible solution is to produce FAIR data, which allows

machine learning algorithms to be trainable across distributed data sources. However, there is a lack of the technological solutions that are able to produce FAIR data, for both clinical or imaging data, in a reliable and efficient way. As we introduced in **Chapter 4**, the O-RAW package was developed for this purpose. O-RAW is able to generate radiomics data complying with the FAIR data principles: (1) radiomics data and extraction details could be published with a Findable(F) and unique identifier; (2) radiomics data and metadata are described with the Radiomics Ontology, which make them accessible(A) and understandable by machines and humans; (3) data uses a formal, standardized and applicable ontology for knowledge representation, which makes interoperability(I) among multiple centers possible; (4) the data offers explicit information on provenance and licenses for reuse (R). FAIR has a broad applicability with medical data. By doing this, it allows machine learning on linked data, which can be seen as an important step forward for federated learning.

There is indeed a difference between traditional federated machine learning and federated deep learning in terms of privacy preservation of data. For traditional federated machine learning, the information shared between local nodes and central sever is usually a score or error that is used to assess whether a federated model meets the criteria of convergence. For example, Deist et al. [52] showed that the shared information is coefficient of each variable in logistic regression. Similarly, a distributed support vector machine (SVM) algorithm was used in another study [31], where the shared information are parameters of SVM (i.e., w and b). These coefficients and parameters are both very high-level abstract representations of the original data. In principle, it is difficult to inverse original data from these abstract representations.

For federated deep learning algorithms, the two commonly used approaches are transfer of gradients and weights [1] of deep neural networks between clients and central server. However, gradients or weights of deep neural networks are just a transformation of the original data, which contain information in the training data. Furthermore, an important hyper-parameter **F** should be set in federated deep learning, that is, frequency of transfer information between local nodes and central server. If **F** is too big, it might influence the performance of the federated model due to over-fitting of training on local data. But if is small (e.g., 1 indicating local training for 1 epoch), it would increase the chance to inverse the original data of each node. A frequent transfer of gradients or weights in continuous iterations has been shown to increase the probability of inversing certain properties of training data from the shared information between local nodes and central servers as described in [53, 54]. For example, if the original data is portrait photography, it is difficult to recover the original images completely, but it is possible to infer the subjects' gender, age, race and so on. This kind of research of federated learning for privacy preservation is still under study.

While federated deep learning cannot guarantee full privacy preservation of client data, it obviously is able to bring higher performance. Comparing to traditional federated machine learning, there is a trade-off between performance and privacy. If certain privacy leakage can be accepted, federated deep learning is still a promising option if it is impossible to share data between centers, because it can achieve comparable performance as central deep learning.

Future perspectives

First, it is clear that data sources including clinical, image, genetic or other types of data play a key role in current AI application development in the medical domain. The question of how to increase the efficient use of "big data" has become a challenge for researchers in this domain. As we described above, one possible solution is to create data which complies to the FAIR data principles via an open and extensible semantic ontology to make data available with metadata and unique identifiers. However, there is a lack of tools that are able to make FAIR data. In this thesis, we proposed O-RAW which can convert radiomic features to FAIR radiomics data. Unfortunately, it is still not sufficient for current quantitative imaging research. There is a need to develop more FAIR tools that can also convert clinical data, genetic data or other types of data, so that we could link data from multiple sources to achieve the goal of using big data to support the development of AI for cancer care.

Second, cancer is a complex disease and usually described by multiple types of data such as clinical data, imaging data, genomic data, pathological data and so on. Previous studies [55] have demonstrated that the developed quantitative imaging biomarkers will achieve a better performance when integrating these features with clinical information compare to just using imaging features alone. Therefore, future work should extend the current studies to integrate multiple types of cancer data to develop prediction/prognostic models for better cancer care. Furthermore, there is strong evidence for quantitative imaging features (i.e., radiomics) that are associated with genomic information [5, 21, 56]. Hence, using the techniques of quantitative imaging to aid the genomic domain to decode tumor phenotypes would be a possible orientation of future work.

Finally, we did not include the studies of federated deep leaning within this thesis, which actually is in progress. Recently, federated learning, in particular federated deep learning, has seen an increasing interest in many domains. In the meantime, there appear many infrastructures enabling federated deep learning among multiple data sources, but most of them with domain design purposes. For instance, the Federated AI Technology Enabler (FATE) [57] led by Webank was developed specifically for financial use cases. To make this possible in the medical domain, we have proposed a federated learning infrastructure called the Personal Health Train [58], which was developed to fit FAIR medical data principles. Therefore, the last possible future work is to perform federated deep learning on the top of FAIR imaging data using the Personal Health Train infrastructure.

Conclusion

There are still much research needed for quantitative imaging analysis to change the routine clinical practice in (radiation) oncology. But there is a positive motivation in the scientific community to develop machine learning programs which support clinical decision-making for cancer treatments. Within this thesis, we have aimed to introduce the current status of data in radiation oncology (**Chapters 2** and **3**), a proposed ontology-guided radiomics analysis workflow (**Chapter 4**), centralized machine learning (**Chapters 5** and **6**), and federated

machine learning (**Chapters 7** and **8**) applications. By doing this, quantitative imaging analysis based on big data is expected to find the multiple clinical, biological and treatment variables that are related to treatment outcomes. This benefits the creation of better predictive models that promote the advance of personalized therapies for each individual patient in future.

References

1. McMahan, H.B., et al., *Communication-efficient learning of deep networks from decentralized data.* arXiv preprint arXiv:1602.05629, 2016.
2. Sheller, M.J., et al. *Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation*. in *International MICCAI Brainlesion Workshop*. 2018. Springer.
3. Gaye, A., et al., *DataSHIELD: taking the analysis to the data, not the data to the analysis.* International journal of epidemiology, 2014. **43**(6): p. 1929-1944.
4. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**.
5. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nature communications, 2014. **5**: p. 4006.
6. Lambin, P., et al., *Radiomics: the bridge between medical imaging and personalized medicine.* Nature Reviews Clinical Oncology, 2017. **14**(12): p. 749.
7. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* European journal of cancer, 2012. **48**(4): p. 441-446.
8. Ou, D., et al., *Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status.* Oral oncology, 2017. **71**: p. 150-155.
9. Parmar, C., et al., *Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer.* Frontiers in oncology, 2015. **5**: p. 272.
10. Coroller, T.P., et al., *CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.* Radiotherapy and Oncology, 2015. **114**(3): p. 345-350.
11. Huang, Y.-q., et al., *Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer.* 2016.
12. Coroller, T.P., et al., *Radiomic phenotype features predict pathological response in non-small cell lung cancer.* Radiotherapy and Oncology, 2016. **119**(3): p. 480-486.
13. Leijenaar, R.T., et al., *External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma.* Acta Oncologica, 2015. **54**(9): p. 1423-1429.
14. Parmar, C., et al., *Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer.* Scientific reports, 2015. **5**: p. 11044.
15. Lambin, P., et al., *Predicting outcomes in radiation oncology—multifactorial decision support systems.* Nature reviews Clinical oncology, 2013. **10**(1): p. 27.
16. Hood, L. and S.H. Friend, *Predictive, personalized, preventive, participatory (P4) cancer medicine.* Nature reviews Clinical oncology, 2011. **8**(3): p. 184.
17. Afshar, P., et al., *From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities.* IEEE Signal Processing Magazine, 2019. **36**(4): p. 132-160.
18. Chalkidou, A., M.J. O'Doherty, and P.K. Marsden, *False discovery rates in PET and CT studies with texture features: a systematic review.* PloS one, 2015. **10**(5).
19. Traverso, A., et al., *Repeatability and reproducibility of radiomic features: a systematic review.* International Journal of Radiation Oncology* Biology* Physics, 2018. **102**(4): p. 1143-1158.
20. Xu, Y., et al., *Deep learning predicts lung cancer treatment response from serial medical imaging.* Clinical Cancer Research, 2019. **25**(11): p. 3266-3275.
21. Hosny, A., et al., *Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study.* PLoS medicine, 2018. **15**(11): p. e1002711.

22. Lao, J., et al., *A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme.* Scientific reports, 2017. **7**(1): p. 1-8.
23. Li, Z., et al., *Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma.* Scientific reports, 2017. **7**(1): p. 1-11.
24. Zhou, L., et al., *A deep learning-based radiomics model for differentiating benign and malignant renal tumors.* Translational oncology, 2019. **12**(2): p. 292-300.
25. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
26. Welch, M.L., et al., *Vulnerabilities of radiomic signature development: the need for safeguards.* Radiotherapy and Oncology, 2019. **130**: p. 2-9.
27. Berenguer, R., et al., *Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters.* Radiology, 2018. **288**(2): p. 407-415.
28. Meyer, M., et al., *Reproducibility of CT Radiomic Features within the Same Patient: Influence of Radiation Dose and CT Reconstruction Settings.* Radiology, 2019. **293**(3): p. 583-591.
29. Zwanenburg, A., et al., *The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping.* Radiology, 2020: p. 191145.
30. Van Griethuysen, J.J., et al., *Computational radiomics system to decode the radiographic phenotype.* Cancer research, 2017. **77**(21): p. e104-e107.
31. 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, 2017. **4**: p. 24-31.
32. 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, 2017. **99**(2): p. 344-352.
33. 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, 2016. **121**(3): p. 459-467.
34. Shi, Z., et al., *Distributed radiomics as a signature validation study using the Personal Health Train infrastructure.* Scientific data, 2019. **6**(1): p. 1-8.
35. Shi, Z., et al., *External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients.* Frontiers in Oncology, 2019. **9**: p. 1411.
36. Reunanen, J., *Overfitting in making comparisons between variable selection methods.* Journal of Machine Learning Research, 2003. **3**(Mar): p. 1371-1382.
37. Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* bioinformatics, 2007. **23**(19): p. 2507-2517.
38. Hawkins, S.H., et al., *Predicting outcomes of nonsmall cell lung cancer using CT image features.* IEEE access, 2014. **2**: p. 1418-1426.
39. Parmar, C., et al., *Machine learning methods for quantitative radiomic biomarkers.* Scientific reports, 2015. **5**: p. 13087.
40. Zhang, B., et al., *Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma.* Cancer letters, 2017. **403**: p. 21-27.
41. Wu, W., et al., *Exploratory study to identify radiomics classifiers for lung cancer histology.* Frontiers in oncology, 2016. **6**: p. 71.
42. Shi, Z., et al. *A Feature-Pooling and Signature-Pooling Method for Feature Selection for Quantitative Image Analysis: Application to a Radiomics Model for Survival in Glioma*. in *International Workshop on Radiomics and Radiogenomics in Neuro-oncology*. 2019. Springer.

43. Justice, A.C., K.E. Covinsky, and J.A. Berlin, *Assessing the generalizability of prognostic information.* Annals of internal medicine, 1999. **130**(6): p. 515-524.
44. Van Soest, J., et al., *Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility.* Medical physics, 2017. **44**(9): p. 4961-4967.
45. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* British Journal of Surgery, 2015. **102**(3): p. 148-158.
46. Sanduleanu, S., et al., *Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score.* Radiotherapy and Oncology, 2018. **127**(3): p. 349-360.
47. Annas, G.J., *HIPAA regulations-a new era of medical-record privacy?* New England Journal of Medicine, 2003. **348**(15): p. 1486-1490.
48. Mercuri, R.T., *The HIPAA-potamus in health care data security.* Communications of the ACM, 2004. **47**(7): p. 25-28.
49. Tikkinen-Piri, C., A. Rohunen, and J. Markkula, *EU General Data Protection Regulation: Changes and implications for personal data collecting companies.* Computer Law & Security Review, 2018. **34**(1): p. 134-153.
50. Voigt, P. and A. Von dem Bussche, *The eu general data protection regulation (gdpr).* A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017.
51. Obermeyer, Z. and E.J. Emanuel, *Predicting the future—big data, machine learning, and clinical medicine.* The New England journal of medicine, 2016. **375**(13): p. 1216.
52. Deist, T.M., et al., *Distributed learning on 20 000+ lung cancer patients–The Personal Health Train.* Radiotherapy and Oncology, 2020. **144**: p. 189-200.
53. Melis, L., et al. *Exploiting unintended feature leakage in collaborative learning*. in *2019 IEEE Symposium on Security and Privacy (SP)*. 2019. IEEE.
54. Hitaj, B., G. Ateniese, and F. Perez-Cruz. *Deep models under the GAN: information leakage from collaborative deep learning*. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.
55. Zhai, T.-T., et al., *Pre-treatment radiomic features predict individual lymph node failure for head and neck cancer patients.* Radiotherapy and Oncology, 2020. **146**: p. 58-65.
56. Wang, S., et al., *Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning.* European Respiratory Journal, 2019. **53**(3): p. 1800986.
57. Liu, Y., et al., *A Communication Efficient Vertical Federated Learning Framework.* arXiv preprint arXiv:1912.11187, 2019.
58. Beyan, O., et al., *Distributed analytics on sensitive medical data: The Personal Health Train.* Data Intelligence, 2020: p. 96-107.

Chapter 10

Appendices

Zhenwei Shi

Appendix I

English summary

There is an increasing interest in the use of Artificial Intelligence (AI) techniques especially machine learning/deep learning to address the challenges in radiation oncology. This thesis focuses on AI-based quantitative imaging techniques (e.g., radiomics) that have the potential to assist doctors and patients to make individualized treatment decisions.

The organization of this thesis is as follows. First, it is necessary to consider the way patient data is collected and stored in order to understand the challenges and opportunities for data exchange and machine-assisted learning. In **Chapter 2**, the concepts pertaining to the organization of big cancer data in the healthcare domain is explored.

The success of modern oncology implies that more and more patients (generally) live longer with the adverse outcomes of their treatment. In **Chapter 3**, we discuss the ways in which clinical value could be derived from data, specifically treatment outcomes prediction which naturally includes side-effects (toxicity) of treatment.

What is clear from our exploration in **Chapters 2** and **3** is that medical imaging represents one of the largest segments of big cancer data and remains hitherto an under-utilized data resource. It is also clear that a number of challenges must be addressed through technology development. Specifically, how to structure imaging data in such a way (i.e. FAIR) that machine algorithms can understand the data well enough to process it automatically with limited human intervention.

In **Chapter 4**. we propose an open-source technology - Ontology-guided Radiomics Analysis Workflow (O-RAW) - that can provide methods to publish radiomic features as a semantically-interoperable data graph object complying to FAIR data principles.

With the ability to make quantitative imaging features amenable to machine learning, we then explored two clinical applications based around the concept of centralized machine learning.

The accurate diagnosis of a lymph node metastasis is crucial for predicting prognosis and guiding treatment decisions. In **Chapter 5**, we perform a TRIPOD type 3 study where we investigated the possibility of machine learning to predict lymph node staging (cN-stage) in esophageal adenocarcinoma. Three prediction models were developed using three types of features: clinical variables only, PET radiomic features only and combined clinical and radiomics predictors. Despite obtaining signal for improved prediction in the development cohort, the models using PET radiomics derived from the primary tumor were found to be not fully replicated in an external validation cohort. This study shows a prediction model based on radiomics may fail on a general population. It also strengthens the necessity of external validation.

Enhanced prognostic models are required to improve risk stratification of patients with esophageal cancer so that treatment decisions can be optimized. In **Chapter 6**, we externally validate a published prognostic model incorporating PET image features for risk stratification of a patient's esophageal cancer after treatment. Transferability of the model was compared using only clinical variables. Although the previous published prognostic model achieved a promising performance in risk

stratification of patients with esophageal cancer after treatment, it did not enable significant discrimination between the validation risk groups even after performing PET harmonization. It again shows the importance of external validation when assessing prediction models based on radiomics.

In the subsequent two chapters, we apply our methodologies to the other major problem introduced in **Chapters 2** and **3**, which is that data tends to be partitioned, particularly over geographical (physical) locations. In **Chapter 7**, we perform a TRIPOD type 4 validation study where two previously-published dyspnea models for lung cancer patients were validated externally on the data of esophageal cancer patients using a distributed learning approach between UK and the Netherlands. According to the results, it shows that toxicity of radiotherapy mainly depends on the details of the organ being irradiated, not directly on a tumor itself. It suggests that a lung toxicity model is not exclusively tied to lung cancers alone, but might also be used in breast, esophagus, stomach cancer and so on. Moreover, we found that the distributed learning approach gave the same answer as local processing, and could be performed without accessing a validation site's patient-level data.

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. In **Chapter 8**, we updated a ground-breaking publication from 2014 by performing a fully decentralized multi-center study to develop a radiomic signature (ZS2019) for non-small cell lung cancer in one institution and validate 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. Although slightly different in terms of data and methods, no statistically significant difference in performance was observed between the new signature and previous work using the same training and validation sets. This study suggests that missing the procedure of external validation for a radiomics model is not a reasonable excuse, if only facing the technical issue of data sharing between centers. We have shown a dataset can be made FAIR in such a way that federated machine learning is possible without having to make data publicly open for privacy reasons.

In **Chapter 9**, the use of quantitative imaging features from big imaging data to predict radiotherapy treatment outcomes and side-effects are discussed. Then the challenges and potential solutions of federated learning are introduced. Finally, future perspectives are described.

Nederlandse samenvatting

Er is een toenemende belangstelling voor het gebruik van kunstmatige intelligentie (AI) - technieken, met name machine learning/deep learning, om de uitdagingen in de radiotherapie aan te pakken. Deze thesis focust op AI gebaseerde kwantitatieve beeldvormingstechnieken (bijv. Radiomics) die artsen en patiënten kunnen helpen om individuele behandelbeslissingen te nemen.

De indeling van dit proefschrift is als volgt. Ten eerste moet worden nagedacht over de manier waarop patiëntgegevens worden verzameld en opgeslagen om de uitdagingen en kansen voor gegevensuitwisseling en machine learning te begrijpen. In **Hoofdstuk 2** worden de concepten verkend die betrekking hebben op de organisatie van en delen van grote hoeveelheden kankergegevens in de gezondheidszorg.

Het succes van de moderne oncologie houdt in dat steeds meer patiënten (over het algemeen) langer leven met de nadelige gevolgen van hun behandeling. In **Hoofdstuk 3** bespreken we de manieren waarop klinische waarde kan worden afgeleid uit gegevens, met name de voorspelling van behandelresultaten inclusief de nadelige gevolgen/bijwerkingen (toxiciteit) van de behandeling.

Wat duidelijk is uit onze verkenning in de **Hoofdstukken 2 en 3** is dat medische beeldvorming een van de grootste segmenten van kankergegevens vertegenwoordigt en tot dusver een onderbenutte gegevensbron blijft. Het is ook duidelijk dat door technologische ontwikkeling een aantal uitdagingen moet worden aangepakt. Specifiek, hoe beeldgegevens zo te structureren (d.w.z. FAIR maken) dat machine-algoritmen de gegevens goed genoeg kunnen begrijpen om ze automatisch te verwerken met beperkte tussenkomt van mensen.

In **Hoofdstuk 4** stellen we een open-source technologie voor - Ontology-guided Radiomics Analysis Workflow (O-RAW) - die methoden kan bieden om radiomic-kenmerken te publiceren als een semantisch interoperabel gegevensgrafiekobject dat voldoet aan de FAIR-gegevensprincipes.

Met de mogelijkheid om kwantitatieve beeldvormingsfuncties vatbaar te maken voor machine learning, hebben we vervolgens twee klinische toepassingen onderzocht die zijn gebaseerd op het concept van gecentraliseerde machine learning.

De nauwkeurige diagnose van een lymfekliermetastase is cruciaal voor het voorspellen van de prognose en het begeleiden van behandelbeslissingen. In **Hoofdstuk 5** voeren we een TRIPOD type 3-studie uit waarin we de mogelijkheid hebben onderzocht van machine learning om de stadia van de lymfeklieren (cN-stadium) bij slokdarm adenocarcinoom te voorspellen. Er zijn drie voorspellingsmodellen ontwikkeld met drie soorten kenmerken: alleen klinische variabelen, alleen PET-radiomics kenmerken en gecombineerde klinische en radiomics-voorspellers. Alhoewel dat het verkrijgen van verbeterde voorspelling in het ontwikkelingscohort mogelijk bleek, bleken de modellen met PET-radiomics afkomstig van de primaire tumor niet volledig te kunnen worden gerepliceerd in een extern validatiecohort. Deze studie toont aan dat een voorspellingsmodel gebaseerd op radiomics soms niet goed werkt bij een algemene populatie. Het versterkt ook de noodzaak van externe validatie.

Verbeterde prognostische modellen zijn vereist om de risicostratificatie van patiënten met slokdarmkanker te verbeteren, zodat behandelbeslissingen kunnen worden geoptimaliseerd. In

Hoofdstuk 6 valideren we extern een gepubliceerd prognostisch model met PET-beeldkenmerken voor risicostratificatie van slokdarmkanker na behandeling. Overdraagbaarheid van het model werd vergeleken met alleen klinische variabelen. Hoewel het eerder gepubliceerde prognostische model na behandeling een veelbelovende prestatie behaalde in risicostratificatie van patiënten met slokdarmkanker, maakte het geen significante discriminatie tussen de validatierisicogroepen mogelijk, zelfs niet na het uitvoeren van PET-harmonisatie. Het toont opnieuw het belang aan van externe validatie bij het beoordelen van voorspellingsmodellen op basis van radiomics.

In de volgende twee hoofdstukken passen we onze methodologieën toe op het andere grote probleem dat in de **Hoofdstukken 2 en 3** is geïntroduceerd, namelijk dat gegevens doorgaans zijn verdeeld, met name over geografische (fysieke) locaties. In **Hoofdstuk 7** voeren we een TRIPOD type 4-validatiestudie uit waarbij twee eerder gepubliceerde dyspnoe-modellen voor longkankerpatiënten extern werden gevalideerd op de gegevens van slokdarmkankerpatiënten met behulp van een gedistribueerde leerbenadering tussen het Verenigd Koninkrijk en Nederland. Uit de resultaten blijkt dat de toxiciteit van radiotherapie voornamelijk afhangt van de details van het bestraalde orgaan, niet rechtstreeks van een tumor zelf. Het suggereert dat een longtoxiciteitsmodel niet uitsluitend is gebonden aan longkankers alleen, maar ook kan worden gebruikt bij borst-, slokdarm-, maagkanker enzovoort. Bovendien ontdekten we dat de gedistribueerde leerbenadering hetzelfde antwoord gaf als lokale verwerking en kon worden uitgevoerd zonder toegang tot de gegevens op patiëntniveau van een validatiesite.

Voorspellingsmodellering met radiomics is een zich snel ontwikkelend onderzoeksthema dat toegang vereist tot enorme hoeveelheden beeldgegevens. Methoden die kunnen werken met decentrale data zijn daarbij dringend nodig vanwege zorgen over de privacy van patiënten. In **Hoofdstuk 8** hebben we een baanbrekende publicatie uit 2014 geüpdatet door een volledig gedecentraliseerde multi-center studie uit te voeren om een radiomic model (ZS2019) voor niet-kleincellige longkanker in één instelling te ontwikkelen en de prestaties in een onafhankelijke instelling te valideren, zonder de noodzaak voor data-uitwisseling en deze methode vergeleken met een analyse waarbij alle data centraal stond. De prestatie van ZS2019 voor 2 jaar algehele overleving gevalideerd in gedistribueerde radiomics was niet statistisch verschillend van de gecentraliseerde validatie. Hoewel enigszins verschillend in termen van gegevens en methoden, werd er geen statistisch significant verschil in prestatie waargenomen tussen het nieuwe model en eerder werk met dezelfde training en validatiesets. Deze studie suggereert dat het niet doen van externe validatie voor een radiomics-model vanwege de technische kwestie van gegevensuitwisseling tussen centra, niet valide is. We hebben laten zien dat een dataset FAIR kan worden gemaakt op ’een manier die federatieve machine learning mogelijk maakt zonder dat gegevens om privacyredenen openbaar moeten worden gemaakt.

In **Hoofdstuk 9** wordt het gebruik van kwantitatieve beeldkenmerken uit beeldgegevens besproken om de resultaten en bijwerkingen van radiotherapiebehandelingen te voorspellen. Vervolgens worden de uitdagingen en mogelijke oplossingen van federatief leren geïntroduceerd. Tenslotte worden toekomstperspectieven beschreven.

Appendix II

Valorization addendum

The prevalence of cancer is an increasing healthcare issue as it is the predominant cause of death worldwide. The growing cancer burden is caused by several factors including population growth, aging, and the changing prevalence of certain causes of cancer during social and economic development [1]. To address the global cancer burden, new technologies, for instance Artificial Intelligence (AI), have been applied in the workflow of cancer care from diagnosis to treatment. For cancer treatment, especially radiotherapy, new innovations are not only useful to provide comprehensive treatment plans, but also able to reduce radiotherapy-induced side-effects which may exist in patients during and (long) after treatment.

The clinical data science (CDS) research group of Maastricht University performs data science with the aim to provide clinical decision aid systems for individualized radiotherapy by the following approaches:

1. Developing global FAIR [2] data sharing infrastructures.
2. Learning personalized prediction models from FAIR data.
3. Applying clinical decision aid systems to improve cancer care.

As a CDS researcher, the work of the thesis has contributed to the investigation of big imaging data and new AI technologies for the first two research aims of CDS.

Knowledge dissemination

Medical imaging represents one of the largest segments of big cancer data and remains hitherto an under-utilized data resource. As data is the basis of machine learning algorithms, it is necessary to figure out the way patient data is collected and stored in hospitals. **Chapter 2** introduced the big radiation oncology data from different angles including data collection from different departments, storage in different and formats and systems, as well as the challenges and opportunities for data exchange and machine-assisted learning.

The success of modern oncology implies that some patients live long with the adverse outcomes of their treatment. The knowledge described in **Chapter 3** might be used to guide cancer treatment, so that the quality of patients' lives can be improved after treatment.

Radiomics on multiple modalities is still under study. The previous findings are that the models developed using CT or PET radiomics indeed had prognostic ability. However, as shown in **Chapter 5 and 6**, they may fail in independently external validation, because of different population between training and validation cohorts, inappropriate feature selection in the phase of model training, and so on. Therefore, appropriate validation is a necessary step to assess the generalization of prediction models based on radiomics.

In order to use big imaging cancer data easily and efficiently, the application of FAIR imaging data was described in this thesis. Furthermore, by integrating distributed learning and FAIR

data, **Chapter 7 and 8** provided a picture about how machine learning can be implemented in multiple centers without sharing data for the aim of data privacy preserving. The introduced studies can be seen as templates for future distributed radiomics studies in the domain.

Societal or commercial relevance

FAIR data is a novel solution to allow machine algorithms can understand the data well enough to process it automatically with limited human intervention. **Chapter 4** introduced an ontology-guided radiomics analysis workflow (O-RAW) [3] that is able to generate FAIR radiomics data, so that multi-center radiomics research is easier. In addition to knowledge sharing, the O-RAW package has been widely used in the CDS group as it provides a pipeline to use DICOM images as the input and produce RDF FAIR data as the output. Furthermore, the cost of data management, retrieval, and interpretation can be reduced by making cancer data FAIR. It is a potential solution to handle the growing big data in the cancer domain.

Distributed learning has shown the feasibility and importance to allow machine learning algorithms on physically federated data sources. Regarding to commercial applications of distributed learning, the reality is that medical companies are often not allowed to access directly the patient data stored in hospitals, because of data privacy and security regulations such as GDPR [4] and HAAP [5]. The work in **Chapter 7 and 8** provided two distributed learning applications, which have shown a potential solution to medical companies by using the concept of the personal health train infrastructure [6] to implement distributed machine learning.

Although the advance of AI-based medical applications is a long and pricy process, the global market is getting larger and larger. This is reflected by the evidence that more and more companies participated in the industry of AI-based medical imaging in the last 5 years including leading enterprises, such as Google, Microsoft, Facebook, Alibaba, Philips, and so on. Due to the phenomena that AI algorithms and computational power are trending to similarity in the market, the properties of data (e.g., volume and quality) seem to be important factors leading to commercial success. Therefore, the study of big cancer data is not only able to improve work efficiency for hospitals and better cancer for individual patient, but to support the advance of AI-medical applications for medical companies and research institutes.

In the current change of era from internet technology (IT) to data technology (DT), new digital technology is developing dramatically. The new digital technology includes Internet of Things, Cloud Computing, Big Data Technology, AI, and Blockchain. These five blocks are building a framework that might work horizontally as follows: (1) Internet of Things can collect real-time data using multi-sensors and transmit this data to the cloud; (2) The cloud can provide data storage space and computing power; (3) big data technology can manage and curate data; (4) AI can process big data, and extract valuable knowledge and information; and (5) Blockchain can provide security support for the knowledge and information in further use. The framework is expected to work in the healthcare domain as well. Hence, there will be many

opportunities for medical enterprises to exploit the healthcare market by using these new technologies.

In conclusion, big imaging data-based prediction models developed by a distributed learning approach have clear valorization potential in cancer care.

References

1. World Health Organization. "*Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018*." International Agency for Research on Cancer. Geneva: World Health Organization (2018).
2. Wilkinson, Mark D., et al. "*The FAIR Guiding Principles for scientific data management and stewardship*." Scientific data 3.1 (2016): 1-9.
3. O-RAW at: https://gitlab.com/UM-CDS/o-raw
4. Annas, George J. "*HIPAA regulations—a new era of medical-record privacy?.*" (2003): 1486-1490.
5. Voigt, Paul, and Axel Von dem Bussche. "*The eu general data protection regulation (gdpr)*." A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017).
6. Personal Health Train at: https://www.health-ri.nl/initiatives/personal-health-train

Appendix III

Acknowledgements

First, I would like to appreciate my supervisors:

Andre Dekker provided this great job for me four years ago. During my PhD, he always supported me and showed me his patient mentorship and valuable advices. I really learnt many merits that will be useful in my life from him, especially how to be a team leader.

As my daily supervisor, Leonard Wee taught me a lot. When I had troubles, he always showed his patience to help me and let me know how to solve issues like a researcher. I really enjoy working with him.

Then, I would also like to thank my dear friends in Cardiff, Kieran, Emiliano, Philip and Craig. Due to their efforts, I can obtain the current research outputs.

Furthermore, I would like to thank my dear friends Chong Zhang, Guangyao Wu, Chang Sun, Cheng Zhu and Tianchen Luo. I enjoyed the time with you. Also, it is my honor meeting you in Maastricht.

Special thanks to Tiantian Zhai who worked as an experienced clinician at UMCG. I learnt a lot from her. Special thanks also to Susu Yan who is working at MGH in US. Thanks for her treat when I was in Boston. Hopefully she will have a great time and career in Boston. Special thanks to Dr Hua Zhang. He gave me many valuable advices about future career.

Then, I would like to thank all of people I met in Groningen, such as Sheng He, Yun Si, Jiawen Chen, Tao Zhang, Guowei Li, Jing Wu and so on. I really had a great time with them, which made me feel not lonely.

I would also like to thank Hugo Aerts for inviting me to Data Farber Cancer Institute as a visiting scholar. The visit to Harvard Medical School was an important and special experience to my life. Also, my thanks to Ahmed Honest. As a machine learning expert, he provided many valuable advices for my research.

During the last four years, I knew many smart people at Maastro Clinic and Maastricht University. Here I would like to thank all of them who are working or worked with me. Also, thanks to my colleagues in India, where they treated me very well. I indeed enjoyed the time there with them.

Most of all, I am grateful to Johan van Soest. He supported me a lot. I still remember the time with him in Anne Arber (Michigan). That was the first time I travelled to US.

Then, I would like to show my most sincere thanks to my parents, in particular to my mother. She always trusts and supports me for all of my decisions.

Finally, I would like to thank people who ever criticized and doubted me and the work I am devoting myself to. These criticisms and doubts always remind me that I must work hard. Hopefully they can see the meanings of the work what we are doing for this world.

Appendix IV

Curriculum vitae

Zhenwei was born on 08 August 1988 in Changchun China. In 2007, he started his bachelor in measurement and control technology and instrument at Changchun University of Science and Technology (China). In 2013, he started his master study of Artificial Intelligence at University of Groningen (The Netherlands). During the period of his master, he mainly investigated machine learning, deep learning and computer vision on images. For his master thesis, he developed a program (DateFinder) using deep learning techniques which can automatically detect date regions on handwritten images.

After obtaining his master's degree, he continued his PhD research at Maastro Clinic/Maastricht University (The Netherlands) in 2016. His main research interest is in big imaging data and machine learning in clinical data science especially in the radiation oncology domain. Between May and July 2017, he visited Dr. Emiliano and Dr. Kieran at Cardiff University School of Engineering for the CORAL international research collaboration between Maastro Clinic and Cardiff University. Between October 2019 and February 2020, he visited Professor Hugo Aerts at Data-Farber hospital/Harvard Medical School (United States) funded by travel grants of the European Society for Radiotherapy and Oncology (ESTRO).

Appendix V

List of manuscripts

Published original research (* contribute equally)

1. **Shi, Z.,** Traverso, A., van Soest, J., Dekker, A., & Wee, L. (2019). Ontology-guided radiomics analysis workflow (O-RAW). *Medical Physics*, *46*(12), 5677-5684.

2. **Shi, Z.*,** Zhovannik, I.*, Traverso, A., Dankers, F. J., Deist, T. M., Kalendralis, P., ... & Dekker, A. (2019). Distributed radiomics as a signature validation study using the Personal Health Train infrastructure. *Scientific data*, *6*(1), 1-8.

3. **Shi, Z**, et al. "External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients." Frontiers in Oncology 9 (2019): 1411.

4. **Shi, Z.,** Zhang, C., Compter, I., Verduin, M., Hoeben, A., Eekers, D., ... & Wee, L. (2019, October). A Feature-Pooling and Signature-Pooling Method for Feature Selection for Quantitative Image Analysis: Application to a Radiomics Model for Survival in Glioma. In *International Workshop on Radiomics and Radiogenomics in Neuro-oncology* (pp. 70-80). Springer, Cham.

5. Foley, K. G.*, **Shi, Z.*,** Whybra, P., Kalendralis, P., Larue, R., Berbee, M., ... & Roberts, S. A. (2019). External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. *Radiotherapy and Oncology*, *133*, 205-212.

6. 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.

7. 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.

8. 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.

9. Kalendralis, P., **Shi, Z.**, Traverso, A., Choudhury, A., Sloep, M., Zhovannik, I., ... & Klein, S. (2020). FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, Interobserver, Lung1 and Head-Neck1 TCIA collections. *Medical Physics*.

Published book chapter

10. **Shi, Z**., Fijten, R., Zhou, Z., Dekker, A., & Wee, L. (2019). Data Sharing and Toxicity Modelling: A Vision of the Near Future. In *Modelling Radiotherapy Side Effects* (pp. 365-399). CRC Press.

11. **Shi, Z.,** Wee, L., & Dekker, A. (2019). Cancer registry and big data exchange. *Big Data in Radiation Oncology* (pp. 155-180). CRC Press.

In preparation/submitted

1. **Shi, Z**., et al. Prediction of Lymph Node Metastases Using Pre-Treatment PET Radiomics of the Primary Tumour in Esophageal Adenocarcinoma: an External Validation Study. (**In preparation**)

2. **Shi, Z**., et al. Lung Organ Segmentation via Privacy-preserving Federated Deep Learning on FAIR Imaging Data. (**In preparation**)

3. Zhu, C., **Shi, Z.,** et al. Towards Federated Transfer Learning in Medical Imaging. (**In preparation**)

4. Zhai, T., Wesseling, F., Langendijk, J., **Shi, Z.**, et al. External validation of nodal failure prediction models including radiomics in head and neck cancer. (**submitted, oral oncology**)

5. Zhang, C., Fonseca, L., **Shi, Z.,** et al. Systematic review of radiomic biomarkers for predicting immunotherapy treatment outcomes. (**submitted, Methods**)

6. Nobel, M., **Shi, Z**., et al. Validation of Artificial Intelligence Tool Diagnosing COVID-19 on Screening Chest CT Scans: How Artificial Intelligence can help in a worldwide pandemic outbreak. (**In preparation**)

International conference abstracts/presentations (first author only)

1. External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patient. ESTRO 37 (**poster viewing**)

2. Ontology-guided Radiomics Analysis Workflow. European Congress of Medical Physics ECMP) 2018 (oral presentation)/Radiomics Conference 2019 (**oral presentation**)

3. A feature-pooling and signature-pooling method for feature selection for quantitative image analysis: application to a radiomics model for survival in glioma. MICCAI-RON-AI workshop 2019 (**oral presentation**)

4. CT-based Radiomics Predicting HPV Status in Head and Neck Squamous Cell Carcinoma. 7th ICHNO Conference (oral presentation) & ICCR 2019 (**oral presentation**)

5. Findable, Accessible, Interoperable, and Reusable (FAIR) Quantitative Imaging Infrastructure. ICCR 2019 (**oral presentation**)

6. Mortality Risk Stratification Model based on Radiomics Only: Analysis of Public Open Access HNC Data. ESTRO 2019 (**poster**)

7. Modelling Head and Neck Radiotherapy outcomes using radiomics biomarkers. ESTRO 2019 (**poster**)

8. Distributed Radiomics - a signature validation study using the Personal Health Train infrastructure. ICCR 2019 (**poster**)

9. Development and External Validation of a Prediction Model Incorporating PET Radiomics for Pathological Lymph Node Metastases in Esophageal Adenocarcinoma. ICCR 2019 (**poster**)