Artifical intelligence in rectal cancer

Artificial Intelligence in Rectal Cancer

Joost J.M. van Griethuysen

ISBN: 978-90-9035-285-5

Lay-out & Coverdesign: Willem Luijendijk
Printed by: Libertas Pascal

Artificial Intelligence in Rectal Cancer

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht, op gezag van de Rector Magnificus, Prof.dr. Rianne M. Letschert volgens het besluit van het College van Decanen, in het openbaar te verdedigen op vrijdag 29 oktober 2021 om 14:00 uur

door:

Joost Johannes Marijn van Griethuysen

Promotores

Prof. dr. R.G.H. Beets-Tan (Universiteit Maastricht / Antoni van Leeuwenhoek, Amsterdam)
Prof. dr. H.J.W.L. Aerts (Universiteit Maastricht / Brigham and Women's Hospital, Harvard Medical School, Boston, USA)

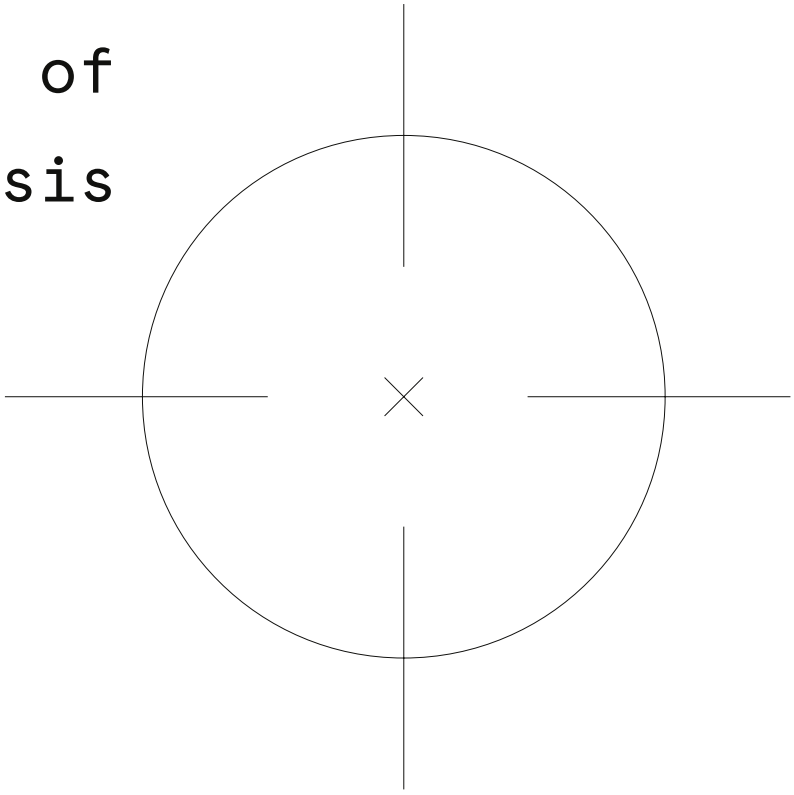
Copromotor

Dr. D.M.J. Lambregts
(Antoni van Leeuwenhoek, Amsterdam)

Beoordelingscommissie

Prof. dr. L.P.S. Stassen (voorzitter)
Prof. dr. P. Lambin
Prof. ir. P.M.A. van Ooijen (UMCG Groningen)
Dr. J.J. Visser (Erasmus MC Rotterdam)
Prof. dr. J.E. Wildberger

Introduction and aim of the thesis



Chapter 1 Introduction and aim of the thesis

Chapter 1

Medical imaging is a crucial element that is interwoven in every step of the diagnosis and treatment of cancer, from early screening and staging, to therapeutic response assessment and post-treatment follow-up, where imaging findings can have a great impact on patient management and consequent outcomes. In current clinical practice, assessment is primarily qualitative, meaning that imaging examinations – e.g. computed tomography (CT), magnetic resonance imaging (MRI) or ultrasound (US) images – are visually interpreted by radiologists who report their findings using text reports. The main tools available to radiologists to "classify" their findings are visual classification systems such as the Tumor Node Metastasis (TNM) guidelines from the American Joint Committee for Cancer (AJCC) and tumor-specific staging systems such as the Breast Imaging Reporting and Data System (BIRADS) for breast cancer and the Prostate Imaging Reporting and Data System (PIRADS) for prostate cancer.

However, there is much more information that can be extracted from medical images than meets the eye. Images are a collection of quantitative measurements, and consist of pixels which reflect the magnitude of some physical property in the small region covered by each pixel. Depending on the imaging modality used and the acquisition settings, multiple physical properties can be imaged, such as the tissue density on CT or the acoustic impedance on ultrasound. There are some simple "quantitative" tools that radiologists already commonly use to supplement their reports. Examples include tumor size measurements used in the Response Evaluation Criteria In Solid Tumors (RECIST) system for oncologic response assessment, or Hounsfield Units (HU) used to measure tissue density and enhancement characteristics on CT. However, these relatively simple methods of image quantification do not at all capture the full extent of information that is present in medical images. The magnitude and complex spatial distribution of pixel values within an image – commonly referred to as the image "texture" – can be used to study the underlying tissue architecture. In oncology, image texture has been shown to have the potential to assess intratumoral heterogeneity which in turn can be correlated to treatment outcomes and prognosis[1,2]. Though initial reports on image texture already date back to the early `70s[3,4], there has recently been a renewed interest for texture analysis in medical imaging, in particular oncologic imaging. The re-introduction of the technique has been boosted by several recent advances. The development of novel imaging techniques such as diffusion-weighted MRI (DWI) and dynamic contrast enhanced (DCE) imaging has opened up a new array of functional imaging parameters reflecting biological tissue properties such as cellularity and perfusion. In addition, the exceptional growth in computational power and data storage has led to the digitization of medical image storage, making huge quantities of data easily available for analysis, which has consequently given rise to widespread use of artificial intelligence (AI) and its subfield, machine learning.

In machine learning the computer uses algorithms to infer or 'learn' correlations from the data it is provided, aiming to predict a desired outcome and making optimum use of the large amounts of data and quantitative information provided by modern-day medical images. The process of extracting this information in the form of "features" to generate a radiographical tumor phenotype, and using this to construct predictive models using machine learning and AI is known as "radiomics"[5].

The process of radiomics comprises of four key steps, which are each associated with their

own specific challenges. The following sections discuss these four steps and highlight which challenges need to be overcome for radiomics to be successfully applied and implemented in oncologic imaging practice.

Key steps and challenges in the radiomics and AI workflow

1. Segmentation

Before quantitative features can be extracted from medical images, the computer needs to 'know' which part of the image represents the region of interest (ROI). In oncology, the ROI will typically be the tumor under investigation or a specific organ harboring a malignancy. The most common method to define the ROI within an image is to ask a reader (often an experienced radiologist) to manually delineate or "segment" it, for example by tracing the boundaries of a tumor lesion on a slice-by-slice level. This is, however, both a labor-intensive and time consuming task which can suffer from substantial inter- and intra-reader variation that can affect extracted feature values[6,7]. There is thus an urgent need for (semi-)automatic tools to reduce the workload for image segmentation and improve segmentation consistency.

2. Image acquisition and quality

Image quality in medical imaging can be influenced by many factors, including patient preparation, hardware/vendors and software, and protocols used for image acquisition. Radiomic studies that have been published so far have mainly been retrospective, meaning that they are based on the analysis of images that have been acquired as part of routine clinical practice. Though this can yield a much larger amount of data compared to prospective studies, the data is less homogeneous because imaging protocols will have typically not been standardized for the purpose of the study. The resulting heterogeneity in imaging can substantially affect the extracted radiomic feature values and the extent to which the images will be usable for automated analysis[8,9]. Even in case of (prospectively acquired) homogenized acquisition, day-to-day noise and artefacts can still severely impact the image quality, and subsequently the quality of extracted features[10].

3. Feature extraction

The next step in the workflow is to extract the radiomic features. Generally, several preprocessing steps are applied first, including resampling of the pixel spacing and normalization of the gray value intensities to reduce acquisition-related variations. Then, a large panel of features is extracted describing the histogram, shape and texture of the ROI[11]. Additional image filtration can be applied to emphasize certain aspects of the image[5,12]. Though many of the mathematical algorithms used to extract radiomics features are well known and documented, there is a lack of standardized implementations. In other words, no standardized tools or software packages exists to apply these algorithms including the necessary pre-processing steps to extract features from the imaging data. As a result, many studies that report on radiomics analysis use in-house developed software that lack transparency, which severely hampers the reproducibility and comparability of the reported results.

4. Feature selection, classification and modeling

The final step in the workflow is the analysis of the extracted features and the development of a model to correlate them to the outcome under investigation. In radiomics studies, this is done using a data-driven analysis, where the correlation of a large panel of tests (the radiomic features) to the desired outcome is inferred through a multistep process of selecting and classifying valid predictors.[13] An important pitfall of radiomics studies is that the number of available features (in the order of hundreds to thousands of features) tends to be much larger than the number of subjects in the dataset. This causes radiomics to suffer from "the curse of dimensionality", where the amount of features or dimensions is so large that combining them into a single model becomes very difficult[14]. To tackle this problem and prevent overfitting of the model, a smaller subset of features must therefore first be selected. This can either be done using "supervised" methods, where features that show good individual performance in univariate analysis are more or less cherry-picked, or "unsupervised", where a subset is selected based merely on the distribution of features in relation to each other. In addition, features can be selected on their "stability", meaning that features that are significantly affected by acquisition or segmentation induced variability (the "unstable features") are excluded. After selection of a final subset of candidate features a predictive model can be trained. To do so, a dataset is generally divided into several subsets, ideally including a training set to fit the model, a validation or "tuning" set to optimize the model settings controlling the training process, and a final test set to evaluate the performance of the developed model. Preferably, this test set should be an independent cohort of cases not used for model development[16], though in many of the small scale and preliminary reports published so far, such an independent test dataset is lacking.

Radiomics in rectal cancer

In this thesis, the concept and stepwise implementation of Radiomics and AI in oncology will be assessed using rectal cancer as a clinical case example. Colorectal cancer ranks 3rd for incidence and 2nd for cancer deaths worldwide, of which over 30% concern cancers of the rectum[17]. Traditionally, the standard treatment of rectal cancer has been radical surgery, preceded by neoadjuvant radiotherapy or chemoradiotherapy (CRT) in case of more locally advanced rectal cancers (LARC) to achieve downstaging and enhance the chance of a curative resection[18]. More recently, the concept of "organ-preservation" was introduced to the treatment landscape for rectal cancer. Small (early) tumors may be managed with less-invasive, local excision surgery and advanced tumors that show a very good response to neoadjuvant treatment may be managed with minimally invasive or even non-operative treatment strategies[19,20]. The latter, commonly referred to as "watch-and-wait" means that patients with a clinical complete response after neoadjuvant treatment are deferred from surgery and instead regularly monitored with clinical examination and imaging. With this strategy the rectum may be spared in up to 20%[21] of patients undergoing neoadjuvant treatment with very promising results in terms of long term oncological outcome and survival[22,23]. These developments in treatment have generated a new array of clinical questions where imaging may play an important role: how can we best select the patients that may benefit from these organ-preserving treatments? In other words, how can we best differentiate the high-risk from low-risk tumors and can we

predict how patients will respond to neoadjuvant treatment, so that the treatment plan may be optimized accordingly? Several studies have investigated the role of quantitative imaging analysis to predict response to chemoradiotherapy in rectal cancer. Parameters derived from dynamic contrast enhanced MRI[24,25] and diffusion weighted imaging[25–28] have shown promise as potential imaging biomarkers of response, albeit with varying results[29]. Only a limited number of studies have investigated the potential of radiomics in rectal cancer, so far mainly focusing on the assessment of response after completion of neoadjuvant therapy[30–32]. In addition to these previous works, this thesis will focus on investigating the potential of radiomics to predict response to neoadjuvant therapy upfront, using MRI data acquired prior to the start of treatment.

Aim of this Thesis

The overall goal of this thesis is to address the key steps and challenges in the radiomics workflow described above, aiming to bring radiomics closer to implementation in clinical practice. We will do so by using the clinical case example of rectal cancer imaging that can serve as a blueprint for future studies investigating applications of radiomics and AI in other (oncologic) imaging fields.

Outline of this Thesis

Chapters 2 and 3 address the challenge of image segmentation. Chapter 2 focuses on the development, training and testing of a deep learning network for fully automatic segmentation of rectal tumors on MRI. In Chapter 3 this algorithm is further optimized and validated in a large multicenter dataset. Additionally, this chapter investigates the influence of scan quality and tumor complexity on automatic segmentation performance.

Chapter 4 focuses on scan-quality – an important prerequisite for radiomics and AI analysis – and evaluates how the quality of diffusion-weighted MR imaging of the rectum may be improved by applying a preparatory micro-enema shortly prior to acquisition to reduce the presence of gas-induced susceptibility artefacts.

In Chapter 5 we developed an open-source python software package for easy, transparent and reproducible radiomics feature extraction.

Finally, Chapter 6 puts the radiomics model to the test in a clinical paper investigating the potential of radiomics for the pre-therapy prediction of response to neoadjuvant chemotherapy in rectal cancer, based on primary staging MRI data.

1. Court LE, Rao A, Krishnan S. Radiomics in cancer diagnosis, cancer staging, and prediction of response to treatment. *Transl Cancer Res.* 2016; 5(4): 337-339.

2. Bae S, Choi YS, Ahn SS, et al. Radiomic MRI phenotyping of glioblastoma: Improving survival prediction. *Radiology.* 2018; 289(3): 797-806.

3. Haralick R, Shanmugan K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973; 3: 610-621.

4. Galloway MM. Texture analysis using gray level run lengths. *Comput Graph Image Process.* 1975; 4(2): 172-179.

5. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5(1): 4006.

6. Velazquez ER, Parmar C, Jermoumi M, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep.* 2013; 3: 3529.

7. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, et al. Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol.* 2016; 94(4): 824-831.

8. Zhao B, Tan Y, Tsai W-Y, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep.* 2016; 6(1): 23428.

9. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol (Madr).* 2017; 56(11): 1544-1553.

10. Tixier F, Hatt M, Cheze-Le Rest C, Le Pogam A, Corcos L, Visvikis D. Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET. *J Nucl Med.* 2012; 53(5): 693-700.

11. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging.* 2012; 30(9): 1234-1248.

12. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology.* 2015; 278(2): 563-577.

13. Maass W, Parsons J, Purao S, Storey VC, Woo C. Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *J Assoc Inf Syst.* 2018; 19(12): 1253-1273.

14. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep.* 2015; 5: 13087.

15. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. *J Mach Learn Res.* 2003; 3: 1157-1182.

16. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017; 14(12): 749-762.

17. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018; 68(6): 394-424.

18. Landelijke Werkgroep Gastro Intestinale Tumoren. Richtlijn Colorectaal Carcinoom (versie 3.0). www.oncoline.nl.

19. Habr-Gama A, Perez RO, Nadalin W, et al. Operative versus nonoperative treatment for stage 0 distal rectal cancer following chemoradiation therapy: long-term results. *Ann Surg.* 2004; 240(4): 711-718.

20. Maas M, Beets-Tan RG, Lambregts DM, et al. Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. *J Clin Oncol*. 2011; 29(35): 4633-4640.

21. Maas M, Nelemans PJ, Valentini V, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol.* 2010; 11(9): 835-844.

22. Martens MH, Maas M, Heijnen LA, et al. Long-term Outcome of an Organ Preservation Program After Neoadjuvant Treatment for Rectal *Cancer. J Natl Cancer Inst.* 2016;108(12): djw171.

23. van der Valk MJM, Hilling DE, Bastiaannet E, et al. Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the International Watch & Wait Database (IWWD): an international multicentre registry study. *Lancet.* 2018; 391(10139): 2537-2545.

24. Martens MH, Subhani S, Heijnen LA, et al. Can perfusion MRI predict response to preoperative treatment in rectal cancer? *Radiother Oncol.* 2015; 114(2): 218-223.

25. De Cecco CN, Ciolina M, Caruso D, et al. Performance of diffusion-weighted imaging, perfusion imaging, and texture analysis in predicting tumoral response to neoadjuvant chemoradiotherapy in rectal cancer patients studied with 3T MR: initial experience. *Abdom Radiol (New York).* 2016; 41(9): 1728-1735.

26. Hötker AM, Tarlinton L, Mazaheri Y, et al. Multiparametric MRI in the assessment of response of rectal cancer to neoadjuvant chemoradiotherapy: A comparison of morphological, volumetric and functional MRI parameters. *Eur Radiol.* 2016; 26(12): 4303-4312.

27. Lambrecht M, Vandecaveye V, De Keyzer F, et al. Value of diffusion-weighted magnetic resonance imaging for prediction and early assessment of response to neoadjuvant radiochemotherapy in rectal cancer: preliminary results. *Int J Radiat Oncol Biol Phys.* 2012; 82(2): 863-870.

28. Chen Y-G, Chen M-Q, Guo Y-Y, Li S-C, Wu J-X, Xu B-H. Apparent Diffusion Coefficient Predicts Pathology Complete Response of Rectal Cancer Treated with Neoadjuvant Chemoradiotherapy. *PLoS One.* 2016; 11(4): e0153944.

29. Pham TT, Liney GP, Wong K, Barton MB. Functional MRI for quantitative treatment response prediction in locally advanced rectal cancer. *Br J Radiol.* 2017; 90(1072): 20151078.

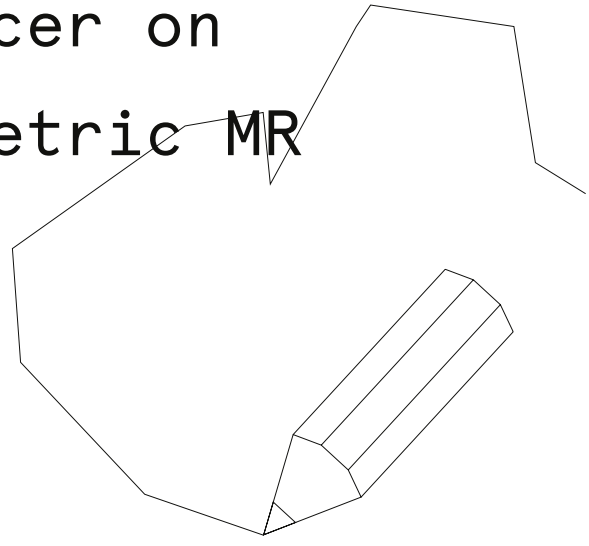
30. Liu Z, Zhang X-Y, Shi Y-J, et al. Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. *Clin Cancer Res.* 2017; 23(23): 7253-7262.

31. Nie K, Shi L, Chen Q, et al. Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. *Clin Cancer Res.* 2016; 22(21): 5256-5264.

32. Li Y, Liu W, Pei Q, et al. Predicting pathological complete response by comparing MRI-based radiomics pre- and postneoadjuvant radiotherapy for locally advanced rectal cancer. *Cancer Med*. 2019; 8(17): 7244-7252.

Segmen-
tation

Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR



Chapter 2 Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR

Stefano Trebeschi*, Joost J.M. van Griethuysen*, Doenja M.J. Lambregts, Max J. Lahaye, Chintan Parmar, Frans C.H. Bakers, Nicky H.G.M. Peters, Regina G.H. Beets-Tan, Hugo J.W.L. Aerts

* Shared first author

Published in:

Scientific Reports 2017; 7:5301 (IF 2019 3.998)

Abstract

Multiparametric Magnetic Resonance Imaging (MRI) can provide detailed information of the physical characteristics of rectum tumors. Several investigations suggest that volumetric analyses on anatomical and functional MRI contain clinically valuable information. However, manual delineation of tumors is a time consuming procedure, as it requires a high level of expertise. Here, we evaluate deep learning methods for automatic localization and segmentation of rectal cancers on multiparametric MR imaging. MRI scans (1.5T, T2-weighted and DWI) of 140 patients with locally advanced rectal cancer were included in our analysis, equally divided between discovery and validation datasets. Two expert radiologists segmented each tumor. A convolutional neural network (CNN) was trained on the multiparametric MRIs of the discovery set to classify each voxel into tumor or non-tumor. On the independent validation dataset, the CNN showed high segmentation accuracy for reader1 (Dice Similarity Coefficient (DSC=0.68) and reader2 (DSC=0.70). The area under the curve (AUC) of the resulting probability maps was very high for both readers, AUC=0.99 (SD=0.05). Our results demonstrate that deep learning can perform accurate localization and segmentation of rectal cancer in MR imaging in the majority of patients. Deep learning technologies have the potential to improve the speed and accuracy of MRI-based rectum segmentations.

Chapter 2

Introduction

Magnetic Resonance Imaging (MRI) is an integral part of the diagnostic work-up of rectal cancer and plays an important role in treatment planning. In addition, MRI can play a role in predicting clinically relevant endpoints, one of the most important ones being the response to neoadjuvant treatment[1–3]. Predicting which patients will show a very good response to treatment can have important clinical implications, since these patients may be considered for organ-preserving treatment strategies (local excision or watchful waiting) as an alternative to standard surgical resection[4]. In carefully selected patients these organ preserving treatments can considerably improve quality of life with a good oncological outcome.

A promising technique to assess response to neoadjuvant treatment is diffusion-weighted MRI (DWI). Various studies have shown that – as an addition to standard morphological MRI – DWI can aid in assessing response to chemoradiotherapy, in particular to differentiate residual tumor within areas of post-radiation fibrosis after CRT. For this purpose use of DWI is now even recommended in international clinical practice guidelines for rectal cancer imaging[4].

Particularly good results have been shown for volumetric measurements derived from diffusion-weighted imaging (DWI)[5–8]. Furthermore, ADC and histogram features derived from DWI-MRI have shown promise as quantitative imaging biomarkers for therapeutic outcome[4,5,9,10]. Most of these measures are calculated from regions of interest (ROI) of the tumor that are typically obtained after manual tumor segmentation by experienced readers. Studies have indicated that whole-volume tumor segmentations, as opposed to single slice or sample measurements, provide the most reproducible and accurate estimates of the true tumor volumes[11,12]. The main problem with manual segmentation approaches is that these are highly time consuming, up to 18 minutes per tumor[13], and as such unlikely to be implemented into daily clinical practice. Previous studies have explored ways to automatically perform segmentations using software algorithms[6,13]. These approaches work best on diffusion-weighted images, as these highlight tumor and suppress background tissues, thereby providing a high tumor-to-background ratio.

Unfortunately, high signal on DWI is not limited to tumor tissue only. Other anatomical structures in the pelvis (e.g. perirectal lymph nodes, prostate and ovaries) as well as artefacts may also show similar hyper-intensity and may not be recognized as such by typical simple segmentation algorithms causing these algorithms to fail to produce sufficiently accurate results[13]. In such cases, the manual input required from an experienced reader will not be limited to a threshold value or a seed point (in case of region growing), but will include manual corrections to adjust the segmentations for these effects[13]. Thus, there is an obvious need for smarter algorithms that can automatically localize and perform accurate segmentations of rectal tumors, which can reduce the need of expert input (Figure 1).

Such fully automatic alternatives would also facilitate the generation of segmentations for large cohort studies, which is beneficial especially in light of new research developments such as Radiomics[14,15], where complex tumor phenotypical characteristics are quantified and correlated to diagnostic or prognostic factors. The computation of these features requires input in the

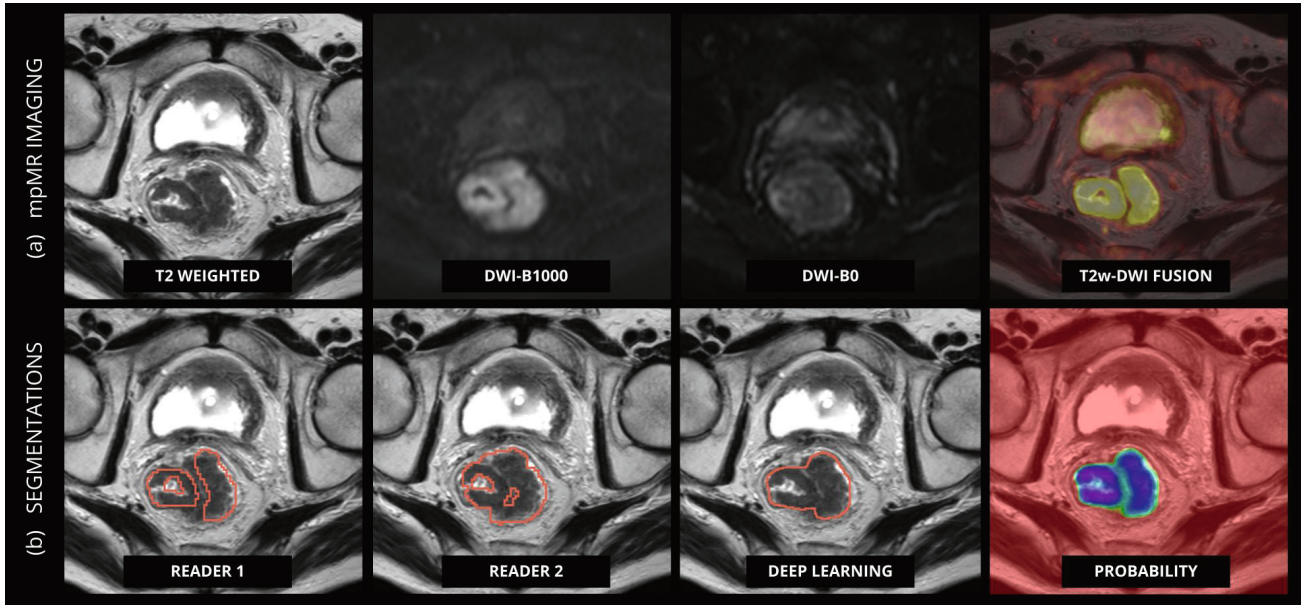


Figure 1. Example of Multiparametric MR in a rectal cancer patient. *mpMR of the pelvis of a male patient with rectal cancer before the start of the treatment. Corresponding slices of different sequences on the transversal plane are shown. (a) The sequences are, in order: T2 weighted, DWI b1000, DWI b0 and fusion imaging between T2 weighted and the DWI b1000. Notice how anatomical structures and tissues surrounding the tumor – such as prostate, bladder, and seminal vesicles – and artefacts in general show the same hyper-intensity on the DWI of the tumor. (b) Delineations of the tumor done by (from the right left hand side): the experienced reader used for the training, the independent reader, the result of the algorithm and the corresponding probability map generated by the algorithm.*

delineation of the region of interest to be described. Artificial intelligence (AI) aims to mimic cognitive, labor intensive tasks via complex computational models trained on top of existing datasets. A computational model trained using the input from expert readers (radiologists) to automatically localize and segment rectal cancer in MR images, could represent a potential solution to this problem.

Novel AI technologies, such as deep learning models, have been exploited in recent years with impressive results. Convolutional neural networks (CNNs) based deep learning approaches can learn feature representations automatically from the training data. The multiple layers of the CNNs aim to process the imaging data with different levels of abstractions, enabling the machine to navigate and explore large datasets and discover complex structures and patterns that can be used for prediction[16]. The advancement of these techniques has been made possible by the availability of large imaging data and the accessibility of dedicated hardware devices such as graphical processing units (GPU)[15,16]. Particularly in the field of biomedical imaging, deep learning has been largely exploited for detection and segmentation purposes, where these methods are proven to systematically outperform traditional machine learning techniques[17–19].

In this study, deep learning methods (CNNs) have been used to fully automatically localize and segment rectum tumors. To evaluate the performance of deep learning based segmentations, we compared them to manual segmentations of two independent expert radiologists. Deep learning technologies have the potential to improve the speed and accuracy of MRI-based rectum segmentations in clinical settings.

Background work

To the best of our knowledge, few investigations were conducted on the automatic localization and segmentation of rectal cancer. Irving et al.[20] proposed an automatic segmentation procedure, based on DCE-MRI, where the authors accounted for the multidimensional nature of DCE signal through a modified version of the supervoxel algorithm corrected by a graphical model producing successful results. Although DCE-MRI tends to give a much clearer and less noisy signal compared to DWI, our method achieved comparable results to the one presented in this study. The most popular semi-automatic approach is region growing. Day et al.[21] used region growing on FDG-PET on phantoms, leading to better results than thresholding of the standardized uptake value (SUV). In this case the intent of the authors was to optimize treatment planning. Region growing was also used by van Heeswijk et al.[13], who concluded that it could represent a more convenient replacement for manual delineation in terms of time. Although the results showed a decrease in the amount of time required, manual input was still required and a DSC > 0.7 could only be achieved when the result of the region growing was adjusted by an experienced radiologist.

Table 1. Patient Characteristics

	Center A	Center B	Both Centers	p-value
N	91	49	140	-
Males / Females	66 / 25	31 / 18	97 / 43	p = 0.498, χ^2 test
Age	66.6 ± 9.3	65.6 ± 9.8	66.2 ± 9.4	p = 0.554, t-test
Tumor Volume [α]	19.0 ± 22.3 cm^3	23.8 ± 29.3 cm^3	20.7 ± 25.0 cm^3	p = 0.321, t-test

(α) according to the segmentation performed by the experienced reader.
No significant difference has been found between the two centers.

Materials and methods

Subjects and Study Dataset

For this study we retrospectively selected 140 patients (97 males, median age 67, range 43 - 87) with biopsy proven locally advanced rectal carcinoma (LARC) from a previously reported bi-institutional study cohort[5]. No significant difference in clinical parameters was observed between the two centers (see Table 1). All patients in this cohort have undergone multiparametric (mp) MRI, consisting of T2 weighted and diffusion weighted imaging (DWI), prior to standard chemo-radiotherapy treatment (CRT), using either an Intera (Achieva) or Ingenia scanner (Philips Healthcare, Best, The Netherlands) (center A, 91 patients) or a Magnetom Avanto system (Siemens Healthcare, Forchheim, Germany) (center B, 49 patients) with a phased array surface coil. Both T2w and DWI sequences were axially angled perpendicular to the tumor axis defined on a sagittal scan. The diffusion sequence was performed using b-values b0, b500, and b1000 (center A) or b1100 (center B). Patients did not receive bowel preparation. As described previously, all research was performed according to guidelines and regulations of The Netherlands[5]. In short, according to the Dutch law, retrospective studies are not subject to the Medical Research Involving Human Subjects Act and informed consent is not required[22]. Detailed parameters of the sequences are specified in Table 2.

Whole-volume tumor segmentations were available for all patients and were done by an experienced reader (Reader 1, DMJL) on the highest b-value (b1000 or b1100) DWI, according to methods previously reported[13], where the reader created an initial segmentation using a simple region growing algorithm and manually adjusted to fit the borders of the tumor. These segmentations were used as ground truth. Additionally, segmentations performed on the same dataset and in the same manner by an independent reader (Reader 2, MJL) were retrieved. These segmentations were used as additional check.

Table 2. Sequence parameters of the diffusion-weighted imaging used during the study period.

	Center A			**Center B**	
Repetition Time	4004 – 4829	4971	4172 – 5241	5100	4300
Echo Time	70	70	68 – 70		79
Number of Slices	50	24	20 – 24	34	34
Slice Thickness (mm)	5	5	5	5	6
Slice Gap (mm)	0.5	0.5	0.5	0.5	0
In-Plane Resolution	2.50 × 3.11 (- 3.18)	1.87 × 2.31	1.82 × 2.27	1.70 × 1.30	2.0 x 2.0
Echo train length	1	1	1	1	1
N. Signal Averages	4	5	5	6	6
b-values	0, (100), 500, 1000	0, 500, 1000	0,(25,50,100),500,1000	0, 500, 1000	0,300,1100
Fat Suppr. Tech.	STIR	SPIR / fatsat	SPAIR	SPIR / fatsat	SPIR / fatsat
Echo Planar Im.	53 – 55	55	61	148	150

Fatsat: Fat Saturation, SPIR: Spectral Attenuated Inversion Recovery,
SPAIR: Spectral Pre-saturation with Inversion Recovery, STIR: Short T1 Inversion Recovery.

Imaging data of 140 patients were included in our analysis. Patients were assigned to discovery or validation dataset depending on their identifier: even numbers were assigned to discovery dataset (N=70), odd numbers to validation dataset (n=70). For the discovery dataset, there were no errors within the imaging data and segmentations, and therefore 70 cases were used for training. For the validation dataset, three cases had to be excluded due to misalignment between DWI and T2 caused by an error in the DICOM metatags, one case where the DWI suffered from severe ghosting artefacts, and one case in which the segmentation file was corrupted. This resulted in 65 cases that could be used for validation.

Pre-processing.

All images underwent standardization of the intensities, namely the intensity distribution was set to have mean zero and standard deviation one. Deformable registration was applied using the elastix toolbox[23,24] to compensate for the anatomical displacement of organs and tissues in different imaging sequences during the acquisition procedure. The DWI-b0 was used as reference image, since it visualizes anatomical structures like the T2w and, at the same time, is well aligned to the DWI-b1000. The deformation field was estimated via adaptive stochastic gradient descent[25] minimizing the advanced mattes mutual information[26]. Transform bending energy[27] was used as penalty measure to correct for anatomically unrealistic transformations. To properly simulate the small, local movements in the bowels, a dense sampling grid of 4 mm together with a strong weight on the penalty measure (1 : 20) was applied.

Deep Learning (CNN) architecture.

In this study, a CNN architecture was implemented to function as voxel classifier. More specifically, for each voxel *v* we (1) extracted a fixed-size patch surrounding *v*, (2) classified the patch via a trained instance of the CNN, (3) collected the resulting probability, and (4) assigned the resulting probability to *v*. By repeating the procedure for each voxel of each image, we could generate a probability map, where *p(v)* is the probability of voxel *v* to represent tumor tissue. The segmentation was generated by thresholding of the probability maps (voxels with *p(v)* ≥ 0.5 were classified as "tumor", and as "not tumor" otherwise) and subsequent selection of the largest connected component. Figure 2 offers a schematic synthesis of the whole process.

Patch extraction

N voxels were randomly sampled from each of the foreground (i.e. tumor region) and background (i.e. non tumor region) regions. This ensured a balanced representation of the two classes during the training procedure. For each voxel, we extracted the surrounding in-plane patch of size M×M in all MR sequences. Each sequence was then fitted in one of the three channels of a standard RGB picture. The ground truth associated with each patch was the label of the central voxel.

Some regions are easier to classify, such as hypointensity on the DWI. Other regions instead are more challenging: for example, the prostate and the tumor regions have similar intensity and heterogeneity on T2w and DWI. Intuitively, the classifier should be able to spend more time learning to how to correctly classify these difficult regions of the image, and less time on regions that are easily classifiable. To translate this concept in implementation, the background class was divided in three regions: I) the area surrounding the tumor (R^B_1) defined by morphological dilation of the tumor segmentation with a spherical structural element of 1 cm radius; II) the regions hyper-intense on the DWI (R^B_2) defined by thresholding on the DWI at $\mu + 2\sigma$. Since all images have been standardized, this operation will result on the thresholding at a value of 2.00; and III) the remaining areas (R^B_3) defined by the voxels not belonging to either R^B_1 or R^B_2. We sampled N/4 voxels from R^B_1 and R^B_3, and N/2 voxels from R^B_2, summing up to total N non-tumor voxels from the background class. Figure 2a shows a schematic representation of the sampling process.

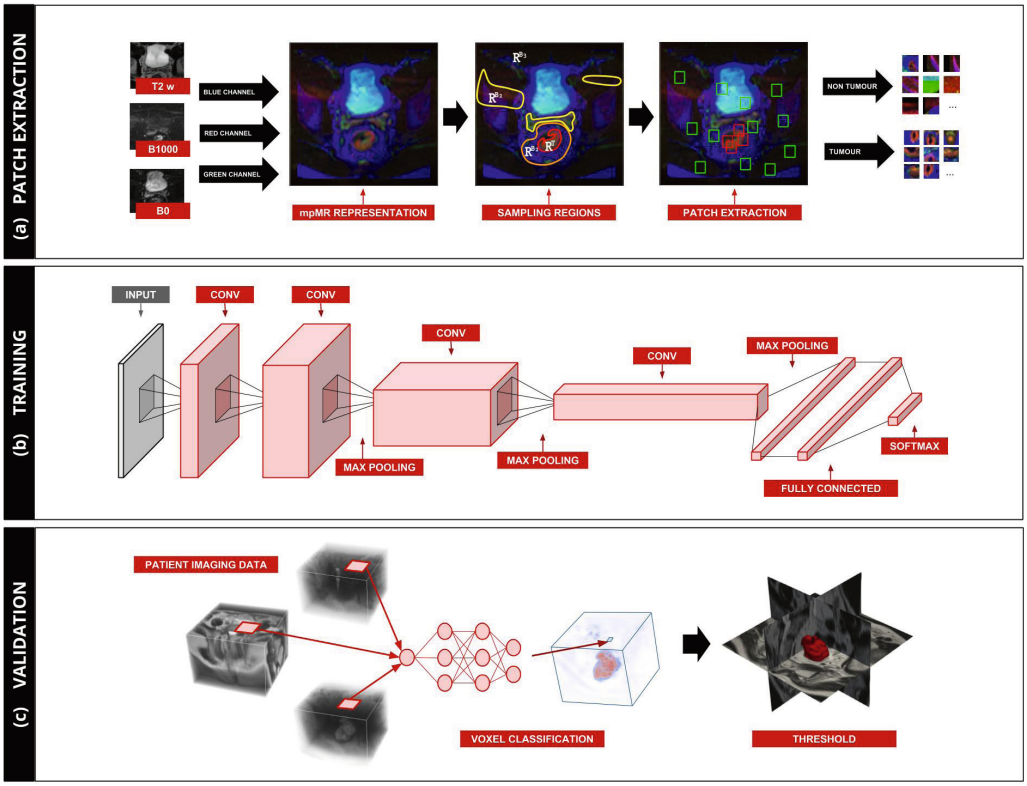


Figure 2. Scheme of the Proposed Solution. *(a) On the left-hand side, a multiparametric representation of the imaging is created via fusion of corresponding slices from different sequences into the color channels of the RGB model. In the center, the label map is first divided in tumor region (R^T) and the background regions (R^B) according to the delineation done by the experienced reader. On the right-hand side, N voxels (together with their surrounding patch) are then randomly sampled from these regions to maintain a balance between number of voxels representing the tumor and number of voxels representing healthy tissue. (b) The architecture of the network, which is trained with the patches of the images in the discovery set. The patches of the images in the test set are used to control for model overfitting. (c) The 3D probability map is generated by classification of each voxel using the trained model. The probability map is thresholded to find the components where the probability of tumor is higher than the probability of healthy tissue. The largest component is selected as segmentation of the tumor.*

Network Definition

The network used for the classification of each patch is composed of a total of nine layers: two subsequent convolutional layers followed by a max pooling layer; a couple of smaller, subsequent convolutional layers, each followed by a max pooling layer; two fully connected layers at the end culminating in the output layer. This architecture is similar to the one proposed in[28], with the addition of a convolutional and max pooling layer. Dropout[29] of 1/2 was used after each fully connected layer and 1/3 after each max pooling layer[30]. Leaky rectified linear unit (ReLu)[31] was used as a nonlinearity in each layer, except on the output layer, where a softmax was used instead. Small filters of 5×5 or 3×3 are used at each convolutional layer; along with stride one and full padding. Stride two was used in the max pooling. Twenty-four features were used in the first convolutional layer, number which doubles in each subsequent layer (i.e. 24, 48, 96, 192), amounting for a total of 360 filters throughout the entire network. Figure 2b shows a schematic representation the network structure. Cross entropy was used as cost function, together with a small L2 regularization on the network parameters. Adadelta[32] with learning rate $\eta = 0.001$ and decay $\rho = 0.9$.

The discovery set was divided into training set (80%) and test set (20%). The training set was used for training the net, the test set was used alongside the training procedure to check for model overfitting. The training procedure was programmed to stop when no improvement on the cost δ of the test set was made for at least five consecutive epochs, where the improvement was defined as $\delta_{EPOCH-1} - \delta_{EPOCH} > 10^{-3}$. The implementation of the algorithm was based on popular Python libraries: Lasagne and Theano[33].

Statistical Analysis

Segmentations were generated by feeding the patch of each voxel of mpMR to the algorithm and assigning the resulting probability to that voxel. The segmentation was generated by thresholding of the probability map at $p = 0.5$. Additional selection of the largest component allowed the exclusion of small isolated voxels, which might pass unseen by most human readers. Figure 2c shows an example of this process.

In stochastic processes where samples are randomly selected, the result might often be not representative, with a significant variance in the final classification results. To evaluate the stability of the algorithm, we repeated the entire sampling, training and testing procedure four times and compared the segmentations generated by the different runs of the algorithm.

Results

Deep Learning (CNN) Training

To develop a deep learning based algorithm for the fully automatic localization and segmentation of rectum tumors, we used independent discovery and validation datasets to develop a CNN-based network and validate its performance. The CNN was trained on multiparametric MR imaging (1.5T, T2-weighted and DWI) of 70 patients, using the segmentations performed by expert reader 1. For each patient, 5000 patches (size M×M = 21×21 voxels) were created by combining T2-weighted, and DWI images, for both tumor and non-tumor areas (Figure 2a). The discovery data consisted of an independent discovery set (totalling 560K patches) and test set (140K patches). The algorithm reached a loss on the discovery set of 0.275 and 0.331 on the test set. The accuracy was 0.895 and 0.871, respectively. Figures 3a, 3b and 3c show the improvement of accuracy, the minimization of the cost function and its improvement over time. Notice from the graph presented in Figure 3c that no major improvement on the cost function has been recorded after the 50th epoch.

Validation of CNN classifier

The performance of the CNN classifier was validated on the validation dataset consisting of multiparametric MR imaging of 65 patients. For each patient volumetric tumor segmentations were generated and compared to both expert readers (Figure 1b). Three cases had to be excluded where there was no agreement between expert readers. Therefore, data of 62 patients were used to validate the performance of the classifier. For all cases, the CNN could successfully generate volumetric tumor segmentations. To evaluate the performance of the CNN on a voxel-by-voxel basis, the area under the curve (AUC) was computed between the CNN probability maps and the segmentations of the experienced readers. The AUC of the resulting probability maps was very high for both readers, AUC = 0.99 (0.05 SD), with no significant difference between readers. Figure 3d shows AUC distributions for both readers.

From the probability maps we then generated volumetric segmentations. To evaluate the performance of the segmentation *the Dice Similarity Coefficient* (DSC) was used. The DSC is a statistical measure of spatial overlap frequently used to compare segmentations. The average DSC between the two expert readers was high (0.83, SD=0.13). The DSC between the algorithm and Reader 1 was 0.68 (0.07 SD) and Reader 2 was 0.70 (0.07 SD), with no significant difference detected between the two distributions ($p = 0.31$, *t-test*). Figure 3e shows DSC distributions between the algorithm and each reader, and between the two readers. Figure 1b shows an example of a tumor successfully delineated by the algorithm (0.99 AUC, 0.85 DCS). Notice that the CNN was trained on segmentations of expert reader 1 in the discovery dataset. However, on the validation dataset the performance of the CNN was similar with both readers (Figure 3e), demonstrating the generalizability of the network.

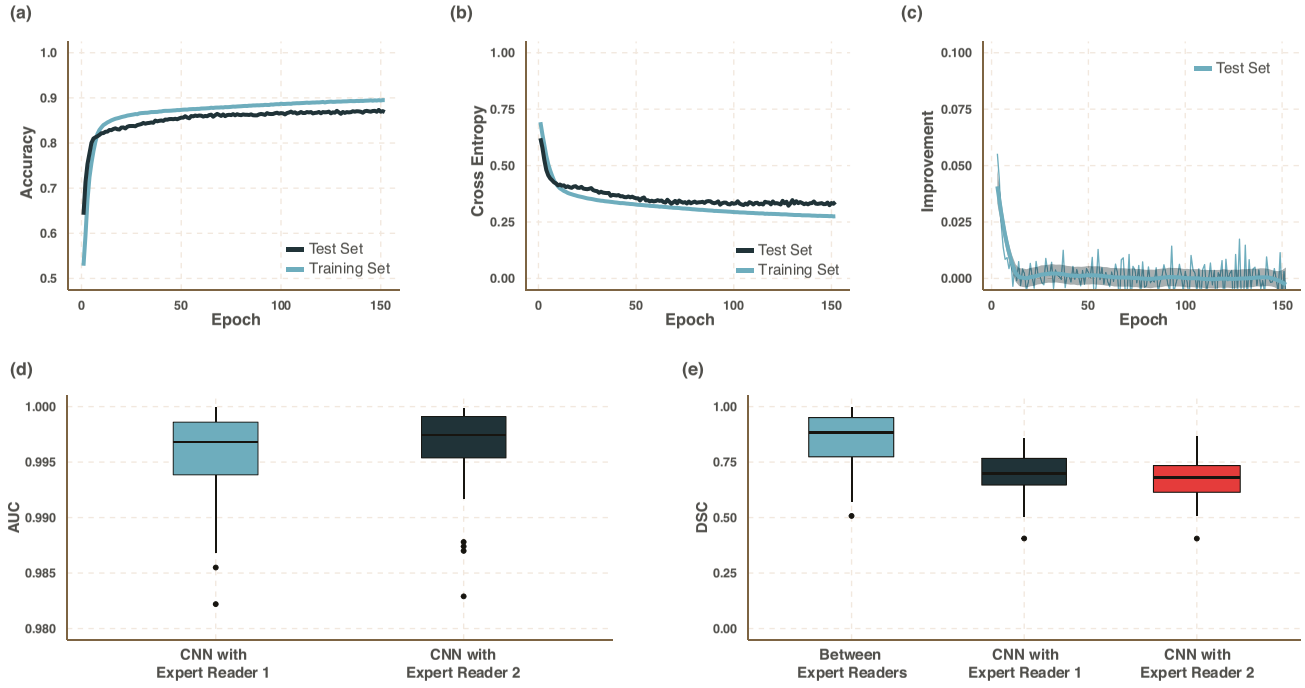


Figure 3. CNN Training and Validation. *Performance of the CNN on the discovery dataset: (a) accuracy, (b) cross entropy and (c) improvement (Δ cross entropy). Improvement shown in panel (c) is computed on the test set only, preventing the model from overfitting. Performance of the CNN on the validation dataset: (d) the Area under the ROC curve (AUC) of the probability map with respect to the reader segmentation, and (e) Dice Similarity Coefficient (DSC) of the generated segmentations.*

The algorithm resulted in a relatively poor result (DSC < 0.50) in ten cases. Figure 4 shows an example of a tumor correctly classified by the algorithm but resulting in a poor DSC after thresholding and selection of the biggest connected component (0.98 AUC, 0 DSC), as the tumor was not the biggest component. In this case, the testicles (visualized in the lower part of the image) were assigned a probability greater than 0.5, enabling them to survive the threshold procedure and be selected as candidate segmentation.

Stability of the Sampling Process.

To access the stability of the sampling procedure and reproducibility of the model across different sampled voxels, the entire discovery and validation procedure was repeated additional four times. The final validation accuracy resulting from each individual training procedure (between 0.875 and 0.895), as well as the cross entropy (between 0.268 and 0.30), was stable. Each trained algorithm was used to generate the segmentations, resulting in four segmentations for each case. The *Intraclass Correlation Coefficient* (ICC) was used to access the agreement across different nets in terms of DSC. The overall agreement was very high (ICC = 0.83, 95% CI 0.77 – 0.88, $p < 0.001$).

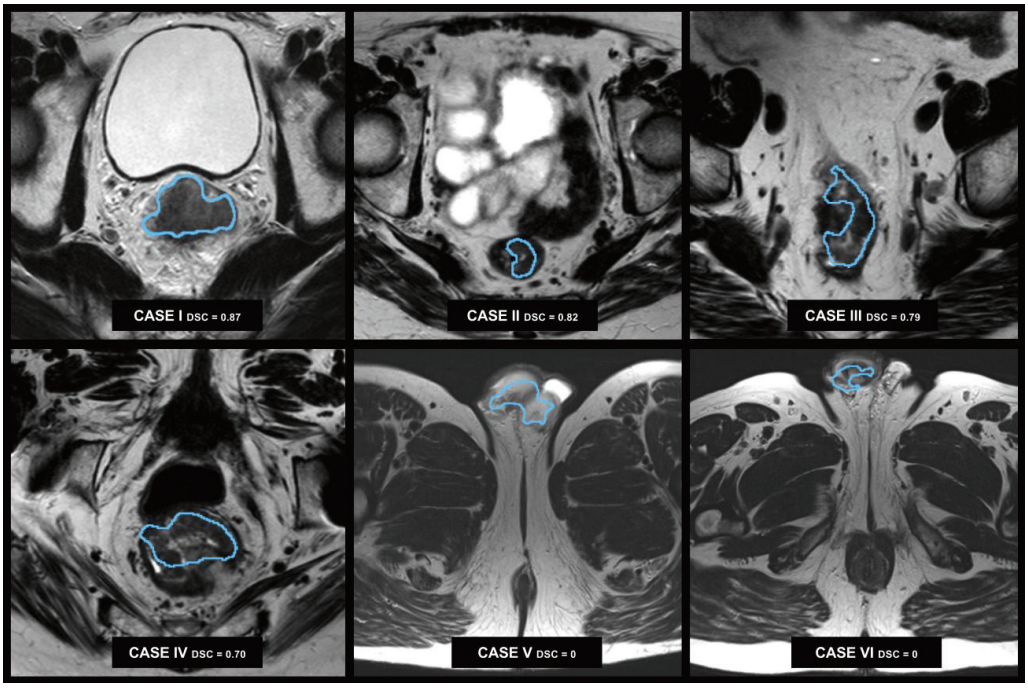


Figure 4. Example cases. *Six example cases of the segmentation performed by the CNN. The algorithm correctly localized and segmented the tumor in case I to IV (small FOV images), but failed in cases with larger FOVs (cases V and VI) where parts of the cavernous bodies of the penis were erroneously included in these examples.*

Discussion

Our aim was to develop a deep learning based network for the fully automatic localization and segmentation of locally advanced rectal tumors. Overall results show good performance of the algorithm, with segmentations comparable to those performed manually by an expert reader with a DSC of 0.70. In terms of classification, the high AUC of 0.99 suggests the ability of the algorithm to properly classify tumor voxels and therefore locate cancer tissue in the image. At visual inspection of the probability maps, one can appreciate how non-malignant hyperintensity of the DWI are attenuated with lower probabilities, whereas the tumor retains higher values.

After thresholding and selection of the largest component as candidate segmentation, the algorithm achieved an overall DSC of 0.70. At first glance, this is lower than the DSC reached between both readers on this dataset. In this case however the reader could rely on a semi-automatic procedure (region growing) known to increase the DSC[13]. Fully manual segmentation is reported to provide a lower DSC of 0.68[13], leaving open the question of whether the region growing algorithm could really provide a much more precise delineation or if the experienced readers were partially influenced by the result of the algorithm. Interesting enough, DSC variability was lower than the variability between experienced readers (0.07 vs 0.13 SD).

During the training of the CNN classifier, the algorithm showed decreased performance in 10 cases of the validation dataset. In each of these cases the AUC was > 0.90 but the DSC was zero, suggesting the algorithm managed to identify the tumor tissue in the image but failed to select the correct candidate. Seven cases out of ten were images acquired at the center B, which applied an imaging protocol with a larger field of view (FOV). The larger FOV inevitably included in the images large chunks of subcutaneous adipose tissue, or anatomical parts – e.g. the testicles –, which were not present in the discovery set (Figure 4). Although the tumor region resulted in higher probability voxels, after thresholding these adipose tissue or anatomical parts were larger than the tumor and therefore selected as candidate. Including more examples from center B will likely enable the network to learn to recognize and remove artefacts in these peripheral areas. Figure 4 shows two example cases of a male patients from center B, where the testicles were misclassified as tumor.

The remaining three cases from center A showed large fat suppression artefacts, rarely present in the rest of the dataset. Most likely, the scarcity of these examples is the reason of misclassification. This indeed represents the main drawback of supervised learning procedures in general, which are often unable to properly classify underrepresented cases. The same effect can be observed on a microscale in Figure 1, where the segmentation generated by the algorithm includes non-tumoral hyperintensities, most likely fecal matter. Given that all patients underwent preparatory enema before image acquisition, fecal matter is rarely present and only in small quantities. Its under representation in the training set, and the vicinity to the tumor leads the algorithm to assign a tumor probability > 0.5 – yet smaller than the probability assigned to the tumor.

Deep learning has been largely used before for segmentation tasks in medical imaging. Out of all possible architectures, we chose this for its straightforwardness and, most importantly, the limited number of images required for training the algorithm. The strategy of using multiple patches from the same patient allows us to generate a large imaging tensors upon which the algorithm can be trained. The patch size chosen allows focusing on a small region without including too much surrounding, but can still generalize textural patterns of specific tissues and organs. Larger patches in fact (e.g. M = 35) as well as smaller patches (e.g. M = 11) resulted in higher training error. These small patches, however, might not provide enough anatomical information needed in some other applications. In Figure 4, for example, we can see how the algorithm selected a group of voxels outside the pelvis. Fully convoluted end-to-end procedures, such as the one presented in SegNet[34] or U-Net[35,36] where 2D slices of MR volumes or entire 3D volumes are fed to a network able to directly generate the target segmentation, would represent an alternative approach worth investigating. Such approach would recognize unlikely tumor locations outside or on the border of the pelvic area, and exclude them automatically. The patch based approach adopted in this study aims to provide the network with an artificially balanced, multiparametric training set from a relatively small dataset via a weighted sampling procedure favoring more challenging regions such as tumor borders and diffusion hyperintensities often found in nearby prostate and seminal vesicles.

Neuroradiology remains the main area of focus of research on segmentation algorithms, and several algorithms and architectures have been proposed for brain tumor segmentation, especially in the context of multiparametric imaging[37–39]. Imaging of the lower abdomen however poses a challenge to automatization, partially because of the bowel movements, which makes it challenging to use voxel-wise mpMRI, and partially because of the high number of artefacts. Although bowel movements can be partially attenuated by deformable registration protocols, such as the one designed in this study, the result is yet suboptimal and needs to be investigated further. Common artefacts are learned as false positive by the algorithm, recognized and removed from the result. Rare artefacts and presence of anatomical parts not seen in the discovery set are still misclassified.

This algorithm represents a preliminary result to support the utilization of deep learning in colorectal MR. We intend to further optimize the protocol, mainly by (1) focusing on alternative architectures which account for anatomical location, and (2) shortening the time needed for the segmentation.

Conclusions

Our results demonstrate that deep learning can perform accurate localization and segmentation of rectal cancer in MR imaging in the majority of patients. Deep learning technologies have the potential to improve the speed and accuracy of MRI-based rectum segmentations, as manual delineation has been shown to be reader dependent and often time consuming, which limits its utility in practice and represents one of the major obstacles in the design of large quantitative imaging studies. Automatic segmentation procedures, such as the one presented in this study, aim to overcome this obstacle by offering a viable alternative to manual delineation.
Further validation of these technologies is warranted before clinical application. If these methods prove reliable, its impact in clinical management of rectal cancer could be significant by providing an efficient and accurate tool to assess residual tumor burden after preoperative treatment with subsequent better stratification of patients for organ preservation resulting in a higher quality of life.

1. Kim YH, Kim DHY, Kim TH, et al. Usefulness of magnetic resonance volumetric evaluation in predicting response to preoperative concurrent chemoradiotherapy in patients with resectable rectal cancer. *Int J Radiat Oncol Biol Phys.* 2005; 62(3): 761-768.

2. Seierstad T, Hole KH, Grøholt KK, et al. MRI volumetry for prediction of tumour response to neoadjuvant chemotherapy followed by chemoradiotherapy in locally advanced rectal cancer. *Br J Radiol.* 2015; 88(1051).

3. Martens MH, van Heeswijk MM, van den Broek JJ, et al. Prospective, Multicenter Validation Study of Magnetic Resonance Volumetry for Response Assessment After Preoperative Chemoradiation in Rectal Cancer: Can the Results in the Literature be Reproduced? *Int J Radiat Oncol Biol Phys.* 2015; 93(5): 1005-1014.

4. Lambregts DM, Rao SX, Sassen S, et al. MRI and Diffusion-weighted MRI Volumetry for Identification of Complete Tumor Responders After Preoperative Chemoradiotherapy in Patients With Rectal Cancer: A Bi-institutional Validation Study. *Ann Surg.* 2015; 262(6): 1034-1039.

5. Carbone SF, Pirtoli L, Ricci V, et al. Assessment of response to chemoradiation therapy in rectal cancer using MR volumetry based on diffusion-weighted data sets: a preliminary report. *Radiol Med.* 2012; 117(7): 1112-1124.

6. Il Ha H, Young Kim A, Sik Yu C, et al. Locally advanced rectal cancer: diffusion-weighted MR tumour volumetry and the apparent diffusion coefficient for evaluating complete remission after preoperative chemoradiation therapy. *Eur Radiol.* 2013; 23(12): 3345-3353.

7. Curvo-Semedo L, Lambregts DMJ, Maas M, et al. Rectal Cancer: Assessment of Complete Response to Preoperative Combined Radiation Therapy with Chemotherapy—Conventional MR Volumetry versus Diffusion-weighted MR Imaging. *Radiology.* 2011; 260(3): 734-743.

8. George ML, Dzik-Jurasz AS, Padhani AR, et al. Non-invasive methods of assessing angiogenesis and their value in predicting response to treatment in colorectal cancer. *Br J Surg.* 2001; 88(12): 1628-1636.

9. Choi MH, Oh SN, Rha SE, et al. Diffusion-weighted imaging: Apparent diffusion coefficient histogram analysis for detecting pathologic complete response to chemoradiotherapy in locally advanced rectal cancer. *J Magn Reson Imaging.* 2016; 44(1): 212-220.

10. Maas M, Beets-Tan RG, Lambregts DM, et al. Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. *J Clin Oncol.* 2011; 29(35): 4633-4640.

11. Lambregts DM, Beets GL, Maas M, et al. Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability. *Eur Radiol.* 2011; 21(12): 2567-2574.

12. Nougaret S, Vargas HA, Lakhman Y, et al. Intravoxel Incoherent Motion-derived Histogram Metrics for Assessment of Response after Combined Chemotherapy and Radiation Therapy in Rectal Cancer: Initial Experience and Comparison between Single-Section and Volumetric Analyses. *Radiology.* 2016; 280(2): 446-454.

13. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, et al. Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol.* 2016; 94(4): 824-831.

14. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5(1): 4006.

15. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine. *JAMA Oncol.* 2016; 2(12): 1636.

16. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015; 521(7553):436-444.

17. Greenspan H, Van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Trans Med Imaging.* 2016; 35(5): 1153-1159.

18. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform.* 2017; 18(5): 851-869.

19. Carneiro G, Mateus D, Loïc P, et al. Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings. *Deep Learning and Data Labeling for Medical Applications*. Vol 10008. Springer International Publishing; 2016.

20. Irving B, Franklin JM, Papież BW, et al. Pieces-of-parts for supervoxel segmentation with global context: Application to DCE-MRI tumour delineation. *Med Image Anal*. 2016; 32: 69-83.

21. Day E, Betler J, Parda D, et al. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. *Med Phys*. 2009; 36(10): 4349-4358.

22. Central Committee on Research Involving Human Subjects. Non-WMO Research. http://www.ccmo.nl/en/non-wmo-research. Accessed March 21, 2017.

23. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: A toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010; 29(1): 196-205.

24. Shamonin D. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform*. 2013; 7.

25. Klein S, Pluim JPW, Staring M, Viergever MA. Adaptive stochastic gradient descent optimisation for image registration. *Int J Comput Vis*. 2009; 81(3): 227-239.

26. Mattes D, Haynor DR, Vesselle H, Lewellen TK, Eubank W. PET-CT image registration in the chest using free-form deformations. *IEEE Trans Med Imaging*. 2003; 22(1): 120-128.

27. Insight Journal (ISSN 2327-770X) - Itk::Transforms supporting spatial derivatives. http://www.insight-journal.org/browse/publication/756. Accessed February 14, 2017.

28. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst*. 2012: 1-9.

29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res.* 2014; 15: 1929-1958.

30. Wu H, Gu X. Max-pooling dropout for regularization of convolutional neural networks. *In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol 9489. Springer Verlag; 2015: 46-54.

31. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. *ICML Work Deep Learn Audio, Speech Lang Process.* 2013; 28.

32. Zeiler MD. ADADELTA: An Adaptive Learning Rate Method. 2012. http://arxiv.org/abs/1212.5701.

33. The Theano Development Team, Al-Rfou R, Alain G, et al. Theano: A Python framework for fast computation of mathematical expressions. 2016: 1-19. http://arxiv.org/abs/1605.02688.

34. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017; 39(12): 2481-2495.

35. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI.* 2015: 1-8.

36. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics).* 2016: 424-432.

37. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging.* 2015; 34(10): 1993-2024.

38. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. *Med Image Anal.* 2017; 35: 18-31.

39. Rao V, Sarabi MS, Jaiswal A. Brain Tumor Segmentation with Deep Learning. *Multimodal Brain Tumor Image Segmentation Challenge, MICCAI 56. 2015.*

Deep learning for fully automated segmentation of rectal tumors on multiparametric MRI in a multicenter setting

Chapter 3 Deep learning for fully automated segmentation of rectal tumors on multiparametric MRI in a multicenter setting

Joost J.M. van Griethuysen, Doenja M.J. Lambregts, Niels W. Schurink, Max J. Lahaye, Stefano Trebeschi, Frans C.H. Bakers, Roy F.A. Vliegen, Remy W.F. Geenen, Vincent Cappendijk, Shira H. de Bie, Hugo J.W.L. Aerts, Regina G.H. Beets-Tan

Submitted for publication in Medical Image Analysis

Abstract

Purpose

To develop and test the performance of a fully automated deep learning segmentation algorithm in a large and heterogeneous clinical multi-center dataset of multiparametric rectal MRIs, taking into account inter-reader variability.

Materials and Methods

This retrospective study was approved by our local institutional review board. All data was collected and stored in compliance with HIPPA guidelines. Baseline staging MRIs including T2W, DWI and ADC maps of 603 patients from 6 institutions were analyzed, randomly split into train, tune, and test cohorts (ratio of 5:1:4). An expert-radiologist manually delineated all rectal tumors to serve as training input and ground truth for analysis of network inferred segmentations using the dice similarity coefficient (DSC). A second expert-radiologist independently re-segmented 142 patients in the test cohort to calculate inter-reader agreement which served as the target level of performance. To evaluate effects of image quality and "case mix" on the network performance, DWI scan quality (SNR and artefacts) and overall morphological tumor complexity (signal heterogeneity, border irregularity) were assessed using 3-5 point Likert-scores. An attention-gated U-Net was trained, with hypertuning via gridsearch. Performance of the network was compared to expert inter-reader agreement using Wilcoxon signed-rank tests.

Results

A deep learning network based on the full dataset of T2W-MRI, DWI and ADC showed good performance with a DSC of 0.669 in the test set, compared to an expert inter-reader agreement of DSC 0.749. Results based on T2W-data only were significantly poorer (DSC 0.314, $p<0.001$). More complex tumors resulted in significantly lower network performance and inter-reader agreement. Low DWI image quality (low SNR) negatively affected the networks' performance, but not inter-reader agreement. In the subgroup of good quality (high SNR) DWI scans, network performance was comparable to inter-reader agreement (DSC 0.708 vs 0.741).

Conclusion

Deep learning can be used for fully-automatic segmentation of rectal tumor segmentation on MRI scans, despite large heterogeneity in clinical data and could thus become an important support tool to allow accurate and fast assessment of quantitative imaging features.

Chapter 3

Introduction

Recent breakthroughs in AI allow for the automation of complex image analysis tasks that up to a few years ago could only be done by humans[1]. These breakthroughs could result in improved prediction of staging, therapeutic response and prognosis through uses of automated extraction of quantitative features from medial image data. To extract quantitative features, ranging from volume to complex radiomics features, typically requires segmentation of a certain region or VOI within the image, such as the primary tumor under investigation. Segmentation methods vary from simply placing a standardized shape (e.g. a circle) around the tumor, to whole-volume segmentation, where the tumor is delineated in detail on a slice-by-slice basis generating a 3D VOI. The latter is generally considered to provide the most accurate approximation of the true tumor volume and has also been shown to yield the most stable and performant quantitative imaging features[2–4]. The main drawback of whole-volume VOIs is that these are generally obtained manually and preferably by experienced readers, which is a highly labor intensive and time-consuming job[5]. For this reason, segmentation remains one of the major bottlenecks that hamper implementation of quantitative medical image analysis in time-constrained day-to-day clinical workflows.

Several studies aimed to address this issue by replacing manual segmentation by semi-automated algorithms, i.e. algorithms that still require some manual user input[5–7]. More recent studies have mainly focused on fully-automated segmentation algorithms based on deep learning networks. An example of such a network is the "U-net", a deep learning architecture that produces segmentation maps with the same size as the input image in a single forward pass. In a recent study by Schlemper et al.[8], this architecture was expanded with attention gates, helping the network to focus on the area of interest at minimal increased computational overhead. In rectal cancer, published reports using deep learning networks have shown encouraging results[9–14]. However, these reports have so far mainly been based on single center study cohorts, with highly curated datasets. While this will yield the best performance in the developed model, it is less representative of the general clinical setting, where such homogeneous, good quality data is often not available, especially when dealing with multi-sequence MRI data where acquisition protocols and image quality are less standardized and more prone to variations between centers compared to for example CT and PET data. For segmentation tools to be generally applicable, they must therefore also be robust to variations in the data, like different vendor systems, MR protocols and large variance in scan quality and tumor complexity.

Another limitation is that published algorithms are generally trained with manual segmentations generated by a single reader that are at the same time used as the ground truth to test the network's performance. This introduces bias and does not take into account effects of inter-reader variations that may occur (even between highly expert readers[5,7]), which may affect the validity of these segmentations when used as a standard of reference.

The aim of this study is to investigate the accuracy of a fully automated deep learning segmentation algorithm in a real-world clinical dataset of rectal MRIs from six institutions, taking into account data heterogeneity and inter-reader variability.

Materials and Methods

Study Population

This retrospective study was approved by our local institutional review board. All data was collected and stored in compliance with HIPPA guidelines. For this study cases were selected from a study dataset collected as part of a larger ongoing multicenter retrospective research project on neoadjuvant treatment prediction in rectal cancer, including patients who underwent rectal MRI including T2W and DWI at baseline, i.e. prior to treatment.
Imaging examinations were performed between January 2012 and January 2017 in 6 institutions; 4 regional non-university centers Noordwest Ziekenhuisgroep (Alkmaar, center 1), Deventer Ziekenhuis (Deventer, center 2), Zuyderland Medical Center (Heerlen, center 3), Jeroen Bosch Ziekenhuis (Den Bosch, center 4), one university hospital Maastricht University Medical Center (Maastricht, center 5) and one third-line referral center The Netherlands Cancer Institute (Amsterdam, center 6). Inclusion criteria for the current study consisted of: [1] biopsy proven adenocarcinoma of the rectum and [2] availability of a baseline MRI including at least a T2W and DWI sequence. Reasons for exclusion (see Figure 1) were mucinous type adenocarcinoma (as these exhibit distinctly different characteristics on both T2W and DWI[15,16]), more than 1 tumor lesion within the field of view (i.e. simultaneous tumor in rectum and sigmoid colon), non-diagnostic image quality (e.g. severe artefacts in case of hip prosthesis), and tumors that were not completely included in the field of view of either the T2W and/or DWI sequence.
Study cases were randomly assigned to train, tune and test sets using a ratio of 5:1:4.

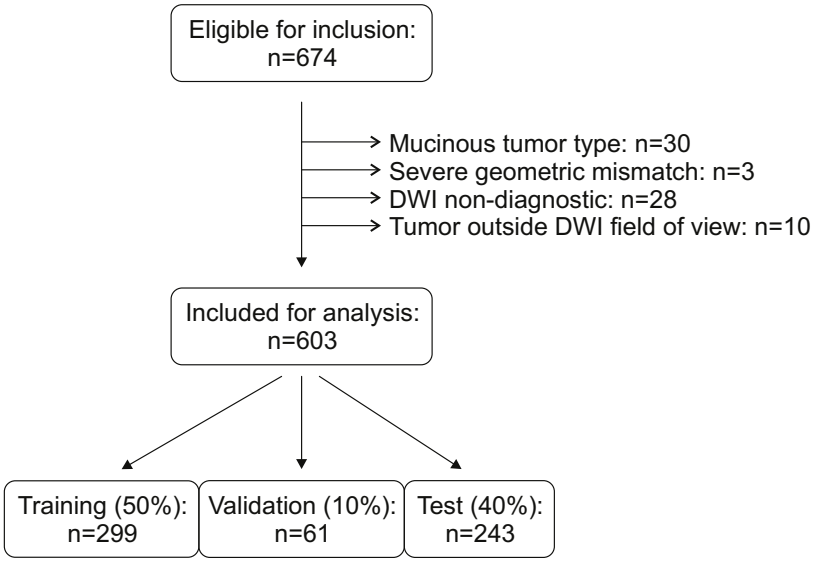


Figure 1. In- and exclusion flowchart

Table 1. Sequence parameters for the 6 different centers

	Center 1		Center 2				Center 3	Center 4		Center 5	Center 6	
General												
Vendor	**Siemens**	**Philips**	**Siemens**	**GE**		**Philips**	**Siemens**	**Siemens**		**Philips**	**Philips**	
Field Strength	1.5	1.5	1.5	1.5	3	3	1.5	1.5	3	1.5	1.5	3
T2W-MRI												
FOV	200x200-230x230	250x250	200x200-362x379	240x240-299x299	160x160-240x240	180x180	240x240-310x310	200x200-280x280	200x200-280x278	180x180-200x200	250x250	140x140-281x281
# Slices	20-48	30	25-55	22-38	16-42	34-42	37-70	23-68	35-41	24-47	40-60	24-77
Slice Thickness	3.0-3.5	4	3.0-3.5	4.0-4.4	3.0-4.3	3	3.0-3.5	3.0-5.0	3.0-5.0	3	3	3.0-4.0
Pixel Spacing	0.34x0.34 - 0.69x0.69	0.49x0.49	0.62x0.62 - 0.99x0.99	0.47x0.47 - 0.59x0.59	0.31x0.31 - 0.47x0.47	0.47x0.47	0.73x0.73 - 0.94x0.94	0.62x0.62 - 0.88x0.88	0.62x0.62 - 0.88x0.88	0.56x0.56 - 0.78x0.78	0.49x0.49	0.27x0.27 - 0.49x0.49
NSA	2-3	2	1-4	1-2	1-5	1	1-2	2	2	2-6	1	1-2
TR	3000-5860	3050	3390-9030	4280-8528	2618-7460	3654-5604	2550-6490	3300-8680	3540-8590	3409-16738	3838-4112	2515-9393
TE	78-102	100	70-101	101-109	82-132	100	102-122	96-108	95-101	90-130	120	110-150
Flip Angle	145-173	90	130-160	90	90-130	90	120-150	150	143-150	90	90	90

DWI												
FOV	200x200-260x260	375x375	260x260-379x379	380x380-480x480	170x170-520x520	350x350	250x250-325x409	308x308-379x409	308x379	180x180-440x440	180x180-375x375	186x186-359x359
# Slices	20-48	25	30-36	23-36	18-48	30-35	24-42	24-40	30	20-50	20-40	20-50
Slice Thickness	3.5-4.0	7	3.5-6.0	8	3.0-8.0	4	3.0-6.0	5.0-6.0	6	5	5	2.73-5.0
Pixel Spacing	0.98x0.98 - 2.03x2.03	1.46x 1.46	0.99x0.99 - 1.98x1.98	1.48x1.48 - 1.88x1.88	0.66x0.66 - 2.03x2.03	1.46x 1.46	1.25x1.25 - 1.98x1.98	1.98x1.98 - 2.05x2.05	1.98x1.98	1.25x1.25 - 1.72x1.72	1.41x1.41 - 1.95x1.95	0.56x0.56 - 1.11x1.11
NSA	2-14	6	2-6	1-6	1-15	2	1-8	4-6	4	4-8	4-8	1-4
TR	4100-11000	1717	4700-6600	4850-6300	2500-10250	3124-3825	4400-7000	4800-7900	7400-7420	2480-5950	2765-5151	4000-10462
TE	56-100	87	61-84	58-83	54-117	94-95	79-88	86-96	83	65-92	65-94	54-96
Flip Angle	90-180	90	90	90	90	90	90	90	90	70-90	90	90
EPI Factor	44-54	53	49-66	N/A	N/A	65	61-73	N/A	N/A	55-95	51-63	55-113
Highest b value	800-1000	800	800-1000	600-1000	600-1200	1500	800-1100	800-1000	800	1000-2000	800-1000	750-1200
# b values	2-3	2	3	2-3	2-4	5	3	2-3	2	2-7	2-4	3-16

FOV: Field Of View, NSA: Number of Signal Averages, TR: Repetition Time, TE: Echo Time, EPI: Echo Planar Imaging

Image sequences

Images were all retrospectively collected and had previously been acquired according to routine practice in the 6 different centers using a variety of MR-system vendors and acquisitions protocols that are summarized in Table 1. For the current study, the transverse DWI and T2W sequences (angled in similar planes) were selected. From each DWI sequence, a high b-value image (b600-1000) was selected for analysis. ADC maps were recalculated for the purpose of this study using a mono-exponential regression model including all available b-values. Voxels with ADC values < 0 and > mean + 6 standard deviations were marked as 'invalid' values and set to 0. In total, 3 imaging datasets (T2W, high b-value DWI, ADC) were thus selected for further analysis. Geometric alignment between sequences (T2W and DWI/ADC) was assessed visually; scans exhibiting severe misalignment were excluded (see Figure 1).

Manual image segmentation and quality assessment

Rectal tumors were segmented using 3D-slicer (version 4.11)[17] using the high b-value DWI combined with the T2W sequences for visual anatomical correlation. Tumors were coarsely segmented using the "Level tracing" algorithm, followed by full manual revision and adaptation by an expert reader (R1, DMJL, board certified radiologist with > 10 years' experience in assessment of rectal cancer on MRI) who traced the tumor boundaries thereby adapting and finalizing the segmentations on a slice-by-slice level. A randomly selected subset of 142 cases from the test set was independently processed using the same methods by a second expert-reader (R2, MJL, board certified radiologist with similar experience level as R1) to analyze inter-reader agreement.

In addition, R1 visually assessed the quality of each DWI scan in the dataset by scoring the overall DWI image quality (SNR and image resolution) and the presence of artefacts (signal pile up and/or geometrical distortions caused by susceptibility effects) at the level of the tumor, using 5-point Likert scales, ranging from 0 (= non-diagnostic / severe artefacts) to 4 (= excellent quality/no artefacts), similar to scoring systems previously reported[18]. Quality assessment was primarily focused on DWI as these sequences are known to be most prone to quality variations and will therefore likely have the most effect on network performance[19]. Finally, the overall level of tumor complexity was scored for each case to estimate if a tumor would be easy or difficult to segment. The latter was assessed using a 3-point scale based on tumor morphology: 0 = difficult to segment (heterogeneous, irregularly shaped and/or poorly demarcated tumor), 1 = moderately difficult to segment (partly heterogeneous and/or irregular) and 2 = easy to segment (homogeneous, regular and well demarcated). Representative examples of the different Likert scores used to assess DWI quality and tumor complexity are shown in Figure 2.

Image pre-processing and data augmentation

Image pre-processing and data augmentation was achieved using the "numpy" (1.17.4) and "SimpleITK" (1.2.3) packages in Python 3.6.8.

For each included case, the 3 input images (T2W, high b-value DWI and ADC-map) were resampled using a B-Spline interpolator and combined into 1 single 3-channel multi-sequence image with (x, y, z) spacing of 1x1x3mm, angled perpendicular to the tumor axis (as copied from the transverse T2W image) and aligned to the T2W image center. Ground-truth segmentations (i.e. manual segmentations from R1) were resampled to the same image space using a Nearest Neighbor interpolator. No restrictions were placed on the output image size.

Additional data augmentation consisting of random small geometric distortions was applied "on-the-fly" for the training cases. Furthermore image intensities were normalized and a fixed size patch approximating the full image size was extracted for both training and tuning cases to allow a forward pass of multiple cases in one minibatch. Full details of the data augmentation and preprocessing steps are provided in supplementary materials S1.

Neural network definition

We trained and evaluated the attention-gated U-Net (AG U-Net) as previously reported by Schlemper et al[8], with an initial feature size of 16. We utilized the code shared by Schlemper et al., with some adaptation to allow multi-channel input and SimpleITK-based data augmentation. The network was trained on a NVIDIA GeForce RTX 2080Ti twice; once using the full multi-sequence dataset (containing the T2W, DWI and ADC maps) and once using only T2W images. The network was optimized using the Sorensen-Dice loss function to address class imbalance between foreground and background voxels[20] for 200 epochs. Hypertuning was applied to determine the optimum batch size, learning rate, optimizer algorithm and learning rate adaptation algorithm. Hypertuning was performed using only the train and tuning sets. Full details on the searched hyperspace are available in supplementary materials S2.

Statistical analysis

Baseline characteristics were compared among centers using one-way ANOVA for continuous variables and Kruskal-Wallis for categorical variables. Segmentations were inferred for the cases in the test set using the networks trained with optimum hyperparameters as determined by the performance in the tuning set. Network performance was defined as the agreement between network-inferred and expert reader's manual segmentations (i.e. for R1 who segmented the whole dataset). Network performance was then compared to the inter-reader agreement between the two expert readers (for the 142 cases that were manually segmented by 2 independent readers) as a "standard of reference". Agreement was calculated using the DSC. Secondary outcome was the network performance based on T2W-only (i.e. without DWI and ADC).

$$DSC = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|}$$

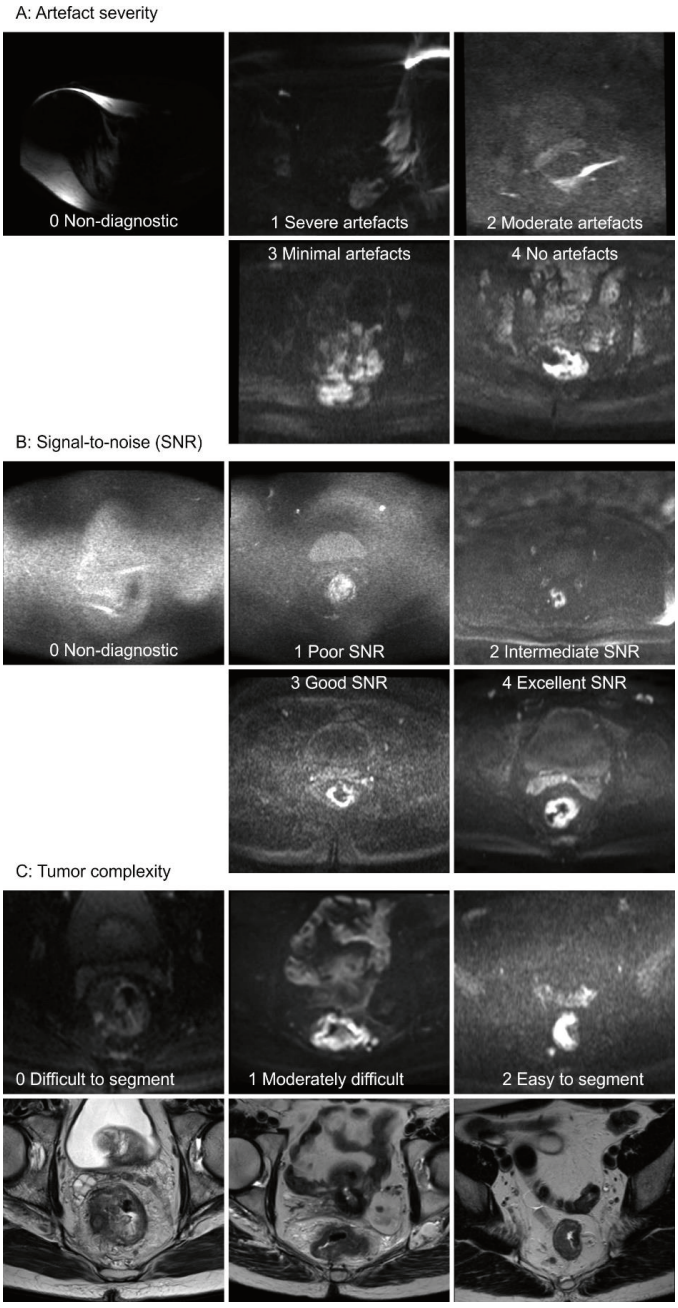


Figure 2. Likert scores used to assess DWI scan quality and tumor complexity. *A) 5-point Likert score used to assess the severity of artefacts related to magnetic susceptibility (signal pile up and geometrical distortions) on DWI. B) 5-point Likert score used to assess the signal to noise ratio on DWI. C) 3 point Likert score used to assess the overall tumor complexity c.q. segmentation difficulty level (taking into account signal heterogeneity and border irregularity) on T2W and DWI.*

DSCs between the trained networks and expert-reader segmentations were compared using Wilcoxon signed rank test. Finally, the influence of DWI scan quality and tumor complexity scores were analyzed using Spearman correlation and by performing subgroup analyses by splitting the test set into poor vs. good scan quality (Likert score 1-3 vs 4), artefacts vs. no artefacts (score 1-3 vs 4) or complex vs easy to segment tumors (score 0-1 vs 2). Differences between subgroups were compared using Mann-Whitney U test.

Results

Study population

In total, 674 patient cases were considered for inclusion. After exclusion of 71 cases (Figure 1), the final study cohort included 603 cases, divided into training (n = 299), tuning (n = 61), and test datasets (n = 243). 376 patients were male, median age was 65 years (range 26-87). Clinical tumor stage at baseline was cT1-2 for 49 patients, cT3 for 480 and cT4 for 74 patients. 499 patients had clinically node-positive disease. There were no significant differences in age, gender and cT stage distribution between the 6 centers (p=0.09-0.90), though cN stage distribution was significantly different (p<0.001). Between train, tune and test cohorts, there were no significant differences in age, gender and TN-stage distribution (p=0.07-0.88).

Network performance

The segmentation performance of the attention gated U-Net, using the segmentation from expert reader 1 as the ground truth, is shown in Table 2 and Figure 3A. The main network performance using the multisequence dataset was DSC 0.710 in the tuning set and DSC 0.669 in the final test set. When compared to the inter-reader agreement between the expert radiologists as a standard of reference (for the 142 cases that were double-read in the test dataset), there was a statistically significant difference in DSC (p<0.001), though overall agreement was good for the network vs expert R1 (DSC 0.661) as well as between the two manual expert readers (DSC 0.749). Evolution of performance during training and effects of attention gating are described in Supplementary Figures S1 and S2.

When trained and tested using only T2W-MRI, network performance was significantly lower (Wilcoxon p<0.001): DSC 0.422 in the tuning set and DSC 0.314 in the test set, indicating a strong positive effect of the DWI/ADC data on the network's performance.

Influence of scan quality and tumor complexity

Effects of scan quality and tumor complexity are shown in Table 3 and Figures 3B-D. The performance of the deep learning network was significantly affected by overall DWI image quality (with lower performance in case of reduced SNR), but not by the severity of artefacts occurring around the tumor. Neither of these factors significantly affected expert inter-reader agreement. When selecting only good-quality scans based on the SNR score, the deep learning network performance was comparable to the expert inter-reader agreement (DSC 0.703 vs 0.741, p=0.092). Tumor complexity had a significant effect on the performance of the deep

learning network as well as on expert inter-reader agreement, with more heterogeneous and irregular tumors (i.e. more complex and difficult to segment tumors) resulting in poorer segmentation performance and lower inter-reader agreement (DSC 0.703 versus 0.649 for the network and DSC 0.801 versus 0.712 between readers). Imaging examples showing the impact of DWI image quality and tumor complexity on the network's performance and inter-reader agreement are shown in Figure 4.

Table 2. Network performance versus expert inter-reader agreement

	DSC (±SD)
Network performance:	
Multi-sequence network (vs. R1)	0.669 (±0.190)
T2W-only network (vs. R1)	0.314 (±0.205)
Expert inter-reader agreement	
R1 vs. R2	0.749 (±0.179)

DSC = Dice Similarity Coefficient, SD = Standard deviation. R1 and R2 are both board certified radiologists with >10 years' experience in assessing rectal MRI. The multi-sequence network included T2W-MRI, high b-value DWI and ADC.

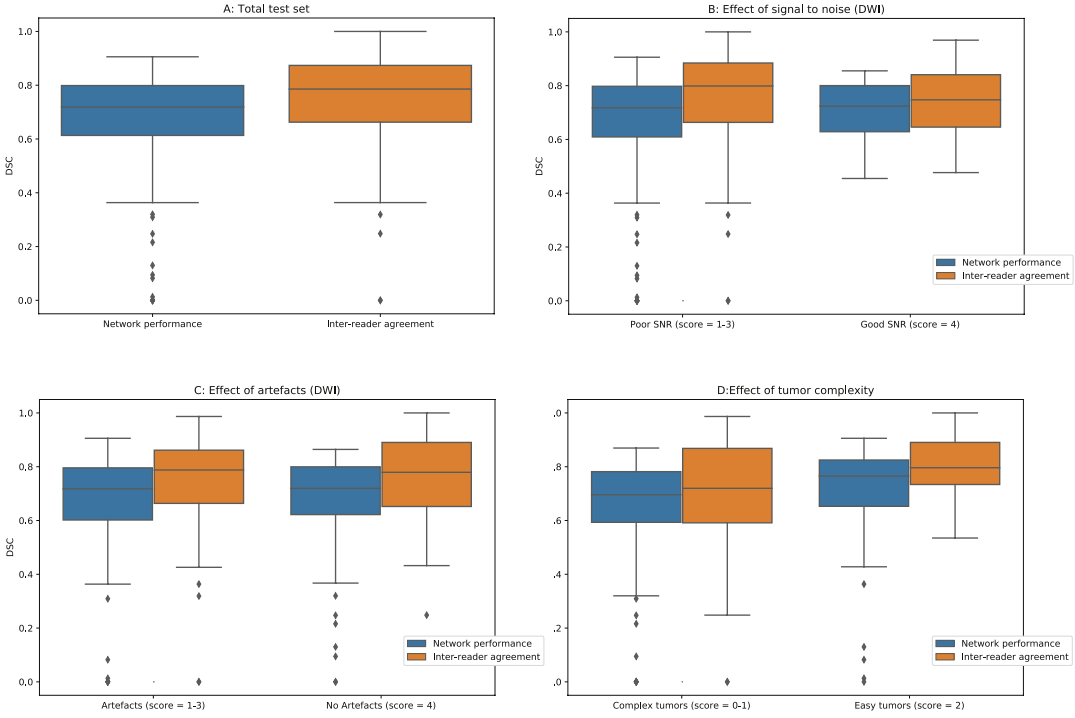


Figure 3. Network performance vs inter-reader agreement.
A) Full dataset. B) Impact of poor vs good SNR on DWI. C) Impact of DWI artefacts. D) Impact of tumor complexity.

Table 3. Effects of DWI scan quality and tumor complexity

	SPEARMAN CORRELATION (ρ)			
DSCS for:	**SNR**		**Artefact severity**	**Tumor complexity**
Multi-sequence network (vs R1)	0.128 (p=0.046)		0.049 (p=0.443)	0.291 (p< 0.001)
Expert inter-reader (R1 vs R2)	-0.005 (p=0.957)		0.021 (p=0.808)	0.177 (p=0.035)

	SUBGROUP RESULTS								
	DWI image quality						**Tumor complexity**		
DSCS for:	Poor SNR	Good SNR	p	Artefacts	No Artefacts	p	Complex	Easy	p
Multi-sequence network (vs R1)	0.661	0.708	0.322	0.667	0.671	0.433	0.649	0.703	<0.001
Expert inter-reader (R1 vs R2)	0.752	0.741	0.153	0.738	0.763	0.303	0.712	0.801	0.010

R1 and R2 are both board certified radiologists with >10 years' experience in assessing rectal MRI. DWI image quality: poor SNR = Likert score 0-3, good SNR = score 4; Artefacts = score 0-3, no artefacts = score 4. Tumor complexity: complex = score 0-1, easy = score 2. For further details of the different Likert scores see Figure 2.

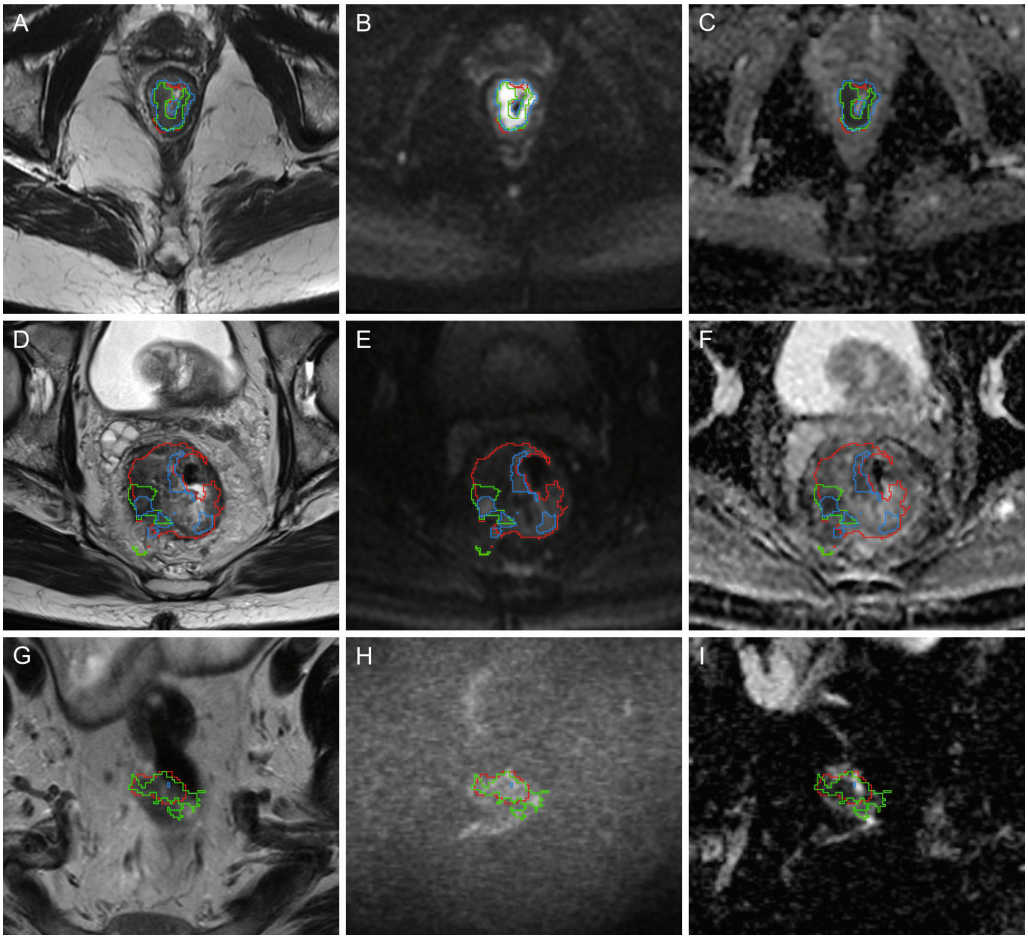


Figure 4. Imaging examples of impact of DWI image quality and tumor complexity on segmentation outcomes. *A-C) Example of an easy to segment (score 2) tumor on a good quality DWI exam: T2-weighted image (A), b1200 DWI (B) and ADC-map (C) of a homogeneous, well-demarcated tumor, resulting in good agreement between the network (blue) and both expert readers (R1 red, R2 green). D-F) Example of a complex tumor (score 0): T2-weighted image (D), b800 DWI (E) and ADC-map (F) showing an irregularly shaped, heterogeneous tumor. There was moderate agreement between R1 (red) and R2 (green). Due to the poor tumor-to-background contrast on DWI, the network largely mis-segmented the tumor (blue). G-I) Example of a DWI scan with poor SNR (score 1): T2-weighted image (G), b750 DWI (H) and ADC-map (I) showing an small tumor in the upper rectum that is very difficult to discern on DWI owing to the low SNR of the image. Taking into account the anatomical T2-weighted images, inter-reader agreement for the two manual readers (red, green) remained good, while the network was unable to produce a valid segmentation (blue).*

Discussion

In this study, we developed and analyzed a deep learning network to fully-automatically segment rectal tumors on MRI in a heterogeneous multiparametric MRI dataset of T2W, DWI and ADC-maps from 6 different centers, intended as a representative clinical sample of MRIs acquired in routine practice. Performance of the network was compared to the inter-reader agreement between two independent expert radiologists, which was defined as the target performance level for the network as a standard of reference.

Our results show that the network attained a good overall segmentation performance with a DSC of 0.669 in the independent test set. Although results remained statistically inferior, this performance level was almost in the range of the expert inter-reader agreement, which resulted in a DSC of 0.749. Both network performance and expert inter-reader agreement were correlated with tumor complexity, with more heterogeneous and poorly demarcated tumors resulting poorer segmentation outcomes and reduced inter-reader agreement. The network's segmentation performance was also significantly influenced by DWI image quality, with poorer quality scans (in particular scans with a low SNR) resulting in poorer segmentation results. This effect was less evident for the two manual readers, probably because these readers relied more on the anatomical T2-weighted images for segmentation in cases where DWI quality fell short (as demonstrated in Figure 4E-F). Interestingly, when selecting only optimal-quality scans the network was able to achieve a performance comparable to that of expert inter-reader agreement, highlighting the need for good quality image acquisition as a prerequisite for automated image analysis. Subgroup analysis using only T2W-MRIs clearly resulted in inferior segmentation performance (DSC 0.314 vs 0.669 for the full multiparametric dataset), indicating that DWI has significant added benefit when training computers to recognize and segment rectal tumors on MRI. Confirming the results of a previous study by Schlemper et al. (using CT data to segment pancreatic tumors), addition of attention gating allowed the network to successfully focus on the area of interest (i.e. the tumor) within the FOV[8].

In a previous study including 140 patients from a bi-institutional dataset, Trebeschi et al.[10] investigated the use of a simple patch-based 2D convolutional network for rectal tumor segmentation on MRI. They achieved a network performance of DSC 0.68-0.70, which is similar to our current report including a more heterogeneous dataset. The use of a patch-based approach in the study by Trebeschi resulted in slower inference and limited contextual information, with misclassification of rare imaging artefacts near the edge of the field of view. This was not observed as a limitation in our current study, as the U-Net only requires one forward pass to generate a segmentation, and better utilizes special context especially when using attention-gating. In another study by Wang et al. from 2018[12], a 2D U-Net showed a similar performance to expert reader agreement, with DSCs of 0.74 and 0.71, respectively, in a relatively small set of 93 cases using 10-fold cross-validation. In a similar study by Men et al.[13], 3 different deep learning architectures showed good performance using 5-fold cross-validation in a small set of 70 cases with DSC ranging from 0.70 to 0.78. However, no comparison was made to inter-reader agreement. Finally, in a recent study by Wang et al. from 2019[11], 568 cases were included from several different centers to fine-tune and test a 2D ResNet with side-output, yielding a high remarkable performance of 0.84 DSC. Unfortunately inter-reader

agreement was again not reported, making it difficult to compare these results directly to ours. Interestingly, in the majority of the abovementioned previous works[11–13], only T2W images were used as segmentation input. Though this eliminates the problem of addressing different spacing and registration errors between different sequences, we found that using only T2W images resulted in significantly poorer results compared to the multisequence dataset including also DWI/ADC. This is likely related to the intrinsic higher tumor-to-background ratio seen in DWI, which has previously been shown to produce lower inter-reader variations when segmenting rectal tumor volumes on MRI compared to T2W sequences[21,22].

To the best of our knowledge, this study is the first to explicitly assess the influence of scan quality and tumor morphology on the segmentation performance of deep learning networks aiming to account for variations in scan quality as well as overall "case mix" between and within centers that are bound to occur when applying such networks in daily clinical practice or on heterogeneous multicenter research datasets. We have learned that both the deep learning network as well as the manual readers struggled with segmenting morphologically complex tumors, i.e. tumors with a heterogeneous signal and ill-defined borders. These tumors (which constituted ±15-20% of our dataset) will likely always remain a bottle neck for segmentation. The segmentation performance of the network also suffered when the quality (SNR) of the DWI was low. This is an important factor to take into consideration when using MRI as input for automated tumor segmentation and stresses the need for protocol optimization. While T2 image acquisition is relatively standardized, there are various methods of DWI image acquisition that may render highly variable quality images. Moreover, patient preparation (such as the reduction of intraluminal gas to avoid susceptibility effects[18]) is a known factor that may affect DWI image quality, in particular for DWI of the gastrointestinal tract.

Our study design contained some limitations, which may warrant further study. First, in the absence of a true gold standard, we used the inter-reader agreement as the target level of performance for the network. Comparing the network's segmentations to for example pathology slides may offer a more accurate ground truth, though this is generally not feasible, especially in retrospective studies. Second, given the highly time-consuming nature of the study, only a subset of cases was segmented by 2 readers (in order to compare the network's results to inter-reader agreement). The network was, however, trained using the segmentation input of only one of these two readers, which can add a bias in the trained networks to the segmentation preference of this reader. In future studies, segmentations by multiple readers should be used as a form of data augmentation, with the aim of training reader-agnostic segmentation networks. Third, the DWI dataset included some cases (<10%) where the highest available b value was <800, which does not meet the current recommendations of the ESGAR consensus guidelines on rectal MRI which state that a clinical DWI protocol should include at least a high b value of ≥b800[23]. This may have resulted in poorer tumor-to-background contrast in these cases and may thus have had a potential negative effect on the network's performance. Finally, cases from all included centers were present in the training, tune and test sets, with the intent to train the network to deal with heterogeneous data originating from different centers, thereby creating a more generalizable segmentation network. Though our final test set was independent in the sense that none of the cases were used to train the network, it did not originate from a fully independent center or dataset.

Conclusions

In conclusion, this study demonstrated that despite large heterogeneity in data, it is possible to train a deep learning network to accurately segment rectal tumors on multiparametric MRI, with performance levels comparable to the overall agreement between expert radiologists, provided that DWI image quality is good. Once further optimized, such networks may significantly reduce the segmentation workload for future quantitative imaging studies and ultimately facilitate the implementation of segmentation tools in routine clinical practice.

1. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018; 18(8): 500-510.

2. Lambregts DM, Beets GL, Maas M, et al. Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability. *Eur Radiol.* 2011; 21(12): 2567-2574.

3. Nougaret S, Vargas HA, Lakhman Y, et al. Intravoxel Incoherent Motion-derived Histogram Metrics for Assessment of Response after Combined Chemotherapy and Radiation Therapy in Rectal Cancer: Initial Experience and Comparison between Single-Section and Volumetric Analyses. *Radiology.* 2016; 280(2): 446-454.

4. Prezzi D, Owczarczyk K, Bassett P, et al. Adaptive statistical iterative reconstruction (ASIR) affects CT radiomics quantification in primary colorectal cancer. *Eur Radiol.* 2019; 29(10): 5227-5235.

5. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, et al. Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol.* 2016; 94(4): 824-831.

6. Van Stiphout RGPM, Valentini V, Buijsen J, et al. Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation. *Radiother Oncol.* 2014; 113(2): 215-222.

7. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One.* 2014; 9(7): e102107.

8. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal.* 2019; 53: 197-207.

9. Larsson R, Xiong JF, Song Y, et al. Automatic Delineation of the Clinical Target Volume in Rectal Cancer for Radiation Therapy using Three-dimensional Fully Convolutional Neural Networks. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.* Vol 2018-July. IEEE; 2018: 5898-5901.

10. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci Rep.* 2017; 7(1): 5301.

11. Wang M, Xie P, Ran Z, et al. Full convolutional network based multiple side-output fusion architecture for the segmentation of rectal tumors in magnetic resonance images: A multi-vendor study. *Med Phys.* 2019; 46(6): 2659-2668.

12. Wang J, Lu J, Qin G, et al. Technical Note: A deep learning-based autosegmentation of rectal tumors in MR images. *Med Phys.* 2018; 45(6): 2560-2564.

13. Men K, Boimel P, Janopaul-Naylor J, et al. Cascaded atrous convolution and spatial pyramid pooling for more accurate tumor target segmentation for rectal cancer radiotherapy. *Phys Med Biol.* 2018; 63(18): 185016.

14. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys.* 2017; 44(12): 6377-6389.

15. Schurink NW, Min LA, Berbee M, et al. Value of combined multiparametric MRI and FDG-PET/CT to identify well-responding rectal cancer patients before the start of neoadjuvant chemoradiation. *Eur Radiol.* 2020; 30(5): 2945-2954.

16. Nasu K, Kuroki Y, Minami M. Diffusion-weighted imaging findings of mucinous carcinoma arising in the ano-rectal region: Comparison of apparent diffusion coefficient with that of tubular adenocarcinoma. *Jpn J Radiol.* 2012; 30(2): 120-127.

17. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging.* 2012; 30(9): 1323-1341.

18. van Griethuysen JJM, Bus EM, Hauptmann M, et al. Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5 T – Effect of applying a micro-enema to improve image quality. *Eur J Radiol.* 2018; 99(0): 131-137.

19. Le Bihan D, Poupon C, Amadon A, Lethimonnier F. Artifacts and pitfalls in diffusion MRI. *J Magn Reson Imaging.* 2006; 24(3): 478-488.

20. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE; 2016: 565-571.

21. Curvo-Semedo L, Lambregts DMJ, Maas M, et al. Rectal Cancer: Assessment of Complete Response to Preoperative Combined Radiation Therapy with Chemotherapy—Conventional MR Volumetry versus Diffusion-weighted MR Imaging. *Radiology*. 2011; 260(3): 734-743.

22. Carbone SF, Pirtoli L, Ricci V, et al. Assessment of response to chemoradiation therapy in rectal cancer using MR volumetry based on diffusion-weighted data sets: a preliminary report. *Radiol Med*. 2012; 117(7): 1112-1124.

23. Beets-Tan RGH, Lambregts DMJ, Maas M, et al. Magnetic resonance imaging for clinical management of rectal cancer: Updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. *Eur Radiol*. 2017; 28(4): 1465-1475.

Image quality

Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5T - effect of applying a micro-enema to improve image quality

Chapter 4 Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5T - effect of applying a micro-enema to improve image quality

Joost J.M. van Griethuysen, Elyse M. Bus, Michael Hauptmann, Max J. Lahaye, Monique Maas, Leon C. ter Beek, Geerard L. Beets, Frans C.H. Bakers, Regina G.H. Beets-Tan, Doenja M.J. Lambregts

Published in:

European Journal of Radiology 2017; 0(0):131-137. (IF 2019 2.687)

Abstract

Purpose:

Assess whether application of a micro-enema can reduce gas-induced susceptibility artefacts in Single-shot Echo Planar Imaging (EPI) Diffusion-weighted imaging of the rectum at 1.5T.

Materials and Methods:

Retrospective analysis of n=50 rectal cancer patients who each underwent multiple DWI-MRIs (1.5T) from 2012-2016 as part of routine follow-up during a watch-and-wait approach after chemoradiotherapy. From March 2014 DWI-MRIs were routinely acquired after application of a preparatory micro-enema (Microlax®; 5 ml; self-administered shortly before acquisition); before March 2014 no bowel preparation was given. In total, 335 scans were scored by an experienced reader for the presence/severity of air artefacts (on b1000 DWI), ranging from 0 (no artefact) to 5 (severe artefact). A score ≥3 (moderate-severe) was considered a clinically relevant artefact. A random sample of 100 scans was re-assessed by a second independent reader to study inter-observer effects. Scores were compared between the scans performed without and with a preparatory micro-enema using univariable and multivariable logistic regression taking into account potential confounding factors (age/gender, acquisition parameters, MRI-hardware, rectoscopy prior to MRI).

Results:

Clinically relevant air artefacts were seen in 24.3% (no micro-enema) vs. 3.7% (micro-enema), odds ratios were 0.118 in univariable and 0.230 in multivariable regression (P=0.0005 and 0.0291). Mean severity score (±SD) was 1.19±1.71 (no-enema) vs 0.32±0.77 (micro-enema), odds ratios were 0.321 (P<0.0001) and 0.489 (P=0.0461) in uni- and multivariable regression, respectively. Inter-observer agreement was excellent (κ0.85).

Conclusion:

Use of a preparatory micro-enema shortly before rectal EPI-DWI examinations performed at 1.5T MRI significantly reduces both the incidence and severity of gas-induced artefacts, compared to examinations performed without bowel preparation.

Chapter 4

Introduction

Diffusion-weighted imaging (DWI) is nowadays increasingly adopted as an integral part of oncologic imaging protocols. In rectal cancer, DWI has mainly shown its value for response evaluation and follow-up of rectal tumors after chemoradiotherapy, specifically for the discrimination of viable tumor within areas of post-radiation fibrosis[1–4].

The most commonly used sequence for abdominal diffusion imaging is a single-shot Echo Planar Imaging (EPI) sequence. The main benefit from an EPI approach is its short acquisition time, which minimizes the risk of motion artefacts. However, an important drawback is that EPI sequences are prone to susceptibility artefacts, particularly at higher field strengths[5,6]. Susceptibility artefacts are changes or distortions in image signal caused by local magnetic field inhomogeneities, for example due to the presence of metal objects (e.g. hip replacements or surgical clips). In bowel imaging, these artefacts are mainly caused by the presence of gas in the rectal lumen. In a study by Caglic et al. (in prostate MRIs) it was reported that increased rectal gas-distension correlates significantly with reduced DWI image quality and increased DWI artefacts[7]. Particularly when the bowel itself is the organ under investigation, gas-induced susceptibility artefacts can severely reduce the diagnostic image quality, in some cases even rendering the images non-diagnostic. In published reports on bowel DWI 4-11% of patients had to be excluded from analyses due to poor DWI scan quality[8–11].

To reduce the influence of these artefacts on image quality, two main strategies can be employed: 1) change the acquisition parameters (i.e. type of DWI sequence) or 2) remove the cause of the artefact.

So far, most published studies have focused on the first approach and tested alternative ways of DWI image acquisition such as parallel imaging[12], smaller Field of View (FOV)[13] or bipolar DWI acquisition[14]. A potential solution to remove the cause of the artefact is to reduce the amount of gas in the rectal lumen by rectal filling, where the rectum is filled with a liquid (such as ultrasound gel) prior to image acquisition, replacing the gas. However, a potential downside of this approach is that it causes distension of rectum and compression of the surrounding mesorectal fat[15,16], potentially hampering correct assessment of the relation between the tumor and mesorectal fascia[17]. Use of endorectal filling is therefore not routinely recommended[18].

An alternative potential solution is the application of a preparatory micro-enema shortly prior to image acquisition. A micro-enema can typically be self-administered by the patient to reduce the amount of gas (and stool) in the rectum.

The aim of this study was to test this hypothesis and investigate to what extend the use of a micro-enema can reduce the amount of gas-induced susceptibility artefacts on EPI-DWI of the rectum.

Materials and methods

The study was approved by the local institutional review board. Due to the retrospective nature of the study, informed consent was not required.

Patients

We retrospectively selected 50 consecutive rectal cancer patients (66% male, mean age 63) who each underwent serial MR imaging including an EPI-DWI sequence of the rectum as part of their routine follow-up during a 'watch-and-wait' policy between January 2012 and February 2016 at Maastricht University Medical Center (MUMC). All patients had previously been treated with long-course chemoradiotherapy and were non-operatively managed due to strong clinical evidence of a clinical complete response. The follow-up protocol included regular MRI performed 3 monthly in the first year and 6 monthly in the second to fifth year of follow-up. In March 2014, use of a rectal micro-enema (Microlax®, McNeil Healthcare, Ireland) was introduced into the routine protocol. Inclusion criteria consisted of: 1) availability of at least 2 consecutive follow-up MRIs including a EPI-DWI sequence, with at least 1 MRI without bowel preparation and 1 MRI after application of a micro-enema , 2) no treatment (radiation or surgery) performed between the various sequential scans, 3) no history of hip replacement surgery (as hip prostheses will results in artefacts on DWI, as a result of which the presence of air artefacts cannot be sufficiently studied).

Image acquisition and patient preparation

All MR images were acquired on a 1.5T MR system (Intera (Achieva) or Ingenia MR system; Philips Healthcare, Best, The Netherlands) using a phased-array body coil. The routine protocol included T2-weighted turbo spin echo sequences in 3 planes (sagittal, axial and coronal) and an axial EPI-DWI sequence with b=1000 being the highest b-factor. The transverse T2-weighted and DW-sequences were angled perpendicular to the former tumor axis (i.e. the fibrotic remnant) as visualized on the sagittal planning scan. Image angulation was consistent over time for the various follow-up scans. Detailed sequence parameters of the DWI-sequences used during the study period are provided in Table 1. For the scans performed with a preparatory micro-enema (from March 2014), the micro-enema consisted of a 5 ml solution, that was self-administered by the patients ±15 minutes prior to acquisition. Apart from the micro-enema no bowel preparation or spasmolytic agents were applied.

Table 1. DWI sequences used during the study period

Parameter	DWI-1	DWI-2	DWI-3	DWI-4
Repetition Time	4652-5727	1240-3721	2624-5308	4186-4549
Echo Time	70-73	81-104	64-79	69-71
No. of Slices	24	20	20-24	20
FOV (mm)	320-370	153-180	247-320	320
In-plane Resolution (mm x mm)	1.25 x 1.25	1.41 x 1.41	1.25 x 1.25	1.25 x 1.25
Slice Thickness (mm)	5	5	5	5
Phase encoding direction	AP	AP	AP	AP
Parallel imaging (SENSE) Factor	2	2	2	2
Number of Signals Averaged (NSA)	5	8	5	5
Flip Angle	70	90	70	70
Acquisition Matrix	256	128	256	256-320
Echo Planar Imaging (EPI) Factor	55-77	55-99	61-91	61-97
Fat Saturation Technique	SPIR	SPIR	SPAIR	SPAIR
b-values*	0,500, **1000**	0, (500), **1000**	0, (25, 50, 100), 500, **1000**, (2000)	0, **1000**

FOV = Field of View, AP = Anterior-Posterior, SPIR = Spectral Presaturation with Inversion Recovery, SPAIR = Spectral Attenuated Inversion Recovery.

** Only the b1000 images were used for image evaluation in the current study.*

Image assessment

In total 335 scans (in 50 patients) were analyzed by an experienced reader (DMJL) who scored the presence and severity of gas-related susceptibility artefacts using a 6-point score (0 = no artefact, 1 = mild artefact, 2 = mild – moderate artefact, 3 = moderate artefact, 4 = moderate – severe artefact, 5 = severe artefact). The scoring system with representative imaging examples is illustrated in Figure 1.

A second experienced reader (JJMVG) independently analyzed a random sample of n=100 scans (50 scans without preparation, 50 scans after application of a micro-enema) using the same scoring system to study inter-observer effects. Both readers were blinded to clinical patient data, whether or not the patient had undergone a preparatory micro-enema and to each other's results.

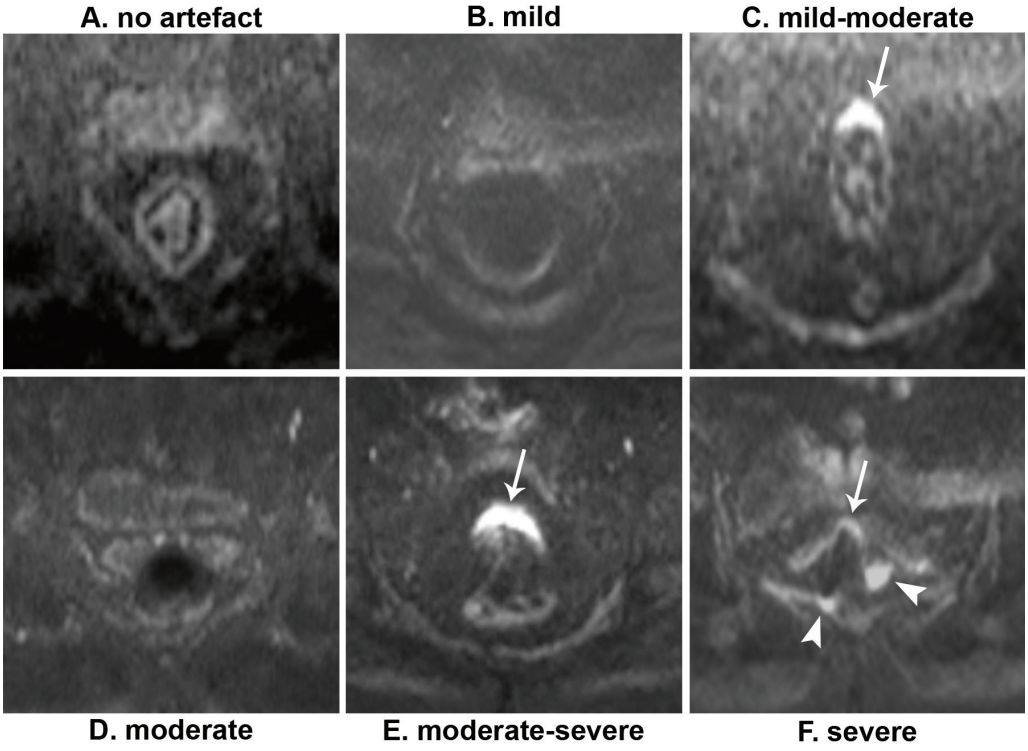


Figure 1. Artefact severity score. *The severity of gas-induced susceptibility artefacts was scored on b1000 DWI using a 6-point scale ranging from 'no artefacts' to 'severe artefacts'. Artefacts with a score of ≥3 were considered clinically relevant artefacts, i.e. artefacts rendering the images to be of insufficient diagnostic quality for adequate clinical image interpretation. (A) Good visualization of the rectal wall and rectal lumen without any visible artefacts, (B) mild distortion of the rectal wall, (C) limited signal pile-up overlapping with the anterior rectal wall (arrow), (D) marked image distortion due to which the rectal wall cannot properly be assessed, (E) marked distortion combined with significant signal pile-up anteriorly (arrow), (F) severe distortion of the rectal wall (arrow) with severe signal pile-up (arrowheads).*

Statistical analysis

The study outline is graphically illustrated in Figure 2: to assess the effect of the introduction of the preparatory rectal micro-enema on DW image quality, the presence and severity of gas-related artefacts were compared between the scans acquired between January 2012 and March 2014 (= 'without micro-enema') and scans acquired between July 2014 and February 2016 (= 'with micro-enema'). Scans acquired between March 2014 and July 2014 (the transit period) were excluded from the analyses as in this period some patients may not yet have received the micro-enema routinely. Primary outcome was the proportion of 'clinically relevant' artefacts in the no micro-enema versus micro-enema scans. Artefacts were considered to be clinically significant if they would considerably hamper clinical DW image interpretation and were defined for the purpose of this study as artefacts with a severity score of ≥3 (moderate, moderate-severe and severe). The artefact severity score itself was assessed as a secondary outcome. The influence of potential confounding factors was assessed using univariable and multivariable logistic regression with a binary outcome for the primary outcome (clinically relevant artefact yes/no) and ordinal logistic regression for the secondary outcome (6-point severity score). Generalized linear models with generalized estimating equations for clustered data were applied, with scans clustered by patient. The working correlation was exchangeable for the binary logistic regression[19] and independent for the ordinal logistic regression. A p-value of <0.05 was considered significant. Results for the binary logistic regression differed very little when an independent working correlation was used. Inter-observer agreement was assessed using quadratic weighted Cohen's Kappa (0-0.20 = poor, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = good and 0.81-1.00 = excellent agreement). Statistical analysis was performed using the Statistical Package for the Social Sciences (SPSS version 23.0, IBM® SPSS® Inc. Chicago, IL) and Statistics Analysis Software (SAS version 9.4, SAS® Institute Inc., Cary, NC).

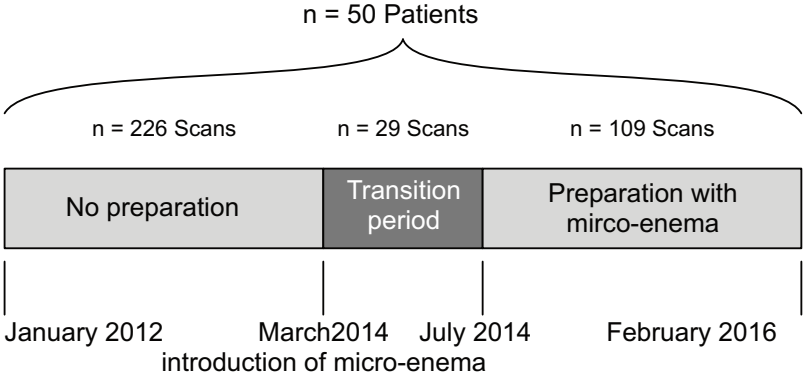


Figure 2. Study outline. *Fifty patients were included. In these patients, 226 scans were acquired from January 2012-March 2014 without any bowel preparation (the no micro-enema group). From July 2014-February 2016, 109 scans were acquired with a preparatory micro-enema (the micro-enema group). Scans made in the transition period after the introduction of the micro-enema (March 2014-July 2014; n=29) were excluded from the analysis.*

Results

Baseline characteristics

Table 2 shows the baseline study characteristics. The 50 study patients together underwent a total of 364 DWI-MRI examinations. Twenty-nine scans from the transit period after the introduction of the micro-enema (March 2014-June 2014) were excluded. This left a total of 335 scans for analysis (mean 6.7 examinations per patient, range 2-10), of which 226 (67.5%) were acquired without a micro-enema (before March 2014) and 109 (32.5%) were acquired with a micro-enema (after June 2014).

Table 2. Baseline characteristics and results for the scans performed without and with a preparatory micro-enema

Baseline characteristics	**No Enema N = 226 scans**	**Enema N = 109 scans**	**Results**	**No Enema N = 226 scans**	**Enema N = 109 scans**
Mean age (±SD)*	63.4 (±11.9)	65.5 (±12.2)	Clinically relevant artefacts		
			Yes	55 (24.3%)	4 (3.7%)
Gender			No	171 (75.7%)	105 (96.3%)
Male	143 (63.3%)	70 (64.2%)			
Female	83 (36.7%)	39 (35.8%)	Severity		
			None	131 (58.0%)	86 (78.9%)
MRI-Hardware			Mild	29 (12.8%)	17 (15.6%)
Philips Intera (Achieva)	113 (50%)	28 (25.7%)	Mild-Moderate	11 (4.9%)	2 (1.8%)
Philips Ingenia	113 (50%)	81 (74.3%)	Moderate	19 (8.4%)	2 (1.8%)
			Moderate-Severe	18 (8.0%)	2 (1.8%)
DWI sequence (see Table 1)			Severe	18 (8.0%)	0 (0.0%)
DWI-1	51 (22.6%)	0 (0.0%)			
DWI-2	60 (26.6%)	77 (70.6%)			
DWI-3	115 (50.9%)	6 (5.5%)			
DWI-4	0 (0.0%)	26 (23.9%)			
Rectoscopy prior to MRI					
Yes	137 (60.6%)	59 (54.1%)			
No	89 (39.4%)	50 (45.9%)			

** At the time of image acquisition.*

Numbers are absolute numbers, unless otherwise indicated percentages are given in parentheses. SD = Standard deviation

Intersobserver agreement

Agreement between the two readers for the use of the 6-point artefact severity score (illustrated in Figure 1) was excellent with a weighted Kappa of 0.85 (95% confidence interval 0.77-0.94).

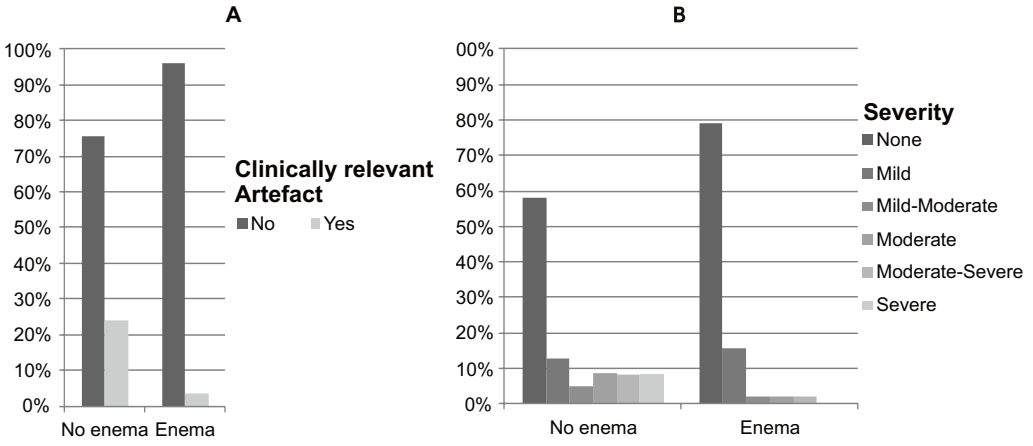


Figure 3. Results. *Boxplots comparing the presence of clinically relevant artefacts (A) and the artefact severity score (B) between the scans acquired without bowel preparation and after application of a preparatory micro-enema.*

Effect of micro-enema on DW image quality

Figure 3 compares the number of clinically relevant artefacts and the severity of artefacts between the group of scans without and with a micro-enema. In the group without a micro-enema, clinically relevant artefacts (severity score ≥3) occurred in 55/226 (24.3%) scans versus 4/109 scans (3.7%) in the group with a micro-enema (Fig 3a). The reduction in the number of clinically relevant artefacts was significant with an odds ratio (OR) of 0.118 (P=0.0005) in univariable binary logistic regression analysis and an OR of 0.230 (P=0.029) in multivariable analysis. The overall artefact severity score was also higher in the group without a micro-enema (Fig 3b), with a mean severity score of 1.19 (± 1.71), compared to 0.32 (± 0.77) in the scans acquired with a micro-enema. The reduction in the severity of artefacts was significant with an OR of 0.321 (*P*<0.0001) in univariable ordinal logistic regression analysis and an OR of 0.489 (*P*=0.046) in multivariable analysis.

Of all available variables to adjust for confounding, b-values, acquisition matrix size, flip angle, Number of Signals Averaged, Field of View (FOV) right-left and anterior-posterior, and water-fat shift were excluded because they were either collinear or highly correlated with another variable (correlation coefficient > 0.8). The remaining variables (gender, age at scan time, repetition time, echo time, FOV cranial-caudal, EPI factor, fat saturation technique, MRI-hardware, and flexible rectoscopy performed <12 hours prior to MRI) were included in all

multivariable models. When these variables, including those excluded due to high correlation, were added to a model with the micro-enema effect one by one, no substantial confounding was observed (Table 3).

Detailed results of the uni- and multivariable analyses are provided in Table 3. A representative example of how gas-related artefacts reduced over time after the introduction of the routine use of a micro-enema is provided in Figure 4 for a patient scanned both without preparation and after application of a micro-enema.

Table 3. Results from uni- and multivariable logistic regression

	Presence of clinically relevant artefacts†		**Severity of artefacts‡**	
	Odds ratio (95% CI)	P-value	Odds ratio (95% CI)	P-value
Effect of preparatory micro-enema				
Univariable	0.118 (0.035-0.396)	0.0005	0.321 (0.193-0.533)	<.0001
Multivariable*	0.230 (0.061-0.861)	0.0291	0.489 (0.242-0.988)	0.0461
Effect of potential confounders				
Age*	0.982 (0.950-1.016)	0.3069	0.995 (0.971-1.020)	0.7110
Gender*	0.687 (0.317-1.489)	0.3413	0.591 (0.349-1.003)	0.0511
Rectoscopy*	1.643 (0.913-2.958)	0.0978	1.478 (0.890-2.454)	0.1309
MRI hardware system*‡	0.971 (0.567-1.661)	0.9143	0.996 (0.647-1.535)	0.9864
Repetition Time*	1.000 (1.000-1.001)	0.1468	1.000 (1.000-1.000)	0.1951
Echo Time*	0.973 (0.941-1.006)	0.1049	0.986 (0.958-1.015)	0.3261
Flip Angle	0.610 (0.334-1.112)	0.1064	0.726 (0.449-1.173)	0.1912
Number of Signals Averaged	0.610 (0.334-1.112)	0.1064	0.726 (0.449-1.173)	0.1912
Water-Fat Shift	1.085 (0.981-1.199)	0.1134	1.075 (0.978-1.183)	0.1347
Echo Planar Imaging (EPI) Factor*	1.007 (0.966-1.050)	0.7272	1.016 (0.982-1.050)	0.3660
Fat Saturation Technique*#	0.861 (0.508-1.462)	0.5801	0.880 (0.542-1.429)	0.6051
Field of View size (mm), anterior-posterior	1.005 (1.000-1.010)	0.0562	1.003 (0.998-1.007)	0.2204
Field of View size (mm), cranial-caudal*	0.994 (0.985-1.003)	0.1605	1.000 (0.994-1.005)	0.8626
Field of View size (mm), right-left	1.004 (0.999-1.008)	0.0908	1.002 (0.999-1.006)	0.2239
Number of b values acquired±	1.057 (0.871-1.284)	0.5715	1.028 (0.858-1.232)	0.7647

NB. CI = Confidence Interval. Results were clustered by patient, since patients contribute multiple scans.

† Binary logistic regression, ‡ Ordinal logistic regression

*Confounders indicated with an * were included in the multivariable analysis*

| All scans were performed on 1.5T using either a Intera (Achieva) or Ingenia MRI system from Philips

 # All scans were performed using either Spectral Presaturation with Inversion Recovery (SPIR) or Spectral Attenuated Inversion Recovery (SPAIR).

± Scans were performed with 2, 3, 6 or 7 b-values, the highest b-value (used for evaluation) always being b1000, see Table 1

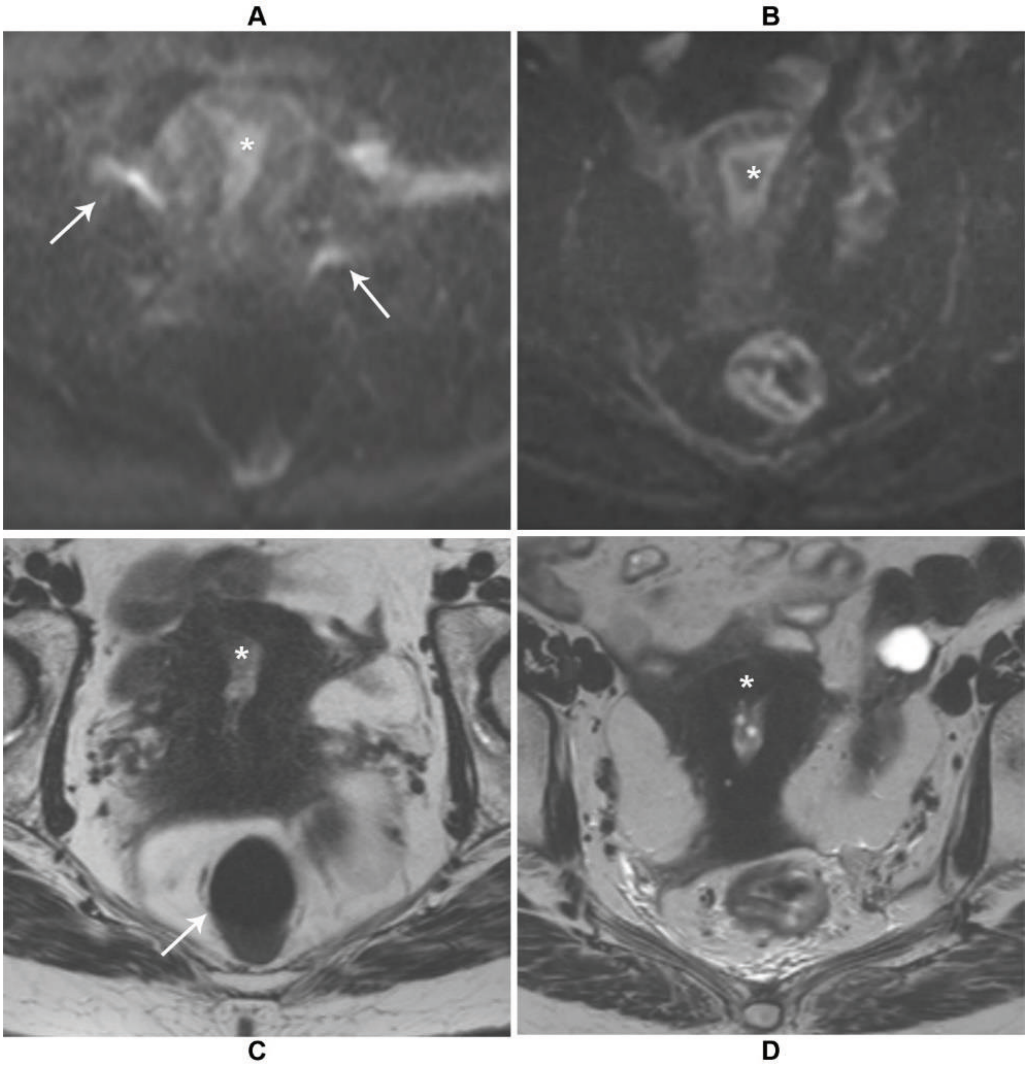


Figure 4. Examples. *Example of the b1000 DWI (A,B) and corresponding T2-weighted images (C, D) of a patient in whom a severe artefact was observed on a scan without bowel preparation (A, C), and no artefact was observed on a later follow-up scan after application of a micro-enema (B, D). All images show the rectum at the level of the uterus (*). On DWI without an enema (A) there is marked signal pile-up (arrows) and pronounced image distortion, due to which the rectal wall cannot be assessed. The corresponding T2W image (C) shows a substantial amount of gas in the rectal lumen (arrow). On the later follow-up DWI scan performed after a preparatory micro-enema (B), there are no artefacts and the rectal wall is clearly visualized. On the corresponding T2W image (D) there is much less gas present within the rectal lumen following the micro-enema.*

Discussion

The results of our study show that the application of a preparatory micro-enema shortly before image acquisition significantly reduces both the incidence as well as the severity of gas-induced susceptibility on rectal DWI performed at 1.5T. The use of a micro-enema reduces the proportion of rectal DWI scans suffering from clinically relevant artefacts – i.e. artefacts that hamper clinical image interpretation – from 1 in every 4 scans to 1 in every 20 scans, thereby offering a substantial potential clinical benefit in terms of improved diagnostic image quality.

To our knowledge only one previous study specifically reported on the use of an enema for the reduction of susceptibility artefacts on DWI. In this study by Lim et al.[20], diagnostic image quality of multiparametric MRI of the prostate was compared at 3.0T in patients without bowel preparation versus patients who were instructed to self-administer an enema on the morning of the day they received their MRI examination. Interestingly, while application of an enema resulted in significantly less stool and gas in the rectum, no significant difference was found in the diagnostic quality of the enema group compared to the no-preparation group.
These seemingly conflicting results may be in part caused by the small number of scans assessed (n= 60, compared to over 300 scans in the current report), with a relatively low number of events (severe artefacts) ranging from 1-7 per group. Another possible explanation is the timing of the enema, which was self-administered in the morning of the day of the examination in the study by Lim et al., compared to 15-30 minutes prior to examination in our study. Finally the target organ under investigation in the study of Lim was the prostate and not the rectum itself. Various alternative methods to reduce artefacts have been reported, most of which focus on making the EPI sequences more robust. An example is the use of parallel imaging techniques. Parallel imaging allows for less phase-encoding steps, thereby reducing both sensitivity to artefacts and acquisition time[12]. In our study parallel imaging was also employed (factor 2.0). An alternative option is reducing the FOV, allowing for a higher spatial resolution in the phase-encoding direction, where the EPI sequence is most influenced by the susceptibility artefacts. Korn et al.[13] assessed the use of reduced FOV-excitation in DWI for prostate cancer detection using a 5 point scale for image distortion and showed that the reduced FOV significantly reduced image distortion scores by 0.48-0.56 points. Thian et al.[21] explored the use of a read-out segmented (rs) EPI Sequence, in which the k-space is filled in several segments (as opposed to a standard single shot technique). In n=30 pelvic DWI examinations, lesion conspicuity was significantly better and geometric distortions significantly less on the rs-EPI sequence. The reported magnitude of effect of these technical alterations was less than that of the use of a preparatory micro-enema (combined with parallel imaging) in our study. Although with this approach we could already reduce the number of clinically relevant artefacts to less than 5%, future research should focus on further optimizing scan quality by combining patient preparation with artefact reduction acquisition techniques such as the examples described above to ultimately offer the best possible DWI protocols with stable and robust diagnostic image quality, on which radiologist may truly rely on for diagnostic decision making. Although we did not perform formal questionnaires to objectively quantify the degree of patient discomfort for the self-administration of the micro-enema, in our experience patients tolerated it very well and considered it a minimal extra burden.

There are some limitations to our study design. First, we fully acknowledge that an important drawback of our study design is that none of the study patients had rectal tumors in situ at the time of image acquisition, making it impossible to study effects of artefact reduction on lesion conspicuity. All MRIs were performed in patients with a clinical complete response who underwent multiple follow-up MRIs (3 to 6-monthly) for a longer time period. We chose this approach as a first exploratory step as it offered the benefit of a consistent clinical setting (in which patients underwent no therapeutic procedures in between scans that may influence image quality) which creates the unique opportunity to perform within patient comparisons, thereby reducing effects of interpatient variability. The obvious next step will be to study the clinical benefit of our approach in the staging and restaging setting to determine effects on lesion conspicuity and staging outcomes. Second, all MRIs were performed at 1.5T. It would be interesting to see how the results of our study translate to 3.0T where susceptibility effects will typically be more severe[22] and the gain of applying a micro-enema may thus be more profound. Third, a substantial number of patients underwent flexible rectoscopy just prior to the MRI, because they were in a follow-up protocol that included MRI and rectoscopy on the same day. During this procedure gas is introduced into the rectum, which may have negative effects on MR image quality when performed shortly afterwards. On the other hand, the endoscopists in our center typically perform a de-sufflation of the rectum, whilst removing the endoscope which will reduce the amount of gas in the lumen. Moreover, the fact that patients underwent rectoscopy just before their MRI did not have a significant confounding effect in our statistical analyses. Finally, the artefacts in this study were assessed using a more or less subjective scoring system focusing on overall diagnostic image quality. In order to reduce subjectivity we however used a standardized 6-point scoring system which led to excellent interobserver agreement between two independent readers (κ0.85).

In conclusion, the use of a preparatory micro-enema shortly prior to image acquisition significantly reduces both the incidence and severity of gas-related susceptibility artefacts in DWI of the rectum performed at 1.5T. A preparatory micro-enema can easily be self-administered just minutes before the MR examination. As such we believe that – when DWI forms an integral part of the imaging assessment of the rectum – a micro-enema should be considered, as it provides a significant benefit to the image quality at a relatively small cost in terms of preparation time and patient discomfort.

1. Rao SX, Zeng MS, Chen CZ, et al. The value of diffusion-weighted imaging in combination with T2-weighted imaging for rectal cancer detection. *Eur J Radiol.* 2008; 65(2): 299-303.

2. Kim SH, Lee JYJMY, Hong SH, et al. Locally advanced rectal cancer: added value of diffusion-weighted MR imaging in the evaluation of tumor response to neoadjuvant chemo- and radiation therapy. *Radiology.* 2009; 253(1): 116-125.

3. Lambregts DMJ, Vandecaveye V, Barbaro B, et al. Diffusion-weighted MRI for selection of complete responders after chemoradiation for locally advanced rectal cancer: a multicenter study. *Ann Surg Oncol.* 2011; 18(8): 2224-2231.

4. Lambregts DMJ, Lahaye MJ, Heijnen LA, et al. MRI and diffusion-weighted MRI to diagnose a local tumour regrowth during long-term follow-up of rectal cancer patients treated with organ preservation after chemoradiotherapy. *Eur Radiol.* 2016; 26(7): 2118-2125.

5. Le Bihan D, Poupon C, Amadon A, Lethimonnier F. Artifacts and pitfalls in diffusion MRI. *J Magn Reson Imaging.* 2006; 24(3): 478-488.

6. Koh D-M, Takahara T, Imai Y, Collins DJ. Practical aspects of assessing tumors using clinical diffusion-weighted imaging in the body. *Magn Reson Med Sci.* 2007; 6(4): 211-224.

7. Caglic I, Hansen NL, Slough RA, Patterson AJ, Barrett T. Evaluating the effect of rectal distension on prostate multiparametric MRI image quality. *Eur J Radiol.* 2017; 90: 174-180.

8. Regini F, Gourtsoyianni S, Cardoso De Melo R, et al. Rectal tumour volume (GTV) delineation using T2-weighted and diffusion-weighted MRI: Implications for radiotherapy planning. *Eur J Radiol.* 2014; 83(5): 768-772.

9. Blazic IM, Lilic GB, Gajic MM. Quantitative Assessment of Rectal Cancer Response to Neoadjuvant Combined Chemotherapy and Radiation Therapy: Comparison of Three Methods of Positioning Region of Interest for ADC Measurements at Diffusion-weighted MR Imaging. *Radiology.* 2017; 282(2): 418-428.

10. Choi MH, Oh SN, Rha SE, et al. Diffusion-weighted imaging: Apparent diffusion coefficient histogram analysis for detecting pathologic complete response to chemoradiotherapy in locally advanced rectal cancer. *J Magn Reson Imaging.* 2016; 44(1): 212-220.

11. Foti PV, Privitera G, Piana S, et al. Locally advanced rectal cancer: Qualitative and quantitative evaluation of diffusion-weighted MR imaging in the response assessment after neoadjuvant chemo-radiotherapy. *Eur J Radiol open.* 2016; 3: 145-152.

12. Nasu K, Kuroki Y, Kuroki S, Murakami K, Nawano S, Moriyama N. Diffusion-weighted single shot echo planar imaging of colorectal cancer using a sensitivity-encoding technique. *Jpn J Clin Oncol.* 2004; 34(10): 620-626.

13. Korn N, Kurhanewicz J, Banerjee S, Starobinets O, Saritas E, Noworolski S. Reduced-FOV excitation decreases susceptibility artifact in diffusion-weighted MRI with endorectal coil for prostate cancer detection. *Magn Reson Imaging.* 2015; 33(1): 56-62.

14. Kyriazi S, Blackledge M, Collins DJ, Desouza NM. Optimising diffusion-weighted imaging in the abdomen and pelvis: comparison of image quality between monopolar and bipolar single-shot spin-echo echo-planar sequences. *Eur Radiol.* 2010; 20(10): 2422-2431.

15. Kaur H, Choi H, You YN, et al. MR imaging for preoperative evaluation of primary rectal cancer: practical considerations. *Radiographics.* 2012; 32(2): 389-409.

16. Dal Lago A, Minetti AE, Biondetti P, Corsetti M, Basilisco G. Magnetic resonance imaging of the rectum during distension. *Dis Colon Rectum.* 2005; 48(6): 1220-1227.

17. Slater A, Halligan S, Taylor SA, Marshall M. Distance between the rectal wall and mesorectal fascia measured by MRI: Effect of rectal distension and implications for preoperative prediction of a tumour-free circumferential resection margin. *Clin Radiol.* 2006; 61(1): 65-70.

18. Beets-Tan RG, Lambregts DM, Maas M, et al. Magnetic resonance imaging for the clinical management of rectal cancer patients: recommendations from the 2012 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. *Eur Radiol.* 2013; 23(9): 2522-2531.

19. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; 73(1): 13-22.

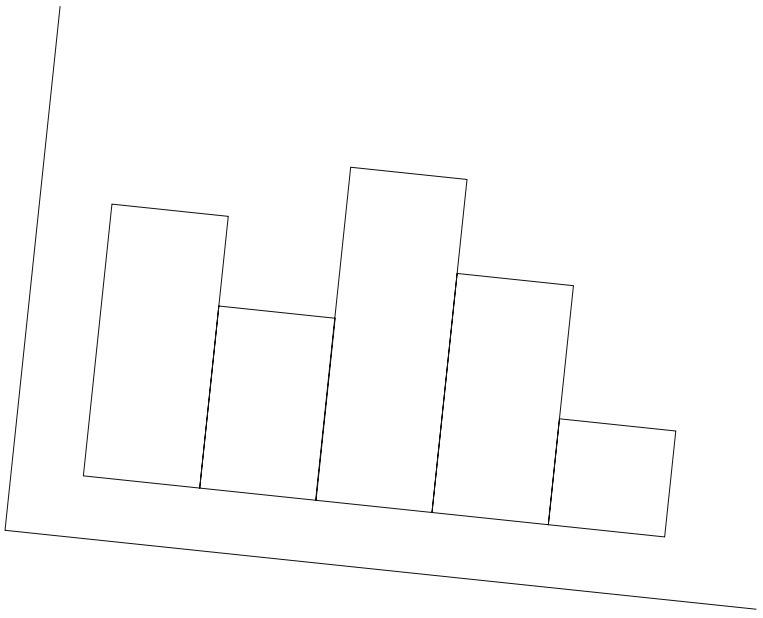
20. Lim C, Quon J, McInnes M, Shabana WM, El-Khodary M, Schieda N. Does a cleansing enema improve image quality of 3T surface coil multiparametric prostate MRI?
J Magn Reson Imaging. 2015; 42(3): 689-697.

21. Thian YL, Xie W, Porter DA, Weileng Ang B. Readout-segmented echo-planar imaging for diffusion-weighted imaging in the pelvis at 3T-A feasibility study. *Acad Radiol*. 2014; 21(4): 531-537.

22. Riffel P, Rao RK, Haneder S, Meyer M, Schoenberg SO, Michaely HJ. Impact of field strength and RF excitation on abdominal diffusion-weighted magnetic resonance imaging.
World J Radiol. 2013; 5(9): 334-344.

Feature extraction and model-ling

Computational Radiomics System to Decode the Radiographic Phenotype



Chapter 5 Computational Radiomics System to Decode the Radiographic Phenotype

Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, Hugo J.W.L. Aerts

Published in:

Cancer Research 2017; 77:e104–e107 (IF 2019 9.727)

Abstract

Radiomics aims to quantify phenotypic characteristics on medical imaging through the use of automated algorithms. Radiomic artificial intelligence (AI) technology, either based on engineered hard-coded algorithms or deep learning methods, can be used to develop non-invasive imaging-based biomarkers. However, lack of standardized algorithm definitions and image processing severely hampers reproducibility and comparability of results. To address this issue, we developed *PyRadiomics*, a flexible open-source platform capable of extracting a large panel of engineered features from medical images. *PyRadiomics* is implemented in Python and can be used standalone or using 3D-Slicer. Here, we discuss the workflow and architecture of *PyRadiomics* and demonstrate its application in characterizing lung-lesions. Source code, documentation, and examples are publicly available at www.radiomics.io. With this platform, we aim to establish a reference standard for radiomic analyses, provide a tested and maintained resource, and to grow the community of radiomic developers addressing critical needs in cancer research.

Chapter 5

Introduction

Medical imaging is considered one of the top innovations that transformed clinical cancer care, as it significantly changed how physicians measure, manage, diagnose, and treat cancer. Imaging is able to noninvasively visualize the radiographic phenotype of a tumor before, during, and after treatment. Radiomics refers to the comprehensive and automated quantification of this radiographic phenotype using data-characterization algorithms[1–3]. Radiomics can quantity a large panel of phenotypic characteristics, such as shape and texture, potentially reflecting biologic properties like intra- and inter-tumor heterogeneities[4].

Radiomic technologies, based on artificial intelligence (AI) methods, are either defined using engineered hard-coded features, which often rely on expert domain knowledge, or on deep learning methods, which can learn feature representations automatically from data[5]. The potential of radiomics has been shown across multiple tumor types, including brain, head-and-neck, cervix, and lung cancer tumors. Furthermore, these data, extracted from MRI, PET or CT images, were associated with several clinical outcomes, and hence, potentially provide complementary information for decision support in clinical oncology[1].

However, there is a lack of standardization of both feature definitions and image processing, which has been shown to have a substantial impact on the reliability of radiomic data[6–8]. Furthermore, many studies use in-house developed software, often not shared with the public, making the reproduction and comparison of results difficult.

To address this issue, we developed a comprehensive open-source platform, called *PyRadiomics*, which enables processing and extraction of radiomic features from medical image data using a large panel of engineered hard-coded feature algorithms. *PyRadiomics* provides a flexible analysis platform with both a simple and convenient front-end interface in 3D Slicer, a free open-source platform for medical image computing[9], as well as a back-end interface allowing automation in data processing, feature definition, and batch handling. *PyRadiomics* is implemented in Python, a language that has established itself as a popular open-source language for scientific computing, and can be installed on any system.

Here, we discuss the workflow and architecture of PyRadiomics and demonstrate its application in characterizing benign and malignant lung lesions. Source code, documentation, instruction videos (see Video 1 and 2), and examples are available at www.radiomics.io/pyradiomics.html. With this resource, we aim to establish a reference standard for radiomic analyses, provide a tested and maintained open-source platform, and raise the awareness among scientists of the potential of radiomics technologies.

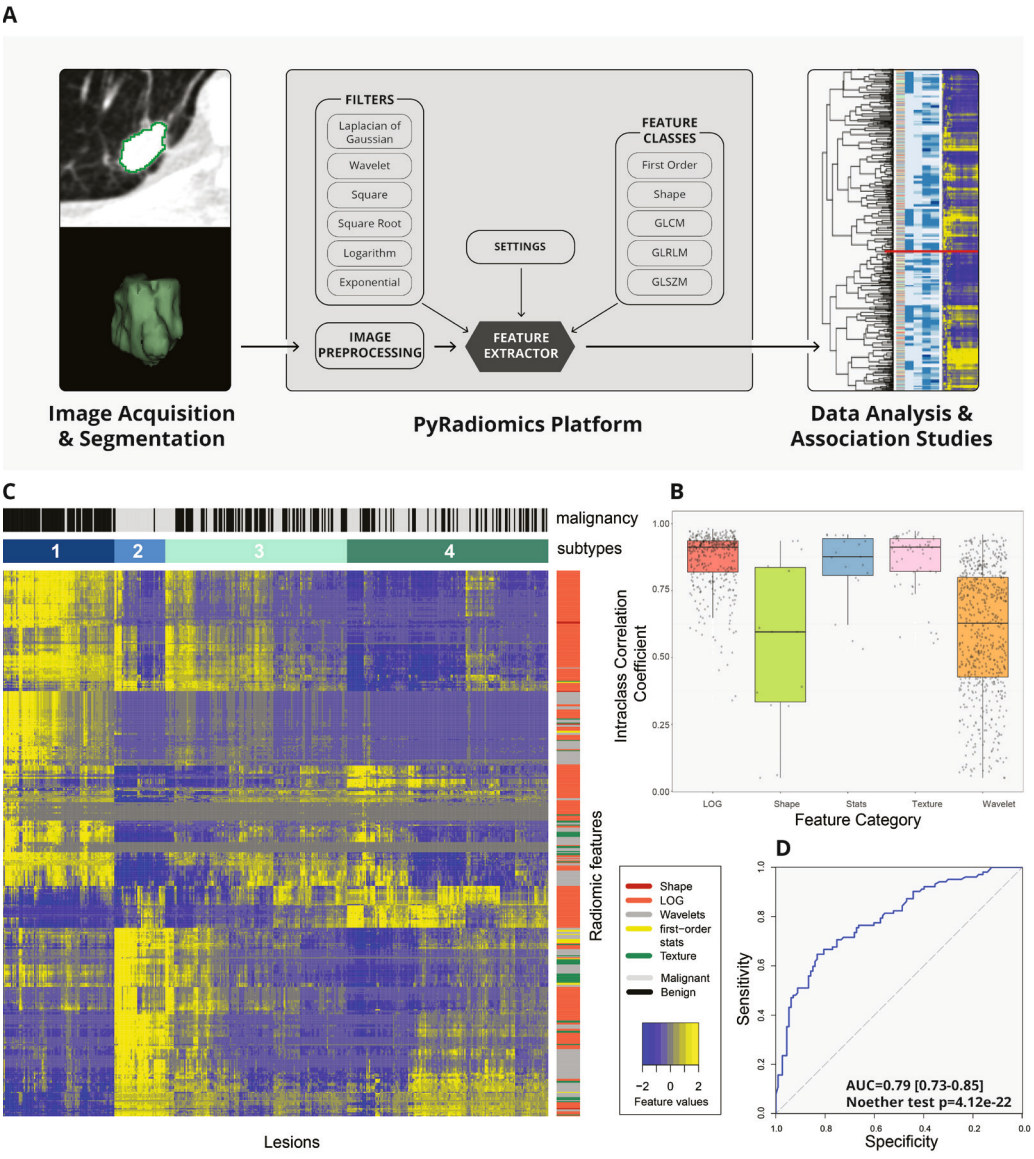


Figure 1. *A Overview figure of the process of PyRadiomics. First, medical images are segmented. Second, features are extracted using the PyRadiomics platform, and third, features are analyzed for associations with clinical or biologic factors. B Stability of radiomics features for variation in manual segmentations by expert radiologists. C Heatmap showing expression values of radiomics features (rows) of 429 lesions (columns). Note the four subtypes that could be identified from the expression values and their associations with malignancy. D Area under curve (AUC) showing the performance of the multivariate biomarker to predict malignancy of nodules.*

Platform

The PyRadiomics platform can extract radiomic data from medical imaging (such as CT, PET, MRI) using four main steps: I) Loading and preprocessing of the image and segmentation maps, II) Application of enabled filters, III) Calculation of features using the different feature classes, and IV) Returning results. See Figure 1A for an illustration of this process.

I) Loading and preprocessing: In this step, medical images (e.g. CT, PET, MRI) and segmentation maps (e.g. performed by radiologist) will be loaded into the platform. The large majority of image handling is done using *SimpleITK*, which provides a streamlined interface to the widely used open-source Insight Toolkit (ITK)[10]. This enables *PyRadiomics* to support a wide variety of image formats, while also ensuring that much of the low-level functionality and basic image processing is thoroughly tested and maintained. For texture and shape features, several resampling options are included to ensure isotropic voxels with equal distances between neighbouring voxels in all directions.

II) Filtering: Features can be calculated on the original image or on images pre-processed using a choice of several built-in filters. These include wavelet and laplacian of gaussian (LoG) filters, as well as several simple filters, including square, square root, logarithm, and exponential filters. For the application of the wavelet and LoG filter, the platform makes use of the *PyWavelets* and *SimpleITK*, respectively. The remaining filters are implemented using *NumPy*.

III) Feature calculation: The platform contains five feature classes: a class for first-order statistics, a class for shape descriptors, and texture classes Gray Level Co-occurrence Matrix[11], Gray Level Run Length Matrix[12,13], and Gray Level Size Zone Matrix[14]. All statistic and texture classes can be used for feature extraction from both filtered and unfiltered images. Shape descriptors are independent from intensity values and therefore can only be extracted from unfiltered images. Feature extraction is supported for both single slice (2D) and whole volume (3D) segmentations.

IV) Results: Calculated features are stored and returned in an ordered dictionary. Every feature is identified by a unique name consisting of the applied filter, the feature class and feature name. Besides the calculated features, this dictionary also contains additional information on the extraction, including current version, applied filters, settings, and original image spacing.

To enhance usability, *PyRadiomics* has a modular implementation, centered around the *featureextractor* module which defines the feature extraction pipeline and handles interaction with the other modules in the platform. All feature classes are defined in separate modules. Furthermore, all are inherited from a base feature extraction class, providing a common interface. Finally, the platform contains two helper modules, *generalinfo* that provides additional extraction information included in the returned result, and the *imageoperations* module that implements the functions used during image preprocessing and filters.

Aside from interactive use in Python scripts through the *featureextractor* module, *PyRadiomics* supports direct usage from the command line. There are two scripts available, *pyradiomics* and *pyradiomicsbatch*, for single image and batch processing, respectively. For both scripts, an additional parameter file can be used to customize the extraction and results can be directly imported into many statistical packages for analysis, including R and SPSS. Additionally, a convenient front-end interface for *PyRadiomics* is provided as the 'radiomics' extension within 3D Slicer. All code, including the Slicer extension, documentation, frequently asked questions, and instruction videos (see Video 1 and 2) are available at www.radiomics.io/pyradiomics.html. In the supplementary information detailed descriptions of feature definitions, dataset, and analyses can be found.

Case Study

In a case study, we demonstrated an application of *PyRadiomics* for lung lesion characterization to discriminate between benign and malignant nodules. We used the publicly available cohort of the Lung Image Database Consortium (LIDC-IDRI)[15], which consists of diagnostic and lung cancer screening CT scans along with marked-up annotated lesions and per-lesion malignancy rating (i.e. if a nodule is benign or malignant) from experienced radiologists (Supplementary Methods 1). From 302 patients, we included 429 distinct lesions in our analysis, each with four volumetric segmentations and malignancy ratings. In total, 1120 radiomic features (14 shape features, 19 first-order intensity statistics features, 60 texture features, 395 LOG features and 632 wavelet features) were extracted from all four delineations of every lesion (Supplementary Methods 2-4).

To assess the effect variations in the manual segmentations on radiomic feature values, we calculated the stability for each of the features extracted from four segmentations performed by expert radiologists. This stability was calculated using intraclass correlation coefficient (ICC) (Figure 1B). High stability (median ± sd: ICC > 0.8) was observed for LOG (ICC= 0.91 ± 0.11), first-order intensity statistics (ICC= 0.88 ± 0.13) and texture features (ICC= 0.91 ± 0.11), whereas shape (ICC= 0.60 ± 0.31) and wavelet (ICC= 0.63 ± 0.23) features showed moderate stability, which indicates their sensitivity towards delineation variability.

Selecting all features with high stability (ICC>0.8), resulted in 535 radiomic features (5 shape features, 14 first-order intensity statistics features, 48 texture features, 310 LOG features and 158 wavelet features). Figure 1C displays unsupervised clustering of the standardized expression values of the 535 stable radiomic features (rows) in 429 nodules (columns). We observed four distinct clusters of lesions with similar expression values. Comparing these clusters with lesion malignancy status, we observed significant difference between them ($P = 2.56e{-}24$, χ^2 test). 92% (n= 81) of the samples of cluster S1 (n = 88) were malignant, whereas 95% (n = 38) of the samples of cluster S2 (n = 40) were benign. For cluster S3 (n = 143) and S4 (n = 158) the proportion of malignant samples were 54% (n = 78) and 34% (n = 53) respectively.
These results demonstrate associations between imaging-based subtypes and malignancy status of lung lesions.

In order to evaluate the performance of a multivariate imaging biomarker, we divided the cohort into training (n=214) and validation (n=215). Using minimum Redundancy Maximum Relevance (mRMR), we selected 25 stable radiomic features from the training cohort (Supplementary Table 1). An multivariate biomarker was developed by fitting selected features into a random forest classifier, based on the training data. The biomarker demonstrated strong and significant performance to characterize lung nodules (AUC=0.79 [0.73-0.85], Noether test p-value=4.12e-22) on the validation cohort (Figure 1D). More details on features extraction and analysis methods are provided in the Supplements.

Conclusion

PyRadiomics provides a flexible radiomic quantification platform, with a simple and convenient front-end interface in 3D Slicer, as well as a back-end interface within Python allowing automation in data processing, feature definition, and batch handling. By providing a tested and maintained open-source radiomics platform, we aim to establish a reference standard for radiomic analyses promoting reproducible science within the quantitative imaging field, raise awareness among scientists of this platform to support their work, and to provide a practical go-to resource. By doing so, we hope to grow the community of radiomic technology developers to address critical needs in cancer research.

1. Aerts HJWL. The Potential of Radiomic-Based Phenotyping in Precision Medicine. *JAMA Oncol.* 2016; 2(12): 1636.

2. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5(1): 4006.

3. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012; 48(4): 441-446.

4. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.* 2012; 12(5): 323-334.

5. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015; 521(7553): 436-444.

6. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol.* 2016; 61(13): R150-R166.

7. Orlhac F, Soussan M, Maisonobe J-A, 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(3): 414-422.

8. Tixier F, Hatt M, Cheze-Le Rest C, Le Pogam A, Corcos L, Visvikis D. Reproducibility of Tumor Uptake Heterogeneity Characterization Through Textural Feature Analysis in 18F-FDG PET. *J Nucl Med.* 2012; 53(5): 693-700.

9. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging.* 2012; 30(9): 1323-1341.

10. Johnson HJ, Mccormick MM. The ITK Software Guide Book 2 : Design and Functionality Fourth Edition Updated for ITK version 4. 2016.

11. Haralick R, Shanmugan K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973; 3: 610-621.

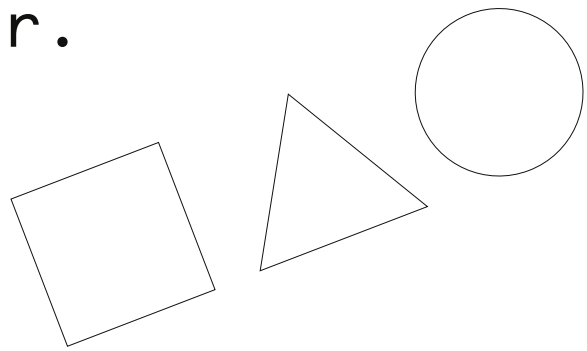
12. Galloway MM. Texture analysis using gray level run lengths. *Comput Graph Image Process.* 1975; 4(2): 172-179.

13. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run lengths for texture analysis.
Pattern Recognit Lett. 1990; 11(6): 415-419.

14. Thibault G, Fertil B, Navarro C, et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification.
Pattern Recognit Inf Process. 2009: 140-145.

15. Armato SG, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans.
Med Phys. 2011; 38(2): 915-931.

Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer.



Chapter 6 Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer.

Joost J.M. van Griethuysen, Doenja M.J. Lambregts, Stefano Trebeschi, Max J. Lahaye, Frans C.H. Bakers, Roy F.A. Vliegen, Geerard L. Beets, Hugo J.W.L. Aerts, Regina G.H. Beets-Tan

Published in:

Abdom Radiol 2020; 632-643 (IF 2019 2.429)

Abstract

Purpose

To compare the performance of advanced Radiomics analysis to morphological assessment by expert radiologists to predict a good or complete response to chemoradiotherapy in rectal cancer using baseline staging MRI.

Materials and methods

We retrospectively assessed the primary staging MRIs (prior to chemoradiotherapy (CRT)) of 133 rectal cancer patients from 2 centers. First, two expert-radiologists subjectively estimated the likelihood of achieving a "complete response" (ypT0) and "good response" (TRG 1-2), using a 5-point score (based on TN-stage, MRF/EMVI-status, size/signal/shape). Next, tumor volumes were segmented on high b-value DWI (semi-automated, corrected by 2 non-expert and 2-expert readers, resulting in 5 segmentations), copied to the remaining sequences after which a total of 2505 radiomic features were extracted from T2W, low and high b-value DWI and ADC. Stability of features for noise due to inter-reader and inter-scanner and protocol variations was assessed using intraclass correlation (ICC) and the Kruskal-Wallis test. Using data from center 1 (n=86; training set), top 9 features were selected using minimum-Redundancy Maximum-Relevance and combined in a logistic regression model. Finally, diagnostic performance of the fitted models was assessed on data from center 2 (n=47; validation set) and compared to the performance of the radiologists.

Results

The Radiomic models resulted in AUCs of 0.69-0.79 (with similar results for the segmentations performed by expert/non-expert readers) to predict response, results similar to the morphologic prediction by the expert-radiologists (AUC 0.67-0.83). Radiomics using semi-automatically generated segmentations (without manual input) did not result in significant predictive performance.

Conclusions

Radiomics could predict response to therapy with comparable diagnostic performance as expert radiologists, regardless of whether image segmentation was performed by non-expert or expert readers, indicating that expert input is not required in order for the Radiomics workflow to produce significant predictive performance.

Introduction

According to current standard of care, patients with very distal and/or locally advanced rectal tumors (≥T3 and/or N+) typically receive neoadjuvant chemoradiotherapy (CRT) aiming to achieve downstaging and thereby increasing the chance of a complete surgical resection. As a result of CRT, approximately 15% of patients undergo a complete tumor response[1]. There is a current paradigm shift in treatment towards considering organ-preservation ('watch-and-wait') for these very good responders[1–3]. In addition to assessing response after completion of CRT to select these patients, there is also an increased clinical interest in the prediction of treatment response before the start of CRT. In patients likely to respond well, neoadjuvant treatment may be intensified, for example with an additional radiotherapy boost, to increase the chance of organ preservation. Patients with smaller tumors have a higher response rate[4,5], but according to current standards are typically treated with direct surgery without CRT. However, with a predicted high response rate chemoradiation might be offered to these small tumors as an alternative with the sole aim to achieve organ preservation, whereas patients with radioresistant tumors remain better off with surgery alone, which is the current standard treatment for these tumors. To date, such an approach is obviously still experimental and offered only in trial settings, for example within the STAR-TREC study, a collaborative phase II trial on CRT + organ-preservation for early rectal cancer running in the UK, Denmark and the Netherlands (ClinicalTrials.gov NCT02945566)[6].

Several studies have shown that imaging may play a role in the pre-treatment prediction of response, with a particular focus on MRI being one of the main imaging modalities used to stage rectal cancer. "Semantic features" including the T-stage, N-stage, Circumferential Resection Margin (CRM), Extra-Mural Venous Invasion (EMVI) and baseline tumor volume have been shown to be associated with the chance of response to varying degrees[7–10]. Promising (though inconsistent) results have also been reported for the use of more novel functional MR imaging sequences such as diffusion-weighted imaging (DWI) and dynamic contrast enhanced (DCE) MRI, that can provide quantifiable information on biological tumor properties such as tumor cellularity and tumor perfusion[11–13].

Another highly interesting recent development is Radiomics, a high-throughput post-processing technique capable of extracting large numbers of quantitative "features" from routinely acquired medical imaging[14]. These features can be used to generate a comprehensive radiologic phenotype and can potentially provide us with new insights into underlying biologic tumor characteristics[15–17]. In rectal cancer, a handful of studies investigating Radiomics for response prediction have shown promising results[18–20], albeit mainly in relatively small single center cohorts. So far, no studies exist that have compared the use of Radiomics to subjective estimation of the likelihood of response by radiologists based on an overall visual interpretation of the local tumor stage at baseline MRI. Such a comparison would be an interesting step to provide at least some preliminary perspective on the potential added benefit from Radiomics in a clinical setting. Moreover, data published so far has mainly been based on relatively small and single center study cohorts.

With this study we aim to add to previous research by investigating the potential of Radiomics to predict treatment response in rectal cancer using the baseline staging MRI data from two institutions (to allow a test and validation dataset and to study effects of acquisition heterogeneity) and by comparing the performance of Radiomics to morphological assessment of the images by expert radiologists to provide a first exploratory estimation of its potential clinical benefit.

Methods and Materials

The study was approved by the local institutional review board (of both institutions). Due to the retrospective nature of the study, informed consent was waived.

Study population

We retrospectively identified 133 patients with rectal cancer who underwent long course chemoradiotherapy at one of two study centers (Maastricht University Medical Center and Zuyderland Medical Center Heerlen) between March 2007 and January 2013. Main inclusion criteria were (a) histologically proven primary non-mucinous type rectal adenocarcinoma, (b) locally advanced disease (≥cT3 and/or N+ disease), (c) neoadjuvant treatment consisting of 28 fractions of 1.8 Gy radiotherapy with concurrent capecitabine 825 mg/m² chemotherapy, (c) availability of a multiparametric pre-treatment MR examination including a T2-weighted sequence, a diffusion-weighted sequence and corresponding quantitative 'apparent diffusion coefficient' (ADC) map, and (d) availability of either histology after surgery or long-term (> 2 years) follow-up in case of a wait-and-see program to establish the final treatment response.

Image acquisition

All patients received a primary staging MRI on a 1.5T MR system (Intera or Ingenia MR system; Philips Healthcare, Best, The Netherlands in center 1; Magnetom Avanto; Siemens in center 2).

The imaging protocol included a T2-weigthed turbo spin echo sequence in sagittal, coronal and transverse plane and a transverse EPI-DWI sequence with 1000 or 1100 s/mm² as the highest b-value. Detailed sequence parameters are provided in Table 1. ADC maps were calculated from the DWI sequences using a mono-exponential model including all available b-values. Oblique transverse T2W and DWI sequences were acquired in identical planes perpendicular to the tumor axis as seen on the sagittal T2W scan. The transverse T2-weighted, low b-value DWI (DWI_{b0}), high b-value DWI ($DWI_{b1000/b1100}$) images and the ADC maps were used for radiomic feature extraction.

Standard of reference / clinical outcome

The main clinical study outcomes were:
(1) the prediction of a complete versus incomplete response after chemoradiotherapy.
(2) the prediction of a good versus poor response after chemoradiotherapy.
The final histopathologic tumor stage after surgery including the tumor regression grade (TRG)

according to Mandard[21] served as the main standard of reference. For the first study outcome, patients with a ypT0 / TRG1 were classified 'complete responders', while patients with residual tumor (ypT1-4, TRG 2-4) were classified as 'incomplete responders'. For the second outcome, patients with a TRG1-2 (indicating predominant fibrosis) were classified as 'good responders' and patients with TRG3-5 as 'poor responders'. For *N*=13 patients who underwent wait-and-see without surgery, a sustained clinical complete response for >2 years follow-up (i.e. no signs of recurrence on follow-up MRI and endoscopy performed 3 monthly in the first year and 6-monthly in the following years) was used as a surrogate endpoint for a complete response. These patients were included in the 'complete response' and 'good response' groups for the two respective outcomes.

Table 1 MR acquisition protocols

	Diffusion-weighted MRI					T2 weighted MRI	
	Center 1			Center 2		Center 1	Center 2
	DWI 1	DWI 2	DWI 3	DWI 1	DWI 2	T2W	T2W
N Slices	50	24	20-24	34	34	22-34	48-60
Repetition Time	3969-5503	3731-5545	4141-5240	5100	4300-5314	3378-9557	3400-4670
Echo Time	70	70-73	65-70	88	79	130-150	118-122
Flip Angle	90	70	70	90	90	90	150
Phase encoding direction	AP	AP	AP	AP	AP	LR	LR
In-plane spacing (mm x mm)	1.7 x 1.7	1.25 x 1.25	1.25 x 1.25	1.25 x 1.25	2.0 x 2.0	0.8 x 0.8 – 0.4 x 0.4	0.8 x 0.8
Slice Thickness	5	5	5	5	6	3-5	3.5
Echo train length	1	1	1	1	1	25-26	23
NSA	4-10	3-5	5	6	6	2-6	2
Fat Saturation	STIR	SPIR	SPAIR	SPIR	SPIR	N/A	N/A
EPI Factor	47-55	55-77	61-83	148	150	N/A	N/A
b-values	0, 500, 1000	0, 500, 1000	0, (25, 50, 100), 500, 1000	0,500,1000	0,300,1100	N/A	N/A

AP: Anterior-Posterior, LR: Left-Right, NSA: Number of Signals Averaged, EPI: Echo Planar Imaging, STIR: Short TI Inversion Recovery, SPIR: Spectral Presaturation Inversion Recovery, SPAIR: Spectral Attenuated Inversion Recovery

Table 2 Likelihood score used by the two radiologists to predict the chance of achieving a good or complete response, respectively, based on visual evaluation of the baseline MRI

	Likelihood of achieving outcome	1: Highly unlikely	2: Unlikely	3: Equivocal	4: Likely	5: Highly likely
Criteria	Size	Large (> 5 cm)			Small (< 3 cm)	
	Signal	heterogeneous			homogenous	
	T-stage	≥ T3cd			≤ T3ab	
	Shape	irregular			regular	
	N-stage	N+			N0	
	EMVI	EMVI+			EMVI -	
	CRM	CRM +			CRM -	
Outcome	Good response	all 7 criteria	≥ 5 criteria	Not meeting the criteria for scores 1-2 or 4-5	≥ 3 criteria	≥ 5 criteria
	Complete response	≥ 5 criteria	≥ 3 criteria		≥ 5 criteria	all 7 criteria

Visual morphologic assessment by expert radiologists

Two independent board-certified abdominal radiologists (DMJL and MJL), with each > 10 years' specific experience in reading rectal MRI, estimated the likelihood of whether a patient would achieve a complete response (outcome 1) or good response (outcome 2), respectively, using a 5-point subjective confidence score (1 = chance to achieve a complete/good response highly unlikely, 2 = good/complete response unlikely, 3 = equivocal, 4 = complete/good response likely, 5 = complete/good response highly likely). The readers based their score on their overall visual morphologic assessment of the size, signal and shape of the tumor, T- and N-stage, circumferential resection margin (CRM) and EMVI, according to the criteria described in Table 2. Tumors with more unfavorable characteristics (e.g., larger size, higher T-stage, positive N-stage, CRM+, EMVI+) were assigned lower scores. Readers were blinded for the patient's outcome and each other's results.

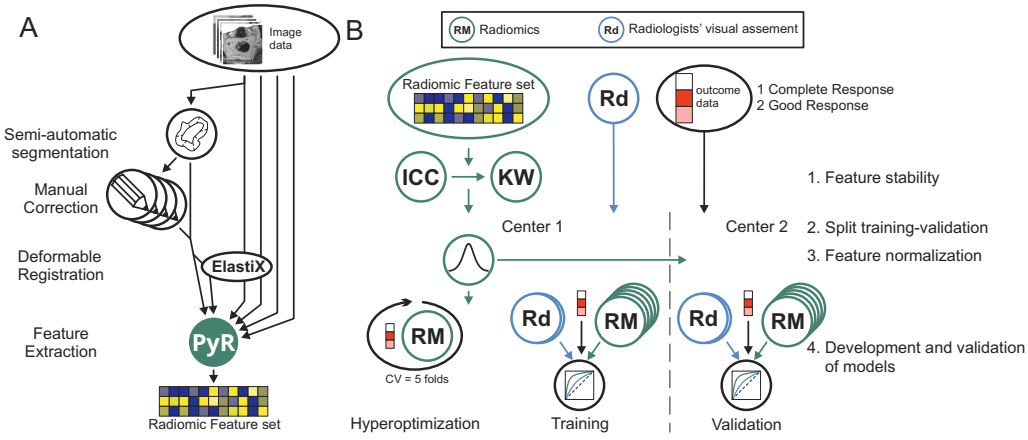


Figure 1. Study Workflow. *Study workflow describing the image segmentation, registration and radiomic feature extraction steps (A) and data analysis steps (B). (A) Tumors were segmented on DWI-b1000. After co-registration of T2W and DWI-b0 images, segmentations were transformed for extraction from T2W images using the transformation map from the registration. Images and segmentation maps were then fed into the PyRadiomics pipeline (PyR) for feature extraction. (B) After exclusion of unstable features (1), data were divided in to training and validation sets by center, with center 1 used for training and 2 for validation (2). Feature values were normalized using mean and standard deviation of features in center 1 (3). Using the training set and 5-fold stratified cross-validation, optimal hyperparameters were determined for the radiomics model. The optimized model was then trained on the full training set for each reader and each outcome separately. Finally, performance to predict response was assess in the training and validation sets (4). ICC: Intraclass Correlation Coefficient, KW: Kruskal Wallis, CV: Cross-Validation*

Radiomics workflow

The Radiomics feature extraction workflow, including image segmentation and radiomic feature extraction as the two main steps, is schematically illustrated in Figure 1A.

I – Image segmentation
The image segmentation comprises the first 3 steps of the workflow:

Step 1: Semi-automatic segmentation
Tumor volumes were semi-automatically segmented on the high b-value diffusion-images using a region-growing algorithm implemented in MANGO (Multi-image Analysis GUI, version 3.8, Research Imaging Institute, University of Texas Health Science Center, San Antonio, TX), according to methods previously reported[22]. The high b-value images were chosen as they provide a good tumor-to-background signal ratio.

Step 2: Manual adjustment
The tumor segmentations derived in step 1 were then checked and manually adjusted where deemed necessary by four independent readers (two resident level non-expert readers (JJMVG and ST) and two expert radiologists (DMJL and MJL)) to allow assessment of effects of interobserver variations and reader experience level (see Figure 2). This resulted in a total of 5 segmentations (1 semi-automated, 2 non-expert, 2 expert) used for radiomic feature extraction. Overlap between segmentations was assessed using the dice similarity coefficient.

Step 3: Registration of different imaging sequences
To correct for organ displacements and deformations, T2W and DWI images were co-registered using deformable B-spline registration implemented in Elastix[23,24]. The resulting deformation maps were then used to adapt the DWI-based segmentations to the T2W images.

II - Radiomic feature extraction
Feature extraction was performed using the *PyRadiomics* toolbox (version 2.1.2)[25]. Prior to feature extraction, images were normalized to 0 mean and 100 standard deviation to reduce influence of differences in MR system vendor and acquisition protocol between the two centers[26] and subsequently interpolated to isotropic voxels with 2mm sides using a B Spline interpolator. To remove outlier intensity values, the five segmentations were re-segmented by excluding voxels which differed > 3σ from the mean. Prior to extraction of texture features and first order Uniformity and Entropy, gray values were discretized using a fixed bin width of 5. For each sequence (T2W, DWI_{b0}, $DWI_{b1000/11000}$, ADC), 623 intensity and texture features were extracted from non-derived, gradient, exponent, logarithm and Laplacian of Gaussian ($\sigma \in$ {1mm, 3mm, 5mm}) filtered images (yielding 4x623=2492 features). In addition, 13 shape descriptors were extracted from non-resegmented DWI-based segmentations, resulting in a grand total of 2505 (2492 + 13) features for each of the five respective segmentations (1 semi-automated, 2 non-expert readers and 2 expert readers). The *PyRadiomics* configuration file used is provided in the supplementary materials.

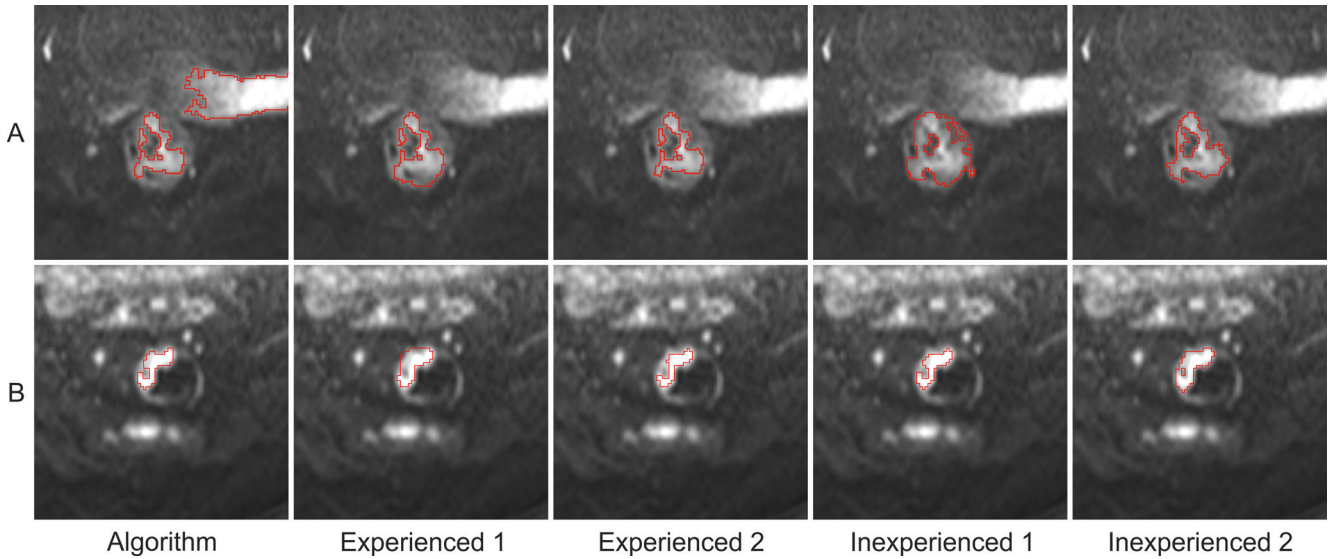


Figure 2. Examples. *Example of results of semi-automatic (algorithm) tumor segmentation performed on high b-value diffusion-weighted images and results after manual correction by two expert and two non-expert readers. (A) Example of a case where the semi-automatic algorithm erroneously included a high signal band-shaped artifact in the tumor segmentation, which was discarded after manual adjustment by both non-expert and expert readers. (B) Example of a case where the semi-automated segmentation by the algorithm performed well and did not require significant changes by either non-expert or expert readers.*

Statistical Analysis

Statistical analysis was performed using the Python (v3.5.3) package Scikit-learn (v0.20)[27] and is schematically illustrated in Figure 1B. Data from center 1 (n=86) were used for training, data from center 2 (n=47) were used for validation. Baseline characteristics were analyzed using χ^2-test for categorical variables and independent samples t-test for continuous variables. Stability of radiomic features for inter-reader variation was assessed using intra-class correlation coefficient (ICC) and stability for differences in MR system vendor and acquisition protocol between the 2 centers was assessed using the Kruskal-Wallis (KW) test. Only features exhibiting sufficient stability (ICC ≥ 0.75 and KW p-value ≥ 0.05) were eligible for selection in the Radiomics prediction model. Stable features were normalized by subtracting the mean and dividing by the standard deviation on a per-reader basis. Mean and standard deviation for each feature is determined using only data from center 1 (training set).

Using the training set, the Radiomics model was trained separately for each of the 5 segmentations in 2 steps: 1) Using minimum Redundancy Maximum Relevance (mRMR), implemented in Python package 'mifs'[28], a set of candidate features was selected from the training set, which 2) were fitted into a logistic regression model with L2 regularization and balanced class weights. To approximate the mutual information between the outcome and continuous features during mRMR selection, we employed the nearest neighbor method as described by Ross et al.[29] Optimum number of features to select [5–10], as well as the k neighbors parameter [5–8] in mRMR and the C regularization parameter [10^{-7}–10^{2}] in the logistic regression model were determined by 5-fold stratified cross-validation on the training set. Finally, the performance of the Radiomics model to predict a 'complete' and 'good' response, respectively, was assessed using the Wilcoxon rank-sum test and by calculating the area under the ROC-curve (AUC). Using the DeLong method[30], AUC for Radiomics was then compared to the AUC calculated for the morphologic prediction of response by the two expert radiologists based on their subjective confidence scores. p-values < 0.05 were considered significant. Interobserver agreement for the subjective scoring by the two radiologists was assessed using quadratic Cohen's kappa.

Table 3 Baseline characteristics

	Total (*N*=133)	Centre 1 (*N*=86)	Centre 2 (*N*=47)	*P*-value
Age	68 [45-87]	69 [48-87]	67 [45-85]	0.1583*
Gender				0.691†
M	92 (69.2%)	61 (70.9%)	31 (66.0%)	
F	41 (30.8%)	25 (29.1%)	16 (34.0%)	
Initial cT stage before treatment				0.931†
1-2	16 (12.0%)	11 (12.8%)	5 (10.6%)	
3	109 (82.0%)	70 (81.4%)	39 (83.0%)	
4	8 (6.0%)	5 (5.8%)	3 (6.4%)	
Initial cN stage before treatment				0.316†
0	11 (8.3%)	8 (9.3%)	3 (6.4%)	
1	40 (30.1%)	22 (25.6%)	18 (38.3%)	
2	81 (60.7%)	56 (65.1%)	26 (55.3%)	
Final treatment after CRT				0.065†
TME	119 (89.5%)	73 (84.9%)	46 (97.9%)	
W&W	13 (9.8%)	12 (14.0%)	1 (2.1%)	
TEM	1 (0.7%)	1 (1.1%)	0 (0%)	
Final yT stage				0.917†
0	28 (21.0%)	18 (20.9%)	10 (21.3%)	
1	11 (8.3%)	8 (9.3%)	3 (6.4%)	
2	30 (22.6%)	19 (22.1%)	11 (23.4%)	
3	59 (44.4%)	37 (43.0%)	22 (46.8%)	
4	5 (3.8%)	4 (4.7%)	1 (2.1%)	
Final yN stage				0.233†
0	94 (70.7%)	65 (75.6%)	29 (61.7%)	
1	29 (21.8%)	16 (18.6%)	13 (27.7%)	
2	10 (7.5%)	5 (5.8%)	5 (10.6%)	
Tumor regression grade (TRG)				0.108†
1	28 (21.0%)	18 (20.9%)	10 (21.3%)	
2	34 (25.6%)	24 (27.9%)	10 (21.3%)	
3	38 (28.6%)	26 (30.2%)	12 (25.5%)	
4	25 (18.8%)	11 (12.8%)	14 (29.8%)	
5	4 (3.0%)	4 (4.7%)	0 (0%)	
Missing	4 (3.0%)	3 (3.5%)	1 (2.1%)	

*TME Total Mesorectal Excision, W&W Watch & Wait (Organ saving treatment), TEM Transanal Endoscopic Microsurgery, cT, cN: clinical T and N stage as assessed on primary MRI, yT, yN: final T and N stage after nCRT as assessed at histopathology after surgery (n=120) or by long-term follow-up in case of wait-and-see treatment (n=13), TRG: Tumor Regression Grade, * t-test, † χ^2 test*

Results

Baseline characteristics

The baseline characteristics of the patients are shown in Table 3. No significant differences were seen between the two centers. In total 28 patients were complete responders (15 after surgery and 13 sustained clinical complete responders undergoing W&S) and 105 patients had residual tumor. For the second clinical outcome, good versus poor response, 62 patients were considered good responders (28 TRG1, 34 TRG2) and 67 poor responders (38 TRG3, 25 TRG4, 4 TRG5). In 4 patients (3 from center 1, 1 from center 2) no TRG stage was available; these patients were therefore excluded from the latter analysis.

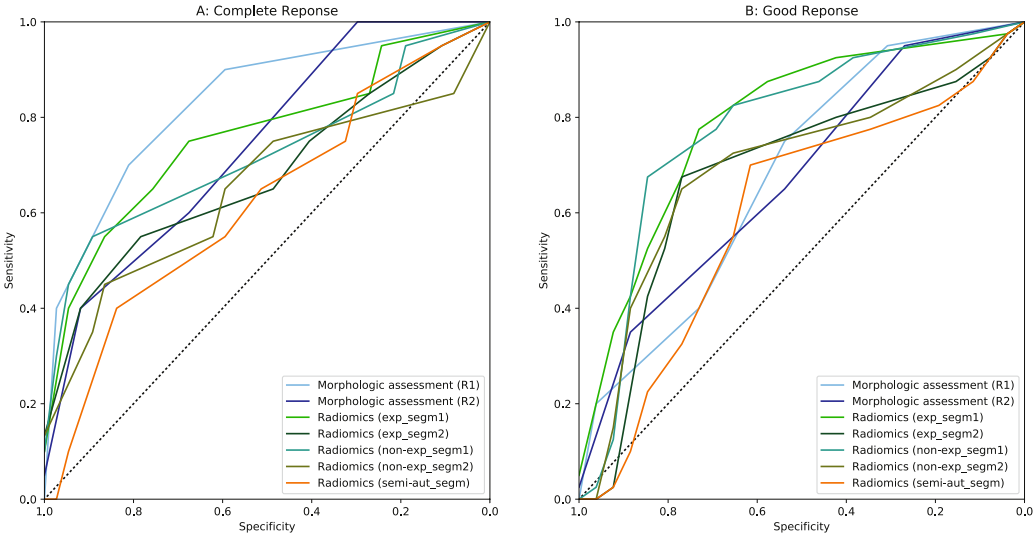


Figure 3. *ROC-curves of morphologic assessment by radiologists and radiomics models to predict the outcome 'complete' (A) and 'good' (B) response. There were no statistically significant differences in diagnostic performance between the 2 radiologist readers and the various radiomics models.*

Performance of radiologists' visual morphologic assessment to predict response

Results for the prediction of response by the two radiologists (compared to the performance of the radiomics models) are provided in Table 4 and illustrated in Figure 3. Overall, AUC to predict a complete response in the validation cohort was 0.83 for the first reader and 0.74 for the second reader. For the prediction of a good response, AUC was 0.68 (reader 1) and 0.67 (reader 2) in the validation cohort. Agreement between the two readers was good with κ=0.64 and κ=0.61 for the prediction of complete and good response, respectively.

Building the radiomics models

1692 out of the in total 2505 radiomic features (68%) showed an ICC ≥ 0.75 (indicating sufficient inter-reader stability), of which only 415 (25%) showed no confounding related to the MR system and acquisition protocol used (i.e. MRI performed in center 1 or center 2). These 415 features were considered stable and available for selection by the radiomics model. Optimum settings for the model, as determined by the hyper-optimization in the training set, turned out to be 9 features, k=8 neighbors and $C=10^{-5}$. Most emphasis was placed on DWI and ADC sequences, with only few features selected from T2W sequences in 8/10 developed models. Further details regarding the selected features per model are provided in Supplementary materials 2.

Table 4 Performance to predict response

	Complete Response					
	Center 1 (Training, n=86)			**Center 2 (Validation, n=47)**		
Reader	AUC (95%CI)	statistic	p-value	AUC (95%CI)	statistic	p-value
Morphologic assessment (R1)	0.77 [0.62-0.91]	3.503	<0.001	0.83 [0.69-0.98]	3.197	0.001
Morphologic assessment (R2)	0.67 [0.51-0.84]	2.272	0.023	0.74 [0.58-0.90]	2.326	0.020
Radiomics (exp_segm1)	0.71 [0.57-0.86]	2.771	0.006	0.77 [0.58-0.96]	2.573	0.010
Radiomics (exp_segm2)	0.74 [0.60-0.88]	3.089	0.002	0.69 [0.47-0.91]	1.820	0.069
Radiomics (non-exp_segm1)	0.71 [0.55-0.86]	2.707	0.007	0.73 [0.51-0.94]	2.183	0.029
Radiomics (non-exp_segm2)	0.74 [0.59-0.88]	3.100	0.002	0.66 [0.42-0.89]	1.508	0.132
Radiomics (semi-aut_segm)	0.73 [0.60-0.86]	3.025	0.002	0.63 [0.42-0.84]	1.248	0.212
	Good Response					
	Center 1 (Training, n=86)			**Center 2 (Validation, n=47)**		
Reader	AUC (95%CI)	statistic	p-value	AUC (95%CI)	statistic	p-value
Morphologic assessment (R1)	0.60 [0.49-0.72]	1.608	0.108	0.68 [0.53-0.83]	2.072	0.038
Morphologic assessment (R2)	0.68 [0.56-0.79]	2.805	0.005	0.67 [0.52-0.83]	2.005	0.045
Radiomics (exp_segm1)	0.77 [0.67-0.87]	4.235	<0.001	0.79 [0.66-0.93]	3.368	0.001
Radiomics (exp_segm2)	0.69 [0.57-0.80]	2.933	0.003	0.69 [0.52-0.86]	2.194	0.028
Radiomics (non-exp_segm1)	0.72 [0.61-0.83]	3.488	<0.001	0.78 [0.64-0.92]	3.235	0.001
Radiomics (non-exp_segm2)	0.71 [0.60-0.82]	3.361	0.001	0.70 [0.53-0.86]	2.260	0.024
Radiomics (semi-aut_segm)	0.65 [0.53-0.77]	2.377	0.017	0.60 [0.43-0.78]	1.197	0.231

Performance of the radiomics models to predict response

In the training set, radiomics models based upon manually-corrected segmentations showed significant performance to predict both 'complete response' (AUC 0.71 to 0.74) and 'good response' (AUC 0.69 to 0.77). In the validation dataset, AUCS ranged between 0.69 and 0.79 (p=0.001-0.028) to predict a good response, with comparable performance for the segmentations performed by the two non-expert and expert readers. For the prediction of complete response, only the radiomics model using the segmentations from 1 expert and 1 non-expert reader retained significant performance, with respective AUCs of 0.77 (p-value 0.010) and 0.73 (p-value 0.029). Performance of the radiomics model using the semi-automated segmentations (without manual reader input) was non-significant for both study outcomes. Average dice coefficients between the semi-automated and different non-expert and expert manual-input segmentations are shown in Table 5.

Comparison between radiomics model and morphologic assessment by radiologists

AUCs for the radiomics models (using segmentations with manual input) were comparable to the AUCs for the visual morphologic assessment by the expert radiologists, both for the prediction of a complete response (0.73-0.77 versus AUC 0.74-0.83, P=0.25-0.88), as well as for the prediction of a good response (AUC 0.69-0.79 versus AUC 0.67-0.68 , P=0.18-0.93).

Table 5 Average (±SD) Dice similarity between reader's segmentations

	exp_segm1	exp_segm2	non-exp_segm1	non-exp_segm2	semi-aut_segm
exp_segm1					
exp_segm2	0.84 (0.15)				
non-exp_segm1	0.73 (0.17)	0.71 (0.19)			
non-exp_segm2	0.77 (0.16)	0.84 (0.16)	0.69 (0.21)		
semi-aut_segm	0.78 (0.22)	0.76 (0.21)	0.64 (0.24)	0.78 (0.21)	

Discussion

In the present study, we compared the performance of advanced radiomics analysis to visual morphologic assessment by experienced radiologists to predict response to neoadjuvant chemoradiotherapy on primary staging MRI. Results show that the radiomics model could predict a good response to therapy upfront with similar diagnostic performance (AUC 0.69-0.79) as highly expert radiologists (AUC 0.67-0.68). Interestingly, the Radiomics models were mainly based on features derived from DWI and ADC, with only few features selected from T2W imaging. This would suggest that DWI plays an important role when building response prediction models based on Radiomics. Moreover, results of the radiomics model were comparable regardless of whether image segmentation was manually adapted by non-expert (young resident level) readers or by experienced radiologists, indicating that expert input is not required in order for the radiomics workflow to produce significant predictive performance. Radiomics models without manual input (using only semi-automated tumor segmentations) did not result in significant predictive performance, despite the fact that the spatial overlap between the semi-automated and manual-input segmentations was quite substantial (Dice 0.64-0.78).

Although in recent years several groups have investigated the potential of Radiomics for rectal tumor response assessment, our current report is one of few to compare Radiomics results to visual radiological assessment in order to put things into a more clinical perspective. To the best of our knowledge, only one previous report by Horvath et al.[31] compared performance of Radiomics to expert reader assessment, though this study focused on response assessment after completion of therapy, rather than prediction upfront. In this study 34 features were extracted from 114 patients, and combined using a random forest classifier. This model showed excellent performance (AUC 0.93) in repeated cross-validation, which was significantly better than consensus scoring by 2 radiologists. A handful of previous studies specifically focused on MR-based Radiomics to predict rectal tumor response prior to the start of treatment using baseline imaging data. Nie K. et al[32] extracted 103 features from primary T1/T2, DWI and dynamic contrast enhanced MRI in 48 patients. Here, an artificial neural network was trained using 4-fold cross validation to address overfitting, with resulting AUCs of 0.84 and 0.89 to predict a complete and good response, respectively. Cusumano D. et al[20] performed a similar study but included an independent validation data cohort from another center. Here, a combination of shape, fractal and LoG-based features were assessed in a cohort of 198 patients, resulting in AUC 0.77 and 0.79 in the training and validation dataset, respectively. Finally, Cui Y. et al[19] assessed performance of radiomic features in 186 patients, achieving a very high AUC of 0.98 in the validation set. However, feature stability for image acquisition variation was not assessed, which may limit clinical applicability. In our current bi-institutional study we used a test and validation dataset from two independent centers in order to investigate potential confounding effects of variations in MR system vendor and acquisition protocols. We found that a large portion (75%) of features were classified as unstable to variations in vendor and image acquisition protocols. This highlights the need for standardized protocols and the importance of assessing feature stability when developing radiomics models. On the other hand, despite these vendor and protocol variations, the radiomics models still achieved performance comparable to expert-reader assessment.

Although, the AUCs of 0.66–0.79 achieved in our current study to predict response are encouraging, they will probably not yet be considered good enough for clinical decision making. As discussed above there is however still room for improvement. Further research should focus on standardization, but also on combining radiomic features with for example other clinical, histopathological, immunohistochemical or genetic biomarkers, which is likely to increase the predictive power, as has also been suggested by previous research[33,34]. An accurate prediction of treatment response upfront, using biomarkers that can already be derived at baseline could impact clinical management in rectal cancer in the future. After completion of CRT, complete responders may already be accurately detected using a combination of simple visual DWI-MRI analysis and endoscopy, which limits the need for advanced imaging analysis tools such as Radiomics in this setting[35]. Tools to predict treatment effects upfront are however not yet available in clinical practice. In locally advanced tumors, where downsizing is desired, but standard CRT is predicted to have little or no effect, one could consider a more intensified regimen or one that relies more on systemic therapy. If lateral resection margins are wide on MRI, one can even consider omitting neoadjuvant therapy altogether, avoiding unnecessary toxicity. In smaller rectal tumors, which are traditionally treated with TME surgery without neoadjuvant therapy, but which also have a higher chance to respond well to radiotherapy, a predictive model can guide treatment decisions towards (chemo)radiotherapy for the predicted responders with the goal to achieve organ preservation.

Our study design contained some limitations. The 95% confidence intervals for the performance of the radiological assessments as well as the radiomics models were large, most likely due to the relatively small size of the dataset, especially the validation set. Moreover, this small size of the training set can make the radiomics models prone to overfitting, as reflected by the fact that the optimum hyperparameters, with a low value for the C parameter and high value for the k parameter, favoring high regularization. Our result will therefore need to be further validated in larger and preferably multicenter cohorts to obtain more stable results. Additionally, the estimated likelihood of achieving a good/complete response by the radiologists remains relatively subjective (despite the criteria provided in Table 2) and is dependent on the experience level of the two readers. We chose this approach to provide some preliminary perspective on how advanced model-based prediction methods would compare to what can potentially be achieved by mere "human" interpretation. We, however, acknowledge that an alternative approach including separate assessment of individual semantic features may allow for better reproducibility. This is a strategy we aim to further explore in future research. Finally, the analyses were all performed in patients with locally advanced rectal tumors. Our results will also need to be tested and validated in smaller tumors before radiomics models can be applied in these cases.

In conclusion, we were able to train radiomics models to reach comparable performance to predict response to chemoradiotherapy on baseline MRI as visual morphologic assessment and staging by highly expert radiologists, even when using tumor segmentations without any expert radiologist input. Furthermore, these results were obtained despite training on a very heterogeneous dataset, where the majority of features had to be excluded due to susceptibility for variations in image acquisition. Although validation in a large multicenter cohort is obviously needed, these results indicate that Radiomics has strong potential to identify meaningful imaging biomarkers that can be included in clinically usable prediction models with the ultimate aim to further optimize and personalize treatment in rectal cancer.

1. Maas M, Beets-Tan RG, Lambregts DM, et al. Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. *J Clin Oncol.* 2011; 29(35): 4633-4640.

2. Martens MH, Maas M, Heijnen LA, et al. Long-term Outcome of an Organ Preservation Program After Neoadjuvant Treatment for Rectal Cancer. *J Natl Cancer Inst.* 2016; 108(12): djw171.

3. van der Valk MJM, Hilling DE, Bastiaannet E, et al. Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the International Watch & Wait Database (IWWD): an international multicentre registry study. *Lancet.* 2018; 391(10139): 2537-2545.

4. Verseveld M, De Graaf EJR, Verhoef C, et al. Chemoradiation therapy for rectal cancer in the distal rectum followed by organ-sparing transanal endoscopic microsurgery (CARTS study). *Br J Surg.* 2015; 102(7): 853-860.

5. Bujko K, Richter P, Smith FM, et al. Preoperative radiotherapy and local excision of rectal cancer with immediate radical re-operation for poor responders: a prospective multicentre study. *Radiother Oncol.* 2013; 106(2): 198-205.

6. Rombouts AJM, Al-Najami I, Abbott NL, et al. Can we Save the rectum by watchful waiting or TransAnal microsurgery following (chemo) Radiotherapy versus Total mesorectal excision for early REctal Cancer (STAR-TREC study)?: protocol for a multicentre, randomised feasibility study. *BMJ Open.* 2017; 7(12): e019474.

7. Maas M, Nelemans PJ, Valentini V, et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol.* 2010; 11(9): 835-844.

8. Curvo-Semedo L, Lambregts DMJ, Maas M, et al. Rectal Cancer: Assessment of Complete Response to Preoperative Combined Radiation Therapy with Chemotherapy—Conventional MR Volumetry versus Diffusion-weighted MR Imaging. *Radiology.* 2011; 260(3): 734-743.

9. Lambregts DM, Rao SX, Sassen S, et al. MRI and Diffusion-weighted MRI Volumetry for Identification of Complete Tumor Responders After Preoperative Chemoradiotherapy in Patients With Rectal Cancer: A Bi-institutional Validation Study. *Ann Surg.* 2015; 262(6): 1034-1039.

10. Mahadevan LS, Zhong JJ, Venkatesulu BP, et al. Imaging predictors of treatment outcomes in rectal cancer: An overview. *Crit Rev Oncol Hematol.* 2018; 129: 153-162.

11. Hötker AM, Tarlinton L, Mazaheri Y, et al. Multiparametric MRI in the assessment of response of rectal cancer to neoadjuvant chemoradiotherapy: A comparison of morphological, volumetric and functional MRI parameters. *Eur Radiol.* 2016; 26(12): 4303-4312.

12. Martens MH, Subhani S, Heijnen LA, et al. Can perfusion MRI predict response to preoperative treatment in rectal cancer? *Radiother Oncol.* 2015; 114(2): 218-223.

13. Chen Y-G, Chen M-Q, Guo Y-Y, Li S-C, Wu J-X, Xu B-H. Apparent Diffusion Coefficient Predicts Pathology Complete Response of Rectal Cancer Treated with Neoadjuvant Chemoradiotherapy. *PLoS One.* 2016; 11(4): e0153944.

14. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018; 18(8): 500-510.

15. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging.* 2012; 30(9): 1234-1248.

16. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology.* 2015; 278(2): 563-577.

17. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5(1): 4006.

18. Liu Z, Zhang X-Y, Shi Y-J, et al. Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. *Clin Cancer Res.* 2017; 23(23): 7253-7262.

19. Cui Y, Yang X, Shi Z, et al. Radiomics analysis of multiparametric MRI for prediction of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Eur Radiol.* 2019; 29(3): 1211-1220.

20. Cusumano D, Dinapoli N, Luca Boldrini ·, et al. Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. *Radiol Med.* 2018; 123: 286-295.

21. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer.* 1994; 73(11): 2680-2686.

22. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, et al. Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol.* 2016; 94(4): 824-831.

23. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: A toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging.* 2010; 29(1): 196-205.

24. Shamonin D. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinform.* 2013; 7.

25. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017; 77(21): e104-e107.

26. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging.* 2004; 22(1): 81-91.

27. Fabian P, Michel V, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011; 12: 2825-2830.

28. Hanchuan Peng, Fuhui Long, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell.* 2005; 27(8): 1226-1238.

29. Ross BC. Mutual Information between Discrete and Continuous Data Sets. Marinazzo D, ed. *PLoS One.* 2014; 9(2): e87357.

30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988; 44(3): 837-845.

31. Horvat N, Veeraraghavan H, Khan M, et al. MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. *Radiology.* 2018; 287(3): 833-843.

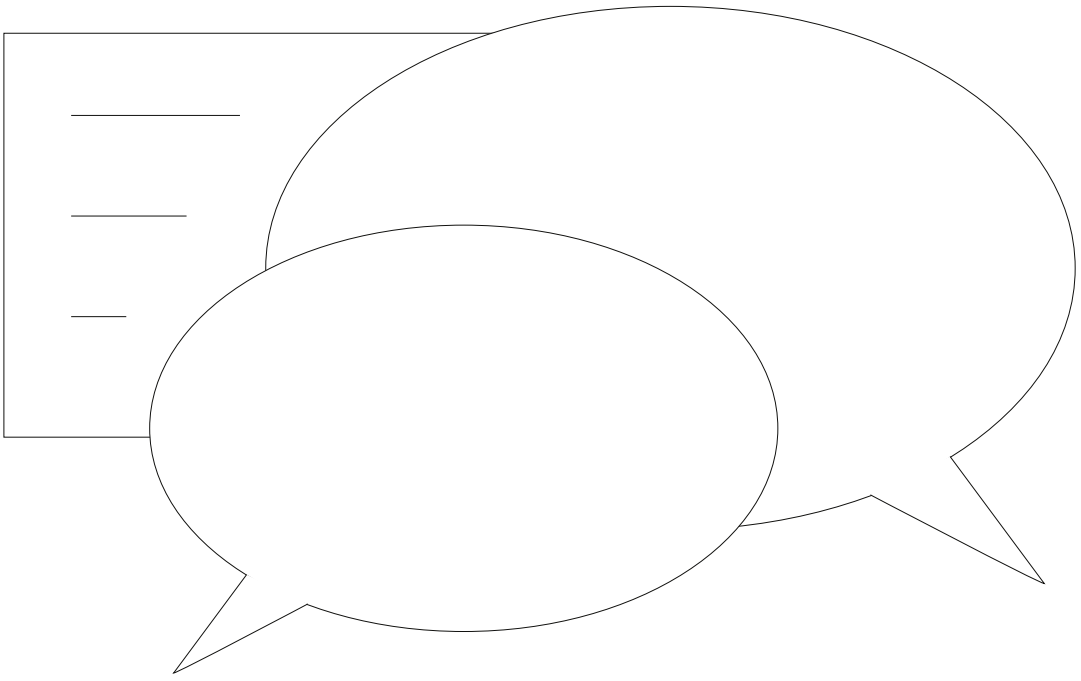
32. Nie K, Shi L, Chen Q, et al. Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. *Clin Cancer Res.* 2016; 22(21): 5256-5264.

33. Huang Y, Liang C, He L, et al. Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. *J Clin Oncol.* 2016;34(18):2157-2164.

34. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol.* 2015; 114(3): 345-350.

35. Maas M, Lambregts DM, Nelemans PJ, et al. Assessment of Clinical Complete Response After Chemoradiation for Rectal Cancer with Digital Rectal Examination, Endoscopy, and MRI: Selection for Organ-Saving Treatment. *Ann Surg Oncol.* 2015; 22(12): 3873-3880.

General Discussion



Using high-throughput radiomics, quantitative information can be extracted from medical imaging that can be used to generate a comprehensive oncologic imaging fingerprint, thereby rendering many potential novel imaging biomarkers of disease. These biomarkers may aid in addressing clinical challenges in oncology through the development of predictive models, which may ultimately be combined with other clinical, histopathological, immunohistochemical and genetic markers to build strong clinical decision-support tools with impact on oncologic treatment management and outcome. Before achieving this ultimate "horizon" goal, there are several more "proximate" goals that will need to be accomplished first. This thesis focuses on the latter and addresses specific challenges in each step of the radiomics workflow that need to be overcome to facilitate high-quality research and ultimately pave the way for application of radiomics in clinical practice.

Addressing the challenge of image segmentation with AI

Before a computer is able to extract radiomics features from medical images, it needs to know what is the region of interest (ROI) within the image. This ROI – which in oncological imaging often entails the tumor lesion under investigation – is typically defined by means of "segmentation", in other words by tracing the boundaries of the tumor lesion within the image. Several previous studies have shown that precise delineation of the whole tumor volume by means of manual segmentation provides the most reproducible results as compared to "simpler" methods such as selecting a single representative slice or sample measurement within the tumor[1–3]. The main drawback of manual whole-volume segmentation is that it can be very labor intensive and time consuming (especially in large-size lesions) and that significant inter-reader variation may occur, especially when segmentations are performed by readers who are less experienced in radiological image interpretation[4]. This dependence on (expert) segmentation input severely hampers the feasibility of implementing radiomics algorithms in daily practice. In 2015 our group took the first steps to addressing this problem and attempted a fairly simple segmentation algorithm to support radiologists in semi-automatically segmenting rectal tumors on MRI. Though this already led to a significant reduction in the time required for segmentation, results were far from optimal and significant correction was required in multiple cases[4]. In this thesis, we expanded our research on tumor segmentation and set out to develop and train a deep learning network to provide fully-automatic segmentation of rectal tumors on MRI aiming to minimize the input required from radiologists to an absolute minimum.
In our pilot study in Chapter 2 we used a simple patch-based convolutional network (CNN) and applied it to a multiparametric image dataset consisting of T2-weighted, high and low b-value diffusion-weighted images, to make optimum use of the high tumor-to-background ratio found in DWI combined with the better visualization of anatomical detail on T2W-MRI. This network predicted pixel-by-pixel which area in the image represented the rectal tumor, based on 2D patches extracted around each pixel. When comparing network-inferred segmentations to manual segmentations generated by an expert-radiologist using dice similarity scores (DSC) – a measure to indicate the spatial overlap between two segmentations on a scale from 0 to 1– results were promising with an overall DSC 0.70. Main limitations of the algorithm from Chapter 2 were its relatively slow pixel-by-pixel inference (meaning that each voxel was independently classified, based on the extracted image patch surrounding the voxel) and the fact that the small patch size resulted in limited spatial context with incorrect classification of structures

near the edges of the images. Also, the algorithm was trained and tested using data from only two centers, while ideally an AI model should be subjected to heterogeneous data from multiple institutions to generate a model that will be broadly applicable to any clinical dataset. In Chapter 3, we addressed these limitations and further optimized our segmentation algorithm by using an attention-gated U-Net, which only requires one forward pass for inference, and has greater spatial context, especially due to the addition of attention-gating. We trained and tested this model in a heterogeneous dataset with rectal MRIs from 6 different centers. In addition, we investigated the influence of DWI scan quality and tumor complexity on the networks' performance. We learned that reduced DWI scan quality negatively impacted the networks' performance, again stressing the importance of optimizing acquisition protocols and taking steps to avoid artefacts as addressed in Chapter 4 and discussed below.
Network performance for segmentation was also reduced in highly complex (i.e. irregularly shaped and heterogeneous signal) tumors, similar to how inter-reader agreement was also lower in these complex cases. All in all, we found that for the majority of more "straightforward" tumors, the network could achieve a performance that was almost similar to the agreement between two expert radiologists, provided that image quality was good. In these cases, AI could really offer a solution to reduce the workload of image segmentation in the radiomics workflow. In the future, the necessity of accurate image segmentation as a requirement for analysis may change as deep learning techniques offer the potential to train models to directly predict the outcome (e.g. response to treatment) from the acquired images, without the need for carefully handcrafted features or image segmentation. In rectal cancer, a few preliminary studies have shown encouraging results for such an approach[5,6], though future studies are needed to compare these new methods to the 'classic' radiomics approach using a predefined ROI and handcrafted features.

Addressing the challenge of image quality: DWI susceptibility artefacts

Diffusion-weighted imaging (DWI) is one of the most commonly used functional MR imaging sequences in oncology. In highly cellular malignant tumors, the movement or "diffusion" of water molecules is restricted and a high signal is retained, whereas in low-cellular tissues, the signal exponentially decreases. This makes high-cellular tumors stand out compared to their background, making DWI a prime candidate for the detection of tumor and the development of (semi-)automated tumor segmentation algorithms[4]. For rectal cancer many centers now include DWI in their MRI protocol, especially in the restaging setting where current clinical guidelines advise the use of DWI because of its superiority to standard MRI in distinguishing residual tumor from post-radiation fibrosis[10–12]. The diffusion characteristics of tissues can also be quantified, most commonly expressed as the apparent diffusion coefficient or "ADC". Several studies in rectal cancer as well as other tumor types have shown that ADC as a measure of tumor cellularity has potential as a biomarker for tumor aggressiveness[7], response[8], and survival[9].

An important potential drawback of DWI is that it can be quite challenging to acquire DW images with consistently good quality, that is reliable enough for clinical interpretation as well as for further image analysis. DWI sequences are generally acquired using echo planar imaging

(EPI) sequences, which allow fast image acquisition but are relatively prone to susceptibility artifacts[15]. Susceptibility artefacts cause a local signal change or distortion due to local magnetic field inhomogeneities. In bowel imaging, these inhomogeneities are often caused by intraluminal gas and the resulting artefacts may seriously hamper the quality of DWI exams of the rectum. In published reports on bowel DWI, up to 11% of cases were excluded due to insufficient quality of the DWI images[16–19]. In our study in Chapter 4 we found that without additional preparation, 24% of the scans contained clinically relevant susceptibility artefacts. To reduce these artefacts there are basically two options: either to change the acquisition parameters (making the scan more robust to the effects causing the artefacts) or alternatively to reduce or remove the cause of the artefacts. This latter approach was investigated in Chapter 4, where we employed a preparatory micro-enema that can be self-administered by patients shortly prior to MR acquisition to reduce the intestinal gas causing the susceptibility artefacts on DWI. We found that this relatively simple intervention resulted in a marked reduction of both the incidence and severity of artefacts, with less than 4% of scans showing clinically relevant artefacts after preparation with the micro-enema, making it a very effective approach to optimize image quality in rectal DWI.

Towards easy and reproducible radiomics feature extraction: PyRadiomics

The features that are typically extracted in a radiomics study can be subdivided into 3 main groups. The first group is the "simple", first-order features, which describe the general distribution of gray values in the ROI, regardless of their spatial location. These features can generally be derived directly from the statistical gray value histogram. The second group is the shape features that describe the 2D or 3D shape of the ROI, including parameters such as the volume, surface area and maximum diameters. This type of features makes no use of the gray values encountered in the ROI, but rather the spatial locations of the segmented voxels. The final group of features, generally referred to as second or higher order features, are texture features that describe the relationship of the gray values and their spatial relationship to one another, thereby highlighting the heterogeneity of the texture within in the ROI[13]. Finally, Aerts et al.[7] defined a 4th group, containing histogram and texture features which have been extracted from the image after applying a filter highlighting certain aspects of the image, such as wavelet and Laplacian-of-Gaussian filters. In addition, several pre-processing steps may be undertaken to improve the reproducibility of the extracted features, such as clustering pixels according to intensity values ("discretization"), geometric resampling to obtain volumes with identical pixel spacing, or gray value normalization to obtain images with comparable ranges of intensity values.

Though many of the principles (i.e. mathematical algorithms) behind radiomic features have long since been described in multiple publications[20–22], the process to extract these features from medical images – including pre-processing steps to prepare the images for data extraction – has lacked standardization prior to the start of this thesis[23–26]. Many published reports on radiomics used in-house developed software for feature extracting, which severely hampers the reproducibility and comparability of the reported results. In Chapter 5 we introduced *PyRadiomics*, an open-source python package developed specifically to address this challenge and provide a transparent, standardized and reproducible tool for radiomics feature extraction.

It was developed with heavy emphasis on ease-of-use and readability of the source code. Special care was taken to provide extensive documentation on both the underlying principles of the radiomics feature extraction process, as well as the specific implementation provided in *PyRadiomics*. To aid researchers without programming experience in performing radiomics analyses, a graphic user interface was provided in the form of *SlicerRadiomics*, an extension module for the medical image analysis software 3D Slicer. Since its publication in 2017, *PyRadiomics* has become a popular package and reference tool for feature extraction in radiomics research with publications of its use in various cancer types[27–33] and applied to multiple imaging modalities, including MRI[27–30], CT[31,32], PET[30] and ultrasound[33]. Though mainly used for research in oncologic imaging, it has also been applied in non-oncologic[34] and even non-medical imaging[35].

Putting it to the test: application of radiomics in rectal cancer

In Chapter 6 we put radiomics, using the *PyRadiomics* software developed in Chapter 5, to the test in a clinical study aiming to predict therapeutic response in rectal cancer. Predicting response to neoadjuvant treatment is one of the main current clinical challenges in rectal cancer, owing to the recent introduction of organ-preserving treatment strategies such as the "watch-and-wait" policy for patients that show a complete regression of their rectal tumor after neoadjuvant chemoradiotherapy (CRT). Identifying those patients who have obtained a complete response after CRT is essential to select the right patients to be treated with a watch-and-wait. In addition, pre-treatment prediction of response could aid in further optimization and personalization of neoadjuvant treatment schemes to increase the chance of obtaining a complete response in the good responders and spare non-responders the exposure to unnecessary morbidity of "futile" CRT prior to major surgery. In Chapter 6 we developed a radiomics model to predict response upfront, i.e. before onset of treatment, using the input of primary staging MRIs. We compared the performance of this model to an approach where treatment response was predicted by 2 expert radiologists based on their visual assessment of the same images, to place the results in a more clinical context.

Excitingly, we found that the radiomics model performed equally compared to these expert-reader predictions, indicating great perspective for AI in developing support tools that can truly benefit clinical decision making in the future. These results are in line with various other reports that have meanwhile been published and similarly suggest a beneficial role for radiomics and AI in this clinical setting, as well as to predict other clinical outcomes such as risk of lymph-node metastasis[36], distant metastasis[37], and survival[38]. Referring to the importance of image segmentation addressed in Chapters 3 and 4, we again found that a sufficient level of segmentation accuracy was a requirement for the model to result in significant predictive performance. Moreover, we found that a significant portion of the features (~75%) was correlated to the acquisition system and had to be excluded from analysis. This highlights the need for standardization of image acquisition and the importance of feature stability analysis in radiomics research.

Future perspectives and recommendations

Though promising, the results of the radiomics model in Chapter 6 are not yet good enough to influence clinical decisions. Apart from the fact that the model should be further optimized and validated using multicentric datasets, the data derived from radiomics should be combined with other clinical information and data. Current clinical management of cancer is not based on imaging alone, but based on multi-disciplinary evaluations incorporating information derived from clinical examinations, imaging, histopathology, and laboratory testing. Similarly, it makes sense that AI models that can offer sufficient power to aid in clinical decision making should also be based on input from several different fields. Future studies should therefore aim to develop comprehensive predictive models, incorporating information from imaging through radiomics, but also "semantic" features such as clinical TNM stage, as well as state of the art histopathologic and genetic biomarkers derived from biopsies and more novel techniques such as liquid biopsies. There have already been some interesting "radiogenomics" studies demonstrating that radiomics features can be directly correlated to genetic biomarkers[39–41] and may even be used to predict mutational status without the need for invasive biopsies. The results of our clinical study in Chapter 6 were still largely based on manual segmentations, performed by human readers. Future research should expand on the use of automatic segmentation methods, such as the ones developed in Chapters 2 and 3, and investigate the performance of radiomics models where these automatically generated segmentations are incorporated as the main source of segmentation input with no or only minimal need for manual corrections. Such approaches are expected to significantly reduce the time required to complete the radiomics workflow and may significantly contribute to the implementation of radiomics in future trials and ultimately into clinical practice.

In addition, deep learning algorithms such as those used for the automatic segmentation of rectal tumors in Chapters 2 and 3 may also be of use in addressing other challenges. For example, deep learning could be used to aid in automatic assessment of image quality and in building workflows to reduce systematic differences between scans related to acquisition protocol and vendor variation. Moreover, deep learning networks may be trained to predict clinical outcomes (such as the likelihood of response) directly from the MRI sequences, without the need for segmentation or carefully handcrafted features.

Finally, there are those who postulate that AI will soon replace the radiologist in the process of image analysis, raising doubt amongst current medicine students whether it is still wise to date to pursue a future career in radiology. However, in my opinion AI will never replace but rather enhance the radiologist profession. As has previously been demonstrated in breast cancer detection, computer aided detection systems ("CAD" a form of AI already used in clinics) did not reduce the need for radiologist input, but instead greatly enhanced the clinical information a radiologist can provide. Therefore, AI will not supplant the radiologist, but the radiologist who embraces AI will supplant radiologists who fear and avoid it.

Chapter 7

Conclusions

In this thesis we have addressed the main challenges that arise in the radiomics workflow within the context of rectal cancer imaging as a clinical case example. We investigated the challenge of creating a dataset suitable for radiomics research, and found that application of a micro-enema can significantly reduce the number and severity of gas-induced susceptibility artefacts on diffusion-weighted MR imaging. We have shown the potential of deep learning to aid in the labor-intensive task of tumor segmentation. Finally we have developed an open-source toolbox for easy and reproducible feature extraction that is now used by research groups worldwide thereby facilitating transparent radiomics research. Using this toolbox ourselves, we have shown the potential of radiomics to build imaging-based prediction models to predict response to neoadjuvant chemoradiotherapy in rectal cancer.

1. Lambregts DM, Beets GL, Maas M, et al. Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability. *Eur Radiol.* 2011; 21(12): 2567-2574.

2. Nougaret S, Vargas HA, Lakhman Y, et al. Intravoxel Incoherent Motion-derived Histogram Metrics for Assessment of Response after Combined Chemotherapy and Radiation Therapy in Rectal Cancer: Initial Experience and Comparison between Single-Section and Volumetric Analyses. *Radiology.* 2016; 280(2): 446-454.

3. Prezzi D, Owczarczyk K, Bassett P, et al. Adaptive statistical iterative reconstruction (ASIR) affects CT radiomics quantification in primary colorectal cancer. *Eur Radiol.* 2019; 29(10): 5227-5235.

4. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, et al. Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry? *Int J Radiat Oncol.* 2016; 94(4): 824-831.

5. Bibault J-E, Giraud P, Housset M, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci Rep.* 2018; 8(1): 12611.

6. Zhang XY, Wang L, Zhu HT, et al. Predicting rectal cancer response to neoadjuvant chemoradiotherapy using deep learning of diffusion kurtosis MRI. *Radiology.* 2020; 296(1): 56-64.

7. Bollineni VR, Kramer G, Liu Y, Melidis C, deSouza NM. A literature review of the association between diffusion-weighted MRI derived apparent diffusion coefficient and tumour aggressiveness in pelvic cancer. *Cancer Treat Rev.* 2015; 41(6): 496-502.

8. Jung SH, Heo SH, Kim JW, et al. Predicting response to neoadjuvant chemoradiation therapy in locally advanced rectal cancer: diffusion-weighted 3 Tesla MR imaging.
J Magn Reson Imaging. 2012; 35(1): 110-116.

9. Moon SJ, Cho SH, Kim GC, et al. Complementary value of pre-treatment apparent diffusion coefficient in rectal cancer for predicting tumor recurrence.
Abdom Radiol (New York). 2016; 41(7): 1237-1244.

10. Schurink NW, Lambregts DMJ, Beets-Tan RGH. Diffusion-weighted imaging in rectal cancer: current applications and future perspectives. *Br J Radiol.* 2019; 92(1096): 20180655.

11. Lambregts DMJ, Vandecaveye V, Barbaro B, et al. Diffusion-weighted MRI for selection of complete responders after chemoradiation for locally advanced rectal cancer: a multicenter study. *Ann Surg Oncol.* 2011; 18(8): 2224-2231.

12. Beets-Tan RGH, Lambregts DMJ, Maas M, et al. Magnetic resonance imaging for clinical management of rectal cancer: Updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. *Eur Radiol.* 2017; 28(4): 1465-1475.

13. Curvo-Semedo L, Lambregts DMJ, Maas M, et al. Rectal Cancer: Assessment of Complete Response to Preoperative Combined Radiation Therapy with Chemotherapy—Conventional MR Volumetry versus Diffusion-weighted MR Imaging. *Radiology.* 2011; 260(3): 734-743.

14. Carbone SF, Pirtoli L, Ricci V, et al. Assessment of response to chemoradiation therapy in rectal cancer using MR volumetry based on diffusion-weighted data sets: a preliminary report. *Radiol Med.* 2012; 117(7): 1112-1124.

15. Bammer R. Basic principles of diffusion-weighted imaging. *Eur J Radiol.* 2003; 45(3): 169-184.

16. Regini F, Gourtsoyianni S, Cardoso De Melo R, et al. Rectal tumour volume (GTV) delineation using T2-weighted and diffusion-weighted MRI: Implications for radiotherapy planning. *Eur J Radiol.* 2014; 83(5): 768-772.

17. Blazic IM, Lilic GB, Gajic MM. Quantitative Assessment of Rectal Cancer Response to Neoadjuvant Combined Chemotherapy and Radiation Therapy: Comparison of Three Methods of Positioning Region of Interest for ADC Measurements at Diffusion-weighted MR Imaging. *Radiology.* 2017; 282(2): 418-428.

18. Choi MH, Oh SN, Rha SE, et al. Diffusion-weighted imaging: Apparent diffusion coefficient histogram analysis for detecting pathologic complete response to chemoradiotherapy in locally advanced rectal cancer. *J Magn Reson Imaging.* 2016; 44(1): 212-220.

19. Foti PV, Privitera G, Piana S, et al. Locally advanced rectal cancer: Qualitative and quantitative evaluation of diffusion-weighted MR imaging in the response assessment after neoadjuvant chemo-radiotherapy. *Eur J Radiol open.* 2016; 3: 145-152.

20. Haralick R, Shanmugan K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973; 3: 610-621.

21. Galloway MM. Texture analysis using gray level run lengths. *Comput Graph Image Process.* 1975; 4(2): 172-179.

22. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognit Lett.* 1990; 11(6): 415-419.

23. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Phys Medica.* 2019; 61: 44-51.

24. Molina D, Pérez-Beteta J, Martínez-González A, et al. Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images. *Comput Biol Med.* 2016; 78: 49-57.

25. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. Fan Y, ed. *PLoS One.* 2019; 14(3): e0213459.

26. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017; 44(3): 1050-1062.

27. Park H, Lim Y, Ko ES, et al. Radiomics signature on magnetic resonance imaging: Association with disease-free survival in patients with invasive breast cancer. *Clin Cancer Res.* 2018; 24(19): 4705-4714.

28. Bae S, Choi YS, Ahn SS, et al. Radiomic MRI phenotyping of glioblastoma: Improving survival prediction. *Radiology.* 2018; 289(3): 797-806.

29. Bonekamp D, Kohl S, Wiesenfarth M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: Comparison to ADC values. *Radiology.* 2018; 289(1): 128-137.

30. Huang S ying, Franc BL, Harnish RJ, et al. Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. *npj Breast Cancer.* 2018; 4(1): 24.

31. Ji GW, Zhang YD, Zhang H, et al. Biliary tract cancer at CT: A radiomics-based model to predict lymph node metastasis and survival outcomes. *Radiology.* 2019; 290(1): 90-98.

32. Tan X, Ma Z, Yan L, Ye W, Liu Z, Liang C. Radiomics nomogram outperforms size criteria in discriminating lymph node metastasis in resectable esophageal squamous cell carcinoma. *Eur Radiol.* 2019; 29(1): 392-400.

33. Yu FH, Wang JX, Ye XH, Deng J, Hang J, Yang B. Ultrasound-based radiomics nomogram: A potential biomarker to predict axillary lymph node metastasis in early-stage invasive breast cancer. *Eur J Radiol.* 2019; 119(August): 108658.

34. Jin C, Chen W, Cao Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat Commun.* 2020; 11(1): 1-14.

35. Singh A, Regenauer-Lieb K, Walsh SDC, Armstrong RT, van Griethuysen JJM, Mostaghimi P. On Representative Elementary Volumes of Grayscale Micro-CT Images of Porous Media. *Geophys Res Lett.* 2020; 47(15): 0-3.

36. Chen L Da, Liang JY, Wu H, et al. Multiparametric radiomics improve prediction of lymph node metastasis of rectal cancer compared with conventional radiomics. *Life Sci.* 2018; 208: 55-63.

37. Liang M, Cai Z, Zhang H, et al. Machine Learning-based Analysis of Rectal Cancer MRI Radiomics for Prediction of Metachronous Liver Metastasis. *Acad Radiol.* 2019; 26(11): 1495-1504.

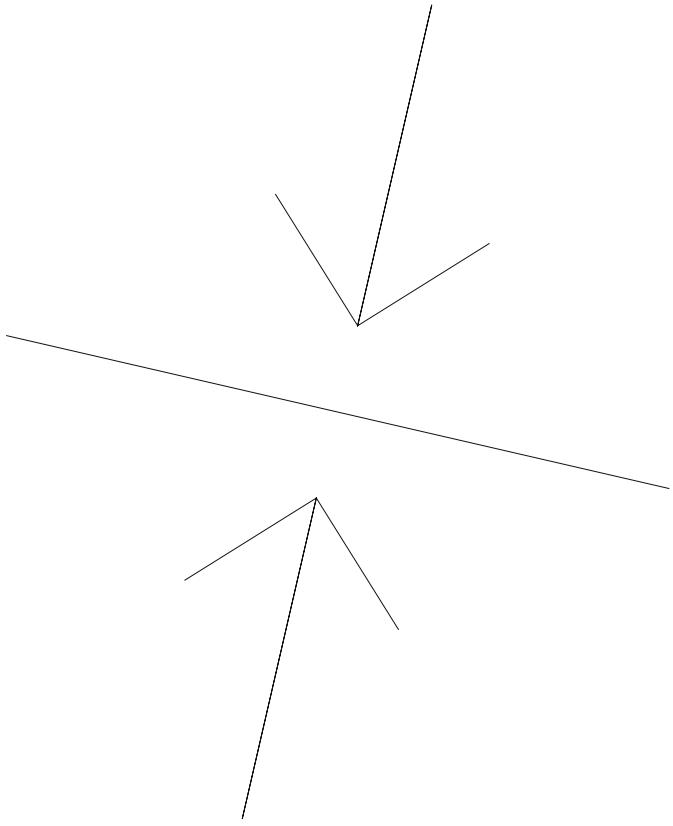
38. Lovinfosse P, Polus M, Daele D Van, et al. FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer. *Eur J Nucl Med Mol Imaging.* 2017; 45: 365-375.

39. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5(1): 4006.

40. Grimm LJ, Zhang J, Mazurowski MA. Computational approach to radiogenomics of breast cancer: Luminal A and luminal B molecular subtypes are associated with imaging features on routine breast MRI extracted using computer vision algorithms. *J Magn Reson Imaging.* 2015; 42(4): 902-907.

41. Stoyanova R, Takhar M, Tschudi Y, et al. Prostate cancer radiomics and the promise of radiogenomics.
Transl Cancer Res. 2016; 5(4): 432-447.

Summary / samenvatting



Summary

There are several key steps in the radiomics workflow, each associated with their own challenges that need to be overcome on the road towards implementing artificial intelligence tools in the clinical workflow of radiology. The goal of this thesis is to investigate these key steps and challenges, using the application of radiomics in rectal cancer as a clinical example.

Segmentation

Chapter 2 focuses on lesion segmentation, an important first step to allow extraction of quantitative features from a given tumor lesion on medical imaging. We investigated the potential of Deep Learning for fully-automated segmentation of rectal tumors on primary staging MRI, using a combination of T2-weighted imaging and Diffusion Weighted Imaging (DWI) sequences. This initial study used a fairly simple patch-based approach, where the network was trained to predict whether a voxel was part of the rectal tumor, based on a small surrounding region. By extracting and classifying patches for all voxels in the input image, a segmentation could be inferred. In the test set we achieved a high performance with Dice Similarity Coefficients (DSC; indicating the spatial overlap between two segmentations on a scale from 0 to 1) between network and manual segmentations of 0.68-0.70, though performance was lower than the agreement between two radiologists performing manual segmentations (0.83 DSC). Moreover, the patch-based approach was hampered by slower inference and low spatial awareness. This was also reflected by the results showing that the network struggled with larger Field-of-View (FOV) acquisitions, erroneously classifying structures near the edge of the FOV as tumor.

More advanced deep learning networks, such as the U-Net are able to process the entire image in a single inference, greatly increasing speed and spatial awareness. In Chapter 3, we optimized the network from Chapter 2 to address these limitations, using a U-Net with additional attention gating, which helps the network to focus on the area of interest. Furthermore, we explicitly investigated the influence of scan quality and tumor complexity on network performance and interreader agreement in a large multicenter dataset with a high variation of acquisition protocols and scan quality. Though performance of the AI model in the test set was still inferior to agreement between radiologists, with DSCs of 0.67 vs. 0.75, the difference was markedly smaller compared to the findings in Chapter 2. Tumor complexity had the largest influence on network performance, with more heterogeneous and irregular tumors resulting in lower DSC scores. Interestingly, tumor complexity had a similar negative effect on the agreement between two radiologists, indicating that these tumors are inherently difficult to segment for computers but also for experienced radiologists. The signal-to-noise on DWI also affected network performance, though to a lesser extent than tumor complexity.

Image Quality

As also highlighted in Chapter 3, image quality is an important prerequisite for Radiomics and AI research. In Chapter 4, we addressed this challenge focusing specifically on the acquisition of DWI of the rectum. Rectal DWI can suffer from significant artefacts caused by the presence of intraluminal gas in the rectum which can have a severe negative impact on diagnostic evaluations. We investigated the application of a preparatory micro-enema shortly prior to acquisition to reduce the amount of intraluminal gas. We observed a marked reduction of significant artefacts from 24% in scans acquired without a micro-enema to <4% in scans performed after application of the micro-enema, showing that even a fairly simple intervention may significantly improve DWI scan quality.

Feature extraction and modelling

In Chapter 5, we addressed the challenge of reproducibility in feature extraction. To this end, we developed PyRadiomics, an easy-to-use open source package for radiomics feature extraction. It is developed in Python, a popular programming language used by many researchers investigating the application of radiomics in medical imaging, coupled with extensions in C for high-performance feature extraction.

Finally, in Chapter 6, we put the radiomics workflow to the test, investigating the application of radiomics to predict the response to neoadjuvant chemoradiotherapy in rectal cancer patients, using only the pre-therapy staging MRI. We compared the performance of a radiomics-based prediction model to the performance of expert radiologists who predicted the response to chemoradiotherapy based on their visual interpretation of the images. The radiomics models achieved a promising performance, with AUCs of 0.69-0.79, which were comparable to the performance of the visual predictions by the expert radiologists. The radiomics model was built using different types of tumor segmentation, performed manually by either expert radiologists or non-expert readers, but also semi-automatically using a basic segmentation algorithm. Interestingly, radiomics performance was similar when using the expert or non-expert manual segmentations, but significantly poorer when using the semi-automatic segmentation algorithm. This highlights the need for better automatic segmentation support tools, such as those addressed in Chapters 2 and 3.

Samenvatting

In het proces van radiomics zijn enkele belangrijke stappen te onderscheiden, elk geassocieerd met uitdagingen die moeten worden overwonnen op de weg naar klinische implementatie van kunstmatige intelligentie (ook wel "artificiële intelligentie"of "AI") in de dagelijkse praktijk van de radiologie. Het doel van deze thesis is het onderzoeken van deze stappen en de bijbehordende uitdagingen, gebruik makende van de toepassing van kunstmatige intelligentie bij endeldarmkanker als klinisch voorbeeld.

Segmentatie

Om kwantitatieve analyse van bijvoorbeeld een tumor laesie mogelijk te maken, is segmentatie van de betreffende laesie op beeldvorming een eerste belangrijke noodzakelijke stap. In Hoofdstuk 2 werd de potentie van AI (Deep Learning) voor het volledig geautomatiseerd segmenteren van tumoren in de endeldarm onderzocht. Hiervoor werd gebruik gemaakt van een combinatie van T2-gewogen en diffusie gewogen imaging (DWI) sequenties die beide onderdeel uit maken het standaard MRI protocol voor endeldarmkanker. In deze initiële studie werd gebruik gemaakt van een vrij eenvoudig netwerk, waarbij per voxel binnen het MRI plaatje werd voorspeld of deze wel of niet tot het tumorgebied behoorde op basis van analyse van een kleine omliggende regio. Een volledige segmentatie van de tumor werd op deze manier verkregen door dit proces te herhalen voor elke voxel in de MRI scan. Toepassing van dit netwerk op een onafhankelijke test set toonde een goed resultaat, met een Dice Similarity Coefficient (DSC; een maat om de spatiele overlap tussen 2 segmentaties weer te geven op een schaal van 0 tot 1) tussen het AI-model en manuele segmentatie door een radioloog van 0.68-0.70. Dit resultaat was echter nog niet geheel vergelijkbaar met de overeenstemming tussen manueel verkregen segmentaties van twee ervaren radiologen (DSC 0.83). Bovendien werd het netwerk gelimiteerd door de regio-gebaseerde methode, wat leidde tot trage segmentatie en weinig ruimtelijk inzicht. Deze nadelen zijn ook herkenbaar in de resultaten, waarbij het netwerk bij beelden met een groter afgebeeld gebied (Field-Of-View, FOV) foutief structuren aan de rand van de afbeeldingen classificeerde als "tumor".

Meer geavanceerde Deep Learning AI netwerken, zoals het U-Net, zijn in staat een afbeelding in zijn geheel te verwerken en zijn zodoende in staat meer ruimtelijk inzicht te tonen. Om de limitaties van het netwerk uit Hoofdstuk 2 te overkomen hebben wij dan ook gebruik gemaakt van een dergelijk U-Net in Hoofdstuk 3, met de toevoeging van "attention-gating" bedoeld om het netwerk te helpen focussen op het belangrijkste doelgebied binnen de scan: de endeldarm en het omliggende vetweefsel. Naast een meer geavanceerde architectuur van het netwerk hebben we ook expliciet de invloed van beeldkwaliteit en tumor complexiteit op de accuratesse van de segmentaties onderzocht. Hoewel de accuratesse van het netwerk ook in dit onderzoek lager uitviel in vergelijking met de overeenkomst tussen ervaren radiologen onderling (DSC 0.67 versus 0.75), was het verschil duidelijk kleiner dan in Hoofdstuk 2. Tumor complexiteit had de grootste invloed op de segmentatie accuratesse, waarbij meer heterogene en grillig gevormde tumoren zowel voor het AI netwerk als voor de radiologen resulteerden in lagere DSCs. Dit leert ons dat deze meer complexe tumoren inherent lastiger zijn om te segmenteren. De beedlkwaliteit (signaal-ruis verhouding) van de DWI-beelden was ook van invloed op de accuratesse van het netwerk, al was dit effect minder uitgesproken in vergelijking met de tumor complexiteit.

Beeldkwaliteit

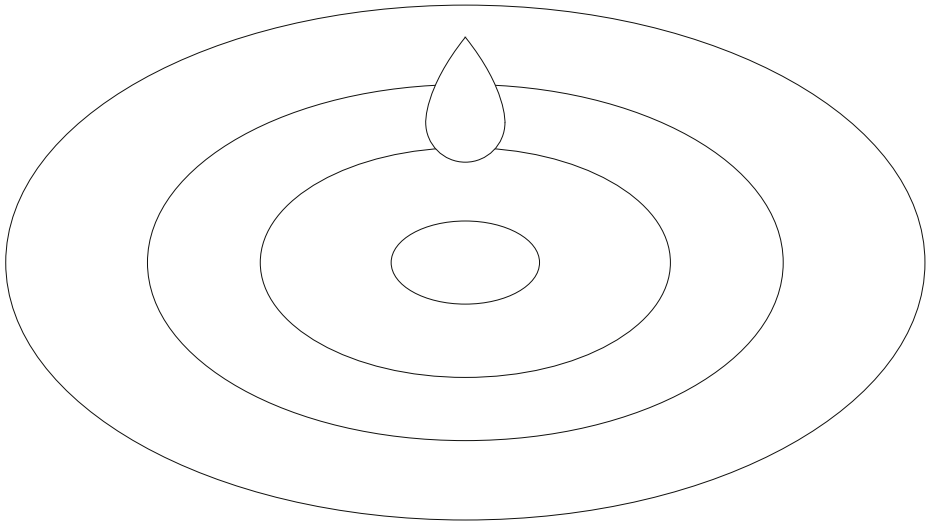
Zoals ook beschreven in Hoofdstuk 3, is goede beeldkwaliteit een belangrijke vereiste voor onderzoek op het gebied van Radiomics en kunstmatige intelligentie. In Hoofdstuk 4 hebben we deze uitdaging onderzocht in het specifieke geval van beeldartefacten op DWI sequenties van de endeldarm. Bij deze beelden kunnen ernstige susceptibiliteitsartefacten optreden die worden veroorzaakt door de aanwezigheid van lucht in de endeldarm, welke kunnen leiden tot substantieel verslechterde beoordeelbaarheid van deze beelden. In deze studie hebben wij onderzocht of het toepassen van een micro-klysma leidt tot een verminderde hoeveelheid intra-luminaal lucht en dus tot een afname in de aanwezigheid van artefacten. Wij hebben kunnen constateren dat deze toepassing inderdaad leidde tot een duidelijke vermindering van klinisch significante artefacten, met een reductie tot <4% na toediening van het micro-klysma versus 24% in scans vervaardigd zonder micro-klysma. Dit toont aan dat zelfs een vrij eenvoudige interventie kan leiden tot een significante verbetering van beeldkwaliteit.

Feature extractie en data modellering

Om de reproduceerbaarheid van radiomics-onderzoek te vergroten hebben wij een open-source software pakket ontwikkeld, genaamd *PyRadiomics*. In Hoofdstuk 5 introduceren wij dit pakket. Het is ontwikkeld in Python, een populaire programmeertaal bij onderzoekers in het veld van medische beeldvorming. Bovendien is het specifiek ontwikkeld met het oog op het gebruik door een gemeenschap van radiomics-onderzoekers, met extra nadruk op de leesbaarheid van de broncode en uitgebreide documentatie. Om een hoge doorvoersnelheid van de feature extractie te garanderen, zijn enkele stukken van het pakket in C-code geschreven.

Tot slot hebben we het proces van radiomics getest in Hoofdstuk 6 in een klinische studie naar endeldarmkanker, waarbij we de applicatie van radiomics voor het voorspellen van respons op neoadjuvante chemoradiotherapie hebben onderzocht met behulp van MRI scans vervaardigd voor start van de behandeling. We hebben de accuratesse van een radiomics-model vergeleken met de accuratesse van ervaren radiologen, welke hun voorspelling baseerden op een visuele interpretatie van dezelfde beelden. Het radiomics model toonde een veelbelovend resultaat met een AUC van 0.69-0.79, vergelijkbaar met de accuratesse van de ervaren radiologen. Het radiomics-model werd ontwikkeld met behulp van segmentaties die manueel werden vervaardigd door respectievelijk ervaren radiologen en meer onervaren lezers, maar ook een door een simpel semi-automatisch segmentatie algoritme. Hieruit bleek dat het radiomics-model een goed en vergelijkbaar resultaat gaf ongeacht het ervarings niveau van de handmatige intekeningen, maar duidelijk slechter presteerde wanneer het semi-automatische algoritme werd gebruikt. Dit benadrukt de noodzaak voor het ontwikkelen van betere segmentatie algoritmen zoals die onderzocht werden in Hoofdstukken 2 en 3.

Impact paragraph



Main aims and outcomes of this thesis

Medical imaging, such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), helps doctors to detect, stage and monitor tumors, which provides them with information that is crucial to help define the best possible treatment for each individual cancer patient. In current clinical practice, medical images are mainly assessed "qualitatively", meaning that the images are assessed visually by trained radiologists, who report their findings in text reports. There are however also several ways to assess medical images quantitatively. Some simple quantitative tools are already being used in daily clinical practice, including for example size or signal intensity measurements. In recent years, the possibilities of quantitative medical imaging analysis have evolved tremendously. An important development in this field has been the introduction of radiomics. In radiomics, the phenotype of a target volume within a medical image (e.g., a cancer lesion on a CT or MRI scan) is captured through extraction of a large panel of "features", calculated using advanced mathematical formulas. This allows for the extraction of much more information from the medical imaging than is visible by the 'naked eye'[1]. Using artificial intelligence (machine learning) computer methods, this information can be correlated to various clinical outcomes such as the response of a tumor to anti-cancer treatment. This way, the radiomics phenotype can be used to render imaging-biomarkers of disease that can be incorporated into clinical prediction models, which may ultimately act as decision-support tools to aid in further personalization of treatment for cancer patients.

In this thesis we addressed some of the key challenges in the radiomics workflow, using rectal cancer as a case example. Radiomics requires segmentation of the volume of interest within an image (e.g. a tumor lesion), which can be a very time consuming task requiring many hours of manual input from radiologists. The results in Chapters 2 and 3 show that fully-automatic segmentation of rectal tumors on MRI is feasible using artificial intelligence (AI) models, which can serve as a starting point and significantly reduce the amount of manual input required from radiologists. Another important prerequisite for successful radiomics analysis is the availability of good quality source images. Diffusion-weighted imaging (DWI) is nowadays an integral part of many oncologic MR imaging protocols and commonly used for quantitative MRI data analysis. In Chapter 4 we have shown that a simple intervention such as a preparatory micro-enema can help to greatly increase DWI scan quality. As shown in Chapter 3, this can contribute to improved performance of automatic segmentation protocols. A final challenge in the radiomics workflow is feature extraction, which is largely dependent on the mathematical formulas used and the implementation of these formulas in feature extraction software algorithms.
In Chapter 5 we introduce *PyRadiomics*, an open-source toolkit for easy and reproducible feature extraction. It has been specifically developed for use in a community of radiomics researchers, aiming to increase the transparency and reproducibility of radiomics research by providing an easy go-to resource for feature extraction. In Chapter 6, we have put the radiomics workflow to the test in a clinical study and have shown that radiomics has potential to render valuable imaging biomarkers for pre-treatment prediction of response to neoadjuvant therapy in locally advanced rectal cancer.

Relevance

The results presented in this thesis may help future research in radiomics, paving the way on the road towards clinical implementation. The segmentation algorithms presented in Chapters 2 and 3 aim to automate the segmentation step of the radiomics workflow. Especially when combined with optimized acquisition protocols, these algorithms can act as support tools to greatly improve the efficiency and reduce the workload of image segmentation.
The *PyRadiomics* toolbox introduced in Chapter 5, with extensive documentation and publicly available source code, is aimed at widespread use by a global community of radiomics researchers. This removes the need for researchers to learn and implement radiomics features in custom-built code and contributes to increased reproducibility and comparability of published work. The success of *PyRadiomics* is reflected in its worldwide use – even serving as the radiomic feature extractor in several commercial products – and high number of citations for the paper[2] introducing this toolbox. Finally, results in Chapter 6 show the potential of radiomics for rendering valuable predictive biomarkers in rectal cancer. In future research, these may be incorporated into clinical prediction models, ultimately aiming to improve patient-tailored treatment planning and outcomes.

Target population

There are several groups who may benefit from the results presented in this thesis. The first are the researchers investigating the application of radiomics. Though this thesis is placed in the context of radiomics in rectal cancer, the workflow and challenges encountered – and the solutions offered in this thesis – can be generalized and applied to radiomics research in other oncological and non-oncological research fields.

Once properly validated in multicenter and prospective clinical studies, radiomics may aid healthcare professionals to build decision support models to better stratify patients, getting the right treatment to the right patient. However, before implementation into clinical workflows is feasible, the challenges described in the thesis need to be overcome.

Activities

The results presented in this thesis have been actively distributed, both through publication of results in peer-reviewed journals and presentation at multiple national and international conferences. The knowledge gained in this thesis is also currently being applied through new collaborations in follow-up research projects investigating radiomics and AI in rectal cancer as well as other tumor types. One such collaboration is an ongoing multicenter trial with participation of ten different centers in the Netherlands, including academic and oncology referral centers, but also several large teaching hospitals. This study aims to further build on the results acquired in chapter 6 by validating the predictive value of radiomics to predict rectal tumor response in a large multicenter dataset with large data heterogeneity, reflecting the regular clinical workflow.

The knowledge gained in this thesis is also incorporated into and distributed via the *PyRadiomics* platform. This is achieved through the public access and open-source nature of this toolbox, which is being used by a continuously growing community of researchers worldwide. *PyRadiomics* is now also part of the "Imaging Biomarker Standardization Initiative" (IBSI)[3], which aims at standardization of radiomic feature extraction, regardless of software implementations.

Ultimately, biomarkers derived from medical imaging using radiomics and AI will most likely only be a part of the prediction. Like multidisciplinary tumor boards incorporating information from multiple aspects of healthcare, a combination of information derived from imaging as well as other clinical, histopathological and genetic sources will most likely result in higher performance than can be achieved with imaging data alone. To this end, collaboration between researchers of multiple disciplines is crucial to build the strongest possible clinical prediction models and decision-support tools that can really have an impact on patient management and ultimately treatment outcomes.

References

1. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-1248.

2. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e107.

3. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338.

4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-577.

List of publications

Primo Loco

1. van Griethuysen JJM, Bus EM, Hauptmann M, Lahaye MJ, Maas M, ter Beek LC, Beets GL, Bakers FCH, Beets-Tan RGH, Lambregts DMJ. Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5T – Effect of applying a micro-enema to improve image quality. *Eur J Radiol.* 2017; 99 (0): 131-137.

2. Trebeschi S, van Griethuysen, JJM, Lambregts DMJ, Lahaye MJ, Parmar C, Bakers, FCH, Peters NHGM, Beets-Tan RGH, Aerts HJWL. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci Rep.* 2017; 7:5301.

3. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin J-C, Pieper S, Aerts HJWL. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017; 77: e104–e107.

4. van Griethuysen JJM, Lambregts DMJ, Trebeschi S, Lahaye MJ, Bakers FCH, Vliegen RFA, Beets GL, Aerts HJWL, Beets-Tan RGH. Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer. *Abdom Radiol.* 2020; 45 (3): 632-643.

5. van Griethuysen JJM, Lambregts DMJ, Schurink NW, Lahaye MJ, Trebeschi S, Bakers FCH, Vliegen RFA, Geenen RWF, Cappendijk VC, de Bie SH, Aerts HJWL, Beets-Tan RGH. Deep learning for fully automated segmentation of rectal tumors on multiparametric MRI in a multicenter setting. *Submitted for publication*.

Alto Loco

1. van Heeswijk MM, Lambregts DMJ, van Griethuysen JJM, Oei S, Rao S-X, de Graaff CAM, Vliegen RFA, Beets GL, Papanikolaou N, Beets-Tan RGH. Automated and semi-automated segmentation of rectal tumor volumes on diffusion-weighted MRI: can it replace manual volumetry? *Int J Radiat Oncol*. 2016; 94 (4): 824-831.

2. Lambregts DMJ, Delli Pizzi A, Lahaye MJ, van Griethuysen JJM, Maas M, Beets GL, Bakers FCH, Beets-Tan RGH. A Pattern-Based Approach Combining Tumor Morphology on MRI With Distinct Signal Patterns on Diffusion-Weighted Imaging to Assess Response of Rectal Tumors After Chemoradiotherapy. *Dis Colon Rectum*. 2018; 61 (3): 328-337.

3. Dinis Fernandes C, Dinh CV, Walraven I, Heijmink, SW, Smolic, M, van Griethuysen, JJM, Simões R, Losnegard A, van der Poel HG, Pos FJ, van der Heide U. Biochemical recurrence prediction after radiotherapy for prostate cancer with T2w magnetic resonance imaging radiomic features. *Phys Imaging Radiat Oncol*. 2018; 7 (June): 9-15

4. Dou TH, Coroller TP, van Griethuysen JJM, Mak RH, Aerts HJWL. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. *PLoS One*. 2018; 13 (11): e0206108.

5. Schwier M, van Griethuysen JJM, Vangel MG, Pieper S, Peled S, Tempany C, Aerts HJWL, Kikinis R, Fennessy FM, Fedorov A. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep*. 2019; 9 (1): 9441.

6. Min LA, Vacher YJL, Dewit L, Donker M, Sofia C, van Triest B, Bos P, van Griethuysen JJM, Maas M, Beets-Tan RGH, Lambregts DMJ. Gross tumour volume delineation in anal cancer on T2-weighted and diffusion-weighted MRI – Reproducibility between radiologists and radiation oncologists and impact of reader experience level and DWI image quality. *Radiother Oncol*. 2020; 150: 81-88.

7. Singh A, Regenauer-Lieb K, Walsh SDC, Armstrong RT, van Griethuysen JJM, Mostaghimi P. On Representative Elementary Volumes of Grayscale Micro-CT Images of Porous Media. *Geophys Res Lett*. 2020; 47 (15): 0-3.

8. Krdzalic J, Beets-Tan RGH, Engelen SME, van Griethuysen JJM, Lahaye MJ, Lambregts DMJ, Bakers FCH, Vliegen RFA, Beets GL, Maas M. MRI predicts increased eligibility for sphincter preservation after CRT in low rectal cancer. *Radiother Oncol*. 2020; 145: 223-228.

9. Schurink NW, Min LA, Berbee M, van Elmpt W, van Griethuysen JJM, Bakers FCH, Roberti S, van Kranen SR, Lahaye MJ, Maas M, Beets GL, Beets-Tan RGH, Lambregts DMJ. Value of combined multiparametric MRI and FDG-PET/CT to identify well-responding rectal cancer patients before the start of neoadjuvant chemoradiation. *Eur Radiol*. 2020; 30 (5): 2945-2954.

10. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020; 295 (2): 328-338.

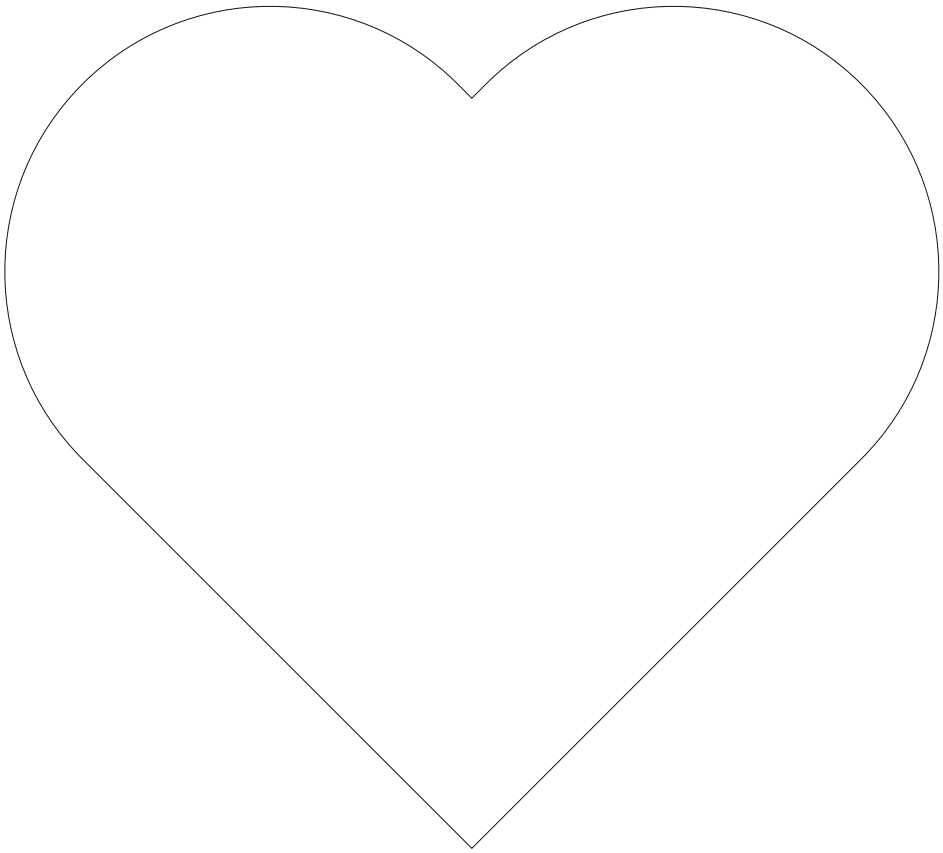
11. Min LA, Ackermans LLGC, Nowee ME, van Griethuysen JJM, Roberti S, Maas M, Vogel, WV, Beets-Tan, RGH, Lambregts, DMJ. Pre-treatment prediction of early response to chemoradiotherapy by quantitative analysis of baseline staging FDG-PET/CT and MRI in locally advanced cervical cancer. *Acta radiol*. 2021; 62 (7): 940-948.

12. Schurink NW, van Kranen SR, Berbee M, van Elmpt W, Bakers FCH, Roberti S, van Griethuysen JJM, Min LA, Lahaye MJ, Maas M, Beets GL, Beets-Tan RGH, Lambregts DMJ. Studying local tumour heterogeneity on MRI and FDG-PET/CT to predict response to neoadjuvant chemoradiotherapy in rectal cancer. *Eur Radiol*. 2021; 31 (9): 7031-7038.

13. Min LA, Castagnoli F, Vogel WV, Vellenga JP, van Griethuysen JJM, Lahaye MJ, Maas M, Beets Tan RGH, Lambregts DMJ. A decade of multi-modality PET and MR imaging in abdominal oncology. *Br J Radiol*. 2021; 94 (1126): 20201351.

14. Bogveradze M, el Khababi N, Schurink NW, van Griethuysen JJM, de Bie S, Bosma G, Cappendijk VC, Geenen RWF, Neijenhuis P, Peterson G, Veeken CJ, Vliegen RFA, Maas M, Lahaye MJ, Beets GL, Beets-Tan RGH, Lambregts DMJ. Evolutions in rectal cancer MRI staging and risk stratification in The Netherlands. *Abdom Radiol*. Accepted 2021

Dankwoord



Het laatste – en mogelijk meest gelezen – hoofdstuk is het dankwoord. Zonder veel directe en indirecte hulp was dit werk er niet geweest. In dit hoofdstuk wil ik graag iedereen bedanken die hebben bijgedragen aan het tot stand komen van dit werk.

Allereerst wil ik graag mijn promotoren bedanken, prof. dr. Beets-Tan en prof. dr. Hugo Aerts. Beste Regina, het is nu bijna 6 jaar geleden dat je naar Amsterdam bent verhuisd en mij hebt gevraagd of ik mee wilde gaan als nieuwe PhD-student bij de afdeling radiologie van het Antoni van Leeuwenhoek ziekenhuis. Hierbij heb je me nog gewaarschuwd dat het wel een hele technische PhD zou worden, waarop ik vrij lacherig reageerde dat dat niet zo'n probleem zou zijn. Ik hoop dat mijn lacherige reactie in de laaste jaren begrijpelijk is geworden. Ik vond het een eer om onder jouw vleugels mijn onderzoek te mogen doen.

Beste Hugo, daar waar Regina de klinische kant van mijn PhD waarborgde met haar uitgebreide ervaring met beeldvorming bij endeldarm kanker, nam jij de technische kant van mijn PhD voor jouw rekening. Ik zal je altijd dankbaar blijven voor alle kansen die je mij hebt geboden, waardoor ik het gevoel heb dat ik echt een impact heb kunnen maken met ons onderzoek. Doenja, mijn co-promotor. Ondanks dat je elke keer weer volhield dat je van al die technische dingen geen bal verstand had, was mijn promotie niet gelukt zonder jouw steun bij al het andere wat ik maar tegenkwam tijdens mijn promotie. Elke eerste revisie was voor zeker de helft rood gekleurd, al kwam dit wel steevast met een opmerking in de mail dat het wel een goed stuk was, en het meer "textuele" aanpassingen waren.

Graag wil ik ook de leden van mijn beoordelingscommissie – bestaande uit prof. dr. Stassen, prof. dr. Lambin, prof. dr. Wildberger, prof. ir. Van Ooijen en dr. Visser – bedanken voor de tijd en moeite die ze hebben genomen om dit werk te beoordelen.

Elk artikel is een team-effort en ik wil dan ook alle co-auteurs bedanken voor hun waardevolle input en de moeite die ze hebben genomen om mee te denken, mee te schrijven en mee te onderzoeken naar radiomics en AI in rectal cancer.

Andrey Federov and Steve Pieper, my fellow main developers of *PyRadiomics*, thank you for an engaging collaboration. I'll fondly remember all the interesting discussions, both on- and off-topic, as well as our face-to-face meetings at the 3D Slicer Project week. You've helped me to grow as a Python software developer. Even though I've not received any official education on software development, you've made me feel appreciated and skilled at Python development.

Het Recteam kan uiteraard niet ontbreken. Vaak hoor je horror verhalen over onderzoeksteams waar het vechten is voor je plekje, maar ik had me geen fijner onderzoeksteam dan het Recteam kunnen voorstellen. Doenja (ja, je krijgt je 2e alinea in het dankwoord), naast je uitstekende begeleiding ben je ook enorm gezellig buiten het werk om. Met plezier denk ik terug aan alle etentjes en borrels, zowel in het buitenland op congres, als even na het werk of zomaar spontaan. Ik hoop dat ik nog van vele chevice etentjes met jou en Max mag genieten. Beste Max, Ik heb intens genoten van je droge humor en alle interessante discussies die we hebben gehad, ook al leidden die meer dan eens tot enig zuchten bij Doenja. Miriam, Rianne en Britt, mijn voorgangers bij het Recteam, onder jullie directe begeleiding begon mijn tijd bij

het recteam. Eerst als WESP-student en later als (beginnend) PhD-student. Met veel plezier denk ik terug aan onze kleine hokjes op de universiteit, met de kanker-bank, koffie sterk genoeg om door metaal te branden en niet te vergeten prinsessen-tape. Ik vond het erg jammer maar begrijpelijk dat jullie in Maastricht bleven om jullie promotie af te maken. Gelukkig hebben we zo nu en dan nog een recteam uitje om even bij te kletsen en was Miriam de afgelopen 2 jaar ook mijn directe collega in Apeldoorn! Ik hoop ook in de toekomst weer met jullie te kunnen samenwerken. Marit, mijn mede pioneer-PhD bij het recteam, afdeling Amsterdam. Uiteindelijk liepen onze onderzoeken vrij parallel, maar zeker in het begin was het gezellig om samen ons plekje bij het AvL te vinden.

Stefano, initially we worked closely together on AI in rectal cancer, but pretty soon you pursued your own research line with AI in immunotherapy! I enjoyed our trips to the United States together and learned a lot from you on how to use deep learning. Whenever I had a question, you were there to help me along. I hope we continue working together in the future.

Members of the Geek Room, sometimes it was a tight fit in our office, but always a pleasure! We started out with Stefano, Paula and Marjaneh, but lost Stefano to "room 10"... I'll fondly remember all the interesting and funny discussions, internet memes on the wall and many, many cups of coffee. Paula, bij je sollicitatiegesprek werd je na je gesprek even in de Geek Room geparkeerd, waar je ondanks het feit dat je een tikkeltje zenuwachtig was meteen een goede indruk achterliet. Dank voor alle gezelligheid en de groene touch die je hebt aangebracht in de Geek Room. Marjaneh, thank you for all the language lessons and fun times! Of course, the "Geek Room" was not the only room in the "tuinhuis". I'd also like to thank all the other members of the "tuinhuis" for all the coffee, Sinterklaas celebrations and many, many lunches.

Niels, niet alleen een latere aanwinst voor de Geek Room, maar ook mijn opvolger en paranimf, wat hebben we een gezellige tijd gehad (en hebben we nog steeds). Ik denk nog met veel plezier terug aan ESGAR 2018. Samen hebben we toen Dublin een beetje verkend, inclusief enkele pubs onder het genot van meerdere pints Guiness. Maar ook in Nederland hebben we al veel gezellige feestjes gehad. Bovendien heb je buiten het werk om mij geholpen bij de verbouwing van mijn huis. Ik vind het erg fijn dat je mijn paranimf wilde zijn en ik ben vereerd dat ik te zijner tijd die rol ook voor jou mag vervullen.

Maud, wat hebben we een hoop meegemaakt sinds we elkaar in het eerste jaar van onze studie leerden kennen. Als enigen van ons zotte groepje zijn wij een PhD begonnen, en hebben maar meteen de pact gesloten om elkaars paranimf te zijn. Zo doende... Ondanks dat jij tijdens onze studie altijd aangaf weer naar het noorden te willen gaan, ben jij degene die in het zonnige zuiden is gebleven, wie had dat ooit gedacht. Ik ben ontzettend blij dat ik jou tot mijn vrienden mag rekenen en ik kijk uit naar het moment dat het mijn beurt is om jou bij te staan bij de verdediging van je thesis.

Ook de stille krachten op de achtergrond kunnen niet ontbreken in dit dankwoord. Jos Slenter, hartelijk dank voor alle moeite die je hebt genomen om mij te helpen met de enorme selecties van scans die nodig waren om mijn onderzoek te doen. Erik-Jan Rijkhorst, dank je wel voor al het vertrouwen dat je hebt in mijn programmeerkunsten. Zonder jouw steun zou het mij niet gelukt zijn wat automatisering in te voeren bij het selecteren van al die scans in het AvL.

Lieve collega's van het Gelre ziekenhuis, dank voor het warme onthaal in Apeldoorn. Na 4 jaar voltijd onderzoek doen twijfelde ik wel een beetje hoe goed de opleiding radiologie zou gaan. Maar in het fijn opleidingsklimaat van Apeldoorn was dat geen probleem, ik kon zelfs meteen part-time beginnen, zodat ik nog genoeg tijd had om aan mijn promotie te kunnen werken.

Willem, al toen ik aankondigde dat ik ging promoveren en dus een thesis boekje ging maken kondigde jij aan dat je het grafisch ontwerp wilde doen. Nu is het dan eindelijk zover, en ligt het boekje er dan. Dank je wel voor alle moeite van het ontwerp maken en al mijn ingewikkelde tekst te zetten. Ik ben enorm trots op het "ontworpen non-ontwerp" van het resultaat.

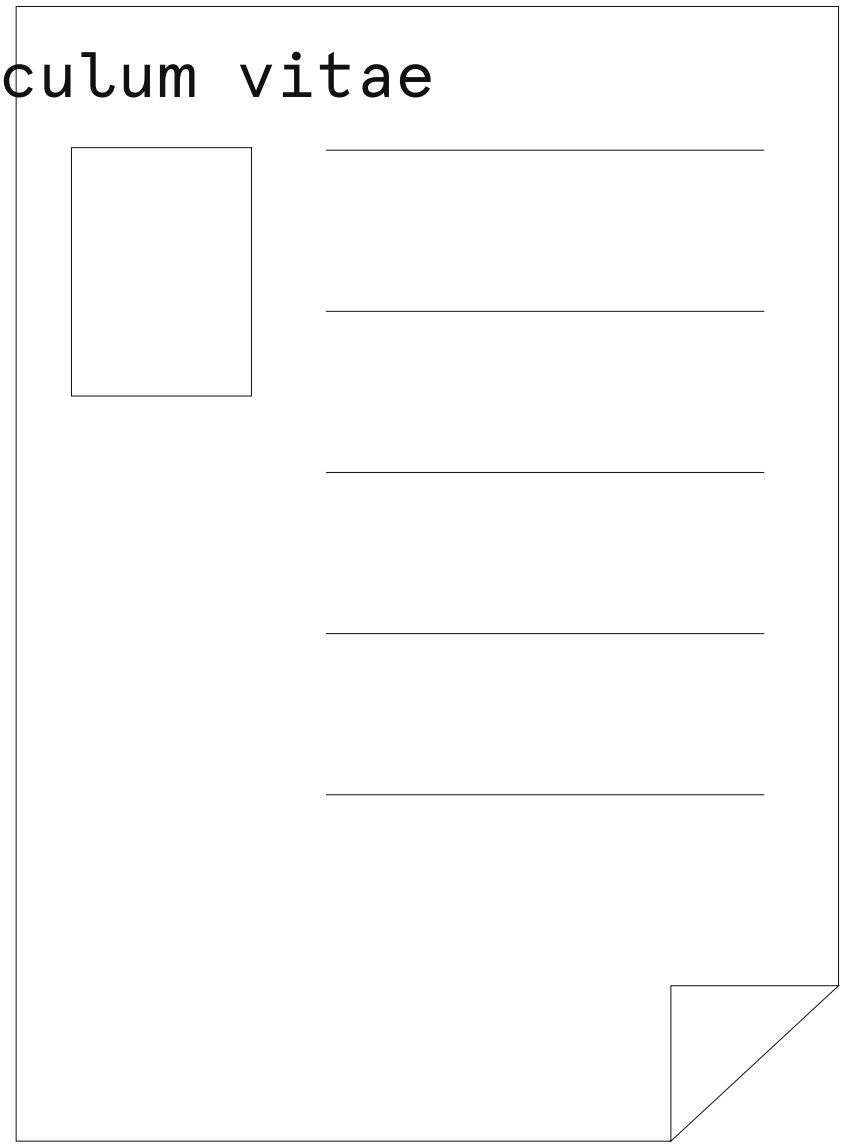
Ook de bourgondiërs mogen niet ontbreken in dit dankwoord. Jullie zorgden voor de broodnodige afleiding tussen al het harde werken door. Het begon als zeilen en feesten in Maastricht tijdens onze studie, maar ondertussen is dat uitgebreid met huizen verbouwen, bruiloften en uiteraard het zeilreisje naar Kroatië! Bij de eetclub leven we ons helemaal uit in de keuken. Begonnen als een leuk ideetje tijdens het eten bij een introweekend, maar nu jaren verder zetten we nog steeds heerlijke 8-gangen diners neer.

Mijn lieve familie, dank je wel voor alle steun door de jaren. Lieve Corry en Jan, zonder jullie wijze lessen was ik nooit zo ver gekomen. Lieve Marjoke, Hessel, Arne, Bernd en Miriam, mijn broers en zussen, wat hebben we toch veel meegemaakt. Ik ben ontzettend trots op onze hechte relatie, ondanks de wat complexe familie samenstelling en leeftijds verschillen. Ons pap en mam zijn er helaas niet meer om mijn verdediging bij te wonen, maar ik herken ze in jullie en weet dat ze trots zouden zijn op wat wij allemaal bereikt hebben.

Ook mijn nieuwe familie wil ik graag bedanken. Lieve Karen en Marco, wat ben ik blij met jullie als schoonouders. Meteen toen ik aankondigde dat ik naar Amsterdam moest verhuizen gaven jullie aan dat als het nodig was ik zonodig bij jullie kon bivakeren. Wageningen is een best eind van Amsterdam, maar nog altijd beter dan Maastricht. Gelukkig was het niet nodig en kon ik op tijd een kamer vinden in Osdorp. Lieve Viv en Bram, samen met Nathalie delen we een voorliefde voor speciaalbier. Met jullie erbij is er nooit een saai moment. Daarnaast vinden wij verduurzaming allebei heel belangrijk en geniet ik van onze discussies over hoe je je impact op het milieu zo klein mogelijk kan houden.

Nathalie, achter elke succesvolle man staat een sterke partner en dat ben jij. Zonder jou had ik dit alles zeker niet gered. Zowel in actieve hulp bij het werk, en dan met name door je uitstekende planinngs-skills (die bij mij soms wel wat hulp kunnen gebruiken), als ook je mentale steun en enorme geduld. Al bij het begin van mijn promotie was je enorm ondersteunend en was je zonder discussie bereid onze plannen voor samenwonen in de koelkast te zetten. Intussen wonen we al weer ruim 4 jaar samen in Utrecht en kijken we met veel plezier en spanning naar onze toekomst samen. Ik had me geen betere levenspartner kunnen voorstellen en ben elke dag ontzettend dankbaar dat jij je leven met mij wil delen.

Curriculum vitae



Joost Johannes Marijn van Griethuysen was born on November 29th 1990 in Casteren, a small village in the Kempen, near the Belgium border in Noord-Brabant. After graduating cum laude from high school in 2009, he started studying medicine in Maastricht. Throughout his studies, Joost had taken a liking to radiology, prompting him to combine his final clinical and scientific internship at the radiology department of the MUMC. It was here that he first met with prof. dr. Regina Beets-Tan, who proposed this PhD-track as part of a new collaboration between her research team and that of prof. dr. Hugo Aerts. After receiving his masters' degree cum laude in 2015, Joost joined prof. dr. Beets-Tan in moving to Amsterdam to work at the Antoni van Leeuwenhoek/The Netherlands Cancer Institute, and supplemented his work with several visits to Boston to work as part of prof. dr. Aerts' team.

In December 2019, Joost finished his PhD contract and started his clinical residency at the radiology department of Gelre hospital in Apeldoorn on a part-time basis. In the additional spare time, Joost continued his research as well as the development of software intended to automate key pre-processing steps in the radiology research workflow at The Netherlands Cancer Institute.