Diffusion MRI as a decision making tool for in-room MRI-guided radiotherapy

Højmark Bisgaard, Anne Louise

DOI:
10.21996/228s-xa44

Publication date:
2023

Document version:
Final published version

Citation for pulished version (APA):
Højmark Bisgaard, A. L. (2023). *Diffusion MRI as a decision making tool for in-room MRI-guided radiotherapy*. [Ph.D. thesis, SDU]. Syddansk Universitet. Det Sundhedsvidenskabelige Fakultet. https://doi.org/10.21996/228s-xa44

August 2023

Ph.D. Thesis

Diffusion MRI as a decision making tool for in-room MRI-guided radiotherapy

by

Anne Louise Højmark Bisgaard

Laboratory of Radiation Physics
Odense University Hospital

Department of Clinical Research
Faculty of Health Sciences
University of Southern Denmark

SDU

Assessment committee

Professor, Ph.D., M.D., Clifton David Fuller
Department of Radiation Oncology
MD Anderson Cancer Center

Professor, Ph.D., Adam Espe Hansen
Department of Diagnostic Radiology
Copenhagen University Hospital

Associate professor, Ph.D., Bo Redder Mussmann
Department of Clinical Research
University of Southern Denmark

Supervisors

Associate professor, M.Sc., Ph.D., Faisal Mahmood (main supervisor)
Laboratory of Radiation Physics
Odense University Hospital

Professor, M.Sc., Ph.D., Carsten Brink (co-supervisor)
Laboratory of Radiation Physics
Odense University Hospital

Professor, M.D., Ph.D., Tine Schytte (co-supervisor)
Department of Oncology
Odense University Hospital

Summary in English

Diffusion MRI as a decision making tool for in-room MRI-guided radiotherapy

Magnetic resonance imaging (MRI) plays an increasing role in radiotherapy (RT) as it provides excellent soft tissue contrast. One of the most recent technical advances in RT is the hybrid MRI linear accelerator (MRI-linac), an integrated MRI scanner and linear accelerator, which enables adaptation of the treatment to match the patient's anatomy-of-the-day. In addition, functional MRI techniques such as diffusion-weighted MRI (DWI) provide biological information of body tissue, and hereby holds great potential for biologically guided RT.

DWI allows derivation of quantitative parameters such as the apparent diffusion coefficient (ADC), which may be used to identify radio-resistant regions within the tumour, and are potential biomarkers for response to RT. DWI might thus be used as a tool to guide treatment decisions in the clinic and hereby help personalizing the treatment to the individual patient.

Before DWI can become a decision making tool in the clinic, a technical validation must be performed to ensure consistency of biomarker measurements in terms of inter- and intra-observer variation, repeatability and reproducibility. Further, a clinical validation is required to establish a relationship between biomarkers and clinical outcomes. With the overall goal of translating DWI into clinical use, this thesis addresses some of the challenges in relation to a technical and clinical validation of DWI.

Study 1: Delineation of regions of interest for ADC measurements

ADC measurements require delineation of a region of interest (ROI), which is observer dependent and time consuming if performed manually. Study 1 presented a simple, threshold-based semi-automatic delineation tool for ADC measurements, and evaluated its performance in patients with rectal cancer. Compared with manual delineation, the tool showed a slightly smaller intra-observer ADC variation. Moreover, it was capable of detecting temporal changes in ADC larger than the repeatability, indicating that the observed ADC changes are true changes and not only a consequence of measurement uncertainty. Thus, the tool may become useful for a technical validation of ADC.

Study 2: Multi-centre ADC reproducibility

Varying methods for ADC derivation between centres (i.e. hospitals or research institutions) lead to a poor multi-centre ADC reproducibility and hamper the technical validation of ADC. Study 2 evaluated the ADC variation between nine MRI-linac centres with respect to two categories: delineation and calculation. MRI scans were shared between the centres, and each centre performed 1) delineation of ROIs and 2) ADC calculation within the delineated ROIs from all centres using the centre's local calculation method. A correlation between delineation variation and ADC variation was observed, underlining the need for and importance of a reduction of the delineation uncertainty. Interestingly, the calculation-related ADC variation was even larger than delineation-related ADC variation, and was mainly driven by different choices of b-values. Other important factors were the software implementation and whether the calculation was performed voxel-wise or based on the mean DWI signal within the ROI. It is the hope that these findings will be useful when designing future studies of ADC as a biomarker, especially in a multi-centre setting.

Study 3: Prediction of response using longitudinal DWI

The MRI-linac has made daily acquisition of DWI feasible, however, the literature investigating the relationship between longitudinal DWI acquired during the RT course and clinical outcome is sparse. Study 3 investigated the value of longitudinal DWI for prediction of overall survival in patients with pancreatic cancer using both a standard, model-based method (ADC) and a model-free decomposition method for parameter derivation. The best prediction model for overall survival based on cross-validation included two DWI parameters, both derived using decomposition of the DWI signal. Moreover, the best model included both baseline information and DWI changes during the RT course, and hence demonstrated potential value of longitudinal DWI for response prediction.

Conclusions

This thesis has addressed important steps in both technical and clinical validation of DWI. It was shown that semi-automatic delineation using a could be a way to improve consistency of ADC measurements and save time compared to manual delineation. Moreover, recommendations for improved multi-centre reproducibility of ADC were provided. These findings can be seen as initiatives to improve the repeatability and reproducibility of ADC measurements, and might be useful in future validation studies of

ADC as a response biomarker. Furthermore, the prognostic value of longitudinal DWI for prediction of overall survival in patients with pancreatic cancer was demonstrated, indicating that longitudinal DWI could potentially be useful for prediction of response.

Perspectives

In the future, DWI could become a tool to guide treatment interventions based on the patients' individual response to the treatmen, or dose escalation to radio-resistant regions within the tumour. This could potentially result in improved probability of tumour control and/or reduced toxicity for the patients.

Dansk resumé

Diffusions-vægtet MR billeddannelse som et værktøj til beslutningstagning i et in-room, MR-vejledt stråleterapiforløb.

Magnetisk resonans billeddannelse (MR) bliver i stigende grad brugt i stråleterapi (RT), da MR giver en god kontrast mellem forskelligt blødt væv i kroppen. En af de nyeste teknologiske fremskridt i RT er udviklingen af en hybrid MR lineær accelerator (MR-linac), en teknologi som muliggør daglig tilpasning af behandlingen til patientens anatomi. Desuden kan funktionelle MR-teknikker som diffusions-vægtet MR (DWI) give information om patientens biologi, og dermed potentielt bruges til at tilpasse behandlingen vejledt af biologisk information.

Det er muligt at udtrække kvantitative parametre fra DWI, som fx parameteren ”apparent diffusion coefficient” (ADC). Sådanne parametre kan bruges til at definere områder i en tumor som er særligt resistente over for RT, og de er mulige biomarkører for respons til RT. Af disse grunde kan DWI potentielt bruges til at vejlede beslutninger i RT-forløbet, og hermed bidrage til at individualisere behandlingen til den enkelte patient.

Før DWI kan bruges i klinisk sammenhæng er det nødvendigt at sikre konsistens i udtrækningen af biomarkører fra DWI i form af inter- and intra-observatør variation, repeterbarhed og reproducerbarhed (teknisk validering). Derudover skal der demonstreres en sammenhæng mellem biomarkører og kliniske endepunkter (klinisk validering). Denne afhandling adresserer nogle af udfordringerne forbundet med en teknisk og klinisk validering af DWI, med det overordnede formål at indføre DWI som et værktøj i RT-behandlingen.

Studie 1: Indtegning af relevante områder for ADC-beregning

ADC beregning kræver for det meste manuel indtegning af relevante områder (ROIs) hvilket både er tidskrævende og forbundet med usikkerhed. Studie 1 præsenterede et semi-automatisk indtegnings-værktøj, og testede det for patienter med endetarmskræft. Sammenlignet med manuel integning havde værktøjet en mindre intra-observatør ADC-variation. Derudover var værktøjet i stand til at udtrække tidslige forandringer i ADC som var større end usikkerheden forbundet med målingerne. Dette indikerer, at forandringerne er ”rigtige” forandringer, og ikke blot et resultat af måleusikkerhed. Dermed kan værktøjet potentielt gavne en teknisk validering af ADC.

Studie 2: Multi-center ADC reproducerbarhed

Varierende metoder til udtrækning af ADC giver en ringe multi-center reproducerbarhed (mellem hospitaler og forskningsinstitutioner) og vanskeliggør en teknisk validering af ADC. Studie 2 undersøgte ADC variationen mellem ni MR-linac-centre i relation til indtegning og beregning. MR scanninger blev delt mellem centrene, og hvert center udførte 1) indtegning af ROIs og 2) beregning af ADC-værdier inden for ROIs fra alle centre ved brug af lokale beregningsmetoder. En sammenhæng blev observeret mellem indtegnings-variationen og ADC-variationen, hvilket understreger vigtigheden af at reducere integnings-usikkerheden. Studiet viste også, at den beregnings-relaterede ADC usikkerhed var større end den indtegnings-relaterede ADC-usikkerhed, og primært skyldtes brug af forskellige b-værdier. Andre vigtige faktorer var forskellige software-implementeringen og hvorvidt ADC blev beregnet ud fra en voxel-vis metode eller ud fra middel-DWI-signalet indenfor den indtegnede ROI. Resultaterne fra dette studie kan forhåbentligt bruges, når fremtidige ADC-studier skal designes, og kan især blive nyttige for multi-center studier.

Studie 3: Forudsigelse af respons ved brug af longitudinal DWI

MR-linac'en har muliggjort daglige DWI scanninger (longitudinal DWI), men der er stadig brug for undersøgelser af sammenhængen mellem longitudinal DWI og kliniske endepunkter, da litteraturen er sparsom på området. Studie 3 undersøgte den prognostiske værdi af longitudinal DWI til at forudsige overlevelsen af patienter med kræft i bugspytkirtlen, ud fra både en model-baseret metode (ADC) og en model-fri dekompositions-metode. Den bedste model for overlevelse baseret på krydsvalidering indeholdt to parametre, begge udledt ved hjælp af dekompositions-metoden. De to parametre repræsenterede henholdsvis baseline-information og tidslig udvikling af DWI henover RT-forløbet, hvilket illustrerer den mulige fordel ved longitudinal DWI i forbindelse med forudsigelse af respons. Studie 3 bidrog til den kliniske validering af DWI ved at vise en sammenhæng mellem longitudinal DWI og et klinisk endepunkt.

Konklusioner

Denne afhandling har adresseret vigtige problemstillinger i forbindelse med en teknisk og klinisk validering af DWI. Studie 1 viste at semi-automatisk indtegning kan bidrage til konsistente ADC-målinger og potentielt spare tid sammenlignet med manuel indtegning. Studie 2 bidrog med anbefalinger for DWI-analyse med det formål at forbedre ADC-reproducerbarheden mellem centre. Begge studier kom med tiltag, som kan blive nyttige for den tekniske validering af ADC. Derudover demonstrerede studie 3 den prognostiske værdi af longitudinal DWI for forudsigelse af overlevelse for patienter med kræft i bugspytkirtlen, hvilket indikerer at longitudinal DWI kan blive et nyttigt redskab til forudsigelse af behandlingsrespons.

Perspektiver

I fremtiden kan DWI potentielt bruges som et redskab til at træffe beslutninger i løbet af RT-behandlingen baseret på patienternes individuelle behandlingsrespons, eller til at definere områder i tumoren, som er mindre følsomme over for RT og derfor kræver en øget dosis. Dette vil potentielt kunne forbedre resultatet af behandlingen og reducere mængden af bivirkninger for patienterne.

Acknowledgements

There are many people who I would like to thank for their help and support throughout this Ph.D. project. First of all, I would like to thank my main supervisor, Faisal Mahmood, for guidance and encouragement during my time as a Ph.D. student. I also owe a great thanks to my co-supervisor, Carsten Brink, who is capable of providing solutions for almost any problem, and who has offered support regarding data analysis and statistics. I would also like to thank my co-supervisor, Tine Schytte, for guidance and advices regarding the clinical aspects of my project.

I would like to thank all my colleagues, both physicists and fellow Ph.D. students, for sharing their knowledge with me through both formal and informal discussions. A special thanks to my Ph.D. colleagues in *barakken*, who have been a great moral support and good company during the long days at the office.

Many people have helped me collect data for my project, for which I owe them thanks. Especially, I would like to thank Maja Lynge Fransen and Frederik Severin Gråe Harbo for performing tumour delineation during one of my studies, and Lars Johnsen, who helped me facilitate the delineation using Pinnacle. Moreover, I would like to thank the staff at the Department of Oncology including doctors, radiographers, physicists, nurses and technicians for making it possible to acquire the necessary MRI scans for my project.

During my change of environment stay at the Netherlands Cancer Institute in Amsterdam, I got to know the research group of Imaging Technology in Radiotherapy lead by Uulke van der Heide. I would like to thank all the group members for welcoming me and for sharing their knowledge and experiences with me. A special thanks to Uulke van der Heide and Petra van Houdt who supervised me during my stay.

I would like to acknowledge the financing of this project by the Danish Cancer Society, the Danish Comprehensive Cancer Center and MANTRA, a frontline centre at Odense University Hospital.

Lastly, I would like to thank my friends and family for their support.

Preface

The thesis is based on three original articles, which are included as individual chapters:

I. *Robust extraction of biological information from diffusion-weighted magnetic resonance imaging during radiotherapy using semi-automatic delineation*

   Anne L.H. Bisgaard, Carsten Brink, Maja Lynge Fransen, Tine Schytte, Claus P. Behrens, Ivan Vogelius, Henrik Dahl Nissen, Faisal Mahmood.
   Physics and Imaging in Radiation Oncology. 2022;21:146-152.

II. *Recommendations for improved reproducibility of ADC derivation on behalf of the Elekta MRI-linac consortium image analysis working group*

   Anne L.H. Bisgaard, Rick Keesman, Astrid L.H.M.W. van Lier, Catherine Coolens, Petra J. van Houdt, Alison Tree, Andreas Wetscherek, Paul B. Romesser, Neelam Tyagi, Monica Lo Russo, Jonas Habrich, Danny Vesprini, Angus Z. Lau, Stella Mook, Peter Chung, Linda G.W. Kerkmeijer, Zeno A. R. Gouw, Ebbe L. Lorenzen, Uulke A. van der Heide, Tine Schytte, Carsten Brink, Faisal Mahmood
   Radiotherapy and Oncology. 2023;186:1-7 .

III. *Prediction of overall survival in patients with locally advanced pancreatic cancer using longitudinal diffusion-weighted MRI*

   Anne L.H. Bisgaard, Carsten Brink, Tine Schytte, Rana Bahij, Mathilde Weisz Ejlsmark, Uffe Bernchou, Anders S. Bertelsen, Per Pfeiffer, Faisal Mahmood.
   Manuscript ready for submission.

The text, figures and tables are identical to the original manuscripts, however the layout and numbering of figures and tables have been changed to match the thesis layout. Further, the numbering of the references have been changed according to the order in which they appear in the thesis. Supplementary materials from the articles can be found in the Appendix I-III.

Abbreviations and acronyms

2D	Two dimensional
3D	Three dimensional
ADC	Apparent diffusion coefficient
AI	Artificial intelligence
BTV	Biological target volume
CI	Confidence interval
CRUK	Cancer Research UK
CT	Computed tomography
CTV	Clinical target volume
CV	Coefficient of variation
Dice	Dice similarity coefficient
DNA	Deoxyribonucleic acid
DWI	Diffusion-weighted magnetic resonance imaging
EORTC	European Organisation for Research and Treatment of Cancer
EPI	Echo planar imaging
GTV	Gross tumour volume
LAPC	Locally advanced pancreatic cancer
LOA	Limits of agreement
MOMENTUM	Multi-outcome evaluation of radiation therapy using the MR-linac
MRI	Magnetic resonance imaging
MRI-linac	Hybrid MRI linear accelerator
MRL	Hybrid MRI linear accelerator
MSD	Mean surface distance
msNMF	Monotonous slope non-negative matrix factorization
NEMA	National electrical manufacturers association
NEX	Number of excitations
OAR	Organs at risk
PTV	Planning target volume
QIB	Quantitative imaging biomarker
QIBA	Quantitative imaging biomarker alliance
RF	Radiofrequency
ROI	Region of interest

RT	Radiotherapy
SADT	Semi-automatic delineation tool
SE	Spin echo
SNR	Signal to noise ratio
TE	Echo time
TR	Repetition time
T2W	T2-weighted
VTV	Viable tumour volume
wCV	Within-subject coefficient of variation
wSD	Within-subject standard deviation

List of Figures

List of Tables

List of Appendices

Table of contents

1 Introduction

1.1 Radiotherapy for cancer treatment

Cancer is one of the main causes of death worldwide, with 19.3 million new cancer cases and 10.0 million deaths occurring each year [1]. The prognosis and the treatment strategies vary depending on the type and stage of the cancer. Treatment can include surgery, chemotherapy, radiotherapy (RT), immunotherapy, hormonal therapy or combinations of these. RT is one of the most common treatments for cancer, and is received by approximately 50% of cancer patients [2]. RT can be used for several purposes, including definitive treatment, down-staging of the tumour before surgery (neo-adjuvant), elimination of remaining microscopic disease after surgery (adjuvant), and palliative treatment. It is a local treatment, meaning that it only affects the irradiated region, and not the entire body (in contrast to systemic treatments such as chemotherapy).

In RT, ionizing radiation (photons or particles) damages the DNA in irradiated cells, which can lead to cell death [3]. Cells have the ability to repair these damages, however, the repair rate is decreased in cancer cells. The total prescribed radiation dose is often delivered in small, daily fractions, which allow the normal cells to repair the damage on their DNA between the fractions. The slight difference in repair rates gives a higher probability of tumour control compared to side effects at a given radiation dose, which is referred to as the "therapeutic ratio" of RT [3].

The aim of RT is to deliver as high a radiation dose as possible to the tumour, while keeping the dose to the surrounding healthy tissue and organs at risk (OAR) as low as possible to minimize the risk of side effects (toxicity) [4] [5]. This requires both an accurate definition of the position of the tumour and the OARs, and a precise dose delivery. Technological advances such as intensity-modulated RT have widened the therapeutic ratio of RT treatment, since it allows shaping the dose distribution to the tumour while sparing OARs [2,6,7]. Moreover, the use of image-guided RT (IGRT) ensures a correct positioning of the patient using pre-treatment imaging and makes it possible to correct for organ motion before each delivered RT fraction [8].

The standard IGRT workflow includes a treatment planning phase and treatment delivery phase, as illustrated in Figure 1.1. In the planning phase, the patient is scanned in order to outline the tumour and the OARs. The standard imaging modality is computed tomography (CT), but often, it is supplemented by magnetic resonance imaging (MRI) or position emission tomography to improve visibility of the tumour. Regarding tumour delineation,

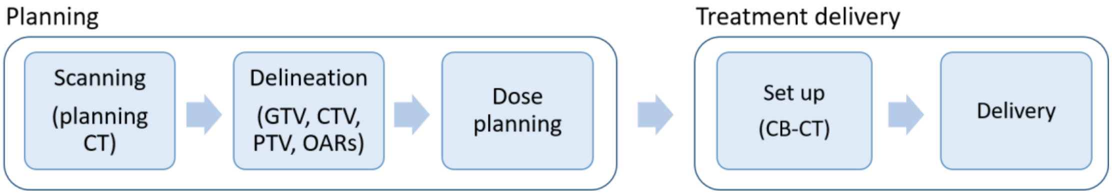


Figure 1.1. Overview of the standard image-guided radiotherapy workflow

there are three important volume definitions in RT: the gross tumour volume (GTV), the clinical target volume (CTV), and the planning target volume (PTV) [9]. The GTV contains the visible tumour, while the CTV comprises the GTV and a margin to account for microscopic disease spread not visible on scans. Another margin is added to the CTV to account for patient motion (including internal target motion) and uncertainties in the treatment planning and delivery, resulting in the PTV [9]. The dose plan is optimized such that the PTV receives the prescribed target dose, while dose constraints to OARs are met. Before each delivered fraction, a cone-beam CT scan is acquired and compared to the planning CT image, to make sure that the same patient positioning is used as during treatment planning.

The poor soft-tissue contrast of CT limits the ability to accurately define the tumour. Further, it makes it difficult to detect movement of OARs between treatment planning and delivery, which can lead to irradiation of normal tissue. MRI offers a much better soft tissue contrast compared to CT, and might thus aid detection of organ motion and reduction of treatment margins. On top of this, it enables so-called functional MRI, which holds great potential for personalization of RT treatment. These advantages have motivated a movement towards in-room MRI-guided RT.

1.2 MRI-guided RT on the MRI-linac

One of the most recent advances in RT is the introduction of the hybrid MRI linear accelerator (MRI-linac), which is an integrated MRI scanner and linear accelerator [10–12]. The MRI-linac technology enables MRI-guided RT, utilizing the superior soft tissue contrast of MRI compared to CT [13]. Due to the high visibility of the tumour and OARs on MRI scans, it is possible to adjust the treatment plan to match the daily anatomy of the patient, while the patient is on the treatment table. This can potentially reduce treatment margins, resulting in smaller irradiated volumes, and consequently less toxicity for the patient. Further, it may allow dose escalation to the tumour to increase the probability of tumour control while meeting OAR constraints [13,14]. Another advantage is the ability to manage intra-fractional motion during beam delivery [13]. The soft tissue contrast makes

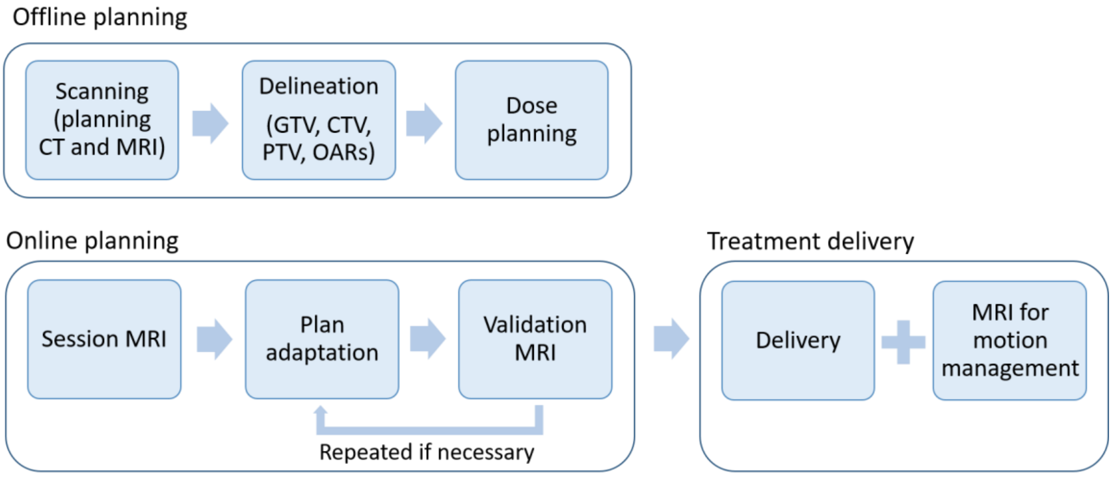


Figure 1.2. Overview of the MRI-guided adaptive RT workflow on the MRI-linac. Inspired by Bertelsen et al. [12].

MRI-guided RT especially advantageous for tumours in soft tissue regions, such as the pelvic and abdominal regions [15–17]. These regions are also prone to motion, due to bladder filling, intestinal movement or gas, and therefore, MRI-guided motion management can be an advantage.

The MRI-guided adaptive RT workflow consists of both offline and online treatment planning, prior to delivery of the dose, as illustrated in Figure 1.2 [11,12]. The main difference from the standard IGRT workflow is the online plan adaptation. This process includes re-delineation of the tumour and OARs on the session MRI to match the daily anatomy of the patient, followed by re-calculation of the dose plan. A validation MRI scan is acquired to validate the adapted plan, e.g. to check if the patient has moved during the online adaptation process. If the adapted plan is satisfactory, the treatment is delivered. During treatment delivery, the MRI scanner can be used for real-time motion monitoring.

The MRI-linac technology is relatively new and has a lot of unrealized potential. The number of patients treated worldwide is still relatively small, and consequently, collaboration between MRI-linac centres is needed to speed up the implementation and technical development of MRI-linac-based RT. One example of such a collaboration is the MOMENTUM study (multi-outcome evaluation of radiation therapy using the MRI-linac)[18], an international partnership which has been formed between the vendor of one of the commercially available MRI-linac systems, Elekta (Elekta AB, Stockholm, Sweden), and several hospitals. Within the MOMENTUM study, technical and clinical patient data is collected from participating hospitals worldwide, to support multi-centre research.

1.3 Biologically guided RT

Currently, advances in MRI-guided RT mainly focus on adaptation of the RT treatment to the patient anatomy based on high-quality anatomical MRI images. However, in addition to anatomical images, MRI also provides information about the tumour biology through functional MRI techniques, which opens new pathways for treatment personalization based on tumour biology [19].

Information on tumour biology is of great relevance for RT as the sensitivity of the tissue to RT may relate to microscopic processes such as hypoxia, perfusion and diffusivity [20]. By escalating the dose to radio-resistant tumour sub-regions identified based on biological information, the probability of tumour control might be improved. This has led to the idea of a biological target volume (BTV) described by Ling et al. in 2000 [21] (Figure 1.3). The use of a heterogeneous dose distribution within the GTV to match the tumour heterogeneity is also referred to as dose painting [22,23].

Another possibility is to use functional MRI to monitor or predict treatment response during the course of RT. RT-induced changes in the tissue microstructure occur at an early stage during the RT course, before becoming visible on anatomical images [24]. The ability to assess the effect of RT in both the tumour and the healthy tissue during the course of RT would potentially allow treatment interventions in order to improve the tumour control and reduce the toxicity [20].

Dose painting and treatment interventions based on treatment response are examples of biologically guided RT. Here, biologically guided RT is defined as the use of biological images to guide treatment decisions before or during the RT course including adaptation of the dose distribution to match the tumour heterogeneity. Both dose painting and response monitoring and prediction have become feasible with the daily acquisition of

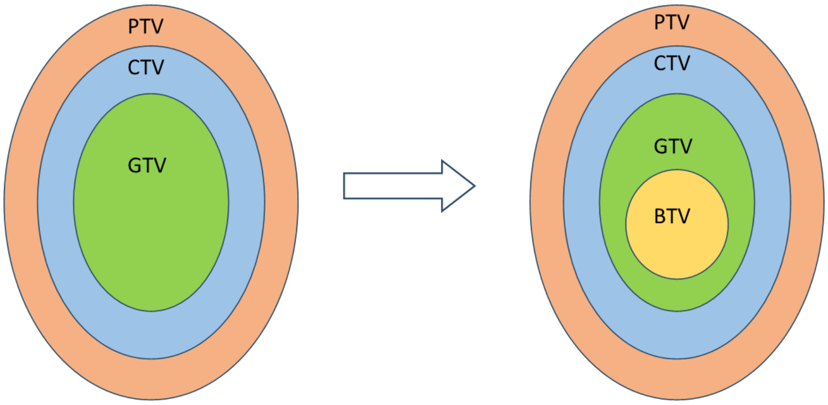


Figure 1.3. Schematic drawing illustrating the biological target volume (BTV). Inspired by Beaton et al. [25].

functional MRI on the MRI-linac. The “dream” scenario would be to integrate these concepts into the MRI-guided adaptive RT workflow, to allow biologically guided RT [20,22,26].

1.4 Diffusion-weighted MRI

Based on the current evidence, the most promising functional MRI technique for biologically guided RT is considered to be diffusion-weighted MRI (DWI) [26]. DWI allows quantitative assessment of the tumour biology in a non-invasive way, and can be acquired on the MRI-linac [27]. Further, it has been well-researched and has shown potential to be used for both dose painting and response prediction [20,26,28].

The DWI signal is sensitive to random motion of water molecules in the body tissue (i.e. diffusion). The motion of water molecules is impeded by obstacles such as cell membranes, and thus, it indirectly provides information of the tissue microstructure. For example, a low cell density will allow the water molecules to move more freely compared to a high cell density, resulting in a change of in diffusivity, as illustrated in Figure 1.4 [29].

DWI allows extraction of quantitative information about the water motion through the application of mathematical models or decomposition-based analysis methods [30,31]. The most commonly derived parameter is the “apparent diffusion coefficient” (ADC) [32,33]. The ADC indirectly provides information about the tissue microstructure such as the cell density [34], and moreover, changes in the ADC have been observed during RT, potentially reflecting RT-induced microstructural changes in the tissue [35]. For these reasons, the ADC is considered a promising quantitative imaging biomarker (see section 1.6) for biologically guided RT [26].

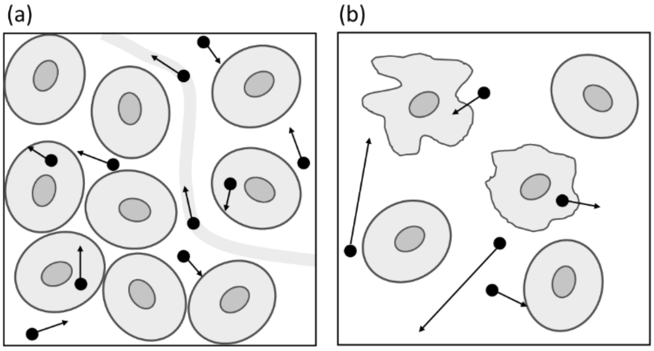


Figure 1.4. Diffusion of water molecules within and between cells in a region of high cell density (a), and a region of low cell density (b). Cells with irregular shapes represent cells with damaged cell membranes. The small, black dots represent water molecules, and the arrows represent their motion, which is hindered by cell membranes. Inspired by: Koh et al. [29]

1.5 The DWI signal and quantitative analysis

1.5.1 The MRI signal

The MRI signal originates from the hydrogen atoms (present in water and fat molecules in the body), or more specifically, from the hydrogen nuclei, i.e. protons. The protons have a "spin" and therefore a magnetic moment, which makes them behave as small magnets. Due to the MRI scanner's strong, homogeneous magnetic field (the B0-field), the magnetic moments of the protons are partly aligned inside the scanner forming a net magnetization, and furthermore, they precess, i.e. rotate at a frequency that is proportional to the magnetic field strength. By applying radiofrequency (RF) pulses at a frequency matching the precession frequency, it is possible to "push" the net magnetization to point in a direction perpendicular to the B0-field (excitation). After excitation, the net magnetization returns to its original state (relaxation). During the relaxation process, the precession causes temporal changes in the magnetic field, which induce a current in the scanner's receiver coils. This signal is used to form the MR images. The spatial information needed to reconstruct MR images is obtained by the use of magnetic field gradients (i.e. spatial variations in the magnetic field) within the scanner, due to which the protons' precession frequency becomes dependent on their position. A detailed description of the principles behind MRI is provided by Nishimura [36].

1.5.2 The DWI signal

The DWI signal reflects the random motion of water molecules in the imaged tissue, also known as Brownian motion. The macroscopic consequence of Brownian motion is the spread of a substance from areas with high concentration of the substance to areas of low concentration, known as diffusion. The 'degree' of diffusion can be described by the diffusion coefficient, D (a proportionality constant between the flux of the substance and the concentration gradient) [37], which depends on the size of the substance particles, the temperature and the surrounding media.

In DWI, the self-diffusion of water within the tissue is measured, as described in detail by Le Bihan et al. [38]. The self-diffusion coefficient of water is approximately $3 \cdot 10^{-3}$ mm^2/s at body temperature. In body tissue, the average displacement of water molecules within a given time interval is typically shorter compared to free diffusion, as the motion is impeded by barriers such as cell membranes [29,39]. More precisely, the motion can be "hindered" or "restricted" as described in detail by White et at. [40]. Thus, the (hindered or restricted) motion of water molecules can be used to probe the microstructure of the tissue.

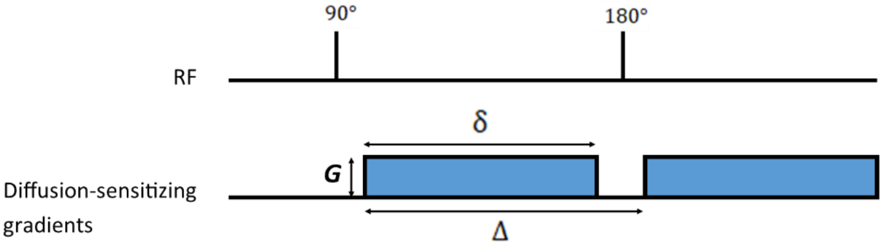


Figure 1.5. Stejskal-Tanner sequence for acquisition of diffusion-weighted MRI (DWI). The figure shows the application of radio-frequency (RF) pulses and diffusion-sensitizing magnetic field gradients to obtain a diffusion-weighted signal. Inspired by Koh et al. [29].

DWI is typically acquired using a Stejskal-Tanner DWI sequence [33]. Here, the signal is sensitized to diffusion using two so-called diffusion sensitizing gradients, as illustrated in Figure 1.5. The first gradient causes a dephasing of the magnetic moments of the protons, which results in a signal loss. The second gradient causes re-phasing of the magnetic moments, which leads to recovery of the signal. Protons that move away from their original position in the time between the two gradients will not be fully re-phased, and therefore, the signal will not be fully recovered. Thus, the motion of the water molecules is linked to signal loss. For this reason, dark regions in DWI scans represent a high degree of diffusion, and bright regions represent a low degree of diffusion.

1.5.3 The ADC model

The sensitivity of the DWI signal to diffusion depends on the strength (G) and duration (δ) of the diffusion-sensitizing gradients, as well as the time between them (Δ) [33]. These factors are included in one factor, referred to as the "b-value":

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right)$$

Here, γ is the gyromagnetic ratio describing the precession frequency of protons at a given field strength. The b-value can be altered on a clinical MRI scanner, to obtain images of different diffusion-weighting by using different gradient strengths, duration and times. With increasing b-values, the DWI signal becomes more sensitive to diffusion, and as a result, the DWI signal decays as a function of b-values (Figure 1.6). By acquiring a series of images with varying b-values, it is possible to use mathematical models to derive quantitative information from the DWI signal.

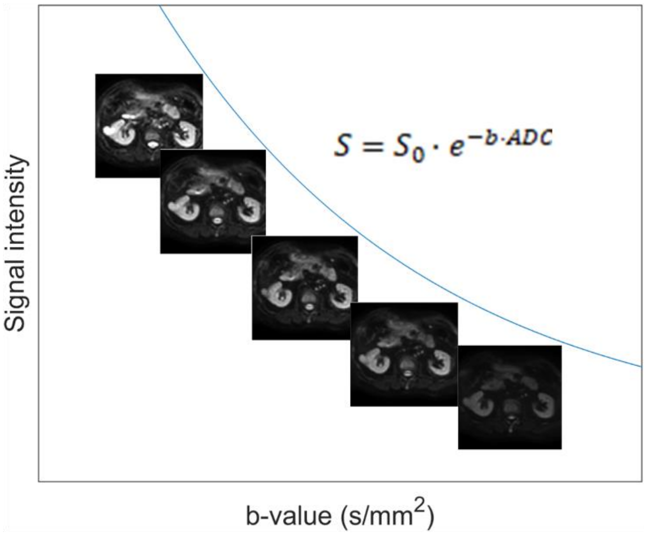


Figure 1.6. Schematic illustration of the DWI signal as a function of b-values. A DWI scan of a patient with locally advanced pancreatic cancer is used as an example.

The most commonly used model is the mon-exponential model, which is used for derivation of the ADC [32]:

$$S = S_0 \cdot e^{-b \cdot ADC}$$

Here, S_0 represents the DWI signal at b=0 s/mm^2. The ADC differs from the diffusion coefficient D, as the ADC is a statistical parameter describing the average motion of all water molecules in a voxel, including those experiencing hindered or restricted motion due to boundaries. Thus, it describes the "apparent diffusion" within a voxel. The ADC can be calculated for each image voxel, resulting in so-called "ADC maps", which may be useful for biological guided RT.

Other mathematical models for DWI analysis include the intra-voxel incoherent motion (IVIM) model [41], and the kurtosis model [42,43]. The IVIM model provides information of both diffusion and perfusion wheras the kurtosis model might provide complementary information about the microstructural organization of the tissue by taking into account the non-gaussian diffusion behaviour resulting from the restricted water motion. However, parameters derived using these models have been less investigated compared to the ADC [30].

1.6 Biomarkers and translational gaps

A quantitative imaging biomarker (QIB) is defined by the quantitative imaging biomarker alliance (QIBA) as "an objective characteristic derived from an in vivo image measured on a ratio or interval scale as an indicator of normal biological processes, pathogenic

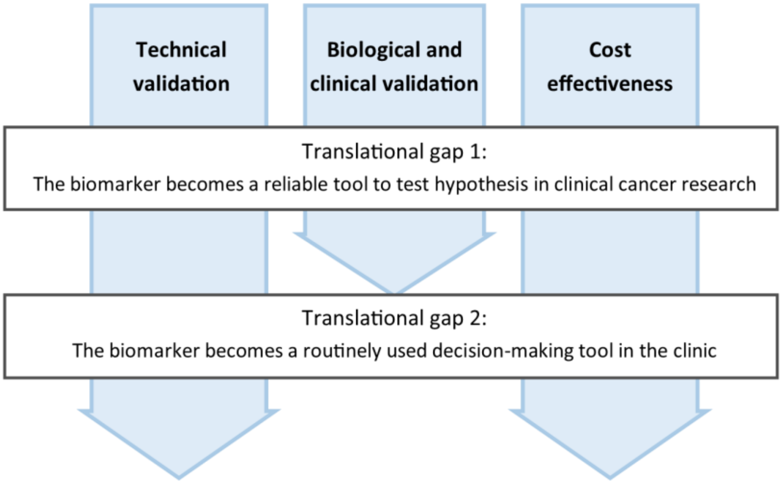

Technical validation
Biological and clinical validation
Cost effectiveness
Translational gap 1:
The biomarker becomes a reliable tool to test hypothesis in clinical cancer research
Translational gap 2:
The biomarker becomes a routinely used decision-making tool in the clinic

Figure 1.7. Overview of the imaging biomarker roadmap for validation of imaging biomarkers. Inspired by: O'Connor et al. [44].

processes, or a response to a therapeutic intervention" [45]. QIBs have many applications in oncology, such as detection and characterization of tumours, delineation of target volumes, and assessment of response to treatment [20]. Biomarkers can be prognostic or predictive, two concepts that are often confused [46]. A biomarker is prognostic if there is a relation between the biomarker and the prognosis of the patients regardless of treatment. A predictive biomarker can be used to predict response to a specific treatment, i.e. it can be used to predict which patients will benefit from an experimental treatment compared to a standard treatment.

All biomarkers must be validated before they can be used to guide decisions in a clinical setting. In 2017, a consensus group assembled by Cancer Research UK (CRUK) and the European Organisation for Research and Treatment of Cancer (EORTC) formulated an imaging biomarker "roadmap" of the validation process [44]. According to this roadmap, imaging biomarkers must cross two so-called "translational gaps" (Figure 1.7). By crossing the first gap, an imaging biomarker becomes a reliable tool to test hypothesis in clinical cancer research, and by crossing the second gap, it becomes a decision making tool in the clinic, meaning that it is used to guide the treatment of the patients.

The roadmap proposes a validation process which runs in three parallel tracks, representing a technical validation, biological and clinical validation and cost effectiveness (Figure 1.7). The technical validation ensures that the biomarker can be measured with sufficient accuracy, repeatability and reproducibility. Accuracy refers to the "true" value of the biomarker, and can be studied by scanning objects (so-called phantoms) with known values of the biomarker. Repeatability refers to the agreement between repeated measurements under identical conditions, while reproducibility refers to the agreement between measurements performed under different conditions, e.g. different observers or different equipment.

The biological and clinical validation is needed to establish a relationship between the biomarker and the underlying biology and clinical outcomes. Importantly, the biomarker must demonstrate value in improving outcomes for the patients, which is referred to as clinical utility. The technical and the biological/clinical validation goes hand in hand, as each of them would be meaningless without the other. The validation process can be resource demanding, and therefore, the cost effectiveness of the biomarker must be taken into account [44].

Initial technical and biological/clinical validation might be performed as single-centre studies, however, large, multi-centre validation studies are needed in order to achieve more statistical power, and to establish multi-centre reproducibility and correlation to treatment outcome. To demonstrate the clinical utility of the biomarker, i.e. proof that the biomarker can lead to an improvement of outcomes for the patients, large, randomized trials are needed with treatment interventions guided by the biomarker.

1.7 Challenges regarding translation of DWI into clinical use

DWI is currently the most promising functional MRI technique for biologically guided RT [26]. However, despite a substantial amount of research, DWI is still not used for this purpose in the clinic. In fact, it has not yet crossed the first translational gap in the imaging biomarker roadmap described in section 1.6. The first steps have been taken with respect to the biological and clinical validation. For example, ADC has been shown to correlate with cell density [34], and ADC values measured before treatment as well as changes in the ADC during the treatment course have been related to treatment response [26,35]. Furthermore, IVIM parameters have shown potential for response prediction [30,47]. However, most studies so far have been small and single-centre (i.e. performed within one single hospital or research institution), and have suffered from a large variation in methods, which makes it difficult to compare results. Hence, there is a need for validation of the findings in larger patient cohorts. With respect to the technical validation, DWI suffer from a poor multi-centre reproducibility due to the large variation of methods, and thus, a standardization of methods is needed. Further, there is a need for investigations of repeatability of DWI parameters using repeated scans (so-called test-retest scans), as the literature on this topic is limited [48].

Since the hardware of the MRI-linac differs from that of a diagnostic MRI scanner, it is necessary to perform a separate technical validation of DWI on the MRI-linac. In fact, a technical validation must be performed on any MRI system including MRI-linacs, since the scanner performance might differ between scanners, as recommended by QIBA [49]. So far, feasibility and accuracy of ADC has been demonstrated using phantoms [27,50], and recommendations for acquisition of DWI for ADC measurements on the MRI-linac has been published [51]. Investigations of the repeatability and reproducibility of DWI parameters measured on the MRI-linac, as well as the use of longitudinal DWI measurements for response prediction are wanted [26,48].

1.8 Outline of this thesis

The aim of the thesis is to investigate methods for a technical and clinical validation of DWI in RT. The long term goal is to use DWI for biologically guided RT, e.g. by defining tumour sub-regions for dose escalation, or by predicting response in order to guide potential treatment interventions.

During the Ph.D. project, three sub studies were conducted, of which the first two addressed the technical validation of DWI and the third addressed the clinical validation of DWI. Specifically, the sub studies aimed at:

1. developing and testing a semi-automatic delineation tool for robust ADC measurements
2. giving recommendations for improved ADC reproducibility
3. investigating the capacity of longitudinal DWI to predict outcome in patients with locally advanced pancreatic cancer

The thesis is structured in three parts. Part I addresses the technical validation of ADC, and includes the first two sub studies. Part II addresses the clinical validation of DWI parameters, and includes the third sub study. Both part I and part II include relevant background information about the topics. Part III includes a discussion of the findings and future perspectives.

PART I

Technical validation

2 Technical validation

2.1 Repeatability and reproducibility

A technical validation of a biomarker implies among other things a determination of the biomarker repeatability and reproducibility. The repeatability describes the agreement of repeated measurements performed in the same patient using the same equipment over a short time period. It is usually reported using the repeatability coefficient (RC), defined as the smallest significant difference between repeated measurements performed under identical conditions (so-called test-retest measurements) [49]. It can be calculated from the within-subject standard deviation (wSD) using a 95% confidence interval (CI), as recommended by QIBA [49]:

$$RC = 1.96 \cdot \sqrt{2} \cdot \sqrt{wSD^2} = 2.77 \cdot wSD$$

Here, wSD^2 is the within-subject variance which can be calculated from test-retest measurements (x_{i1} and x_{i2}) performed for N subjects [52]:

$$wSD^2 = \frac{1}{2N}\sum_{i=1}^{N}(x_{i1} - x_{i2})^2$$

Alternatively, if the measurement agreement varies with the magnitude of the measurement, the percentage wise RC can be calculated using the within-subject coefficient of variance (wCV) [49]:

$$\%RC = 1.96 \cdot \sqrt{2} \cdot \sqrt{wCV^2} = 2.77 \cdot wCV$$

Here, wCV^2 is calculated by [52]:

$$wCV^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(x_{i1} - x_{i2})^2}{2\,\bar{x}_i^2}\right)$$

Where $\bar{x}_i$ denotes the mean of the replicate measurements for the i'th subject.

It is important to know the RC of a biomarker in order to detect patient-related differences or changes over time. When following a patient using longitudinal DWI, the RC must be known in order to determine whether DWI changes reflect biological changes or merely measurement uncertainty. For example, if the %RC of ADC is 20%, a change in ADC of more than 20% between two time points can be considered a true change.

Reproducibility is somewhat similar to repeatability, except that there are differences in the measurement procedures between the “repeated” measurements. Regarding DWI, the measurements could for example be obtained using different MRI scanners or by different observers (see section 2.2). The reproducibility coefficient is expected to be larger compared to the repeatability coefficient [49]. Like the repeatability coefficient, the reproducibility coefficient represents the smallest significant difference between measurements. For example, the reproducibility of DWI must be known in order to compare DWI studies performed at different centres, where the measurements may have been performed using different MRI scanner types and DWI sequences.

2.2 Challenges regarding repeatability and reproducibility of ADC

ADC measurements require acquisition of DWI with at least two, well-separated b-values, delineation of a region of interest (ROI) in which the ADC should be measured, and ADC calculation using either commercial or in-house software. Several factors related to both image acquisition, delineation and ADC calculation impact the repeatability and reproducibility of ADC measurements [35].

First of all, the repeatability reflects the image-related uncertainty, which is related to both image noise, patient movement and organ deformation between and during DWI acquisitions. Motion may cause misalignment of voxels between b-value images or between DWI acquisitions, and thus impact ADC measurements. Moreover, geometric distortions in DWI images can contribute to the image-related uncertainty, due to deformation of the images as well as false high-intensity or low-intensity regions [53].

Another important source of ADC uncertainty is the delineation of a ROI. Delineation is normally performed manually, and gives rise to inter- and intra-observer ADC variability [54] [55]. Hence, the delineation process impacts both repeatability and reproducibility. Furthermore, the choice of ROI strategy has been shown to impact the ADC measurements, and is therefore relevant in relation to ADC reproducibility [56,57]. A ROI may be delineated on T1- or T2-weighted MRI which have a high level of anatomical details, or it can be delineated directly on DWI images [35]. In many investigations of ADC, delineation has been performed on high b-value DWI images, as they have a high tumour-to-background ratio [35]. This type of ROI is generally considered to represent the region with the highest cell density within the tumour [34], and some studies have described it as the "viable tumour volume" (VTV), as necrotic and cystic regions are excluded [35,58]. The VTV differs from the GTV, which normally includes necrotic and cystic regions. From a theoretical point of view, since the VTV represents the region with the highest cell density, the mean ADC is expected to be lower within the VTV compared to the GTV. Thus, the use of different delineation strategies may give rise to ADC variation between studies or centres. Currently, there is no standard way of performing delineation for ADC measurements, and the optimal ROI strategy for RT purposes (e.g. response prediction) remains to be determined.

DWI acquisition protocols vary between centres, which impacts the multi-centre reproducibility of ADC. For example, the ADC is sensitive to the choice of b-values, since inclusion of low b-values might lead to an overestimation of ADC due to perfusion effects [41]. Further, inclusion of high b-values might lead to an underestimation of ADC both due to the kurtosis effect, i.e. non-gaussian diffusion behavior [42], and the so-called noise-floor present in DWI due to the rician noise distribution in magnitude MR images [59]. The noise floor might dominate the DWI signal at high b-values since the signal-to-noise-ratio (SNR) decreases with increasing b-values. Besides the choice of b-values, parameters such as the voxel size and the diffusion time also affect ADC measurements [53,60].

Lastly, the ADC calculation process itself may lead to ADC variation. As described above, the choice of b-values plays an important role. The b-values used for calculation may be a subset of those acquired on the scanner, e.g. one could exclude low b-values to avoid the impact of perfusion. One important point is that the highest acquired b-value determines other acquisition parameters (e.g. Δ, see Figure 1.5) in a DWI sequence which may impact the ADC [51]. For this reason, different DWI sequences may cause ADC variation even if the same b-values are used for ADC calculation. For example, an ADC value calculated using the b-values 0 and 500 s/mm^2 may differ between two DWI scans acquired using b-values of [0, 500, 1000] s/mm^2 and [0, 500] s/mm^2, respectively.
In addition to the choice of b-values, different calculation approaches may cause ADC variation. Typically, calculation is performed in a voxel-wise manner, i.e. by calculating the ADC within each voxel and subsequently calculating the mean or median within a ROI. Another approach is ROI-based calculation, i.e. calculation based on the mean or median DWI signal within a ROI. Different centres have different softwares available for ADC calculation, which may differ with respect to both ROI-/voxel-based calculation, fitting methods (e.g. least squares fitting) and filtering (e.g. removal of low SNR voxels). The literature investigating the impact of software choice on ADC measurements is limited [61,62].

2.3 Contribution from study 1 and study 2

Study 1 presents a semi-automatic delineation tool (SADT) for ADC calculation as an alternative to manual delineation. The tool represents one possible approach for automated delineation of the VTV, using a simple-to-implement strategy. The tool's performance is tested in patients with rectum cancer, and it is compared to manual delineation by a radiologist. Potential benefits of semi-automatic delineation are improved inter-and intra-observer agreement, as well as time saved for the medical doctors performing delineations.

Study 2 provides recommendations for improved ADC reproducibility, based on an evaluation of delineation- and calculation-related ADC variation between MRI-linac centres. Potentially, these recommendations can become useful when designing future investigations of ADC, especially in multi-centre studies.

3 Paper I

Robust extraction of biological information from diffusion-weighted magnetic resonance imaging during radiotherapy using semi-automatic delineation

Anne Louise Højmark Bisgaard[a,b*], Carsten Brink[a,b], Maja Lynge Fransen[c], Tine Schytte[b,d], Claus P. Behrens[e], Ivan Vogelius[f,g], Henrik Dahl Nissen[h**], Faisal Mahmood[a,b**]

[a]Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense C, Denmark;
[b]Department of Clinical Research, University of Southern Denmark, Odense C, Denmark;
[c]Department of Radiology, Odense University Hospital, Odense C, Denmark;
[d]Department of Oncology, Odense University Hospital, Odense C, Denmark;
[e]Department of Oncology, Herlev and Gentofte Hospital, Herlev, Denmark;
[f]Department of Clinical Oncology, Rigshospitalet, Copenhagen, Denmark;
[g]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark;
[h]Department of Oncology, Vejle Hospital, Vejle, Denmark

***Corresponding Author**

****Co-last-authors**

Published in: Physics and Imaging in Radiation Oncology, 21 (2022), 146-152.

3.1 Abstract

Background and Purpose

Diffusion-Weighted Magnetic Resonance imaging (DWI) quantifies water mobility through the Apparent Diffusion Coefficient (ADC), a promising radiotherapy response biomarker. ADC measurements depend on manual delineation of a region of interest, a time-consuming and observer-dependent process. Here, the aim was to introduce and test the performance of a new, semi-automatic delineation tool (SADT) for ADC calculation within the viable region of the tumour.

Materials and Methods

Thirty patients with rectal cancer were scanned with DWI before radiotherapy (RT) (baseline) and two weeks into RT (week 2). The SADT was based on intensities in b=1100 s mm^{-2} DWI and derived ADC maps. ADC values measured using the SADT and manual delineations were compared using Bland-Altman- and correlation analyses. Delineations were repeated to assess intra-observer variation, and repeatability was estimated using repeated DWI scans.

Results

ADC measured using the SADT and manual delineation showed strong and moderate correlation at baseline and week 2, respectively, with the SADT measuring systematically smaller values. Intra-observer ADC variation was slightly smaller for the SADT compared to manual delineation both at baseline, [-0.00; 0.03] vs. [-0.02; 0.04] 10^{-3} mm^{2} s^{-1}, and week 2, [-0.01; 0.00] vs. [-0.04; 0.07] 10^{-3}mm^{2} s^{-1} (68.3% limits of agreement). The ADC change between baseline and week 2 was larger than the ADC uncertainty ($\pm 0.04 \cdot 10^{-3}$mm^{2} s^{-1}) in all cases except one.

Conclusion

The presented SADT showed performance comparable to manual expert delineation, and with sufficient consistency to allow extraction of potential biological information from the viable tumour.

Keywords: Diffusion-weighted MRI, apparent diffusion coefficient, automatic delineation, MRI guided radiotherapy, imaging biomarker

3.2 Introduction

Magnetic Resonance imaging (MRI) is used in radiotherapy (RT) planning as a supplement to computed tomography (CT), primarily because it provides better soft-tissue contrast. Furthermore, advanced MRI techniques can provide information on tumour biology. For example, Diffusion-Weighted MRI (DWI) probes the micro-environment of the tissue by measuring local water mobility [33]. Low water mobility correlates with high cell density, which is often indicative of high tumour viability [29].

DWI is acquired as a set of images with different diffusion weighting, defined through so-called b-values. With higher b-values, the acquisition becomes more sensitive to random motion of water molecules, inflicting large signal loss in regions with high diffusivity. Vice versa, in regions with reduced diffusivity, e.g. due to high cell density, signal loss is small. The water mobility can be quantified if at least two b-values are acquired, as the "Apparent Diffusion Coefficient" (ADC) [29].

ADC is a promising biomarker for treatment response both when derived from pre-treatment imaging and during the course of RT [63–65]. ADC has also shown potential clinical value in treatment planning and adaptation [66]. However, ADC has yet to be translated into widespread clinical use in RT planning and response evaluation. Lacking randomized trials and lack of consistency in ADC measurement are some of the challenges that need to be addressed before such translation can happen [20,44].

Recent introduction of the hybrid MRI linear accelerator (MR-linac) has allowed MRI at every treatment fraction, making frequent ADC measurements more accessible [10,11,67]. This gives a unique opportunity to collect large data sets to investigate ADC for clinical use. However, to perform large multi-center studies, consistency in ADC calculation and workflow feasibility is important [20].

An important aspect of ADC calculation is the delineation of a region of interest (ROI), which, if done manually, is time-consuming, requires a high level of expertise, and suffers from intra- and inter-observer variation [54,68]. A potential solution is automated delineation, which is already widely employed in medical images, including standard MRI [69,70], but not adequately developed for advanced imaging such as DWI. Simple, threshold-based delineation has been tested on DWI, exploiting the large tumour-to-background ratio on raw, high b-value DWI images. Despite promising results, manual inspection was required to avoid non-tumour regions [71]. Fully automated delineation may be achieved using more advanced methods based on artificial intelligence (AI), e.g.

convolutional neural networks [72], however, this approach requires training data and suffers from a lack of transparency.

In this study, we present a simple, semi-automatic tool for delineation of the viable tumour volume (VTV), a recommended ROI for ADC measurement, which is not directly identifiable from raw, high b-value DWI images [35,56,58]. Instead, the VTV was in this study defined using combined information from both raw, high b-value DWI images and derived ADC maps [35]. The study aim was to test the performance of the tool in terms of robustness of ADC measurements, compared to manual delineation. Also, the capacity of the tool to measure temporal changes in ADC was tested using longitudinal DWI data.

3.3 Materials and methods

3.3.1 Patients

This prospective study included thirty patients with biopsy-proven locally advanced adenocarcinoma of the rectum. All patients received long-course chemoradiotherapy and were treated with a daily fraction five times a week. Tumour was prescribed a dose of 60 Gy in 30 fractions, and the elective lymph node volumes were prescribed 50 Gy in 30 fractions using a concomitant boost intensity-modulated radiation therapy (IMRT) technique. All patients received a 5 Gy brachytherapy boost. The Regional Committee on Health Research Ethics for Southern Denmark has approved the study (study ID S-20110021 and S-20130030), and informed consent was obtained from all patients.

3.3.2 MRI protocol

Patients were MRI scanned before RT (baseline) and two weeks into RT (week 2) with a 1.5 T clinical MRI scanner (Philips Ingenia, Philips Healthcare, Best, The Netherlands). The imaging protocol consisted of T2-weighted imaging (T2W) and DWI. T2W was acquired using a turbo spin-echo sequence (repetition time (TR)/echo time (TE): 7161/100 ms) with an in-plane resolution of (0.8×0.8) mm^2, a slice thickness of 2.4 mm, and a slice gap of 1.0 mm. Scan duration was 5 minutes and 43 seconds. DWI was implemented as a single-shot spin-echo echo-planar imaging sequence (TR/TE: 2860/82 ms) with fat suppression (spectral presaturation with inversion recovery) and with b-values ranging from 0 to 1100 s mm^{-2} (0 (2), 7 (2), 20 (2), 40 (2), 90 (2), 170 (2), 300 (2), 500 (4), 700 (4), 900 (4) and 1100 (6) s mm^{-2}); with the number of image averages for each b-value given in parenthesis. In this study, only b=0, b=170 and b=1100 s mm^{-2} were used. In-plane resolution was (1.82×1.82) mm^2, slice thickness 4.6 mm and slice gap 0.4 mm. Sequence duration was 4 minutes and 23 seconds. In each imaging session, DWI was performed twice in succession (test-retest), while the patient remained in the same position to assess repeatability.

3.3.3 Included data

Thirty patients were MRI scanned at baseline, and out of these patients, twenty-nine were MRI scanned at week 2, resulting in a total of 59 scan sets of images (each set including T2W and DWI for both test and retest). Of these 59 sets, five sets of DWI images were excluded either due to artefacts in DWI (2), severe bulk motion (1), or the tumour being partly outside the field of view (2). In total, full scan data from 27 image sessions at baseline and 27 image sessions at week 2 were used to evaluate repeatability and intra-observer variation at baseline and at week 2. Of these, 25 patients had all scans available at both time-points and were used to assess ADC change between baseline and week 2.

3.3.4 ADC calculation

A set of b-values (170 and 1100 s mm^{-2}) were selected for ADC calculation; the high b-value was selected to obtain high diffusion sensitivity and the non-zero, low b-value was selected to avoid perfusion effects [41,51]. ADC-maps were calculated voxel-wise by applying linear regression to the logarithm of the signal intensity S to get $\ln(S_{high})=\ln(S_{low})-[b_{high}-b_{low}] \cdot ADC$.

3.3.5 Semi-automatic delineation tool

A semi-automatic delineation tool (henceforward named 'SADT') was developed with the purpose of delineating VTVs for ADC calculation. VTV was defined as viable tumour, excluding necrotic regions [35,58]. The SADT was implemented using in-house developed software (Matlab R2019a, Mathworks ab, Sweden), following a 3-step process (Figure A1.1 in Appendix I)[1].

In step 1, a rough delineation of the relevant 3D region was given as manual input (manual mask). The manual mask indicates the 'relevant' region, and the algorithm described below is, except for the very last step, only performed on voxels within this region.

In step 2, 3D thresholding in b=1100 s mm^{-2} DWI and ADC-maps was used to define two masks that fulfilled criteria of high DWI intensity and low ADC values, respectively. In the b=1100 s mm^{-2} DWI image, a binary DWI mask for the bright voxels was created using a threshold value obtained from the Otsu algorithm [73]. Furthermore, an ADC-mask indicating the low ADC values was created from the ADC map using a threshold value equal to the median ADC value plus 0.5 times the standard deviation of the measured ADC values

[1] The code is available upon request to the corresponding author.

within the manual mask. Half the standard deviation was used to obtain a stable delineation.

In step 3, the overlap between the DWI mask and the ADC mask was created and used as the raw VTV, fulfilling both criteria (high DWI intensity and low ADC values).

The final VTV was obtained by a 2-step post-processing of the raw VTV: First, only the largest volume was retained if non-connected sub-regions were included in the raw VTV. Second, to account for the possibility that the manual mask (Step 1) might have excluded target voxels, the VTV was allowed to expand beyond the boundaries of the manual mask iteratively while respecting threshold criteria (Step 2) and a criterion of connectivity.

3.3.6 Delineation of viable tumour volumes

A radiologist (author MLF) performed manual VTV delineation on b=1100 s mm^{-2} DWI images guided by T2W and ADC-maps on test-retest data for all included patients. The same radiologist re-contoured the images after two months for intra-observer variation assessment. T2W and DWI were rigidly registered prior to delineation using clinical software (MIM, MIM Software Inc., Cleveland, Ohio).

A non-radiologist (author ALHB, physicist, experienced with DWI) used the SADT to delineate the same cases as the radiologist including test-retest data. The manual mask used for input to the SADT included rectum, mesorectum, and some surrounding areas respecting anatomical boundaries to other anatomical structures, e.g. prostate, and was restricted to the same slices as included by the radiologist. Delineation was performed on b=0 s mm^{-2} DWI images using b=1100 s mm^{-2} DWI images for guidance. The same observer created the manual mask input for the SADT twice with a time interval of at least one day to assess intra-observer variation.

3.3.7 Statistics

Measured ADC values were median ADC within the VTV. Bland-Altman analysis [74] was used both at baseline and week 2 to compare delineation methods, and to evaluate intra-observer ADC variation (repeated delineations on the same scan) and ADC repeatability (test-retest difference). As data were non-normally distributed, non-parametric statistics (median, 15.9% and 84.2% percentiles) were used in the Bland Altman analysis to describe bias and 68.3% limits of agreement (LOA) for the observed ADC differences; mimicking a one standard deviation confidence interval (CI) for normal distributed data. Correlation between ADC values from the two delineation methods was assessed using Pearson's correlation coefficient, including the 95% CI.

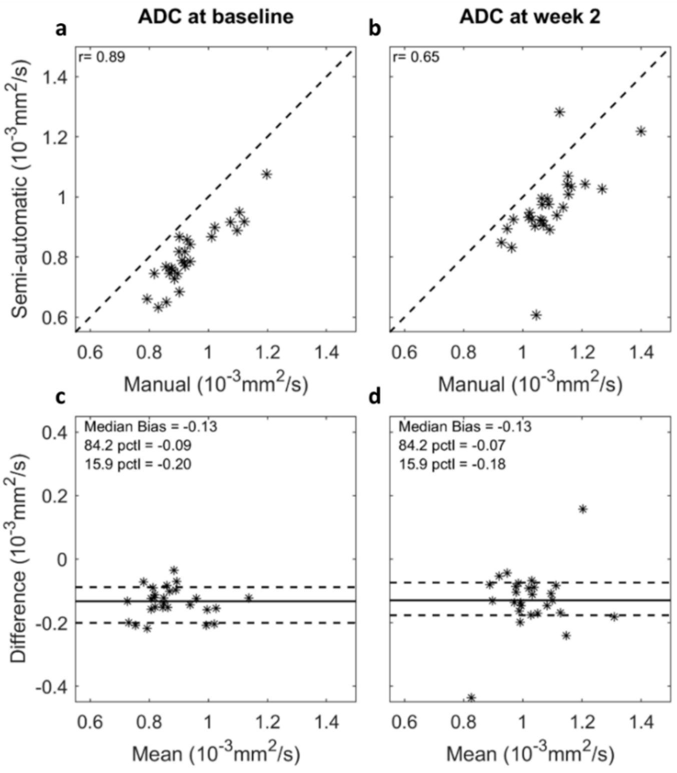


Figure 3.1. Comparison of delineation methods: Correlation plots (a-b) and Bland-Altman plots (c-d) comparing ADC values measured using semi-automatic delineation by a non-radiologist and manual delineation by a radiologist, at baseline and week 2 in RT. Pearson's correlation coefficient (r) is shown on correlation plots. The Bland-Altman plots show the ADC difference (semi-automatic minus manual) against the mean ADC; the solid and dashed lines represent median ADC difference and 68.3% limits of agreement, respectively. Two extreme measurements were observed at week 2 ($-0.437 \cdot 10^{-3}$ and $0.158 \cdot 10^{-3}$ mm^2 s^{-1}); the first may be explained by the fact that tumour volume was very small, making ADC calculation sensitive to delineation, and the second by a sub-optimal SADT delineation.

To obtain an uncertainty estimate of the ADC values measured with the SADT, all test-retest values at baseline and week 2 were analysed together, bearing in mind that the test-retest differences represented a combined imaging and delineation uncertainty. The uncertainty was estimated as the range between the 15.9% and 84.2% percentiles of the distribution of the differences in median ADC between test and retest. For each individual test-retest scan, ADC differences were calculated as both ‘test minus retest’ and ‘retest minus test’ to get a symmetric distribution with zero mean. The obtained uncertainty estimate was used to evaluate whether a significant change in ADC values between baseline and week 2 could be observed.

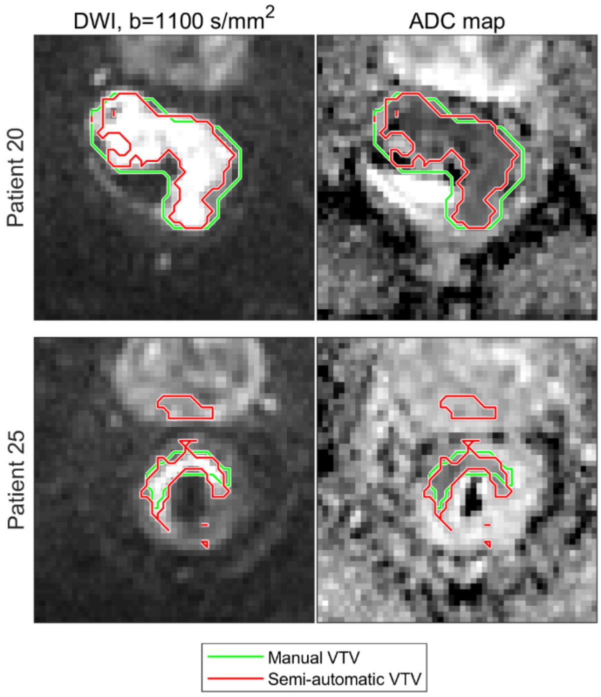


Figure 3.2. Delineation agreement: Example of a good (Patient 20) and a bad (Patient 25) agreement between manual (green) and semi-automatic (red) delineations for two patients. The images are transaxial and have been cropped, such that an area of (92.8x92.8) mm2 is shown. Rectum and part of the prostate are visible. Although the VTV defined by the semi-automatic delineation tool (SADT) appears as several separated regions when presented in 2D, it is in fact one connected 3D region.

3.4 Results

3.4.1 Comparison of semi-automatic and manual delineation

Median ADC values were systematically smaller when derived using the SADT compared to manual delineation by a radiologist. This difference was observed both at baseline and week 2 (Figure 3.1). The observed median differences and LOA as given in the method section were -0.13 [-0.20; -0.09] 10^{-3} mm^2 s^{-1} and -0.13 [-0.18; -0.07] 10^{-3} mm^2 s^{-1} for baseline and week 2, respectively . The Bland Altman plots in Figure 3.1 did not indicate any association between ADC differences and ADC values. The correlation [95% CI] of ADC between the two delineation methods was 0.89 [0.78, 0.95] at baseline and 0.65 [0.36, 0.83] at week 2.

By visual inspection, representative examples of good (patient 20) and poor (patient 25) delineation agreements between the SADT and manual delineation were identified (Figure 3.2). VTVs delineated with the SADT were typically smaller than manually delineated VTVs (mean volume was 42% smaller at baseline and 39% smaller at week 2).

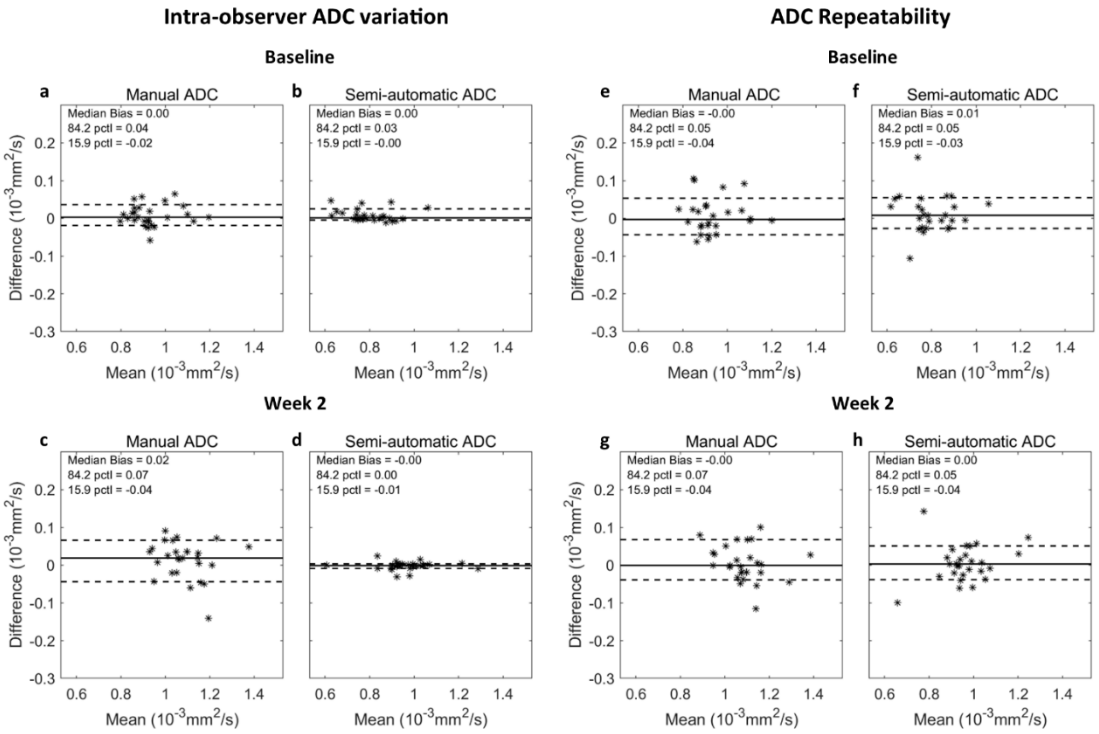


Figure 3.3. ADC variation: Bland-Altman plots showing intra-observer ADC variation (a-d) and ADC repeatability (e-h) at baseline and week 2 for delineation with the semi-automatic delineation tool (SADT) by a non-radiologist and manual delineation by a radiologist. The solid and dashed lines represent median ADC difference and 68.3% limits of agreement, respectively. The limits of agreement is defined as the 15.9% and 84.2% percentiles (pctl) of the ADC differences.

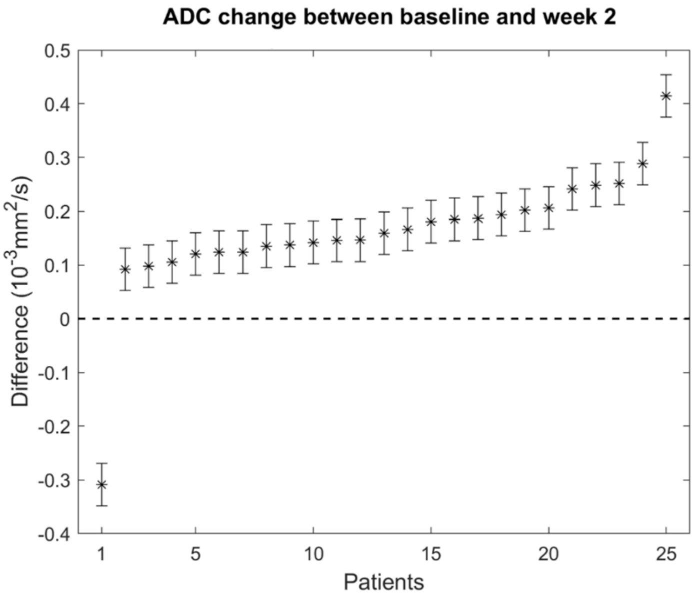


Figure 3.4. Temporal ADC changes: ADC change between baseline and week 2 measured using the semi-automatic delineation tool (SADT). The error bars represent the estimated ADC uncertainty described in section 3.3.7 (±0.04 mm2 s-1). The ordering of patients on the x-axis is arranged to show increasing ADC change from left to right.

The intra-observer ADC variation (repeated delineations on the same scan) was compared between the manual delineation and the SADT at baseline (Figure 3a-b) and week 2 (Figure 3c-d). For manual delineation, the median difference and LOA were 0.00 [-0.02; 0.04] 10^{-3} mm^2 s^{-1} at baseline and 0.02 [-0.04; 0.07] 10^{-3} mm^2 s^{-1} at week 2. For the SADT, the median difference and LOA were 0.00 [-0.00; 0.03] 10^{-3} mm^2 s^{-1} at baseline and -0.00 [-0.01; 0.00] 10^{-3} mm^2 s^{-1} at week 2; hence, the tool demonstrated a smaller intra-observer ADC variation compared to manual delineation.

The ADC repeatability (test-retest difference) was compared between manual delineation and the SADT at baseline (Figure 3.3e-f) and week 2 (Figure 3.3g-h). For manual delineation, the median difference and LOA were -0.00 [-0.04; 0.05] 10^{-3} mm^2 s^{-1} at baseline and -0.00 [-0.04; 0.07] 10^{-3} mm^2 s^{-1} at week 2, and for the SADT 0.01 [-0.03; 0.05] 10^{-3} mm^2 s^{-1} at baseline and 0.00 [-0.04; 0.05] 10^{-3} mm^2 s^{-1} at week 2. Thus, the ADC repeatability was comparable between the two delineation methods.

3.4.2 ADC changes and related uncertainty

For each patient, the ADC change between baseline and week 2 was evaluated when using the SADT. The obtained differences showed an increase between baseline and week 2 for all patients except one, with a mean ADC increase of $0.159 \cdot 10^{-3}$ mm^2 s^{-1} (Figure 3.4). The error bars in Figure 3.4 represent the ADC uncertainty estimate derived in section 3.3.7. The ADC increase was larger than the uncertainty ($\pm 0.04 \cdot 10^{-3}$ mm^2 s^{-1}) in all cases.

3.5 Discussion

The SADT for ADC calculation of tumours was implemented to improve consistency of ADC measurements and increase feasibility in the clinical workflow of MR guided RT. The observed correlation between the SADT and manual delineation was strong at baseline and moderate at week 2, with the SADT measuring systematically smaller ADC values. A clearly smaller intra-observer ADC variance was seen for the SADT compared to manual delineation, and ADC repeatability was comparable between the delineation methods. In all patients but one, an ADC change larger than the uncertainty was observed between baseline and week 2 using the SADT. This supported the assumption that ADC might carry potential biological information which can be detected using the presented SADT. The SADT was simple, transparent, easy to implement, and may be used in other types of cancer.

There has been debate as to which ROI strategy to use for ADC calculation [35,56,75]. Using the VTV is a recommended strategy, which has been explored in recent studies [35,56,58].

One study found that ADC within the VTV was a potential response biomarker and that the VTV was more relevant for response prediction than the GTV [58]. It would be highly relevant to validate the VTV approach against histology in future studies. The presented SADT delineated the VTV based on the underlying assumption that hyperintense areas on high b-value DWI correspond to high cellularity. This assumption has been widely supported [34], although a lower degree of correlation has been observed in patients [76]. A higher correlation might have been found if delineations were better confined to viable tumour sub-regions, such as the VTV. To prevent the SADT from capturing false high cellularity regions, due to so-called T2-shine-through, ADC maps were included in the delineation process, to impose the low-ADC criterion. The SADT made use of Otsu's method of thresholding to determine a threshold automatically based on the intensity distribution within the manual mask. This approach allowed the threshold to be tailored to a specific ROI in a particular image volume; hence, the threshold differed between delineations. This method was preferred over a constant threshold level due to the arbitrary intensities in MRI and the possibility of intensity variation across images and MRI scanners. A drawback of this method, however, was a slight sensitivity to the manual input.

Some differences between the manually and semi-automatically segmented VTVs were observed (Figure 3.2). The manual delineation by the radiologist included more of the periphery of the tumour region, leading to a systematically higher ADC (Figure 3.1) since the diffusion usually is less restricted in the surrounding normal tissue. Furthermore, the manual delineations did not have as concave shapes as those created from the SADT, as seen in the upper part of Figure 3.2. In general, the SADT seemed to perform as intended, and the observed differences in size and shape between delineation methods were not considered a failure of the SADT. However, in a few cases, VTVs delineated with the SADT erroneously included healthy tissue, e.g. prostate tissue, as illustrated in Figure 3.2, patient 25. This error arose due to the expansion of the VTV in the post-processing step of the SADT's algorithm. However, rejecting the expansion-step might lead to exclusion of relevant regions. This implied that the resulting contours should be reviewed for obvious errors. Overall, the two delineation methods correlated well, indicating that the same tendency could be captured. Since the purpose of the SADT was not to mimic manual delineation but rather to deliver a reproducible measure of a representative tumour ADC, the observed offset between semi-automatic and manual delineation was considered acceptable.

Manual delineation showed larger intra-observer ADC variation at week 2 compared to baseline (Figure 3.3). This may be explained by the fact that the tumour volume decreased, and the tumour outline became less clear, as it is often seen during the course of RT. No information of tumour size and position was available to the radiologist during delineation, which might have caused the manual intra-observer variation to be larger than in normal clinical situations. In comparison, the intra-observer variation was smaller for the SADT, showing its potential to improve the consistency of ADC measurements. Potentially, the SADT may also reduce the inter-observer ADC variation, which should be investigated in a follow-up study.

ADC repeatability was affected by imaging-related uncertainty and intra-observer variation. The test-retest scans were acquired in quick succession while the patient remained positioned on the treatment table. Re-positioning the patient between the scans might give a better estimate of the actual clinical ADC repeatability. Though, even without re-positioning, tumour motion (due to bulk motion, peristalsis, gas etc.) between test and retest was evident, and may therefore be a good estimate of the true repeatability. Due to observed differences in tumour position and shape between test and retest, it was found more appropriate to re-contour on test and retest, than to apply structure propagation between scans. This was also representative of a clinical setup where delineations are made on the data by hand.

In all patients except one, ADC increased between baseline and week 2 (Figure 3.4). An increase in ADC during RT is in correspondence with earlier ADC studies and might be explained by radiation-induced cellular changes [63,77]. The observed ADC change was larger than the ADC uncertainty in all cases, indicating a potential biological change. In this study, ADC changes were not compared to treatment response, as the aim was to evaluate the capacity of the SADT to extract potential biological changes and not response prediction in this particular patient cohort.

The manual mask defined by the non-radiologist was restricted to the same slices as included in the radiologist´s delineation. Hence, only the in-plane delineations were compared between the SADT and the expert manual delineation. This was done to allow a more fair comparison of the delineation methods, since the non-radiologist was inexperienced in recognizing rectal tumours, although it limits the use of the SADT as a standalone tool for ADC calculation. Nevertheless, in a clinical workflow, it may be preferable to use the GTV as manual input to the SADT since the VTV is by definition a sub-region of the GTV. The GTV could be obtained by AI based delineation if this became robust and commonly available. In all cases, the GTV as manual input is appealing since GTVs are already available from the normal workflow, and it may facilitate treatment adaption

based on ADC in an MRI guided RT workflow on the MR-linac.
In conclusion, the presented SADT showed performance comparable to manual expert delineation, and demonstrated potential to improve consistency of ADC measurements. The SADT was able to detect temporal ADC changes larger than the uncertainty associated with ADC measurements, which implies capability of measuring ADC changes attributed to change in tumour biology during the course of RT. The SADT may therefore prove useful for validation of ADC as a treatment response biomarker.

Disclosure of interest: The authors report no conflicts of interest.

Funding: This work was supported by Danish Cancer Society (Grant no. R231-A13852), Danish Comprehensive Cancer Center RT (Danish Cancer Society grant) (Grant no. R191-A11526), and by MANTRA (New MAgNetic resonance Technology for Response Adapted radiotherapy), a Frontline research center based at Odense University Hospital, Denmark. The funding sources had no role in the study design, collection, analysis and interpretation of data, writing of the report or in the decision to submit the article for publication.

Acknowledgements

We thank Frederik Severin Gråe Harbo, Department of Radiology, Odense University Hospital, for useful insight in the initial phase of the study, and Lars Johnsen, Department of Oncology, Odense University Hospital, for assistance with handling data in Pinnacle and MIM.

4 Paper II

Recommendations for improved reproducibility of ADC derivation on behalf of the Elekta MRI-linac Consortium Image Analysis Working Group

Anne L.H. Bisgaard*[a,b], Rick Keesman[c], Astrid L.H.M.W. van Lier[d], Catherine Coolens[e], Petra J. van Houdt[f], Alison Tree[g], Andreas Wetscherek[h], Paul B. Romesser[i], Neelam Tyagi[j], Monica Lo Russo[k], Jonas Habrich[l], Danny Vesprini[m], Angus Z. Lau[n], Stella Mook[d], Peter Chung[o], Linda G.W. Kerkmeijer[c], Zeno A. R. Gouw[f], Ebbe L. Lorenzen[a], Uulke A. van der Heide[f], Tine Schytte[b,p], Carsten Brink[a,b], Faisal Mahmood[a,b]

[a]Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Kløvervænget 19, 5000 Odense, Denmark
[b]Department of Clinical Research, University of Southern Denmark, J.B. Winsløws Vej 19.3, 5000 Odense, Denmark
[c]Department of Radiation Oncology, Radboud University Medical Centre, P.O. Box 9101 , 6500 HB Nijmegen, The Netherlands
[d]Department of Radiotherapy, University Medical Centre Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands
[e]Department of Medical Physics, Princess Margaret Cancer Centre, University Health Network, 610 University Avenue, M5G 2M9, Toronto, ON, Canada
[f]Department of Radiation Oncology, the Netherlands Cancer Institute, Postbus 90203, 1006 BE Amsterdam, The Netherlands
[g]Department of Urology, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, Downs Road, Sutton, Surrey, SM2 5PT, London, UK
[h]Joint Department of Physics, The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, SM2 5NG London, UK
[i]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center 1275 York Avenue, Box 22, NY 10065, New York, USA
[j]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, 545 E. 73rd street, NY 10021, New York, USA
[k]Department of Radiation Oncology, University Hospital and Medical Faculty, Eberhard Karls University, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany
[l]Section for Biomedical Physics, Department of Radiation Oncology, University of Tübingen, Hoppe-Seyler-Str. 3, 72076 Tübingen, Germany
[m]Department of Radiation oncology, Sunnybrook Odette Cancer Centre, University of Toronto, 2075 Bayview Avenue, M4N 3M5 Toronto, ON, Canada
[n]Physical Sciences Platform, Sunnybrook Research Institute. Department of Medical Biophysics, University of Toronto, 2075 Bayview Avenue, M4N 3M5 Toronto, ON, Canada
[o]Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network. Department of Radiation Oncology, University of Toronto, 610 University Avenue, M5G 2M9 Toronto, ON, Canada.
[p]Department of Oncology, Odense University Hospital, Kløvervænget 19, 5000 Odense, Denmark

***Corresponding author**

Published in: Radiotherapy and Oncology, 186 (2023) 1-7.

4.1 Abstract

Background and purpose

The apparent diffusion coefficient (ADC), a potential imaging biomarker for radiotherapy response, needs to be reproducible before translation into clinical use. The aim of this study was to evaluate the multi-centre delineation- and calculation-related ADC variation and give recommendations to minimize it.

Materials and methods

Nine centres received identical diffusion-weighted and anatomical magnetic resonance images of different cancerous tumours (adrenal gland, pelvic oligo metastasis, pancreas, and prostate). All centres delineated the gross tumour volume (GTV), clinical target volume (CTV), and viable tumour volume (VTV), and calculated ADCs using both their local calculation methods and each of the following calculation conditions: b-values 0-500 vs. 150-500 s/mm^2, region-of-interest (ROI)-based vs. voxel-based calculation, and mean vs. median. ADC variation was assessed using the mean coefficient of variation across delineations (CV_D) and calculation methods (CV_C). Absolute ADC differences between calculation conditions were evaluated using Friedman's test. Recommendations for ADC calculation were formulated based on observations and discussions within the Elekta MRI-linac consortium image analysis working group.

Results

The median (range) CV_D and CV_C were 0.06 (0.02-0.32) and 0.17 (0.08-0.26), respectively. The ADC estimates differed 18% between b-value sets and 4% between ROI/voxel-based calculation (p-values <0.01). No significant difference was observed between mean and median (p=0.64). Aligning calculation conditions between centres reduced CV_C to 0.04 (0.01-0.16). CV_D was comparable between ROI types.

Conclusion

Overall, calculation methods had a larger impact on ADC reproducibility compared to delineation. Based on the results, significant sources of variation were identified, which should be considered when initiating new studies, in particular multi-centre investigations.

4.2 Introduction

Quantitative imaging biomarkers (QIBs), derived from in-vivo imaging, are useful in oncology, as they non-invasively provide quantitative information on tissue characteristics [44,45,78]. Development of QIBs has the potential to improve precision and reduce subjectivity of image analysis, and hereby enable a more robust association between image-derived parameters and biological and clinical parameters [79,80]. QIBs may provide spatially and temporally resolved information linked to tumour biology, which in radiotherapy may be used for improved target delineation, dose painting and prediction and monitoring of response. Hence, QIBs may improve personalization of the treatment [20].

The advanced magnetic resonance imaging (MRI) technique, diffusion-weighted MRI (DWI), is a potential QIB for the above-mentioned radiotherapy purposes [20,23,26,29]. In standard DWI, strong magnetic gradients are applied to sensitize the MRI signal to the random motion of water molecules within the scanned object. The amount of diffusion weighting is defined by the b-value, and if at least two appropriately selected b-values are acquired, the quantitative parameter, the apparent diffusion coefficient (ADC), can be derived. ADC correlates with tissue cellularity, and have been shown to identify *radio-resistant* regions [10] [11]. DWI and derived ADC maps are used in the clinic to guide target delineation for some tumours, and may be a future tool for dose painting [28,82]. Further, baseline ADC and ADC changes during treatment have shown potential to predict response, although lack of consistency is preventing translation to the clinic [26,63,65,66,83]. Specifically, varying acquisition protocols and analysis methods reduce ADC reproducibility, potentially hindering validation of ADC as a QIB. To overcome this problem, a standardization of measurements is needed, and large multi-centre validation trials are warranted [35,44].

Hybrid MRI linear accelerators (MRI-linac) allows daily measurement of ADC, with no or limited prolongation of the radiotherapy fractions [10,11]. As such, an MRI-linac provides an ideal platform for clinical validation of potential QIBs such as ADC. Accuracy of ADC on MRI-linac has been demonstrated using phantoms, and feasibility has been demonstrated in patients [27,35,50]. Furthermore, recommendations for MRI protocols to acquire DWI on an Elekta MRI-linac have been published [51]. The current study focused on the analysis of the acquired DWI scans to obtain an ADC value.

Different approaches to DWI analysis may introduce a variation across centres/studies. Within the Elekta MRI-linac consortium image analysis working group [84], two expected sources of variation were identified: The delineation of a region of interest (ROI), and the calculation method. Delineation uncertainty is a well-known source of uncertainty in radiotherapy and propagates as ADC variation as well [55,85]. The impact of calculation methods on ADC reproducibility has been investigated to a lesser extent [61]. The current study investigated the impact of variations in both delineations and calculation methods on the ADC reproducibility utilizing the same data, which enabled assessment of their relative contributions. The aim was to give vendor-neutral recommendations to improve ADC reproducibility, based on an evaluation of the observed ADC variation between MRI-linac centres and discussions within the working group.

4.3 Methods

4.3.1 Study design

Nine MRI-linac centres participated in the study using anonymized patient MRI data from four different clinical cases, acquired at one of the participating centres. At each centre, two steps were performed (Figure 4.1). In step 1, an oncologist performed delineations. In step 2, each centre calculated ADC for delineations made at all centres using their local calculation method. This resulted in a 9x9 table of ADC values for each clinical case and delineation type.

4.3.2 Clinical cases

The study included four patients with different cancerous tumours.

Adrenal gland (76 year old male with oligo progression after systemic treatment for non-small cell lung cancer)

Pancreas (68 year old male with recurrent pancreas cancer, consolidative radiotherapy after systemic treatment)

Oligo metastasis in the pelvis (54 year old woman with recurrent ovarian cancer, consolidative radiotherapy after systemic treatment)

Prostate and adjacent seminal vesicles (74 years old man with low volume metastatic prostate cancer)

All patients received treatment on the same 1.5 T MRI-linac (Unity by Elekta, Stockholm, Sweden) at one of the participating centres. The patients were included in the MOMENTUM study (clinicaltrials.gov NCT04075305) [28] . Informed consent was obtained from all patients, and DICOM-data was anonymized and stored adhering to ethics standards.

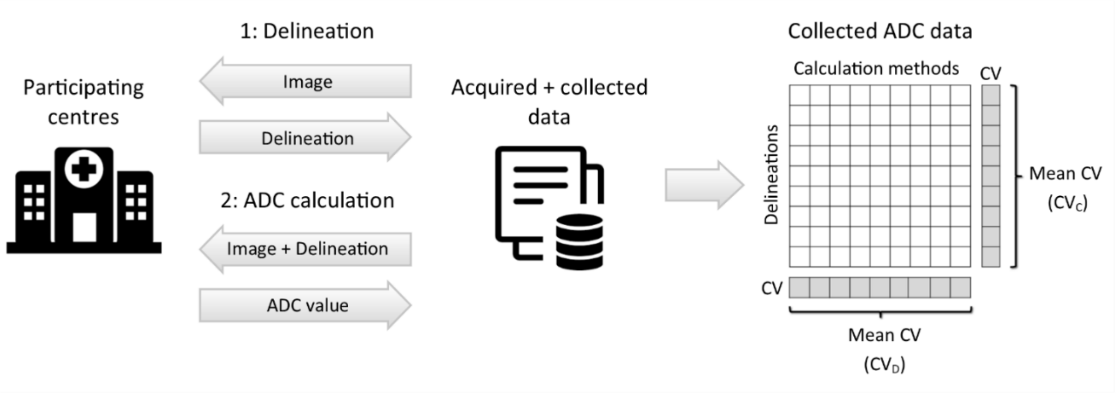


Figure 4.1. Study design. Each of the nine participating centres performed delineation and ADC calculation. The collected ADC values were organized in a table as illustrated to the right, where rows and columns represent the delineations and calculation methods from the nine centres. Tables were made for each combination of cancer diagnosis and delineation types (GTV, CTV, VTV). The ADC variation across delineations and calculation methods were assessed using the mean coefficient of variation (CV), as indicated on the table.

4.3.3 MRI data

MRI data were acquired at fraction one, prior to beam delivery and included T2-weighted images (T2W) and DWI with the b-values 30, 80, 150, 300 and 500 s/mm^2 (adrenal gland and pancreas), and 0, 30, 80, 150 and 500 s/mm^2 (oligo metastasis and prostate) adhering to the normal MRI-linac workflow [29]. Sequence details are listed in Table A2.1 in Appendix IIa. DWI were acquired twice in succession while the patient remained in position, to obtain test-retest data for repeatability estimation.

4.3.4 Delineation

T2W images and DWI images with b-values 150 and 500 s/mm^2 were available for delineation. Provided with brief clinical case descriptions, the oncologists delineated the gross tumour volume (GTV), clinical target volume (CTV) (prostate only) and the viable tumour volume (VTV) (except for prostate) in a mutually blinded manner using the ProKnow platform (Version 1.32.0, Elekta Solutions AB, Stockholm, Sweden). The VTV was defined as the GTV excluding cystic and necrotic parts. A description of the technical data preparation is given in Appendix IIb.

4.3.5 ADC calculation

Each centre provided a brief description of their local calculation method, including software implementation, choice of b-values, and whether a ROI- or voxel-based calculation was used. The ROI-based method refers to ADC calculation using the mean or median ROI signals of DWIs, whereas the voxel-based method refers to calculating ADC

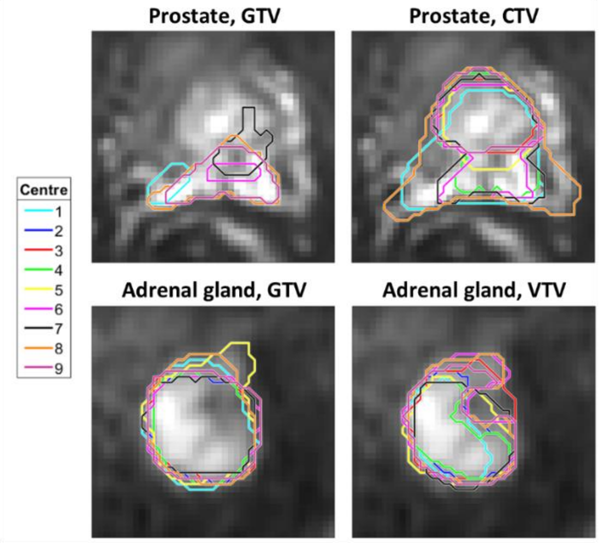


Figure 4.2. Examples of delineations. Delineations made by the nine participating centres for prostate and adrenal gland, shown on b=500 /mm^2 DWI images, cropped to an area of 7.7×7.7 cm^2 (prostate) and 4.9×4.9 cm^2 (adrenal gland) around the tumour. For the prostate, not all delineated contours included the shown slice, thus, only five contours are visible.

within each voxel, after which the mean or median value is calculated within the ROI. If a centre's standard approach was to use the scanner software for ADC calculation, that centre was provided with ADC maps calculated with the scanner software using all b-values, the lowest and the highest value, and b≥150 s/mm^2, respectively. They were asked to choose the set best representing their normal choice.

Each centre provided ADC values for both their own and other centre's delineations. The calculation was based on 1: the centre's own calculation method, and 2: all combinations of the following calculation conditions: all b-values vs. b≥150 s/mm^2, ROI-based vs. voxel-based and mean vs. median (referred to as pre-specified calculation conditions).

4.3.6 Data analysis and statistics

Delineations were compared pairwise to calculate the Dice similarity coefficient (Dice) and mean surface distance (MSD). ADC variation across delineations and calculation methods was assessed using the mean coefficient of variation (CV), calculated in the following way (Cf. Figure 4.1): The CV describing variation across calculation methods was calculated for each of the nine delineations, and the average of these nine values was used as a measure of variation across calculation methods (CV_C). Likewise, the CV describing variation across delineations was calculated for each of the nine calculation methods, and the average was used as a measure of variation across delineations (CV_D).

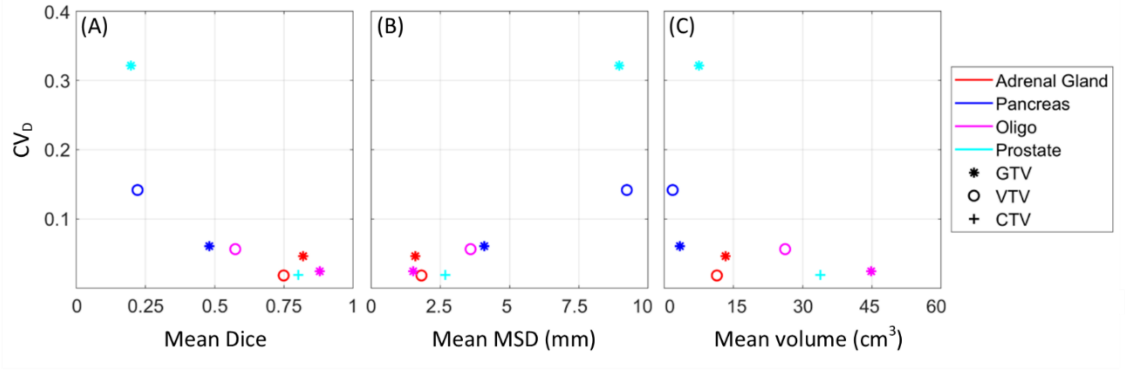


Figure 4.3. Delineation-related ADC variation. Delineation-related ADC variation (mean coefficient of variation, CV$_D$) as a function of mean Dice Similarity Coefficient (A), Mean Surface Distance, MSD (B), and volume (C), for the different clinical cases (marker colors) and types of ROIs (marker types).

Retest ADC values were calculated using rigid contour propagation of GTVs between test- and retest-scans. Median ADC values within the GTVs were extracted from ADC maps calculated with the scanner software using b≥150 s/mm^2. The within-subject coefficient of variation (wCV) was calculated as a measure of test-retest ADC variation (ADC repeatability), as recommended by the Quantitative Imaging Biomarkers Alliance (QIBA) [30].

The ADC difference between the sets of b-values, ROI/voxel-based analysis and mean/median values, respectively, were evaluated using Friedman tests with a 5 % significance level and with Bonferroni correction for multiple testing. Only GTVs were used for this purpose.

4.4 Results

A total of 69 out of 72 expected delineated volumes (9 centres x 8 volumes) were available for the analysis. Within these volumes, a total of 4483 ADC values were obtained out of 5589 (69 delineation x 9 centres x 9 combinations of calculation conditions). The reasons for the reduced number were the following: One centre omitted calculation within two prostate volumes and two centres omitted the ROI-based calculations due to technical difficulties or limitations of their local software. One centre omitted ADC calculation using the pre-specified calculation conditions due to limited time and resources. One centre used software that reported only one decimal, which in some cases led to CV's of zero. CV's of zero were excluded before calculating the mean CV.

Representative delineations are presented in Figure 4.2. The delineation variation was large for pancreas VTV and prostate GTV (Dice: 0.20-0.22 and MSD: 9.09-9.23 mm) compared to

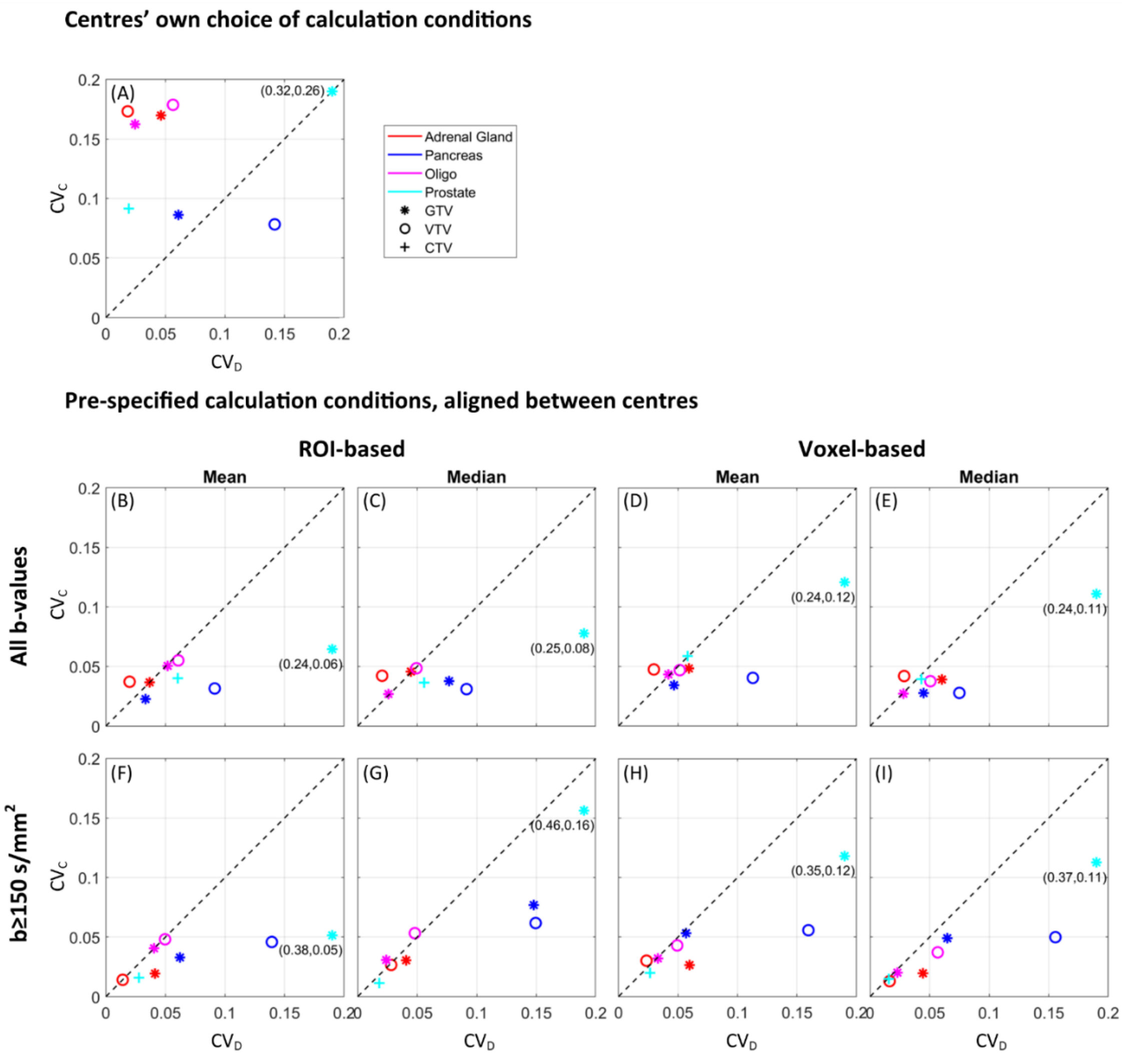


Figure 4.4. ADC variation. ADC mean coefficient of variation across delineations (CV_D) and calculation methods (CV_C) from the nine MR-linac centres, with the centres' own choice of calculation conditions (A), and with pre-specified calculation conditions (B-I). Median ADC values were used in (A). The marker colours and types represent the different clinical cases and types of ROIs. The dotted line at x=y represents the points where delineation- and calculation-related ADC variation are the same. For the prostate GTV, CV_D is outside the axis range, and therefore, the true coordinates are indicated next to the marker.

the remaining cases (Dice: 0.48-0.88 and MSD: 1.52-4.09 mm) (Figure 4.3.A-B). A closer inspection of the prostate delineations revealed that some GTV delineations did not overlap (Figure 4.2). The prostate CTV delineation variation was smaller (Dice: 0.80, MSD: 2.68 mm), despite not all centres included the seminal vesicles in the delineation. The CV_D was comparable between GTV and VTV, although the delineation variation was slightly smaller for GTV compared to VTV (Figure 4.3.A-B). There was a clear correlation between delineation variation and ADC variation (Figure 4.3.A-B).

	All b-values				b ≥ 150 s/mm^2				Mean difference (%)		
	ROI-based		Voxel-based		ROI-based		Voxel-based				
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	b-sets	ROI/ vox	Mean/ median
Adrenal gland											
GTV	1.27	1.37	1.21	1.24	0.90	0.97	0.88	0.90	32.86	5.84	-4.91
VTV	1.26	1.34	1.23	1.24	0.88	0.96	0.87	0.90	33.69	4.40	-4.47
Pancreas											
GTV	0.95	0.95	0.88	0.96	0.88	0.85	0.81	0.82	11.05	4.30	-1.59
VTV	1.08	1.08	1.03	1.07	1.03	1.06	0.98	1.00	4.59	3.90	-2.18
Oligo metastasis											
GTV	1.51	1.41	1.44	1.34	1.33	1.33	1.29	1.19	10.35	5.99	5.48
VTV	1.49	1.39	1.43	1.32	1.33	1.31	1.29	1.18	9.58	5.45	6.33
Prostate											
GTV	1.41	1.37	1.35	1.34	1.07	0.99	1.03	1.02	28.41	1.85	3.09
CTV	1.56	1.52	1.49	1.46	1.31	1.29	1.27	1.29	15.62	3.00	1.13
Mean value	1.32	1.30	1.26	1.25	1.09	1.09	1.05	1.04	18.27	4.34	0.36

Table 4.1. Mean ADC values (x 10^{-3} mm^2/s) across nine centres for different combinations of calculation conditions. The mean %-wise ADC differences between b-sets (all-b-values minus b ≥ 150 s/mm^2), ROI/voxel-based analysis (ROI-based minus voxel-based) and mean/median values (mean minus median) are shown.

All centres used a voxel-based approach as their local calculation method. One centre used ADC maps generated by the scanner software, while remaining centres used in-house software for ADC calculation with a mono-exponential Stejskal-Tanner model [31]. The main differences between the local calculation methods were the choice of b-values, fitting method, and applied filtering. A full comparison of the centres' local calculation methods is presented in Table A2.2 in Appendix IIc.

With the centres' own calculation methods, the median (range) CV_D and CV_C were 0.06 (0.02 – 0.32) and 0.17 (0.08 – 0.26), respectively (Figure 4.4.A). The delineation-related variation was larger for pancreas VTV and prostate GTV (CV_D: 0.15-0.32) compared to the remaining cases (CV_D: 0.02-0.06). In comparison, the ADC repeatability (wCV) based on test-retest scans was estimated to 4.0% (adrenal gland), 6.6% (pancreas), 1.3% (oligo metastasis), and 15.2% (prostate). A detailed overview of the ADC variation for each delineation and calculation method is shown in Figure A2.1-3 in Appendix IId.

When centers aligned their calculation methods according to any of the pre-specified calculation conditions, the calculation-related ADC variation was clearly smaller than when centres used their own choice of calculation conditions (Figure 4.4.B-I compared to Figure4.4.A), with a reduction of median (range) CV_C to 0.04 (0.01 - 0.16) (or 0.04 (0.01 - 0.08) with the low-agreement prostate GTV excluded).

In terms of absolute ADC, there was a trend towards larger values for calculation methods that included b-values below 150 s/mm^2, (calculation methods no. 1, 4 and 9 in Figure A2.1 and A2.3 and Table A2.2 in Appendix IIc-d). Averaged across all combinations of the pre-specified calculation conditions, ADC estimates were 18% larger for the full b-set compared to $b \geq 150$ s/mm^2 ($p<0.01$) and 4% larger for ROI-based analysis compared to voxel-based ($p<0.01$) (Table 4.1). There was no significant difference between mean and median values ($p = 0.64$).

4.5 Discussion

This study evaluated the ADC variation related to differences in delineation and calculation methods between centres. The calculation-related variation was generally larger than delineation-related variation (Figure 4.4.A), and was primarily driven by different choices of b-values. When calculation conditions (all b-values vs. $b \geq 150$ s/mm^2, ROI-based vs. voxel-based, and mean vs. median) were aligned between centres, the calculation-related variation was reduced to about the same level as the delineation-related variation. Furthermore, the delineation- and calculation-related ADC variation was comparable to the ADC repeatability, indicating that acquisition and post-processing of the images contribute equally to the ADC variation. The GTV and VTV performed comparably with respect to ADC reproducibility.

Overall, the observed delineation-related ADC variation largely agreed with other studies, showing CV of 0.1 and inter-observer coefficient of repeatability of 1.9-14% in pancreas [57] [86], and 9.5-13.7% in prostate [34], although not directly comparable due to differences in methods. The large delineation variation of the pancreas VTV was likely due to the higher sensitivity to delineation of small volumes (Figure 4.3.C). For the prostate GTV, the large delineation variation could arise from the GTV not being a standard delineation type. In fact, large variation in definitions of intra-prostatic lesions has been reported in earlier studies [88] [89]. Potentially, the use of a higher b-value would have improved the conspicuity of the intra-prostatic lesions. To comply with the MRI-linac recommendations, a maximum b-value of 500 s/mm^2 was used [51]. The delineation variation in prostate may also have been overestimated as not all centres included the vesicles in the CTV (as case descriptions indicated).

Other studies have shown that the type of ROI influences both absolute ADC values, relative ADC changes during treatment, and the reproducibility of delineations [35,55,56,85]. Therefore, this study included two types of ROIs. The GTV, because it has the advantage of being available before the start of treatment in both the standard and MRI-linac radiotherapy workflow. The so-called VTV was included because it excludes non-viable parts of the tumour and may be relevant for probing the cellular response directly and assessing treatment response, as suggested by Padhani et al. [35]. Further, one study showed that ADC based on VTV was superior to GTV in stratifying between responding and non-responding patients [58]. An advantage of the VTV is that, by definition, it contains only high signal-to-noise-ratio (SNR) voxels. For tumours with no significant necrotic/cystic components, e.g. prostate, the VTV corresponds to the GTV.

Since the choice of ROI type did not influence the ADC reproducibility in the current study, selection of ROI type depends on its application in radiotherapy. While the VTV may define radio-resistant regions and be relevant for dose painting, it is not obvious which ROI is best suited for response prediction. The literature investigating the potential benefits of using GTV vs. VTV is limited [58,65]. In general, the results of the current study advocate improving delineation consistency (Figure 4.3.A-B), which underlines the importance of having as precise consensus guidelines as possible. In the future, delineation variation may be reduced by automatic delineation tools including AI models, as indicated in several studies [69–71,83].

The DWI-signal is sensitive to perfusion at low b-values (below ~100 s/mm^2), and therefore, including low b-values in the analysis is expected to overestimate ADCs [35,41] as observed in this study also (all b-values compared to b≥150 s/mm^2) (Table 4.1). Therefore, a previous publication by the Elekta MRI-linac working group, recommended that the lowest b-value should be 100-150 s/mm^2 [51]. A maximum b-value of 500 s/mm^2 was also recommended to ensure sufficient SNR and a diffusion time comparable to that of a diagnostic scanner. Moreover, if notably higher b-values are included in the calculation (b>1000 s/mm^2), non-Gaussian diffusion effects may result in an underestimation of ADC, as the mono-exponential model assumes a Gaussian diffusion behaviour [42].

The ROI- and voxel-based approach have been used in previous studies and are therefore relevant from a reproducibility point-of-view [55,56,86,90,91]. It should be noted that strictly speaking, the average ADC across voxels within a ROI cannot be derived using the ROI-based approach, which is based on the mean DWI signal within the ROI. I.e. the ROI-based method is mathematically inconsistent with the exponential model of ADC calculation (when more than one voxel is present within a ROI). However, using the ROI-based method may lead to better estimates of ADC as it is expected to be more robust to motion induced misalignment of individual DWI acquired at different b-values, which if not properly corrected can lead to invalid ADC values. Further, it may improve SNR which may give a better goodness of fit of data, as was confirmed using the current data (not shown) [20]. In the current study, the ROI-based approach led to larger ADC values compared to the voxel-based approach (Table 4.1), while the two approaches performed comparable with respect to ADC reproducibility (Figure 4.4.B-I).

The residual calculation-related ADC variation present after aligning the pre-specified calculation conditions between centres (Figure 4.4.B-I) may be accounted for as use of different software implementations, including different fitting and filtering methods (Table A2.2 in Appendix IIc). Specifically, five centres used linear least squares fitting of ln(S) as a function of b-values to estimate the ADC (Table A2.2 in Appendix IIc). Since the SNR decreases with increasing b-value, the uncertainty of ln(S) also increases with b-values, if not accounted for by averaging signals from multiple excitations. Thus, a better approach will be to use weighted linear least squares fitting (see Appendix IIe) [92]. For the voxel-based approach, five centres used filtering by excluding voxels containing non-physical values, i.e. values outside a certain range (Table A2.2 in Appendix IIc). Alternative to this, voxels may be removed based on low SNR or poor quality of the fit, which is a more objective criterion. Contributions from fitting and filtering were not determined individually, nevertheless, in combination, they resulted in calculation-related ADC variations comparable to the delineation-related variations (points close to the dotted line in Figure 4.4.B-I). This stresses the importance of excluding sources of variation whenever possible, especially if the aim is to establish common ADC cut-off values, e.g. for response prediction. Making a platform-independent software available for public download might be a way to proceed such that in-house developed software can be validated against a common software.

The SNR has also been shown to play a role in estimation of the ADC [51,93]. Although not specifically investigated in this study, it is worth mentioning a few implications. Low SNR levels lead to an underestimation of the ADC, due to the so-called noise floor present in magnitude reconstructed DWI-images [20,59]. Therefore, to allow a comparison between studies, the SNR should always be reported based on defined standards, e.g. published by the National Electrical Manufacturers Association (NEMA) [94] or QIBA [95]. For practicality, it may be sufficient to measure SNR once, if patient and coil positioning is consistent between scans [20]. Applying noise correction has been shown to reduce the ADC bias [93].

Other specific points of attention when calculating ADC include pre-processing of the image data. For example, to minimize the effect of motion, registration between b-values is recommended [96], and is available on most MRI scanner software, including the Unity MRI-linac. As a minimum, b-value images should visually be inspected for motion and artefacts. Further, as the intensity-histogram of DWI images may be "stretched" to fully utilize the storage bit depth, the stored pixel values should as a general rule be "unscaled" prior to quantitative analysis as described by Chenevert et al. [97].

A main limitation of this study is that only one patient was included per tumour type. This was deemed a necessary compromise to increase the realizability of the investigation. However, by including four tumour types instead of e.g. four tumours of the same type, we were able detect differences in the analysis-related ADC variation between tumour types. Minor limitations include that no re-positioning of the patient was performed between the test and retest scans, which may underestimate the true repeatability. ADC reproducibility may also be affected by the sequence used to acquire the images (turbo-spin-echo (TSE) vs. echo planar imaging (EPI) [98]) and the diffusion time [60], but investigations of this was outside the scope of the current study where EPI based readout was used. Moreover, as EPI is notorious for low geometric accuracy [53], a high ADC reproducibility can still lead to a misinterpretation of the extent of the GTV. The effect of geometric distortions on ADC reproducibility and GTV misalignment should be investigated in a future study.

4.6 Conclusion

This investigation provides recommendations for improving reproducibility of ADC calculations, based on observations and discussions within the Elekta MRI-linac consortium image analysis working group. These recommendations are focused towards future investigations of ADC as a potential imaging biomarker in radiotherapy. Investigations of other potential quantitative imaging biomarkers using a similar setup, and the geometric accuracy of these, are warranted.

In summary, the calculation-related ADC variation was larger than the delineation-related ADC variation. Specifically, the calculation-related ADC variation can be attributed to the choice of b-values, ROI-based/voxel-based calculation, and software implementation including fitting and filtering method. Therefore, it is recommended to align these factors in multi-centre studies, and to report details of the ADC calculation method within a study to allow comparison between studies. In general, delineation variation correlates with ADC variation, and should therefore be reduced as much as possible. Selection of GTV vs. a dedicated volume for ADC derivation seems less critical for ADC reproducibility, and should depend primarily on feasibility and the radiotherapy purpose.

Role of funding source

AB, FM acknowledges the support of the Danish Cancer Society (Grant no. R231-356 A13852), Danish Comprehensive Cancer Center RT (Danish Cancer Society grant) (Grant no. R191-A11526), and by MANTRA (New MAgNetic resonance Technology for Response Adapted radiotherapy), a Frontline research center based at Odense University Hospital, Denmark.

AT acknowledges the support of Cancer Research UK grant numbers C7224/A28724 and C33589/A28284. Further, AT acknowledges NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and The Institute of Cancer Research. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

MLR acknowledges the support of the German Research Council (DFG, Grant no. ZI 736/2-1; PAK997/1), the University Hospital Tübingen and the Medical Faculty Tübingen.

None of the above funding sources had any role in the study design, collection, analysis and interpretation of data, writing of the report or in the decision to submit the article for publication.

Acknowledgements

We thank Marco Luzzara, Senior Director of Medical Affairs and Clinical Research, Elekta, for his support in facilitating data sharing and analysis in ProKnow, Elekta. We thank Liam Lawrence, Physical Sciences Platform, Sunnybrook Research Institute, Toronto, Canada, and Edward Taylor, Department of Medical Physics, Princess Margaret Cancer Centre, University Health Network, Toronto, Canada, for their contributions to the data analysis.

PART II

Clinical validation

5 Clinical validation

5.1 Survival analysis and the Cox proportional hazards model

In order to do a clinical validation of DWI for RT purposes, it is essential to establish a relationship between DWI and clinical endpoints. An important step is to investigate whether DWI can be used to predict the prognosis of patients with cancer, and for this purpose, it is relevant to investigate the survival time of the patients, e.g. the time between diagnosis (starting point) and death (or other types of events such as recurrence of a disease or disease progression). In medical research, this is referred to as survival analysis. [99].

When investigating the overall survival (OS) of a patient cohort, some patients are typically alive at the cut-off date of the study, meaning that the event time (time of death) is unknown for these patients. In survival analysis, this is referred to as censoring. Since it is known that the censored patients have had no event until their censoring time point (e.g. the time point of follow up), they contribute to determining the probability of survival up until this time point. As a note, censoring should be "uninformative", meaning that the risk of a subsequent event should be the same for censored and uncensored patients, otherwise, censoring may introduce a bias [99].

Typically, the survival probability, $S(t)$, describing the probability that a patient survives from the starting point to the time t, is estimated using the Kaplan Meier (KM) method [100]:

$$S_{KM}(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

Here, j is an index running over the observed survival times in increasing order, both censored and uncensored, d_j is the number of patients with an event at time t_j, and n_j is the number of patients who have had no event just before t_j. This survival probability estimate, also referred to as the KM estimator, is a step function that is constant between the event time points. In medical research, the survival of a patient cohort is often described using the median survival time based on the KM estimator [99].

The survival time of patients with cancer depends on multiple parameters related to the patient, the tumour and the treatment. Hence, in order to investigate the prognostic value of DWI, it is therefore relevant to study the correlation of both DWI parameters and clinical parameters with OS. The Cox proportional hazards model, a commonly used method for survival analysis, is suited for this purpose, as it can be used to describe the relationship between event incidence and a set of parameters [101,102]. According to the Cox model, the hazard, h, defined as the risk of an event per time at a given time, t, is expressed by:

$$h(t) = h_0(t) \cdot \exp(\boldsymbol{x}_i \cdot \boldsymbol{\beta})$$

where $\boldsymbol{x}_i = (x_1, x_2, \dots, x_n)$ contains the a set of parameter values for a patient and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)$ contains coefficients reflecting the impact of each parameter. The baseline hazard, $h_0(t)$ describes the time-dependent hazard for a patient with all parameters equal to zero. Positive values of the coefficients, $\boldsymbol{\beta}$, mean that higher values of the corresponding parameters are associated with an increased hazard, and therefore a poorer survival, while the opposite is the case if the coefficients are negative.

The term "proportional hazard" refers to the proportional relationship between the hazards for different sets of parameter values ($\boldsymbol{x_i}$ and $\boldsymbol{x_j}$):

$$\frac{h(t|\boldsymbol{x_i})}{h(t|\boldsymbol{x_j})} = \frac{h_0(t) \cdot \exp(\boldsymbol{x_i}\boldsymbol{\beta})}{h_0(t) \cdot \exp(\boldsymbol{x_j}\boldsymbol{\beta})} = \exp\left((\boldsymbol{x_i} - \boldsymbol{x_j}) * \boldsymbol{\beta}\right)$$

$$\Updownarrow$$

$$h(t|\boldsymbol{x_i}) = \exp\left((\boldsymbol{x_i} - \boldsymbol{x_j}) * \boldsymbol{\beta}\right) \cdot h(t|\boldsymbol{x_j}) = constant \cdot h(t|\boldsymbol{x_j})$$

This proportionality means that the hazard in one patient or patient group with the parameter values $\boldsymbol{x_i}$ is a constant times the hazard in any other patient or patient group with the parameter values $\boldsymbol{x_j}$ [102]. The time dependence cancels out, and thus the relationship will remain proportional at all times. As a result of the proportionality, if one parameter, e.g. x_1, increases by one unit while the other parameters are constant, the hazard increases with at factor of $\exp(\beta_1)$. For this reason, the exponential function of the coefficients are interpreted as so-called hazard ratios, i.e. $\exp(\beta_1)$ is the hazard ratio associated with increasing the parameter, x_1, by one unit.

The coefficients, $\boldsymbol{\beta}$, are determined by maximizing the partial log likelihood, given by [101]:

$$\log\left(L(\boldsymbol{\beta})\right) = \sum_i \log\left(L_i(\boldsymbol{\beta})\right)$$

where $L_i(\boldsymbol{\beta})$ denotes the partial likelihood for the i'th patient:

$$L_i(\boldsymbol{\beta}) = \frac{h(Y_i|\boldsymbol{x_i})}{\sum_{j:Y_j \geq Y_i} h(Y_i|\boldsymbol{x_j})}$$

Here, Y_i is the survival time for patient i, and j is an index running over all patients who are still at risk when the event occurs for patient i. The partial likelihood, $L_i(\boldsymbol{\beta})$, describes the likelihood of an event for the i'th patient at the observed event time point, Y_i, according to the model. This term is maximized when the hazard predicted by the model is highest for the patient who actually experience an event at the time Y_i. Hence, the partial log likelihood can be interpreted as a measure of how well the Cox model fits the data for a given set of coefficients, $\boldsymbol{\beta}$.

The Cox model can be used to estimate the survival probability as a function of time by using the cumulative hazard, $H(t)$ [99]:

$$S_{Cox}(t) = S_0 \cdot \exp\left(-\int_0^t h(t)\right) = S_0 \cdot \exp(-H(t))$$

5.2 Contribution from study 3

Study 3 contributes to the clinical validation of DWI by investigating the prognostic value of longitudinal DWI in patients with locally advanced pancreatic cancer (LAPC). For this purpose, a multivariable Cox proportional hazards model for overall survival is made based on best-subset selection using cross validation. To address the potential benefit of longitudinal DWI over pre-treatment DWI, the analysis includes both DWI at baseline and DWI changes during the RT course. Moreover, a model-free decomposition method for derivation of DWI parameters is investigated as an alternative to the standard, model-based ADC.

6 Paper III

Prediction of overall survival in patients with locally advanced pancreatic cancer using longitudinal diffusion-weighted MRI

Anne L. H. Bisgaard*[a,b], Carsten Brink[a,b], Tine Schytte[b,c], Rana Bahij[c], Mathilde Weisz Ejlsmark[b,c], Uffe Bernchou[a,b], Anders S. Bertelsen[a], Per Pfeiffer[b,c], Faisal Mahmood[a,b].

[a]*Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense, Denmark*
[b]*Department of Clinical Research, University of Southern Denmark, Odense, Denmark*
[c]*Department of Oncology, Odense University Hospital, Odense, Denmark*

***Corresponding author**

Manuscript ready to be submitted for publication.

6.1 Abstract

Introduction

Biomarkers for prediction of outcome in patients with pancreatic cancer are wanted in order to personalize the treatment. This study investigated the value of longitudinal diffusion-weighted magnetic resonance imaging (DWI) for prediction of overall survival (OS) in patients with locally advanced pancreatic cancer (LAPC) treated with stereotactic radiotherapy (SBRT) on an MRI-linac.

Materials and methods

The study included 45 patients with LAPC who received five fractions of 10 Gy on a 1.5 T MRI-linac. DWI was acquired before each fraction. The analysis included the apparent diffusion coefficient (ADC) and DWI parameters obtained using a decomposition method within the gross tumour volume (GTV). Both the DWI parameters' baseline values and time-trends during the treatment course were investigated. A multivariable Cox proportional hazards model for OS was made based on best-subset selection, using cross-validation based on Bootstrap. The model's discriminating power was assessed using the C-Harrel index.

Results

The median OS from the first day of SBRT was 15.5 months (95% CI: 13.2-20.6), and the median potential follow-up time was 19.8 months. The best-performing multivariable model for OS included two decomposition-based parameters: one baseline and one time-trend parameter. The C-Harrel index was 0.754. An association between high baseline ADC values and reduced OS was observed, possibly due to the inclusion of necrotic regions in the GTVs. No association between the ADC time-trend and OS was observed.

Conclusion

Decomposition-based DWI parameters demonstrated prognostic value in LAPC. Results indicated that time-trends of DWI parameters are useful for response prediction, underlining the benefit of acquiring longitudinal DWI during the SBRT course. The investigated parameters may be potential predictive biomarkers for treatment response, and may thus be used to guide treatment interventions to improve the outcome in patients with LAPC.

6.2 Introduction

Pancreatic cancer is the fourth most common cause of cancer-related death in Europe, with a 5-year survival rate of less than 10% and an increasing incidence [103,104]. Twenty percent of the patients are eligible for curative-intend surgery, which increases the survival rate to about 20 % [105,106]. RT can be used to downstage tumours, making them eligible for surgery or as a definitive treatment. The development of IGRT, especially the recent introduction of hybrid MRI linear accelerators (MRI-linacs), have made hypofractionated stereotactic RT (SBRT) well-tolerated due to the ability to see and adapt the treatment to the position and daily shape of the sensitive organs at risk (OAR) surrounding the pancreas [11,16,107]. Still, the best RT treatment for patients with pancreatic cancer remains to be settled. A possible problem in interpreting the result from pancreatic cancer studies can be the considerable variation within the studied cohorts, which might relate to different responses to RT. Thus, biomarkers may allow a more informed treatment choice [108]. The impact of such biomarkers could be either dose escalation or de-escalation to obtain an optimal balance between survival and toxicity.

One promising candidate for such a biomarker is the apparent diffusion coefficient (ADC) derived from diffusion-weighted MRI (DWI). ADC quantifies the motion of water molecules, which indirectly reflect the tissue microstructure [29]. Several groups have investigated the value of ADC in detecting pancreatic cancer [109–113] and the relation between ADC and pathological response [114–116]. The correlation between overall survival (OS) and ADC has also been investigated [106,117,118], with findings indicating that pre-treatment ADC might contain more predictive information than standard clinical parameters such as age and tumour size. However, the predictive power of these studies was limited. A potential improvement can be sought through the MRI-linac, where DWI can be naturally integrated into the workflow in each treatment fraction. This setup allows investigation of whether changes in DWI parameters during the treatment course could add independent information to the prediction of the outcome for the patients.

The current study utilized longitudinal DWI data to investigate how baseline DWI parameters and changes in DWI parameters during the treatment course impact survival prediction in patients with locally advanced pancreatic cancer (LAPC). DWI parameters were derived using a data-driven method recently proposed by Rahbek et al. [31,119]. The potential advantage of data-driven approaches is that there are no initial model assumptions which could lead biased parameters. The standard ADC based on a mono-exponential model was included as comparison.

The overall aim was to create a survival model for patients with non-resectable locally advanced pancreatic tumours treated with SBRT on an MRI-linac using DWI parameters and clinical parameters, with the long-term goal of potentially adapting the treatment based on the response of the individual patient.

6.3 Materials and Methods

A multivariable Cox survival model was made to predict OS in patients with LAPC treated on a high-field MRI-linac (Elekta Unity). The model was based on both clinical parameters and parameters derived from DWI (the variable selection method is described in the statistics section below). The DWI parameters were extracted from the GTV volume using both a model-based method (ADC) and a model-free decomposition method (msNMF) (see the image analysis section below).

Before data analysis of the outcome, a statistical analysis plan (SAP) was created (see Appendix IIIc). The SAP states time to local progression as a second endpoint. However, due to too few local progression events (8 events), it was impossible to derive stable models; thus, time to local progression is not included in the analysis.

6.3.1 Patients and endpoints

The current study included patients diagnosed with LAPC, treated with SBRT for downstaging or definitive (consolidation) SBRT. No tumours were resectable at the time of diagnosis. Three patients underwent surgery after SBRT. All patients had either primary tumour or recurrence in the pancreas, with no tumour-involved lymph nodes. All patients received induction chemotherapy prior to SBRT and were treated with five fractions of 10 Gy on a 1.5 T MRI-linac (Unity by Elekta, Stockholm, Sweden). Only patients without visual artefacts in the GTV regions for at least two treatment fractions were included in the study. The follow-up after SBRT, including clinical examination and CT scan, was scheduled for every three months for two years. Only patients that attended the first three-month follow-up were included in the study (i.e., they were alive three months after SBRT), due to the initial intention of using local progression as a second endpoint. OS was defined as the time from the first SBRT fraction to death for any cause. All patients were prospectively included in the MOMENTUM study (an international partnership between several hospitals and the industy partner Elekta (Elekta AB, Stockholm, Sweden) [18]), although the current investigation was not initially planned. Inclusion of data in MOMENTUM was approved by the research ethics committee at the local institutional board (26/68031) and storage of data was approved by the Region of Southern Denmark (20/35211). Informed consent was obtained from all patients.

6.3.2 MRI protocols

Before treatment, all patients had T2-weighted MRI scans acquired on a 1.5 T diagnostic scanner (Ingenia, Phillips). Imaging before each treatment fraction at the MRI-linac included T2-weighted MRI scans and DWI scans. The majority of the patients (n=39) were DWI scanned using the b-values 0, 30, 80, 150 and 500 s/mm^2 (sequence 1), while six patients were scanned using b-values of 0, 20, 60, 100, 300, 800 and 1000 s/mm^2 (sequence 2). Acquisition details are presented in Table A3.1 and A3.2 in Appendix IIIa.

All DWI scans were visually inspected for artefacts. In some patients, an alternation of DWI signal between neighboring slices was observed, likely cross-talk due to magnetization saturation effects. The alternation was reduced by a pre-processing step in which all slices were convoluted with the neighboring slices with a weight of 0.25 on neighboring slices and 0.5 on the current slice. All evaluations were done by persons without knowledge of the clinical data.

6.3.3 Region of interest

GTV delineations from the treatment pre-plan (based on T2-weighted images from the diagnostic scanner) were used to measure the volume of the GTV at baseline. The GTV delineations from the daily adapted treatment plans (based on T2-weighted images from the MRI-linac) were the basis for extracting all DWI parameters.

The GTVs were transferred to DWIs without translational adjustment. A 5 mm margin was added to the GTVs to account for delineation uncertainty and possible misalignment of the GTV between T2-weighted images and DWIs. In the following, all DWI information obtained from the "GTV" refers to the GTV with the 5 mm margin. All image analysis was performed in Matlab (R2020b, Mathworks ab, Sweden).

6.3.4 ADC calculation

ADC was included in the current study as a standard, model-based method. ADC values were calculated using a mono-exponential Stejskal-Tanner model [33], using the b-values 150 and 500 s/mm^2 (DWI sequence 1) and 300 and 800 s/mm^2 (DWI sequence 2), to adhere as closely as possible to recommendations from the Elekta MRI-linac Consortium [51]. The b-value of 1000 s/mm^2 in DWI sequence 2 was excluded from the analysis, as the noise floor present for high b-values (b>=1000 s/mm^2) may lead to an underestimation of the ADC. The median ADC within the GTV was extracted.

6.3.5 Data-driven DWI decomposition

The monotonous slope non-negative matrix factorization (msNMF) method was included as a data-driven alternative to the standard model-based ADC analysis described above. A detailed description of msNMF is provided by Rahbek et al. [31]. The main idea is to extract the "typical" behaviour of the DWI signals using the observed data, which is somewhat similar to principal component analysis. However, in the msNMF, the aim is to obtain "typical" physically meaningful components, i.e., non-negative, monotonous, and with a monotonous slope as a function of the b-values of the DWI scans. The DWI signals over all the b-values can be described as a combination of the msNMF components (linear combination with only positive weights). Thus each DWI voxel (across all b-values) is represented by weights, one for each component. Mathematically, the components can be represented by a matrix, $\boldsymbol{W}$, and the corresponding weights by a matrix, $\boldsymbol{H}$. The decomposition can be written as:

$$\boldsymbol{X}_{mxn} = \boldsymbol{W}_{mxk}\boldsymbol{H}_{kxn} + \boldsymbol{E}_{mxn}$$

Here $\boldsymbol{X}$ contains the DWI signal for m b-values and n voxels, $\boldsymbol{W}$ contains k msNMF components, $\boldsymbol{H}$ describes the weights associated with each component for each voxel, and $\boldsymbol{E}$ is the residual left in case the number of components is less than the number of b-values (which is the typical case to perform data reduction). The decomposition can be used to reduce data since k components can describe m b-values.

The components ($\boldsymbol{W}$) were determined based on data from all patients scanned with DWI sequence 1 (n=39). For this purpose, all GTV voxels from all patients and all fractions were pooled to create the matrix $\boldsymbol{X}$. In the current study, a two, three and four components described 86%, 94%, and 96% of the initial data variance, respectively. Based on these values, the current study used three msNMF components (k=3).

The DWI signal within each voxel was normalized by dividing by the signal intensity in b=0 s/mm^2, as the signal behavior is of interest and not the absolute signal intensity. Overall, DWI signals should decrease as a function of the b-value, however, image noise can disturb the measured signal between b-values, particularly for voxels with low signal values. To reduce the impact of noise, voxels with normalized signal values above 1 for b≥80 s/mm^2 were excluded before calculating the msNMF components. Further, voxels were excluded if a linear fit to the logarithm of the signal had a positive slope (voxels with a non-decreasing signal). As a result, 22% of the voxels were excluded based on this noise reduction approach. Details regarding the use of msNMF in the current study are provided

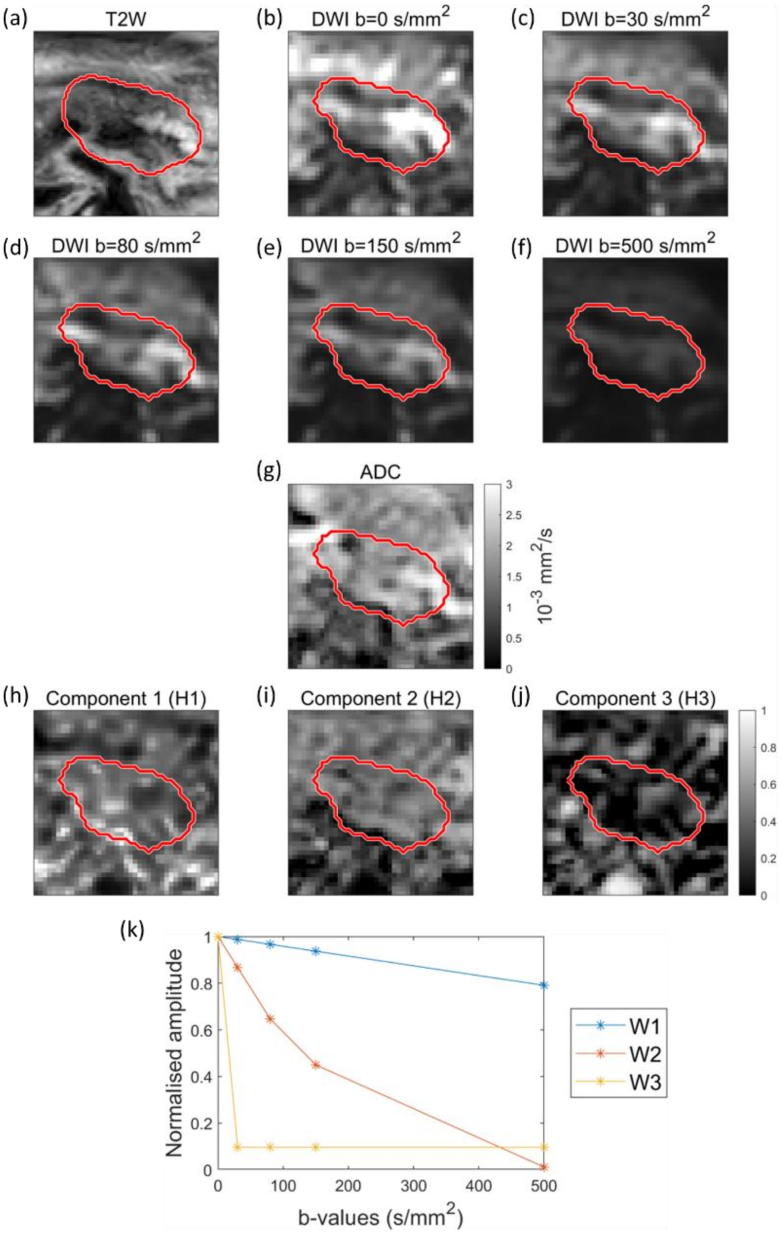


Figure 6.1. Example of T2-weighted image (a), DWI images (b-f) and ADC map (g) for a patient. Example of weight maps (**H**) corresponding to the three components (**W**) (h-j). Components (**W**) resulting from decomposition analysis (msNMF) based on a total of 190 DWI scans from 39 patients (k).

in Appendix IIIb. In the following, H1, H2, and H3 refer to the weights associated with the components W1, W2 and W3, respectively. The extracted components can be seen in Figure 6.1 together with example scans.

6.3.6 Derivation of DWI parameters

As a result of msNMF analysis, each voxel within a DWI scan can be represented by three weights, one weight per component. These weights can be thought of as the "amount" of

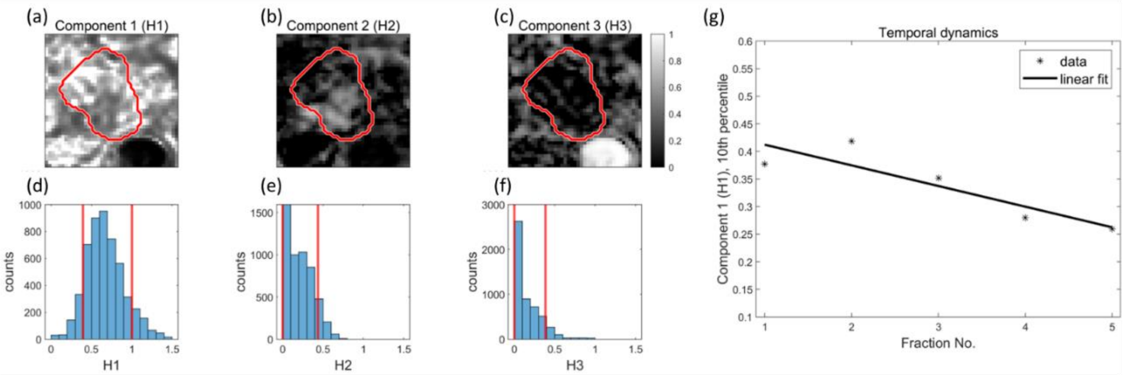


Figure 6.2. Derivation of decomposition-based DWI parameters for outcome prediction. Examples of weight maps for a patient (a-c). The red contour represents the GTV. The distribution of weights within the GTV (3D volume) are presented using histograms (d-f). The red, vertical lines represent the 10th and 90th percentiles of the distributions. Temporal dynamics are extracted from each percentile using a linear fit to the data as a function of fraction number (g).

each component within the voxel. The spatial distribution of weights can be presented in so-called weight maps, as presented in Figure 6.2 (a-c) (corresponding to H1, H2, and H3).

The weight distributions within the GTV were extracted, and the 10th and 90th percentiles were calculated (Figure 6.2, d-f). The 10th and 90th percentiles were chosen in order to characterize changes in the shape of the weight distributions. The process was repeated for each of the five fractions, and a time-trend of each percentile was extracted using a linear fit, as illustrated in Figure 6.2, g. The slope of the linear fit and the value at fraction one were used as parameters in the statistical analysis. This resulted in a total of 12 decomposition-based parameters (3 components x 2 percentiles x (slope + value at fraction one)). These decomposition-based parameters were given names referring to the component (H1, H2, or H3), the percentile (10th or 90th percentile), and whether the value represents the slope or the value at fraction one. For example, "H1_prc10_slope" refers to the slope of the 10th percentile of weights for component one.

Likewise, a time-trend was extracted for the median ADC. The slope of the fit as well as the ADC value at fraction one were included in the statistical analysis.

6.3.7 Statistical analysis

The study aimed to create a multivariable Cox proportional hazards survival model with OS as the endpoint. Patients were censored if they were still alive at the cut-off date for data collection (30th March 2023). In total, 14 DWI-derived parameters (12 decomposition-based and 2 ADC) and 6 clinical variables were included in the analysis. The clinical parameters at the time of SBRT were age, GTV volume at baseline, time between diagnosis and SBRT, sex, performance status (PS), and primary tumour/recurrence, out of which the three latter were categorical variables. The continuous variables were standardized by subtracting the mean and dividing by the standard deviation.

The parameter selection was based on best-subset selection using bootstrap-based cross-validation. Thus, for each combination of potential variables for the multivariable model ($2^{\text{No. of parameters}}=2^{20} \approx 1$ mio. models), a bootstrap of the initial data was made. Those patients included in the boot (in-boot patients) were used to calculate the regression constants, and the cross-validated model likelihood was subsequently calculated using the patients not included in the Bootstrap (out-of-boot patients). The number of out-of-boot patients will vary per boot, and the computed likelihood was divided by the number of out-of-boot patients, as suggested by Schemper [120]. The entire bootstrap process was iterated 50 times, and the mean cross-validated likelihood was obtained. The selected multivariable model was the model performing the best during the cross-validation. For the best model, model regression constants (β) and corresponding 2-sided 95% CI were established using 2000 bootstraps defined as the range of the 95% most central values (2.5 % removed at each tail).

The multivariable Cox model was validated by dividing the patients into low-, medium- and high-risk groups. Kaplan-Meier survival curves and model estimates were compared for each risk group. The risk groups were defined based on the value of the linear predictor for each patient ($\sum_i \beta_i x_i$). The division was made such that the high- and low-risk groups each contained 25% of the patients, and the remaining 50% were in the medium-risk group. Besides such calibration plots, the model's ability to discriminate between patients with high and low risk (i.e. the linear predictor) was reported as the C-Harrell index, which measures the fraction of patient pairs in which the one with the lower risk survives the longest.

The median survival time for the entire cohort was assessed using the Kaplan-Meier estimator, and the median potential follow-up time was calculated using the inverse Kaplan-Meier method. Although not needed for the multivariable model selection, univariable Cox models were performed to provide an overview of the entire data set.

Statistical analysis was performed using R (Version 4.3.1).

6.4 Results

6.4.1 Included patients

In total, 50 patients fulfilled the inclusion criteria (see section 6.3.1). Out of these, five patients were excluded before statistical analysis, as their DWI sequences differed substantially from those of the remaining patients (DWI sequence 1 and 2). No patients were excluded due to artefacts in the GTV region. For two patients, DWIs were missing for two and three fractions, respectively, however, they were kept as part of the analysis. The analysis was thus based on 45 patients (demographics shown in Table 6.1).

6.4.2 Decomposition

The decomposition analysis (msNMF) resulted in three well-separated components ($\boldsymbol{W}$): one long lifetime component (W1), one intermediate lifetime component (W2), and one component mainly driven by b=0 s/mm^2 (W3) (Figure 6.1, k). An example of weight maps ($\boldsymbol{H}$) associated with each component is shown for one patient (Figure 6.1, h-j). The weight maps reveal tissue heterogeneity, highlighting different regions within the GTV. Visually, the weight maps provide spatial information about the weight variation of the individual msNMF component extracted from the DWI images (Figure 6.1, h-j). It should be noted that the weight maps do not correlate 1:1 with the DWI images or the ADC map and thus provide complementary information (Figure 6.1, b-g).

6.4.3 Overall survival

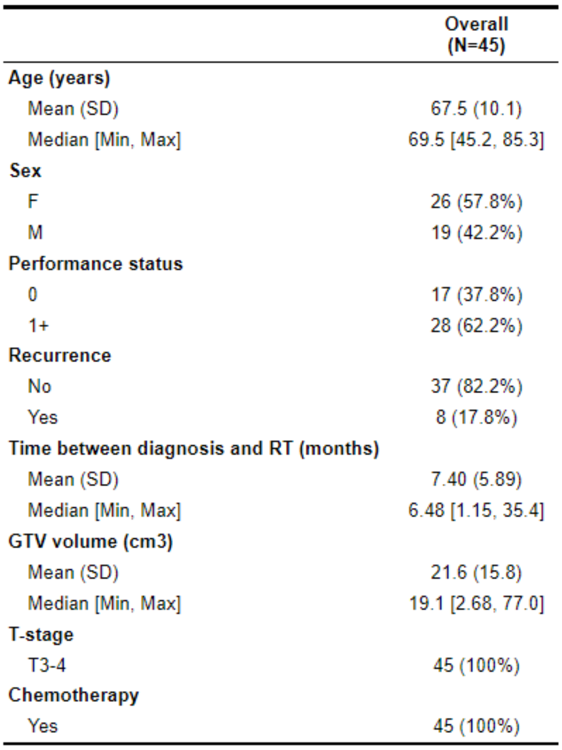
The median OS was 15.5 months (95% CI: 13.2-20.6) (see Kaplan-Meier plot in Figure A3.1 in Appendix IIIa). The median potential follow-up time was 19.8 months.

A univariable analysis was performed to provide an overview of the association between each variable and OS (Figure 6.3). None of the clinical parameters showed a statistically significant association on a 5% level; however, the association between age and OS was borderline significant. Several of the DWI-derived parameters, including ADC at the first treatment fraction, showed a statistically significant association with OS. It might be noted that for the decomposition-based parameters H1 slope and H2 at fraction 1, the hazard ratio was similar for the 10% and 90% percentile values.

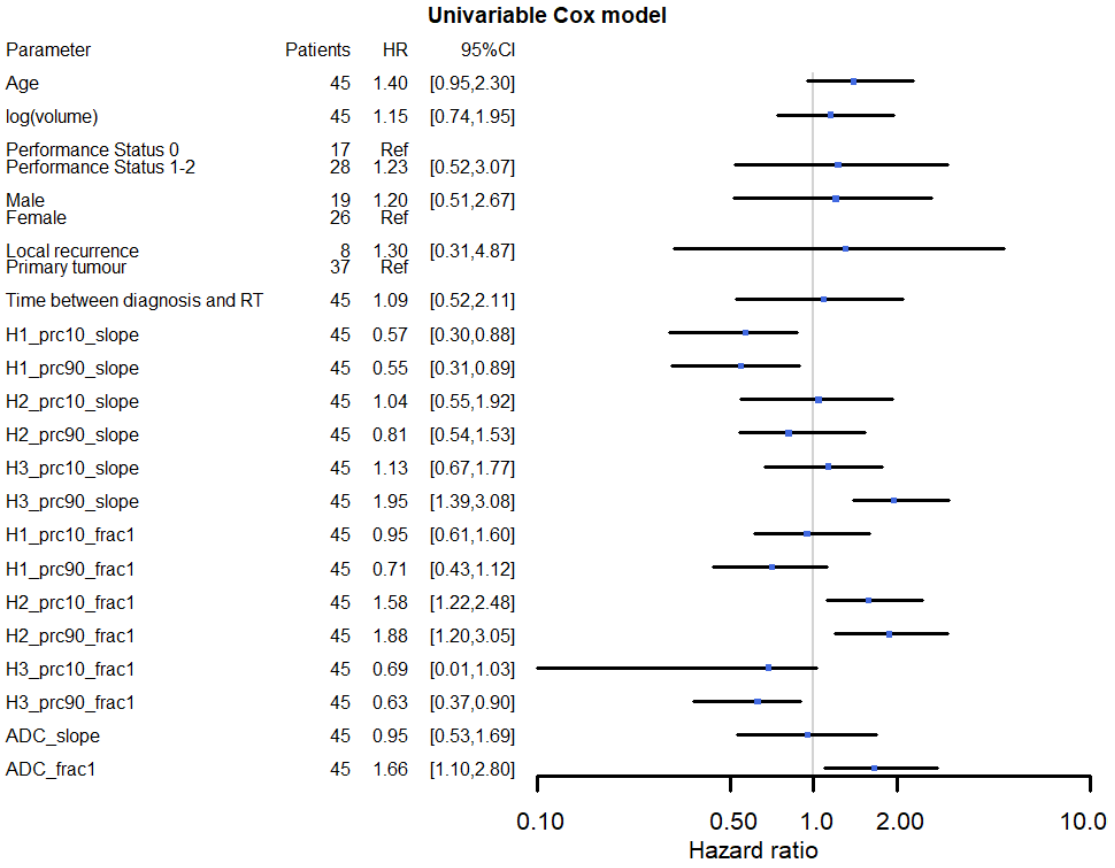
The best-performing cross-validated multivariable Cox model for OS included two decomposition-based parameters ("H1_prc10_slope" and "H2_prc90_frac1") (Figure 6.4). The C-Harrell conformance index for the best model was 0.754. A calibration plot showing the Kaplan-Meier survival curve and the model performance divided into risk groups is presented in Figure 6.5, demonstrating good agreement between the model and data.

Clinically it can be of interest to make outcome predictions as early as possible (e.g., at fraction one and not after fraction five); thus, the individual performance of the two predictors in the best-performing model is shown in Figure A3.2 in Appendix IIIa. The model's performance is obviously reduced when the parameters are used separately and have C-Harrel indexes of 0.633 and 0.688 for H2_prc90_frac1 and H1_prc10_slope, respectively.

Table 6.1. Patient demographics for 45 patients receiving SBRT for localized panctreatic cancer. The reported T-stage is at the time of diagnosis. No re-staging was performed at the time of recurrence.



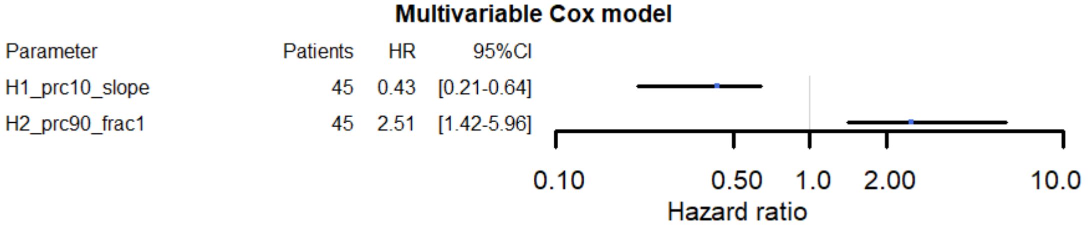
	Overall (N=45)
Age (years)	
Mean (SD)	67.5 (10.1)
Median [Min, Max]	69.5 [45.2, 85.3]
Sex	
F	26 (57.8%)
M	19 (42.2%)
Performance status	
0	17 (37.8%)
1+	28 (62.2%)
Recurrence	
No	37 (82.2%)
Yes	8 (17.8%)
Time between diagnosis and RT (months)	
Mean (SD)	7.40 (5.89)
Median [Min, Max]	6.48 [1.15, 35.4]
GTV volume (cm3)	
Mean (SD)	21.6 (15.8)
Median [Min, Max]	19.1 [2.68, 77.0]
T-stage	
T3-4	45 (100%)
Chemotherapy	
Yes	45 (100%)



Univariable Cox model

Parameter	Patients	HR	95%CI
Age	45	1.40	[0.95,2.30]
log(volume)	45	1.15	[0.74,1.95]
Performance Status 0	17	Ref	
Performance Status 1-2	28	1.23	[0.52,3.07]
Male	19	1.20	[0.51,2.67]
Female	26	Ref	
Local recurrence	8	1.30	[0.31,4.87]
Primary tumour	37	Ref	
Time between diagnosis and RT	45	1.09	[0.52,2.11]
H1_prc10_slope	45	0.57	[0.30,0.88]
H1_prc90_slope	45	0.55	[0.31,0.89]
H2_prc10_slope	45	1.04	[0.55,1.92]
H2_prc90_slope	45	0.81	[0.54,1.53]
H3_prc10_slope	45	1.13	[0.67,1.77]
H3_prc90_slope	45	1.95	[1.39,3.08]
H1_prc10_frac1	45	0.95	[0.61,1.60]
H1_prc90_frac1	45	0.71	[0.43,1.12]
H2_prc10_frac1	45	1.58	[1.22,2.48]
H2_prc90_frac1	45	1.88	[1.20,3.05]
H3_prc10_frac1	45	0.69	[0.01,1.03]
H3_prc90_frac1	45	0.63	[0.37,0.90]
ADC_slope	45	0.95	[0.53,1.69]
ADC_frac1	45	1.66	[1.10,2.80]

Figure 6.3. Hazard ratios and 95% confidence intervals for univariable Cox proportional hazard models for overall survival. In the graphical representation, the confidence interval for the parameter "H3_prc10_frac1" was cut off at 0.10 in order to make a more well-balanced figure.



Multivariable Cox model

Parameter	Patients	HR	95%CI
H1_prc10_slope	45	0.43	[0.21-0.64]
H2_prc90_frac1	45	2.51	[1.42-5.96]

Figure 6.4. Hazard ratios and 95% confidence intervals for the multivariable Cox proportional hazard model for overall survival. The variables included in this model were selected by the cross-validation process.

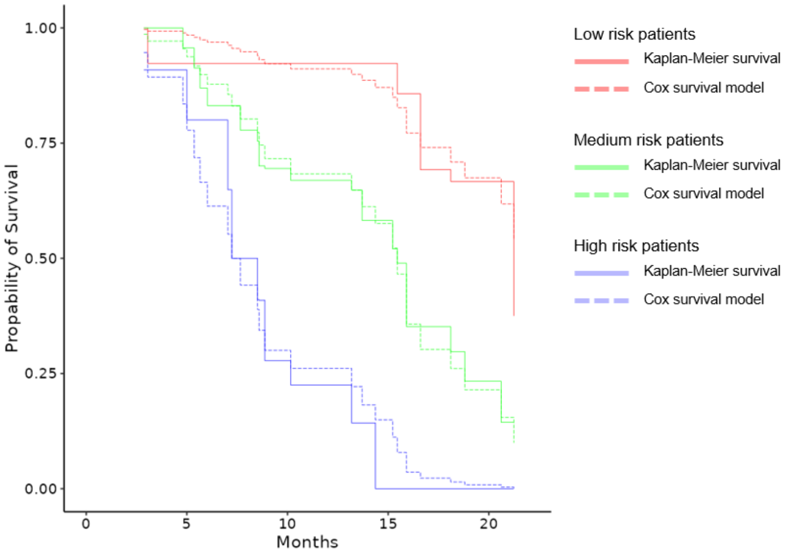


Figure 6.5. Comparison of the multivariable Cox model and the Kaplan-Meier estimator for the best-performing model for overall survival. The model included the decomposition-based parameters "H1_prc10_slope" and H2_prc90_frac1" (see section 6.3.6). Patients were split into high, medium, and low-risk groups based on the 25% and 75% percentiles of the calculated linear predictors (-0.67 and 0.74), i.e. the high- and low-risk groups each contained 25% of the patients, and the medium-risk group contained 50% of the patients. The confidence intervals for each of the risk groups were overlapping (not shown).

6.5 Discussion

This study aimed at predicting OS in 45 patients with LAPC utilizing parameters derived from longitudinal DWI using both a standard, model based approach (ADC) and a model-free decomposition approach (msNMF). To our knowledge, this study was the first to investigate the value of longitudinal DWI in predicting OS in patients with LAPC.

The best model for OS prediction was found to include only two parameters (H1_prc10_slope and H2_prc90_frac2), both of which were based on DWI decomposition. None of the clinical parameters were selected by the cross-validation process. The best model reached a C-Harrell index of 0.754 indicating that the model is good at determining which of two patients will survive the longest (a value of 0.5 corresponds to a random guess, and a value of 1 means that the model can perfectly rank the patients' survival times). The best model represented both information from the baseline (H2_prc90_frac1) and the longitudinal change of the DWI signals (H1_prc10_slope). Moreover, the C-Harrell index was markedly reduced if the model was based on just one of the parameters compared to the best model (Figure A3.2 in Appendix IIIa), indicating that both baseline and temporal information is important for the best survival prediction.

The fact that decomposition-based parameters were selected by the cross-validation process instead of the ADC information indicates that the decomposition-based parameters might be more stable than the standard ADC values. Increased stability could be related to the fact that the extracted components are based on data from all patients and fractions, hereby reducing the impact of noise. The C-Harrell index for a model based on ADC and ADC-slope was 0.623 (data not shown), indicating that the discriminating power was reduced if the model was based only on the standard ADC value. That being said, it can be seen in Figure A3.4 in Appendix IIIa that the Pearson correlation between ADC_frac1 and H2_prc90_frac1 is 0.78 showing that most of the information in H2_prc90_frac1 is also present in the ADC signal. Thus the main difference between the standard ADC approach and the msNMF approach is mainly within the longitudinal changes in which msMNF predict survival much better than the temporal ADC signals (ADC_slope).

It is seen that H2_prc90_frac1 is a risk factor indicating that increasing values of H2_prc90_frac1 will reduce the expected lifetime. This is in line with the univariable result in Figure 6.3, showing that a high baseline ADC value also is a risk factor. Initially, these results seem to contradict the finding in some other studies, in which an association between low ADC and poor survival was reported [117,118,121]. However, the explanation is likely related to the GTV region from which the DWI parameters were derived. The current study aimed to use the clinically available GTV so that no additional delineation was needed. In contrast, the previous studies focused on the delineation of the "viable" part of the tumour, excluding necrotic and cystic parts. The different results regarding outcome might be due to inclusion of necrotic regions in the GTV used in the current study. This explanation is supported by a study by Lyng et al. [122], which shows that some necrotic regions are related to increased ADC values. Further, necrosis has been related to worse outcomes in patients with LAPC [123]. Thus, the initial counterintuitive behavior might be related to the volume used for the DWI analysis.

It may be important to note that the univariable results of the two parameters in the best-performing model have very similar hazard ratios. Furthermore, the correlation of the 10% and 90% percentile values for H2 frac1 and H1 slope is 0.70 and 0.61, respectively (Figure A3.4 in Appendix IIIa). This indicates that they contain quite similar information. The average of the 10% and 90% percentile values for symmetrical distributions will be very close to the median. Therefore, it is likely that it was unnecessary to focus on the tails of the distribution and that the median values could have been used instead. To test this assumption, a calibration plot of a model based on the related median values of the best model is shown in Figure A3.5 in Appendix IIIa (i.e. a model based on "H1_prc50_slope" and H2_prc50_frac1"). This model had a C-Harrell index of 0.757, similar to the C-Harrel index of the best model (0.754). For future studies, it is likely as good (and simpler) to focus on the median value and not include the tails of the weight distributions.

Finally, it should be noted that none of the two variables in the best model correlated with any of the six included clinical parameters. The largest Pearson correlation between the model variables and the continuous clinical variables was 0.24, and the lowest p-value of a Mann–Whitney U test of best model variables between the levels of the categorical clinical variables was 0.53. So there is no indication that the DWI parameters could be a proxy for any of the clinical variables.

Although a careful variable selection based on cross-validation was performed to avoid overfitting, it is important to underline that the entire cohort consists of only 45 patients. This is considered a relatively large cohort compared to other RT studies within LAPC, but it is still a poor representation of the overall cohort of patients with LAPC treated worldwide. We thus hope that people with access to independent LAPC patient cohorts will be interested in performing validation of the presented findings.

6.6 Conclusion

A model predicting survival after SBRT for patients with LAPC based on only two DWI parameters has been developed. The model contains both baseline information and longitudinal DWI changes during the SBRT course and can quite clearly separate the patients into high- and low-risk groups. It is the hope that in the future, the model can assist in stratifying patients for individual treatment (e.g. dose escalation) based on their individual DWI responses. Furthermore, the model might help identify differences between LAPC cohorts, which might be needed to get a common understanding of the best treatment option for specific groups of patients with LAPC.

6.7 Acknowledgements

We thank Sofie Rahbek for providing Matlab code for the msNMF decomposition method code and for the assistance with implementation of the prediction framework.

PART III

Discussion, perspectives and conclusion

7 Discussion and perspectives

In the three papers presented in this thesis, steps have been taken towards an integration of DWI into the MRI guided RT workflow. Each paper investigated a different aspect of the technical and clinical validation of DWI, with the end goal of translating DWI into clinical use. This chapter includes a discussion of the findings under different topics regarding the technical and clinical validation, as well as future perspectives.

7.1 Delineation

Delineation of a ROI is an important aspect of ADC measurements, however, manual delineation gives rise to both inter-and intra-observer ADC variation, hampering both repeatability and reproducibility [54,55,57,124]. Further, manual delineation is time-consuming and must be performed by someone with a high level of expertise, e.g. a radiologist, making ADC measurements less feasible in a busy clinic. Study 1 investigated semi-automatic delineation as a potential solution to these challenges using an easy-to-implement, threshold-based approach. Automated delineation based on artificial intelligence (AI) was also considered an option, since it might allow fully automated delineation, however, it would require training on large data sets [69,70]. A threshold-based tool has the advantages that it does not require training, and that it is more transparent in in terms of the criteria for inclusion of voxels in the ROI, which might help translation of the tool to other tumour sites. Thus, a threshold-based approach was preferred.

The semi-automatic delineation tool (SADT) presented in study 1 performed comparable to manual delineation by a radiologist in terms of ADC repeatability and slightly better in terms of intra-observer ADC variation, demonstrating its potential for a more robust derivation of ADC values. Further, it did not require the same level of expertise as manual delineation, and might thus save time for the medical doctors in future investigations of ADC as a response biomarker.

One of the strengths of Study 1 is the use of test-retest DWI scans, which allowed comparison between the ADC repeatability and temporal ADC changes. The SADT demonstrated the ability to detect temporal ADC changes larger than ADC repeatability, which is crucial if the tool should prove useful for detection of RT-induced biological changes. An important question that remains unanswered is whether the observed ADC values and temporal changes of these correlate with clinical outcome. Further, although previous studies have investigated the use of temporal ADC changes for response prediction [26,67,125–127], the literature is limited regarding ADC changes during SBRT treatments on the MRI-linac which have a shorter duration compared to conventional RT [128].

Another important aspect of delineation which impacts ADC measurements is the delineation strategy, i.e. which type of ROI is delineated. The SADT delineated the so-called viable tumour volume (VTV) defined as the GTV excluding necrotic and cystic parts [35,58]. The VTV has been recommended for ADC calculation [35], however, it still remains unanswered if it is the best ROI for detecting RT-induced biological changes. For example, RT leads to cell death e.g. via necrosis, which is normally associated with an increase in the ADC value [122]. By excluding necrotic parts from the delineation, this ADC increase might not be detected, and hence information on RT-induced biological changes is lost [35]. Another option might be to use the GTV which is already available from the clinical RT workflow. However, since the GTV is usually based on a T2-weighted image, this requires registration between T2-weighted images and DWI, which introduces a registration error and uncertainties related to the geometric distortions in MRI (especially in DWI). Therefore, it might be preferable to delineate the ROI directly on the DWI. So far, very few studies have investigated the potential benefits of using the VTV in comparison to the GTV, thus, investigations of both are relevant [35,58].

It is the hope that the SADT (or similar tools) can be implemented and aid ADC measurements at other centres. However, the SADT has not yet been tested using other data sets, and thus, external validation of the SADT is needed. One concern in this regard is that the tumour-to-background ratio might differ between DWI sequences and tumour sites. It should be noted that the current data were acquired on a diagnostic scanner with a maximum b-value of 1100 s/mm^2, whereas the maximum b-value recommended for the MRI-linac (Unity, Elekta) is 500 s/mm^2 due to limitations of the MRI-linac hardware. Consequently, the tumour-to-background ratio is probably reduced on MRI-linac data, which might impact the ability of the SADT to distinguish between tumour and normal tissue. If the tumour-to-background ratio is too low, a threshold-based approach is unfit, and other delineation approaches should be considered instead, e.g. based on AI [129].

In the future, the SADT might become useful for dose painting purposes [22]. The idea of using a focal boost to improve chances of local control has already been investigated in prostate cancer, where the dose was escalated to a prostate sub-region defined using multi-parametric MRI, including DWI and dynamic contrast-enhanced MRI [28]. In rectum cancer, a homogeneous dose is normally prescribed to the entire rectum and mesorectum, however, chances of local control might be improved by using a focal boost to VTV, as it is represents a region of high tumour load [22,130].

Repeatable and reproducible ADC measurements are essential for detecting differences between time-points or patients, and are thus critical if ADC should be used for response prediction purposes. The SADT (or similar tools) may improve both repeatability and reproducibility of ADC measurements by reducing the delineation-related uncertainty, and therefore aid technical and clinical validation of ADC. It may also save time for the medical doctors who would normally perform the manual delineation, and hence make ADC measurements more feasible, both for research purposes and future clinical purposes.

7.2 Multi-centre reproducibility

Currently, there is a large variation in the methods used for ADC measurements across studies, which results in a poor reproducibility and hampers technical validation of ADC [20]. Many factors are known to impact ADC reproducibility, such as acquisition, delineation, and ADC calculation method, however, the relative importance of these factors is less clear. Study 2 aimed at providing recommendations for improved reproducibility of ADC measurements by evaluating the ADC variation between centres with respect to two categories: delineation and calculation method. The study was focused towards multi-centre ADC validation studies for the MRI-linac, and hence included nine MRI-linac centres across different countries. Recommendations regarding DWI acquisition for ADC measurements published by the image analysis working group within the Elekta MRI-linac consortium were followed in Study 2 [51].

The study design of Study 2 allowed a direct comparison of the different sources of variation in a multi-centre setting. Interestingly, it was found that the variation in calculation methods between centres had a larger impact on the ADC reproducibility compared to the delineation variation. A closer look revealed that the choice of b-values was the most important factor, but also choice of voxel-based versus ROI-based calculation and factors related to the software implementation such as filtering and fitting methods were important. Hence, the findings advocate reducing the calculation-related ADC variation between centres in order to improve reproducibility.

One could argue that differences in calculation methods between centres are less important for multi-centres studies, as long as the data analysis is performed in a centralized manner. However, it is important to be aware of calculation-related ADC variation if findings should be compared between studies. Further, if ADC should be used in the future as a decision making tool in a clinical setting, it might be necessary to establish common cut-off values to discriminate between good and poor responders, which may be difficult to do with varying methods across centres. Since the tools available for ADC calculation differs between centres, it might be useful to validate local tools against a commonly available algorithm in order to detect biases [19,35,131].

Besides reducing calculation-related ADC variation, Study 2 indicated that it is worthwhile to reduce the delineation variation in order to improve ADC reproducibility. One way to achieve this is to adhere strictly to delineation guidelines, and to ensure that delineation procedures are aligned between centres [124,132]. Automated delineation might be another way to reduce the delineation-related ADC variation, which also has the potential benefit of saving time compared to manual delineation. The potential benefit of using delineation guidelines and semi-automatic delineation was investigated as a sub-analysis, utilizing the same data as in study 2, although the results were not included in the publication (see Figure A4.1 in Appendix IV). Here, a slight improvement in delineation agreement was observed when using the SADT developed in Study 1, supporting semi-automatic delineation as a potential solution. No clear benefit was observed when manual delineation was performed using a set of instructions ("guidelines") for VTV delineation. However, these instructions did not represent detailed delineation guidelines, which might still be useful [35].

The main limitation of Study 2 is the small number of patients (only one per cancer diagnosis). As a result, the findings may not be directly translatable to other patients with the same diagnoses. However, the underlying trend regarding the relative contribution from delineation and calculation methods to the ADC variation is still informative.

Study 2 did not provide a specific recipe for ADC calculation, but rather sought to bring awareness of the factors, which are important for ADC reproducibility. This might aid comparison between studies, as well as future multi-centre validation studies of ADC as a response biomarker. As described by van Houdt et al., important steps towards a technical and clinical validation of imaging biomarkers are the integration of a technical validation into clinical trials as well as the collection of multi-centre data (e.g. the MOMENTUM database [18]) [48]. In this context, the recommendations provided in study 2 could be used as a checklist to obtain reproducible ADC measurements across centres.

7.3 Longitudinal DWI for response prediction

The MRI-linac has made acquisition of longitudinal DWI feasible, which holds a great potential for detecting biological changes. Both pre-treatment ADC values and changes in ADC values have shown potential for response prediction [26], and but only a few studies included ADC measurements from more than two time-points during the RT course [67,125–127]. Hence, the potential benefit of using longitudinal DWI for response prediction still needs to be investigated.

Study 3 showed that longitudinal DWI has prognostic value for overall survival in patients with LAPC treated with stereotactic RT on the MRI-linac. Interestingly, both DWI parameters derived before RT and time trends during the RT course seemed to be important to obtain the best prediction, suggesting that a combination of pre-treatment and longitudinal DWI is useful for response prediction, compared to only pre-treatment DWI.

Although the ADC has been widely studied as a biomarker, it does not necessarily represent the most optimal way to derive information from DWI data for response prediction purposes. Study 3 investigated a data-driven alternative to the model-based ADC, namely a method based on decomposition of the DWI signal [31,119]. The considerations behind the choice of the decomposition method were the following: The use of models to derive parameters generally introduces a risk of violating the model assumptions, which might lead to biased parameters. For example, the mono-exponential model used for ADC derivation assumes a mono-exponential DWI signal decay as a function of b-values, however, in reality, the DWI signal within a voxel often does not exhibit a mono-exponential decay. This is both due to partial volume effects, i.e. the fact that each voxel may contain sub-compartments with different tissue types [133], perfusion effects which causes a more rapid decay [41,51], and non-gaussian diffusion behaviour due to restricted diffusion [42]. Therefore, it might be desirable to decompose the DWI signal into signal components, reflecting the different sub-compartments within a voxel, in order to better characterize the tissue. A decomposition method might be less sensitive to partial volume effects compared to the model-based ADC, and moreover, it might be more robust to image noise within individual DWI acquisitions, as the components are extracted based on a pooled dataset consisting of DWI images from all patients and all RT fractions.

Interestingly, Study 3 found that the best prediction of survival was obtained using only two parameters, both of which were based on decomposition of the DWI signal. This indicates that the decomposition-based parameters carry information that is useful for response prediction, and may even be a better choice compared to the ADC. Moreover, none of the clinical parameters showed a statictically significant association with overall survival, suggesting that DWI could be a more useful predictor compared to standard clinical parameters. This is in agreement with a previous study reporting no association between age, sex or tumour size and overall survival [118]. A negative correlation between ADC and overall survival was observed, whereas previous studies have found a positive correlation [117,118]. A possible explanation of this difference might be the inclusion of necrotic regions in the ROI used for ADC measurements in Study 3, as necrosis has been associated with high ADC values and poor overall survival [122] [123]. In contrast, the previous studies used ROIs delineated on DWI images excluding necrotic regions.

One of the strengths of the study was the use of cross validation for selection of parameters for the best prediction of overall survival, which was performed to avoid the risk of overfitting. It should be kept in mind however, that the patient cohort was relatively small (45), and may not be representative of all patients with LAPC. Therefore, the findings need external validation.

Overall survival depends on several factors related to both the patient, the tumour-type and -stage and the treatment. The correlation found between DWI parameters and overall survival could indicate a relation between the tumour microstructure and the aggressiveness of the tumour. Indeed, previous studies have shown that DWI provide information of cell density and necrosis, factors that could be related to the aggressiveness of the tumour and response to RT [122]. Moreover, one study found that ADC is correlated with the development of distant metastasis in patients with LAPC [117], indicating a possible relation between DWI and the aggressiveness of the disease.

Since RT is a local treatment with the purpose of achieving local control, it would be interesting to study if early changes in DWI during RT could be used to predict time to local progression. In the initial statistical analysis plan (Appendix IIIc), time to local progression was included as an endpoint in addition to overall survival. However, due to a small number of events (8 out of 45), it was not possible to develop a prediction model for time to local progression. Hence, it could be a topic for future investigations.

Study 3 demonstrated the prognostic value of longitudinal DWI, which is an important step towards a clinical validation of longitudinal DWI for response prediction in patients with LAPC. It still remains unanswered if longitudinal DWI has predictive value, e.g. if it can be used to predict the patients' response to RT. If this is the case, longitudinal DWI could potentially be used in the future to discriminate between poor and good responders, and thereby help select patients for intensified RT treatment to improve outcome.

7.4 Resources

Besides the challenges of standardizing methods for biomarker measurements, as described in section 7.1-2, it is worth mentioning a few practical challenges in performing biomarker studies. Investigations of quantitative imaging biomarkers require time and resources, both for the acquisition and the analysis of functional MRI images and the collection of clinical data. Although daily acquisition has become much easier with the MRI-linac, the image analysis pipeline itself can make it time-consuming to derive quantitative imaging biomarkers. The analysis steps may include anonymization of data, data storage, image registration, visual inspection of the images, delineation of ROIs performed by experts and processing using mathematical models or other analysis tools. The analysis may be partly or completely performed using in-house software, which it requires resources to develop. Commercial softwares can be used for analysis purposes, however, they may be less transparent, and may not allow for advanced analysis methods. A more streamlined analysis pipeline would be a great help for biomarker investigations, as it would allow a more efficient use of the researchers' time. Based on experiences during the current Ph.D., investigations of ADC (or other DWI parameters) might benefit from a more systematic approach with respect to data storage and processing. An integration of semi-automatic delineation tools in the image analysis pipeline, such as the tool presented in study 1, might also be useful in saving time and resources, and thus make biomarker research more feasible.

7.5 Geometric accuracy

An important topic which has not been addressed in the three papers is the geometric accuracy of ADC (and MRI biomarkers in general) [30]. While "accuracy" often refers to the true ADC value, geometric accuracy refers to the spatial information of the ADC measurements. The geometric accuracy may be impacted by geometric distortions present in DWI, which might lead to a misalignment of voxels in relation to the patients anatomy. Hence, besides repeatability and reproducibility, an evaluation of the geometric accuracy of ADC is important for a successful translation of ADC into the clinic.

Geometric distortions in MR images are due to inhomogeneities of the magnetic field which arise from the scanner itself, e.g. inhomogeneities of the so-called B0-field, and from magnetic susceptibility differences within the patient. DWI is acquired using echo planar imaging (EPI) readout scheme, which allows fast acquisition, but suffers from a large sensitivity to field inhomogeneities [53,134]. This makes DWI is more prone to geometric distortions compared to e.g. T2-weighted MRI. Geometric distortions can be expressed as compression, dilation, shift and shear of the images. Further, due to the different magnetic susceptibility of tissue and air, signal may be “displaced” near tissue/air interfaces, which can result in signal voids and signal pile-up, i.e. false dark and bright regions in the image. Susceptibility artefacts are often observed in the rectum or the intestines due to the presence of gas [135].

The geometric distortions of DWI can cause a misalignment of the tumour between T2-weighted images and DWI. As a result, propagations of ROIs from a T2-weighted image to a DWI image may lead to missing tumour voxels or inclusion of non-tumour voxels in the ADC calculation. On the other hand, ROIs delineated on DWI might not have the correct position in relation to the patient’s anatomy. This could cause problems if the idea is to use DWI to delineate regions for dose escalation. Geometric inaccuracies may be less important for ADC measurements for response prediction, however, since geometric distortions can vary between DWI acquisitions, they might impact the repeatability and reproducibility of ADC.

It is possible to reduce the geometric distortions DWI acquired using EPI to some degree, by using appropriate acquisition settings (e.g. parallel imaging or segmented k-space acquisition [136]), and the use of post-processing steps such as B0-field map correction [137], correction of susceptibility distortions [138], and deformable image registration [19,139]. Nevertheless, studies investigating the impact of geometric distortions on ADC measurements are wanted, as the literature on this topic is still limited.

7.6 The next step

Study 2 has contributed to a technical validation of ADC by providing recommendations regarding the image analysis of DWI in order to improve the multi-centre ADC reproducibility. A natural next step would be to move towards multi-centre validation studies with the aim of determining the ADC repeatability and reproducibility in larger cohorts, and correlating ADC with clinical outcomes to validate its prognostic value. This requires collaboration and sharing of data between centres. An example of such a collaboration is the MOMENTUM study [18], an international partnership which facilitates the sharing of both clinical and technical data between institutions, and thus might aid a technical and clinical validation of ADC and other potential MRI biomarkers.

One way to obtain the necessary DWI data for assessment of repeatability and reproducibility could be to integrate a quality assurance step and acquisition of test-retest DWI data in clinical trials, as suggested by van Houdt et al. [48]. Such an approach could also be used to establish the DWI reproducibility between diagnostic MRI scanners and MRI-linacs, and hereby improve the generalizability of the findings.

Further down the line, the predictive value of DWI must be validated in relation to RT, since only predictive biomarkers can be used to predict which patients will benefit from specific treatments [48]. For this purpose, large interventional clinical trials are needed. Such trials could be designed to test the potential benefit of selecting patients for dose escalation, dose de-escalation or online treatment adaptation (i.e. dose painting) based on DWI. When designing such trials, there are a some challenges that must be solved. One is to develop a strategy for translation of biomarker measurements into treatment interventions. E.g. for dose painting, quantitative biomarker maps need to be translated into dose-plans [23,48]. Secondly, the time-point at which biomarker changes can be detected during the treatment course should be determined, in order to choose the optimal time point for stratification of patients based on treatment response [48,65].

In order to ensure scalability of the findings from investigations of DWI biomarkers, it would be an advantage to use DWI sequences available on clinical MRI-guided RT systems, and to make a common image analysis available, to make biomarker measurements feasible at both expert and non-expert centres [19,35,131].

7.7 Bridging the gap – the future role of DWI

In the future, DWI might play a role both before, during and after RT [48]. Pre-treatment DWI might provide prognostic and predictive information, and hereby aid decisions regarding the treatment strategy for the patients. During the on-line plan-adaptation in the daily MRI-linac workflow, DWI could be used as a tool for dose painting, e.g. by using ADC maps to define biological target volumes within the tumour. Thus, a heterogeneous dose distribution might be shaped to match the tumour biology, to obtain the most optimal tumour control. Potentially, early changes in the tumour ADC might be used to predict response to RT, and might thus be used to guide treatment interventions, such as escalation or de-escalation of the dose. Lastly, DWI might help guide the next treatment steps after RT. It is the hope that DWI will help personalize the treatment, in order to improve tumour control and reduce toxicity for the patients.

8 Conclusions

This thesis addressed some of the challenges regarding a technical and clinical validation of DWI. One challenge is the manual delineation, which is observer-dependent and requires time and expertise. Study 1 showed that the use of semi-automatic delineation can reduce the intra-observer ADC variation compared to manual delineation in patients with rectal cancer. Moreover, it can be used to detect temporal ADC changes larger than the uncertainty associated with the measurements, indicating that changes in the tumour biology can be captured. These findings suggest that simple, threshold-based delineation tools could aid future ADC investigations by improving the consistency of ADC measurements and by saving time during the delineation process.

Study 2 investigated the delineation-related and calculation-related ADC variation in a multi-centre setting, in order to provide recommendations for improved multi-centre ADC reproducibility. Interestingly, the largest contributor to the ADC variation was differences between calculation methods. The calculation-related ADC variation was mainly driven by different choices of b-values, but also choice of voxel-based vs. ROI-based calculation and software-related factors such as fitting and filtering methods were important. The results show how important it is to carefully consider the ADC calculation procedure when planning multi-centre studies, and that a detailed description of the calculation procedure is needed in order to compare results across studies. It is the hope that awareness of the factors causing ADC variation can help improve ADC reproducibility, and hereby aid future multi-centre investigations of ADC as a biomarker for biologically guided RT.

Study 3 demonstrated the prognostic value of longitudinal DWI in patients with LAPC treated with SBRT on an MRI-linac. The best prediction of overall survival was obtained when both a baseline parameter and a time-trend DWI parameter were included in the prediction model, indicating that longitudinal DWI is beneficial compared to pre-treatment DWI only. Moreover, study 3 also indicated a potential benefit of deriving parameters using decomposition of the DWI signal, since the decomposition-based parameters had a larger discriminating power compared to the standard ADC parameter.

All in all, study 1 and 2 have contributed to the technical validation of ADC, study 1 by providing a semi-automatic delineation tool for robust extraction of ADC values from DWI images, and study 2 by providing recommendations for improving the ADC reproducibility in a multi-centre setting. Study 3 demonstrated a relation between parameters derived from longitudinal DWI and clinical outcome, which is a step towards clinical validation of DWI.

If DWI crosses the translational gaps to the clinic, it may become a useful tool for decision-making in the MRI-guided RT workflow. Potential uses include dose painting and prediction of response to RT, which could be used to guide treatment interventions. It is the hope that DWI might help personalizing the RT treatment, in order to improve the outcome for patients with cancer.

References

[1] Deo SVS, Sharma J, Kumar S. GLOBOCAN 2020 Report on Global Cancer Burden: Challenges and Opportunities for Surgical Oncologists. Ann Surg Oncol. 2022;29:6497–6500.

[2] Baskar R, Lee KA, Yeo R, et al. Cancer and radiation therapy: Current advances and future directions. Int J Med Sci. 2012;9:193–199.

[3] Minniti G, Goldsmith C, Brada M. Radiotherapy. Handb Clin Neurol. 2012. p. 215–228.

[4] Emami B, Lyman J, Brown A, et al. Tolerance of normal tissue to irradiation. Int J Radiat Oncol Biol Phys. 2013;21:109–122.

[5] Thwaites DI, Tuohy JB. Back to the future: the history and development of the clinical linear accelerator. Phys Med Biol. 2006;51.

[6] Feng FY, Kim HM, Lyden TH, et al. Intensity-Modulated Radiotherapy of Head and Neck Cancer Aiming to Reduce Dysphagia: Early Dose-Effect Relationships for the Swallowing Structures. Int J Radiat Oncol Biol Phys. 2007;68:1289–1298.

[7] Wang-Chesebro A, Xia P, Coleman J, et al. Intensity-modulated radiotherapy improves lymph node coverage and dose to critical structures compared with three-dimensional conformal radiation therapy in clinically localized prostate cancer. Int J Radiat Oncol Biol Phys. 2006;66:654–662.

[8] Jaffray DA, Siewerdsen JH, Wong JW, et al. Flat-panel cone-beam computed tomography for image-guided radiation therapy. Int J Radiat Oncol Biol Phys. 2002;53:1337–1349.

[9] Burnet NG, Thomas SJ, Burton KE, et al. Defining the tumour and target volumes for radiotherapy. Cancer Imaging. 2004;4:153–161.

[10] Raaymakers BW, Lagendijk JJW, Overweg J, et al. Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept. Phys Med Biol. 2009;54:N229--N237.

[11] Raaymakers BW, Jürgenliemk-Schulz IM, Bol GH, et al. First patients treated with a 1.5 T MRI-Linac: Clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment. Phys Med Biol. 2017;62:L41–L50.

[12] Bertelsen AS, Schytte T, Møller PK, et al. First clinical experiences with a high field 1.5 T MR linac. Acta Oncol (Madr). 2019;58:1352–1357.

[13] Ng J, Gregucci F, Pennell RT, et al. MRI-LINAC: A transformative technology in radiation oncology. Front Oncol. 2023;13:1–16.

[14] van Herk M, McWilliam A, Dubec M, et al. Magnetic Resonance Imaging–Guided Radiation Therapy: A Short Strengths, Weaknesses, Opportunities, and Threats Analysis. Int J Radiat Oncol Biol Phys. 2018;101:1057–1060.

[15] Werensteijn-Honingh AM, Kroon PS, Winkel D, et al. Feasibility of stereotactic radiotherapy using a 1.5 T MR-linac: Multi-fraction treatment of pelvic lymph node oligometastases. Radiother Oncol. 2019;134:50–54.

[16] Hal WA, Straza MW, Chen X, et al. Initial clinical experience of Stereotactic Body Radiation Therapy (SBRT) for liver metastases, primary liver malignancy, and pancreatic cancer with 4D-MRI based online adaptation and real-time MRI monitoring using a 1.5 Tesla MR-Linac. PLoS One. 2020;15:1–10.

[17] Winkel D, Bol GH, Kroon PS, et al. Adaptive radiotherapy: The Elekta Unity MR-linac concept. Clin Transl Radiat Oncol. 2019;18:54–59.

[18] de Mol van Otterloo SR, Christodouleas JP, Blezer ELA, et al. The MOMENTUM Study: An International Registry for the Evidence-Based Introduction of MR-Guided Adaptive Therapy. Front Oncol. 2020;10.

[19] Datta A, Aznar MC, Dubec M, et al. Delivering Functional Imaging on the MRI-Linac: Current Challenges and Potential Solutions. Clin Oncol. 2018;30:702–710.

[20] Gurney-Champion OJ, Mahmood F, van Schie M, et al. Quantitative imaging for radiotherapy purposes. Radiother Oncol. 2020;146:66–75.

[21] C.Clifton Ling, John Humm, Steven Larson, Howard Amols, Zvi Fuks, Steven Leibel JAK. Towards multidimensional radiotherapy (MD-CRT): biological imaging and biological conformality. Int J Radiat Oncol. 2000;47:551–560.

[22] Heide UA van der, Houweling AC, Groenendaal G, et al. Functional MRI for radiotherapy dose painting. Magn Reson Imaging. 2012;30:1216–1223.

[23] Lips IM, van der Heide UA, Haustermans K, et al. Single blind randomized Phase III trial to investigate the benefit of a focal lesion ablative microboost in prostate cancer (FLAME-trial): Study protocol for a randomized controlled trial. Trials. 2011;12.

[24] Azzam EI, Jay-Gerin JP, Pain D. Ionizing radiation-induced metabolic oxidative stress and prolonged cell injury. Cancer Lett. 2012;327:48–60.

[25] Beaton L, Bandula S, Gaze MN, et al. How rapid advances in imaging are defining the future of precision radiation oncology. Br J Cancer. 2019;120:779–790.

[26] van Houdt PJ, Yang Y, van der Heide UA. Quantitative Magnetic Resonance Imaging for Biological Image-Guided Adaptive Radiotherapy. Front Oncol. 2021;10:1–9.

[27] Kooreman ES, van Houdt PJ, Nowee ME, et al. Feasibility and accuracy of quantitative imaging on a 1.5 T MR-linear accelerator. Radiother Oncol. 2019;133:156–162.

[28] Kerkmeijer LGW, Groen VH, Pos FJ, et al. Re: Focal Boost to the Intraprostatic Tumor in External Beam Radiotherapy for Patients with Localized Prostate Cancer: Results from the FLAME Randomized Phase III Trial. J Clin Oncol. 2021;39:787–796.

[29] Koh DM, Collins DJ. Diffusion-weighted MRI in the body: Applications and challenges in oncology. Am J Roentgenol. 2007;188:1622–1635.

[30] Leibfarth S, Winter RM, Lyng H, et al. Potentials and challenges of diffusion-weighted magnetic resonance imaging in radiotherapy. Clin Transl Radiat Oncol. 2018;13:29–37.

[31] Rahbek S, Madsen KH, Lundell H, et al. Data-driven separation of MRI signal components for tissue characterization. J Magn Reson. 2021;333:107103.

[32] Le Bihan D, Breton E, Lallemand D, et al. MR imaging of intravoxel incoherent motions:

application to diffusion and perfusion in neurologic disorders. Radiology. 1986;161.

[33] Stejskal EO, Tanner JE. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys. 1965;42:288–292.

[34] Surov A, Meyer HJ, Wienke A. Correlation between apparent diffusion coefficient (ADC) and cellularity is different in several tumors: a meta-analysis. Oncotarget. 2017;8:59492–59499.

[35] Padhani AR, Liu G, Mu-Koh D, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: Consensus and recommendations. Neoplasia. 2009;11:102–125.

[36] Nishimura DG. Principles of Magnetic Resonance Imaging. Stanford University; 1996.

[37] Einstein A. Investigations on the theory of the Brownian movement. (Collection of papers translated from the German). In: Furthe R., Cowper A. D., editors. New York, Dower. Dover Publ. Inc. 1956.

[38] Le Bihan D, Iima M. Diffusion magnetic resonance imaging: What water tells us about biological tissues. PLoS Biol. 2015;13:1–13.

[39] Neil JJ. Measurement of water motion (apparent diffusion) in biological systems. Concepts Magn Reson. 1997;9:385–401.

[40] White NS, McDonald CR, Farid N, et al. Diffusion-weighted imaging in cancer: Physical foundations and applications of restriction spectrum imaging. Cancer Res. 2014;74:4638–4652.

[41] Le Bihan D, Breton E, Lallemand D, et al. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology. 1988;168:497–505.

[42] Rosenkrantz AB, Padhani AR, Chenevert TL, et al. Body diffusion kurtosis imaging: Basic principles, applications, and considerations for clinical practice. J Magn Reson Imaging. 2015;42:1190–1202.

[43] Jensen JH, Helpern JA, Ramani A, et al. Diffusional kurtosis imaging: The quantification of non-Gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med. 2005;53:1432–1440.

[44] O’Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14:169–186.

[45] Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory. Stat Methods Med Res. 2015;24:9–26.

[46] Ballman K V. Biomarker: Predictive or prognostic? J Clin Oncol. 2015;33:3968–3971.

[47] Noij DP, Martens RM, Marcus JT, et al. Intravoxel incoherent motion magnetic resonance imaging in head and neck cancer: A systematic review of the diagnostic and prognostic value. Oral Oncol. 2017;68:81–91.

[48] van Houdt PJ, Saeed H, Thorwarth D, et al. Integration of quantitative imaging biomarkers in clinical trials for MR-guided radiotherapy: Conceptual guidance for multicentre studies from the MR-Linac Consortium Imaging Biomarker Working

Group. Eur J Cancer. 2021;153:64–71.

[49] Shukla-Dave A, Obuchowski NA, Chenevert TL, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. J Magn Reson Imaging. 2019;49:e101–e121.

[50] Lewis B, Guta A, Mackey S, et al. Evaluation of diffusion-weighted MRI and geometric distortion on a 0.35T MR-LINAC at multiple gantry angles. J Appl Clin Med Phys. 2021;22:118–125.

[51] Kooreman ES, van Houdt PJ, Keesman R, et al. ADC measurements on the Unity MR-linac – A recommendation on behalf of the Elekta Unity MR-linac consortium. Radiother Oncol. 2020;153:106–113.

[52] Obuchowski NA. Interpreting Change in Quantitative Imaging Biomarkers. Acad Radiol. 2018;25:372–379.

[53] Le Bihan D, Poupon C, Amadon A, et al. Artifacts and pitfalls in diffusion MRI. J Magn Reson Imaging. 2006;24:478–488.

[54] Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - Are they relevant and what can we do about them? Radiol Oncol. 2016;50:254–262.

[55] Lambregts DMJ, 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;21:2567–2574.

[56] Mahmood F, Johannesen HH, Geertsen P, et al. The effect of region of interest strategies on apparent diffusion coefficient assessment in patients treated with palliative radiation therapy to brain metastases. Acta Oncol (Madr). 2015;54:1529–1534.

[57] Ma C, Guo X, Liu L, et al. Effect of region of interest size on ADC measurements in pancreatic adenocarcinoma. Cancer Imaging. 2017;17:1–7.

[58] Mahmood F, Hjorth Johannesen H, Geertsen P, et al. Diffusion MRI outlined viable tumour volume beats GTV in intra-treatment stratification of outcome. Radiother Oncol. 2020;144:121–126.

[59] Prah DE, Paulson ES, Nencka AS, et al. A Simple Method for Rectified Noise Floor Suppression: Phase- Corrected Real Data Reconstruction With Application to Diffusion-Weighted Imaging. Magn Reson Med. 2010;64:418–429.

[60] Reynaud O. Time-dependent diffusion MRI in cancer: Tissue modeling and applications. Front Phys. 2017;5:1–16.

[61] Ghosh A, Singh T, Singla V, et al. Comparison of absolute Apparent Diffusion Coefficient (ADC) values in ADC maps generated across different postprocessing software: Reproducibility in endometrial carcinoma. Am J Roentgenol. 2017;209:1312–1320.

[62] Schurink NW, van Kranen SR, Roberti S, et al. Sources of variation in multicenter rectal MRI data and their effect on radiomics feature reproducibility. Eur Radiol. 2022;32:1506–1516.

[63] Schurink NW, Lambregts DMJ, Beets-Tan RGH. Diffusion-weighted imaging in rectal cancer: Current applications and future perspectives. Br J Radiol. 2019;92:20180655.

[64] King AD, Chow KK, Yu KH, et al. Head and neck squamous cell carcinoma: Diagnostic performance of diffusion-weighted MR imaging for the prediction of treatment response. Radiology. 2013;266:531–538.

[65] Mahmood F, Johannesen HH, Geertsen P, et al. Repeated diffusion MRI reveals earliest time point for stratification of radiotherapy response in brain metastases. Phys Med Biol. 2017;62:2990–3002.

[66] Tsien C, Cao Y, Chenevert T. Clinical Applications for Diffusion Magnetic Resonance Imaging in Radiotherapy. Semin Radiat Oncol. 2014;24:218–226.

[67] Yang Y, Cao M, Sheng K, et al. Longitudinal diffusion MRI for treatment response assessment: Preliminary experience using an MRI-guided tri-cobalt 60 radiotherapy system. Radiol Oncol. 2016;43:1369–1373.

[68] Min LA, Vacher YJL, Dewit L, et al. 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.

[69] Hesamian MH, Jia W, He X, et al. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. J Digit Imaging. 2019;32:582–596.

[70] Lenchik L, Heacock L, Weaver AA, et al. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. Acad Radiol. 2019;26:1695–1706.

[71] 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 Biol Phys. 2016;94:824–831.

[72] 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;7:1–9.

[73] Otsu N. Threshold Selection Method From Gray-Level Histograms. IEEE Trans Syst Man Cybern. 1979;SMC-9:62–66.

[74] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8:135–160.

[75] Maja Bruvo; Faisal Mahmood. Apparent diffusion coefficient measurement of the parotid gland parenchyma. Quant Imaging Med Surg. 2021;11:3812–3829.

[76] Rasmussen JH, Olin AB, Lelkaitis G, et al. Intratumor heterogeneity is biomarker specific and challenges the association with heterogeneity in multimodal functional imaging in head and neck squamous cell carcinoma. Eur J Radiol. 2021;139.

[77] Amodeo S, Rosman AS, Desiato V, et al. MRI-based apparent diffusion coefficient for predicting pathologic response of rectal cancer after neoadjuvant therapy: Systematic review and meta-analysis. Am J Roentgenol. 2018;211:W205–W216.

[78] Atkinson AJ, Colburn WA, DeGruttola VG, et al. Biomarkers and surrogate endpoints:

Preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69:89–95.

[79] Sullivan DC. Imaging as a quantitative science. Radiology. 2008. p. 328–332.

[80] Abramson R, Burton K, Yu J, et al. Methods and Challenges in Quantitative Imaging Biomarker Development Richard. Acad Radiol. 2015;22:25–32.

[81] Henning EC, Azuma C, Sotak CH, et al. Multispectral tissue characterization in a RIF-1 tumor model: Monitoring the ADC and T2 responses to single-dose radiotherapy. Part II. Magn Reson Med. 2007;57:513–519.

[82] Moffat BA, Chenevert TL, Lawrence TS, et al. Functional diffusion map: A noninvasive MRI biomarker for early stratification of clinical brain tumor response. Proc Natl Acad Sci U S A. 2005;102:5524–5529.

[83] Bisgaard ALH, Brink C, Fransen ML, et al. Robust extraction of biological information from diffusion-weighted magnetic resonance imaging during radiotherapy using semi-automatic delineation. Phys Imaging Radiat Oncol. 2022;21:146–152.

[84] Kerkmeijer LGW, Fuller CD, Verkooijen HM, et al. The MRI-linear accelerator consortium: Evidence-based clinical introduction of an innovation in radiation oncology connecting researchers, methodology, data collection, quality assurance, and technical development. Front Oncol. 2016;6:1–6.

[85] Song M, Yue Y, Jin Y, et al. Intravoxel incoherent motion and ADC measurements for differentiating benign from malignant thyroid nodules: Utilizing the most repeatable region of interest delineation at 3.0 T. Cancer Imaging. 2020;20:1–9.

[86] Barral M, Soyer P, Ben Hassen W, et al. Diffusion-weighted MR imaging of the normal pancreas: Reproducibility and variations of apparent diffusion coefficient measurement at 1.5-and 3.0-Tesla. Diagn Interv Imaging. 2013;94:418–427.

[87] Ueno Y, Tamada T, Sofue K, et al. Do the variations in ROI placement technique have influence for prostate ADC measurements? Acta Radiol Open. 2022;11:205846012210865.

[88] Dinh C V., Steenbergen P, Ghobadi G, et al. Magnetic resonance imaging for prostate cancer radiotherapy. Phys Medica. 2016;32:446–451.

[89] Steenbergen P, Haustermans K, Lerut E, et al. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. Radiother Oncol. 2015;115:186–190.

[90] Vidić I, Egnell L, Jerome NP, et al. Modeling the diffusion-weighted imaging signal for breast lesions in the b = 200 to 3000 s/mm2 range: quality of fit and classification accuracy for different representations. Magn Reson Med. 2020;84:1011–1023.

[91] Iima M, Partridge SC, Le Bihan D. Six DWI questions you always wanted to know but were afraid to ask: clinical relevance for breast diffusion MRI. Radio. 2020;30:2561–2570.

[92] De Deene Y, Van De Walle R, Achten E, et al. Mathematical analysis and experimental investigation of noise in quantitative magnetic resonance imaging applied in polymer gel dosimetry. Signal Processing. 1998;70:85–101.

[93] Dietrich O, Heiland S, Sartor K. Noise correction for the exact determination of apparent diffusion coefficients at low SNR. Magn Reson Med. 2001;45:448–453.

[94] NEMA. NEMA Standards Publication MS 1–2008: Determination of Signal-to-Noise Ratio (SNR) in Diagnostic Magnetic Resonance Imaging. Rosslyn, Va Natl Electr …. 2021;2008:1–21.

[95] DWI MR Biomarker Committee B. QIBA Profile: Diffusion-Weighted Magnetic Resonance Imaging (DWI) [Internet]. Quant. Imaging Biomarkers Alliance. 2019. Available from: https://qibawiki.rsna.org/index.php/Profiles.

[96] Pathak R, Tian J, Thacker NA, et al. Considering tumour volume for motion corrected DWI of colorectal liver metastases increases sensitivity of ADC to detect treatment-induced changes. Sci Rep. 2019;9:1–10.

[97] Chenevert TL, Malyarenko DI, Newitt D, et al. Errors in quantitative image analysis due to platform-dependent image scaling. Transl Oncol. 2014;7:65–71.

[98] Tyagi N, Cloutier M, Zakian K, et al. Diffusion-weighted MRI of the lung at 3T evaluated using echo-planar-based and single-shot turbo spin-echo-based acquisition techniques for radiotherapy applications. J Appl Clin Med Phys. 2019;20:284–292.

[99] Clark TG, Bradburn MJ, Love SB, et al. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 2003;89:232–238.

[100] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–481.

[101] Cox DR. Regression Models and Life-Tables Authors (s): D . R . Cox Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 34 , No . 2 Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.org/stable. J R Stat Soc. 1972;34:187–220.

[102] Bradburn MJ, Clark TG, Love SB, et al. Survival Analysis Part II: Multivariate data analysis- An introduction to concepts and methods. Br J Cancer. 2003;89:431–436.

[103] Kirkegård J, Bojesen AB, Nielsen MF, et al. Trends in pancreatic cancer incidence, characteristics, and outcomes in Denmark 1980–2019: A nationwide cohort study. Cancer Epidemiol. 2022;80:102230.

[104] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71:209–249.

[105] Mizrahi JD, Surana R, Valle JW, et al. Pancreatic cancer. Lancet. 2020;395:2008–2020.

[106] Noda Y, Tomita H, Ishihara T, et al. Prediction of overall survival in patients with pancreatic ductal adenocarcinoma: histogram analysis of ADC value and correlation with pathological intratumoral necrosis. BMC Med Imaging. 2022;22:1–10.

[107] Ejlsmark MW, Schytte T, Bernchou U, et al. Radiotherapy for Locally Advanced Pancreatic Adenocarcinoma-A Critical Review of Randomised Trials. Curr Oncol. 2023;30:6820–6837.

[108] Le N, Sund M, Vinci A. Prognostic and predictive markers in pancreatic

adenocarcinoma. Dig Liver Dis. 2016;48:223–230.

[109] Ichikawa T, Erturk SM, Motosugi U, et al. High-b value diffusion-weighted MRI for detecting pancreatic adenocarcinoma: Preliminary results. Am J Roentgenol. 2007;188:409–414.

[110] Matsuki M, Inada Y, Nakai G, et al. Diffusion-weighed MR imaging of pancreatic carcinoma. Abdom Imaging. 2007;32:481–483.

[111] Inan N, Arslan A, Akansel G, et al. Diffusion-weighted imaging in the differential diagnosis of cystic lesions of the pancreas. Am J Roentgenol. 2008;191:1115–1121.

[112] Seung SL, Jae HB, Beom JP, et al. Quantitative analysis of diffusion-weighted magnetic resonance imaging of the pancreas: Usefulness in characterizing solid pancreatic masses. J Magn Reson Imaging. 2008;28:928–936.

[113] Kim B, Lee SS, Sung YS, et al. Intravoxel incoherent motion diffusion-weighted imaging of the pancreas: Characterization of benign and malignant pancreatic pathologies. J Magn Reson Imaging. 2017;45:260–269.

[114] Niwa T, Ueno M, Ohkawa S, et al. Advanced pancreatic cancer: The use of the apparent diffusion coefficient to predict response to chemotherapy. Br J Radiol. 2009;82:28–34.

[115] Dalah E, Erickson B, Oshima K, et al. Correlation of ADC With Pathological Treatment Response for Radiation Therapy of Pancreatic Cancer. Transl Oncol. 2018;11:391–398.

[116] Cuneo KC, Chenevert TL, Ben-Josef E, et al. A pilot study of diffusion- weighted mri in patients undergoing neoadjuvant chemoradiation for pancreatic cancer. Transl Oncol. 2014;7:644–649.

[117] Garces-Descovich A, Morrison TC, Beker K, et al. DWI of pancreatic ductal adenocarcinoma: A pilot study to estimate the correlation with metastatic disease potential and overall survival. Am J Roentgenol. 2019;212:323–331.

[118] Kurosawa J, Tawada K, Mikata R, et al. Prognostic relevance of apparent diffusion coefficient obtained by diffusion-weighted MRI in pancreatic cancer. J Magn Reson Imaging. 2015;42:1532–1537.

[119] Rahbek S, Mahmood F, Tomaszewski M, et al. Decomposition-based framework for tumor classification and prediction of treatment response from longitudinal MRI. Phys Med Biol. 2023;68.

[120] Schemper M. Further Results on the Explained Variation in Proportional Hazards Regression. Biometrica. 1992;79:202–204.

[121] Robertis R De, Martini PT, Demozzi E, et al. Diffusion-weighted imaging of pancreatic cancer. World J Rdiology. 2015;7:319–329.

[122] Lyng H, Haraldseth O, Rofstad EK. Measurement of Cell Density and Necrotic Fraction in Human Melanoma Xenografts by Diffusion Weighted Magnetic Resonance Imaging. 2000;836:828–836.

[123] Hiraoka N, Ino Y, Sekine S, et al. Tumour necrosis is a postoperative prognostic marker for pancreatic cancer patients with a high interobserver reproducibility in histological

evaluation. Br J Cancer. 2010;103:1057–1065.

[124] van Schie MA, Dinh C V., Houdt PJ va., et al. Contouring of prostate tumors on multiparametric MRI: Evaluation of clinical delineations in a multicenter radiotherapy trial. Radiother Oncol. 2018;128:321–326.

[125] Sun YS, Cui Y, Tang L, et al. Early evaluation of cancer response by a new functional biomarker: Apparent diffusion coefficient. Am J Roentgenol. 2011;197:23–29.

[126] van Schie MA, van Houdt PJ, Ghobadi G, et al. Quantitative MRI Changes During Weekly Ultra-Hypofractionated Prostate Cancer Radiotherapy With Integrated Boost. Front Oncol. 2019;9.

[127] Shaverdian N, Yang Y, Hu P, et al. Feasibility evaluation of diffusion-weighted imaging using an integrated MRI-radiotherapy system for response assessment to neoadjuvant therapy in rectal cancer. Br J Radiol. 2017;90.

[128] Gao Y, Kalbasi A, Hsu W, et al. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. Phys Med Biol. 2020;65.

[129] Shah V, Turkbey B, Mani H, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. Med Phys. 2012;39:4093–4103.

[130] Hearn N, Bugg W, Chan A, et al. Manual and semi- - automated delineation of locally advanced rectal cancer subvolumes with diffusion- - weighted MRI. Br J Radiol. 2020;93.

[131] Coolens C, Driscoll B, Foltz W, et al. Unified platform for multimodal and voxel-based analysis to evaluate tumour perfusion and diffusion characteristics before and after radiation treatment evaluated in metastatic brain cancer. Br J Radiol. 2019;92:1–8.

[132] Joye I, Lambrecht M, Jegou D, et al. Does a central review platform improve the quality of radiotherapy for rectal cancer? Results of a national quality assurance project. Radiother Oncol. 2014;111:400–405.

[133] Jones DK, Cercignani M. Twenty-five Pitfalls in the Analysis of Diffusion MRI Data. NMR Biomed. 2010;23:803–820.

[134] Mansfield P. Multi-planar image formation using NMR spin echoes. J Phys C Solid State Phys. 1977;10.

[135] Plodeck V, Radosa CG, Hübner HM, et al. Rectal gas-induced susceptibility artefacts on prostate diffusion-weighted MRI with epi read-out at 3.0 T: does a preparatory micro-enema improve image quality? Abdom Radiol. 2020;45:4244–4251.

[136] Robson MD, Anderson AW, Gore JC. Diffusion-weighted multiple shot echo planar imaging of humans without navigation. Magn Reson Med. 1997;38:82–88.

[137] Jezzard P, Balaban RS. Correction for geometric distortion in echo planar images from B0 field variations. Magn Reson Med. 1995;34:65–73.

[138] Hasler SW, Bernchou U, Bertelsen A, et al. Tumor-site specific geometric distortions in high field integrated magnetic resonance linear accelerator radiotherapy. Phys

Imaging Radiat Oncol. 2020;15:100–104.

[139] Li Y, Xu N, Fitzpatrick JM, et al. Geometric distortion correction for echo planar images using nonrigid registration with spatially varying scale. Magn Reson Imaging. 2008;26:1388–1397.

Appendix I: Supplementary materials for paper I

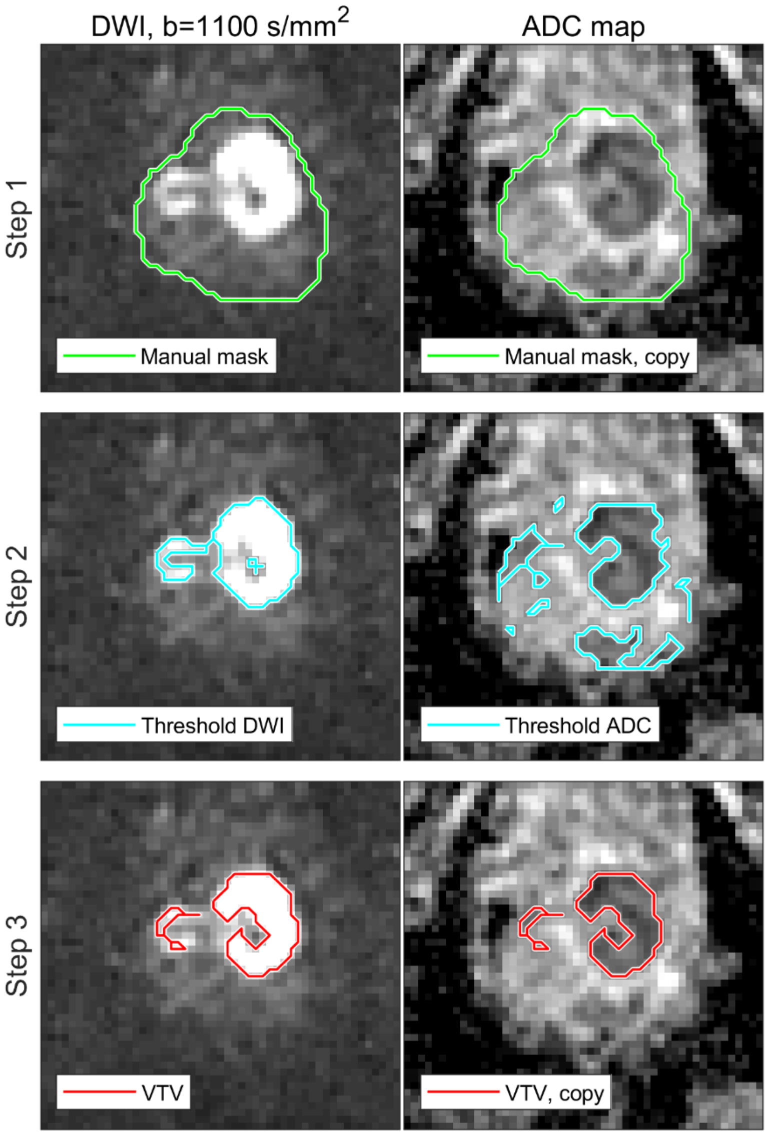


Figure A1.1. 3-step process of the semi-automatic delineation tool (SADT): Step 1: manual input delineates roughly the region of interest, Step 2: two masks are created based on criteria of high DWI intensity and low ADC, Step 3: overlap between the masks from step 2 defines the resulting VTV. In this example, the manual mask was defined by the non-radiologist. The image is transaxial and has been cropped to a size of (92.8x92.8) mm^2. Note that semi-automatic delineation is performed on 3D images.

Appendix II: Supplementary materials for paper II

Appendix IIa: MRI sequence details

Table A2.1. Details on DWI and T2W sequences. Abbreviations: Spin echo (SE), echo planar imaging (EPI), spectral presaturation with inversion recovery (SPIR), spectral attenuated inversion recovery (SPAIR)

	Adrenal gland	Pancreas	Oligo metastasis	Prostate
T2W:				
Sequence	3D SE	3D SE	3D SE	3D SE
Fat saturation	No	No	No	No
TE/TR (ms)	137/1400	137/1400	151/1400	151/1400
Parallel imaging (SENSE factor)	3.7	3.7	3.7	3.7
No. of excitations (NEX)	3	3	2	2
In-plane resolution (mm)	1x1	1x1	0.81x0.81	1x1
Field of view (mm)	448x448	448x448	544x544	448x448
Slice thickness (mm)	2	2	2	2
Slice gap (mm)	0	0	0	0
Scan duration	6 min and 4 s	6 min and 4 s	5 min and 32 s	3 min and 51 s
DWI:				
Sequence	Monopolar diffusion encoding, 2D SE with single shot EPI readout	Monopolar diffusion encoding, 2D SE with single shot EPI readout	Monopolar diffusion encoding, 2D SE with single shot EPI readout	Monopolar diffusion encoding, 2D SE with single shot EPI readout
Diffusion gradient encoding	Three orthogonal directions along the imaging plane axes	Three orthogonal directions along the imaging plane axes	Three orthogonal directions along the imaging plane axes	Three orthogonal directions along the imaging plane axes
Fat saturation	SPIR	SPIR	SPAIR	SPAIR
b-values (No. of excitations (NEX))	30 (2) 80 (2) 150 (4) 300 (4) 500 (16)	30 (2) 80 (2) 150 (4) 300 (4) 500 (16)	0 (2) 30 (2) 80 (2) 150 (4) 500 (16)	0 (2) 30 (2) 80 (2) 150 (4) 500 (16)

Gradient duration (ms) (δ)	22.47	22.47	20.22	20.22
Effective diffusion time (ms) $\left(\Delta - \frac{\delta}{3}\right)$	27.95	27.95	34.40	34.40
TE/TR (ms)	70.90/559.43	70.90/559.43	82.30/4538.00	82.30/3354.17
Parallel imaging (SENSE factor)	2	2	2.3	2.3
In-plane resolution (mm)	1.22x1.22	1.22x1.22	1.92x1.92	1.92x1.92
Field of view (mm)	351x351	351x351	430x430	430x430
Slice thickness (mm)	6	6	4	4
Slice gap (mm)	0.6	0.6	0	0
Scan duration	2 min and 23 s	2 min and 23 s	5 min and 40 s	4 min and 12 s

Appendix IIb: Technical preparation of data

DWI images were split into individual b-value-images with unique series UID's using in-house software (Matlab R2020b, Mathworks ab, Sweden) to accommodate delineation in ProKnow. For each b-value, a rigid registration between the DWI and T2W image was performed using MIM (MIM Software Inc., Cleveland, Ohio).

The original, non-split DWI images were used for ADC calculation. A registration between b-values was performed using the built-in function of the scanner software (R5.7.1, Philips Ingenia, Philips Healthcare, Best, The Netherlands). All delineations were transferred to the original DWI images using the transfer matrix from the rigid registration between b=500 s/mm^2 DWI and T2W images in MIM.

Appendix IIc: Calculation method details

Table A2.2. Details on calculation methods used by the nine participating centres.

Calculation method No.	b-values (s/mm^2)	Fitting method	Filtering	Normally reported ADC metrics
1	All	Linear least squares	Values below 10^{-6} mm^2/s were excluded	Histogram distribution. Median (range). Mean
2	≥ 150	Linear least squares	No	Median + 25th and 75th percentiles
3	≥ 150	Linear least squares	Values below 0 mm^2/s were excluded	Median
4	Adrenal gland + pancreas: 30, 80, 150, 300, 500 Oligo metastasis + prostate: 0, 80, 150, 500	Linear least squares	Values below 0 mm^2/s were excluded	Mean
5	150 and 500	Direct solution (no fit): ADC[mm^2/s]= -ln(S(500)/S(150))/(500-150)	Limiting ADC values to range (0-4) × 10^{-3} mm^2/s. Values were not excluded.	Mean + SD
6	≥ 150	Scanner software	No	Mean+SD (or Median + 5th/95th percentile if clinical software provided those)
7	≥ 150	Linear least square	Values below 0 mm^2/s were excluded	Median ADC + 25th and 75th percentiles
8	≥ 150	Non-linear least squares	Lower bound of 0 and upper bound of 3.1×10^{-3} mm^2/s	Mean ADC for GTV.
9	All	Weighted linear regression on function Weighting function: 1/S(b)	No	Change in median ADC + 75th percentile

Appendix IId: ADC variation boxplots

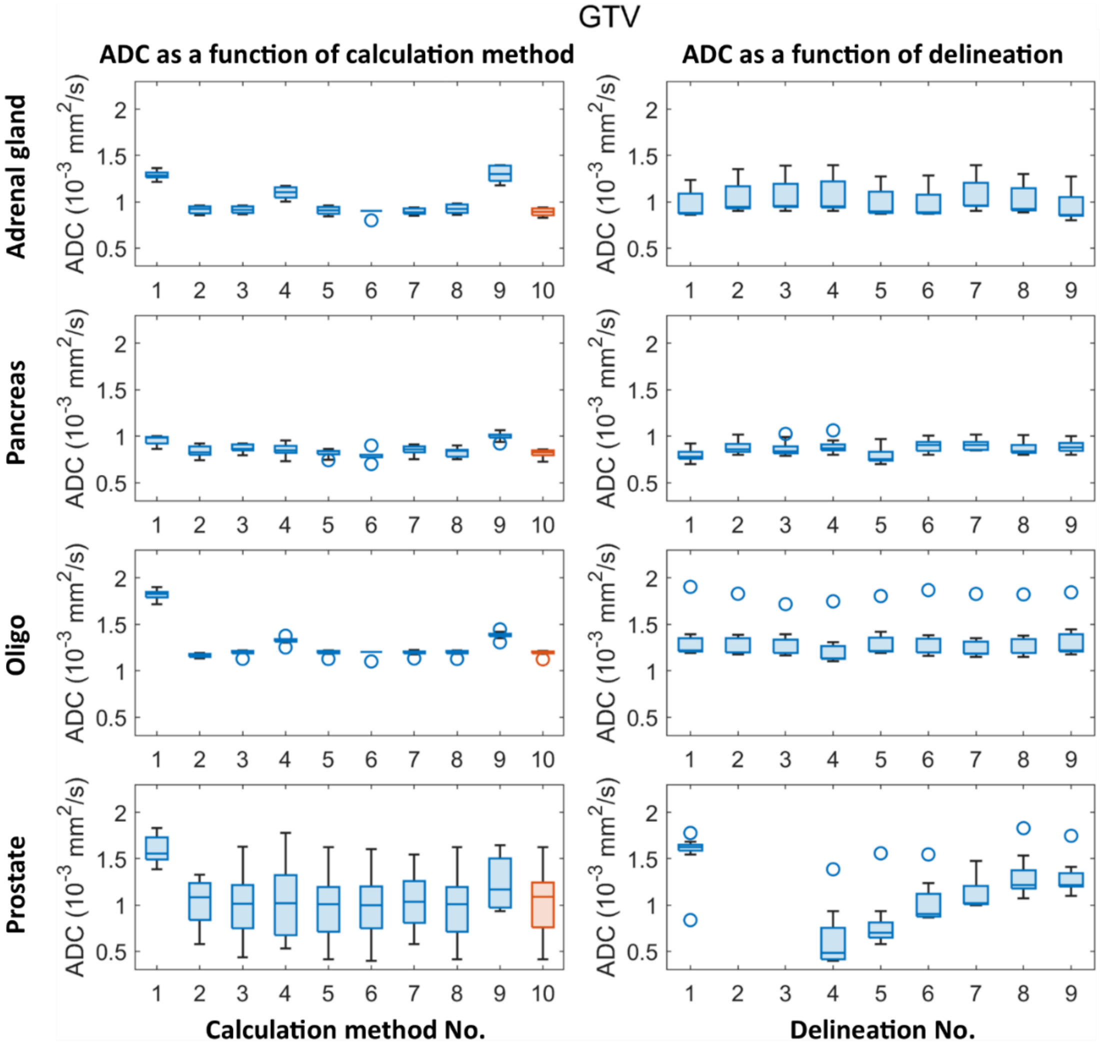


Figure A2.1. ADC values within GTVs as a function of calculation method (left) and delineation (right) for the four clinical cases. The boxes represent variation across delineations and calculation methods, respectively. For comparison, calculation method no. 10 (red) represent the scanner software using b-values≥150mm/s^2. The 'o' marker indicates outliers, defined as more than 1.5 times the interquartile range away from the bottom or top edges of the box.

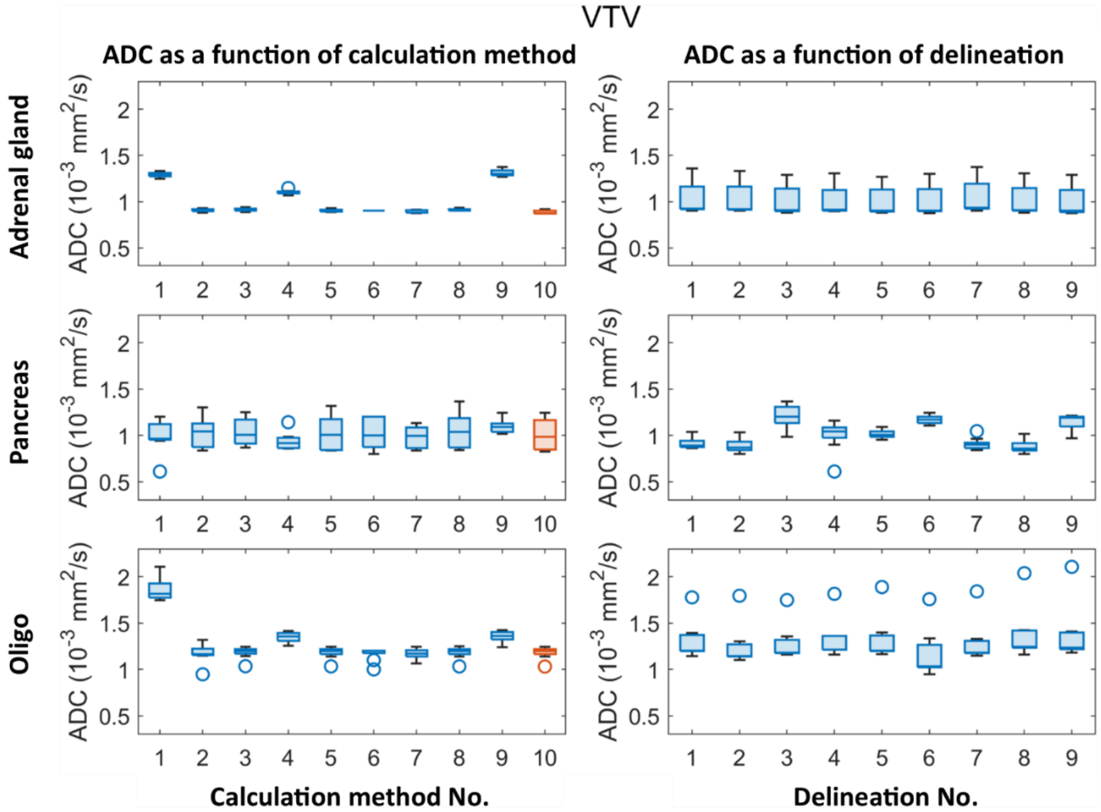


Figure A2.2. ADC values within VTVs as a function of calculation method (left) and delineation (right) for the four clinical cases. The boxes represent variation across delineations and calculation methods, respectively. For comparison, calculation method no. 10 (red) represent the scanner software using b-values≥150mm/s². The 'o' marker indicates outliers, defined as more than 1.5 times the interquartile range away from the bottom or top edges of the box.

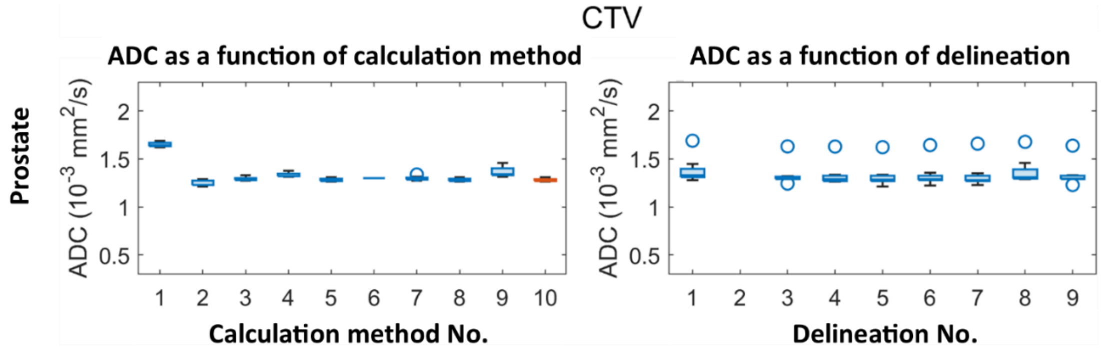


Figure A2.3. ADC values within CTVs as a function of calculation method (left) and delineation (right) for the prostate. The boxes represent variation across delineations and calculation methods, respectively. For comparison, calculation method no. 10 (red) represent the scanner software using b-values≥150mm/s². The 'o' marker indicates outliers, defined as more than 1.5 times the interquartile range away from the bottom or top edges of the box.

Appendix IIe: Weighted least squares fitting

The Stejskal-Tanner model for the DWI signal, S, is given by [1][33]:

$$S = S_0 \cdot e^{-b \cdot ADC}$$

By taking the logarithm on both sides, we get:

$$\ln(S) = \ln(S_0) - b \cdot ADC$$

For one voxel, measurements are performed for N different b-values of the MRI signal magnitude. S_i denotes the signal at the i-th b-value, measured with n_i measurements (number of measurements = number of excitations (NEX)).

For each single measurement, we assume Gaussian noise, σ, that is independent on signal and b-value. Linear fitting is performed for the logarithm to the signal: $y_i = \ln(S_i)$. To determine the variation associated with y_i, we use uncertainty propagation. For a function of a single variable, f(x), the uncertainty propagation equation is given by:

$$\sigma_f^2 = \left(\frac{df}{dx}\right)^2 \sigma_x^2$$

Applying this to y_i, we get:

$$\sigma_{y_i}^2 = \left(\frac{d_{y_i}}{d_{S_i}}\right)^2 \sigma_{S_i}^2 = \frac{\sigma_{S_i}^2}{S_i^2}$$

Assuming that the measurements are uncorrelated,

$$\sigma_{S_i} = \frac{\sigma}{\sqrt{n_i}}$$

Combining the above expressions,

$$\sigma_{y_i}^2 = \frac{\sigma^2}{n_i S_i^2}$$

Thus, the weights used for the weighted least squares fitting are:

$$w_{y_i} = \frac{1}{\sigma_{y_i}^2} = \frac{n_i S_i^2}{\sigma^2}$$

Since we assume that σ is constant over all measurements, it is left out:

$$w_{y_i} = n_i S_i^2$$

Appendix III: Supplementary materials for paper III

Appendix IIIa: Figures and Tables

MRI acquisition parameters

Table A3.1. Acquisition parameters for T2-weighted images acquired on the MRI-linac and the diagnostic scanner.

T2W:	MRI-linac	Diagnostic scanner
Sequence	3D SE	3D SE
Fat saturation	No	No
TE/TR (ms)	137/1400	137/1400
No. of excitations (NEX)	3	3
In-plane resolution (mm)	1x1	1x1
Field of view (mm)	448x448	488x488
Slice thickness (mm)	2	2
Slice gap (mm)	0	0
Scan duration	6 min and 4 s	6 min and 4 s

Table A3.2. Acquisition parameters for DWI-sequences on the MRI-linac

DWI:	Sequence 1 (39 patients)	Sequence 2 (6 patients)
Sequence	Monopolar diffusion encoding, 2D SE with single shot EPI readout	Monopolar diffusion encoding, 2D SE with single shot EPI readout
Diffusion gradient encoding	Three orthogonal directions along the imaging plane axes	Three orthogonal directions along the imaging plane axes
Fat saturation	SPIR	SPIR
b-values (s/mm^2)	0 30 80 150 500	0 20 60 100 300 800 1000
TE/TR (ms)	82.30/3354.17	84.72/628.58
In-plane resolution (mm)	1.92x1.92	1.22x1.22
Field of view (mm)	224x224	288x288
Slice thickness (mm)	4	6
Slice gap (mm)	0.00	0.60
Scan duration	4 min and 11 s	3 min and 56 s

Kaplan-Meier plot for overall survival

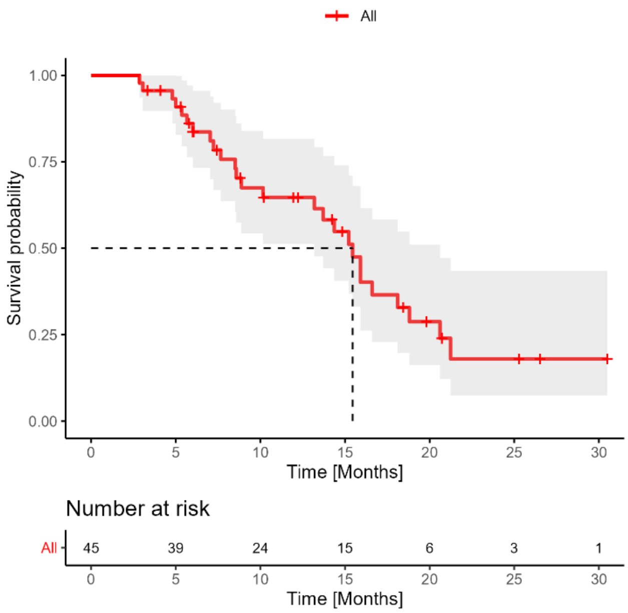


Figure A3.1. Kaplan-Meier plot for overall survival. The red curve represents the Kaplan-Meier estimator, while the shaded area represents the 95% confidence interval. The dotted line represents the median survival time.

Univariable models for the parameters included in the best-performing model

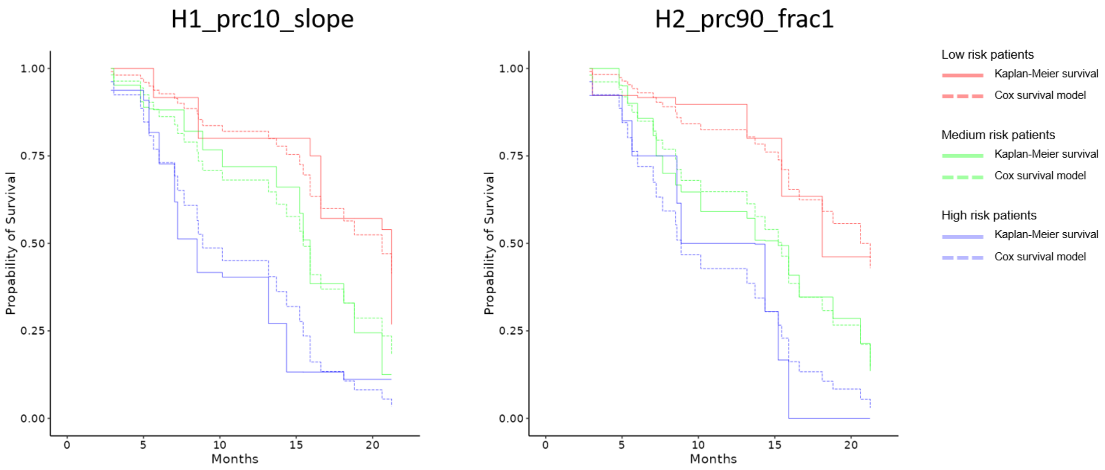


Figure A3.2. Comparison of the univariable Cox model and the Kaplan-Meier estimator for the variables "H1_prc10_slope" and "H2_prc90_frac1" respectively. Patients were split into high, medium and low-risk groups based on the 25% and 75% percentiles of the calculated linear predictors, i.e. the high and low-risk groups each contained 25% of the patients, and the medium-risk group contained 50% of the patients. The division-coefficients of the linear predictor were -0.34 and 0.36 for "H1_prc10_slope" and -0.36 and 0.41 for "H2_prc90_frac1".

Univariable model for ADC

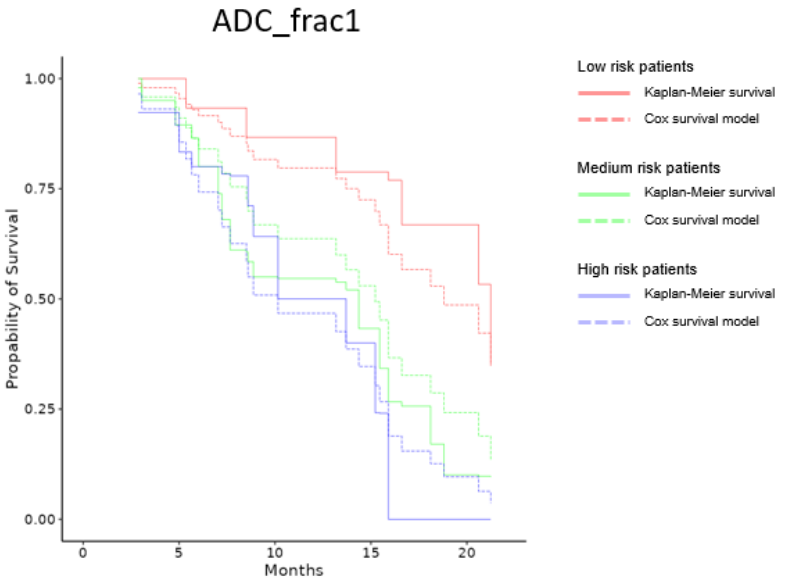
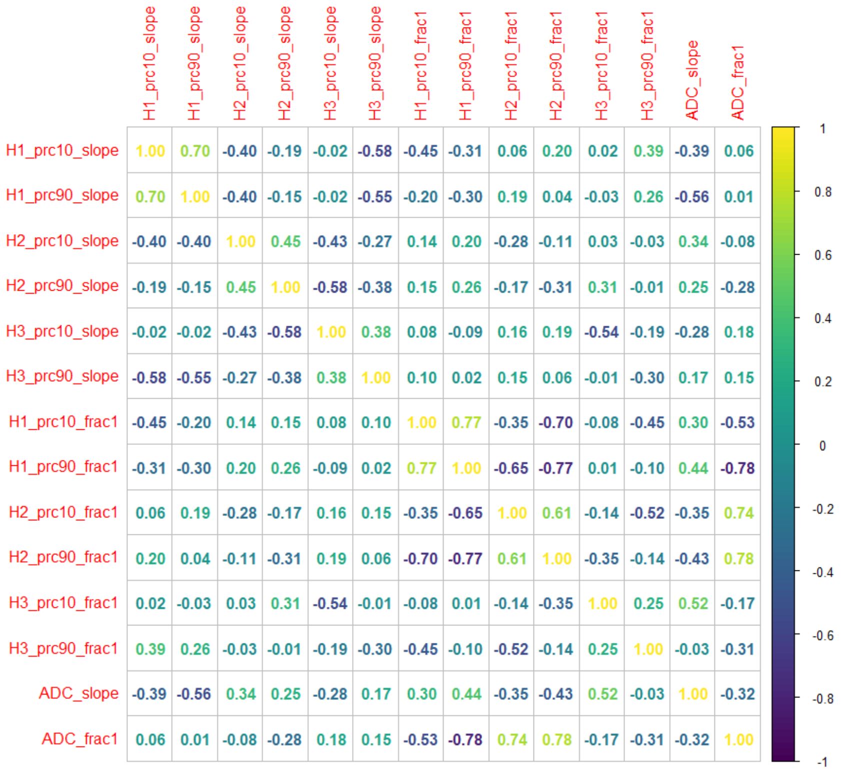


Figure A3.3. Comparison of the univariable Cox model and the Kaplan Meier estimator for ADC at fraction 1. Patients were split into high, medium and low-risk groups based on the 25% and 75% percentiles of the calculated linear predictors, i.e. the high and low risk groups each contained 25% of the patients, and the medium-risk group contained 50% of the patients. The division-coefficients were -0.40 and 0.35.

Correlation between DWI parameters



	H1_prc10_slope	H1_prc90_slope	H2_prc10_slope	H2_prc90_slope	H3_prc10_slope	H3_prc90_slope	H1_prc10_frac1	H1_prc90_frac1	H2_prc10_frac1	H2_prc90_frac1	H3_prc10_frac1	H3_prc90_frac1	ADC_slope	ADC_frac1
H1_prc10_slope	1.00	0.70	-0.40	-0.19	-0.02	-0.58	-0.45	-0.31	0.06	0.20	0.02	0.39	-0.39	0.06
H1_prc90_slope	0.70	1.00	-0.40	-0.15	-0.02	-0.55	-0.20	-0.30	0.19	0.04	-0.03	0.26	-0.56	0.01
H2_prc10_slope	-0.40	-0.40	1.00	0.45	-0.43	-0.27	0.14	0.20	-0.28	-0.11	0.03	-0.03	0.34	-0.08
H2_prc90_slope	-0.19	-0.15	0.45	1.00	-0.58	-0.38	0.15	0.26	-0.17	-0.31	0.31	-0.01	0.25	-0.28
H3_prc10_slope	-0.02	-0.02	-0.43	-0.58	1.00	0.38	0.08	-0.09	0.16	0.19	-0.54	-0.19	-0.28	0.18
H3_prc90_slope	-0.58	-0.55	-0.27	-0.38	0.38	1.00	0.10	0.02	0.15	0.06	-0.01	-0.30	0.17	0.15
H1_prc10_frac1	-0.45	-0.20	0.14	0.15	0.08	0.10	1.00	0.77	-0.35	-0.70	-0.08	-0.45	0.30	-0.53
H1_prc90_frac1	-0.31	-0.30	0.20	0.26	-0.09	0.02	0.77	1.00	-0.65	-0.77	0.01	-0.10	0.44	-0.78
H2_prc10_frac1	0.06	0.19	-0.28	-0.17	0.16	0.15	-0.35	-0.65	1.00	0.61	-0.14	-0.52	-0.35	0.74
H2_prc90_frac1	0.20	0.04	-0.11	-0.31	0.19	0.06	-0.70	-0.77	0.61	1.00	-0.35	-0.14	-0.43	0.78
H3_prc10_frac1	0.02	-0.03	0.03	0.31	-0.54	-0.01	-0.08	0.01	-0.14	-0.35	1.00	0.25	0.52	-0.17
H3_prc90_frac1	0.39	0.26	-0.03	-0.01	-0.19	-0.30	-0.45	-0.10	-0.52	-0.14	0.25	1.00	-0.03	-0.31
ADC_slope	-0.39	-0.56	0.34	0.25	-0.28	0.17	0.30	0.44	-0.35	-0.43	0.52	-0.03	1.00	-0.32
ADC_frac1	0.06	0.01	-0.08	-0.28	0.18	0.15	-0.53	-0.78	0.74	0.78	-0.17	-0.31	-0.32	1.00

Figure A3.4. Pearson's correlation coefficient for all pairs of DWI parameters.

Multivariable model including median values

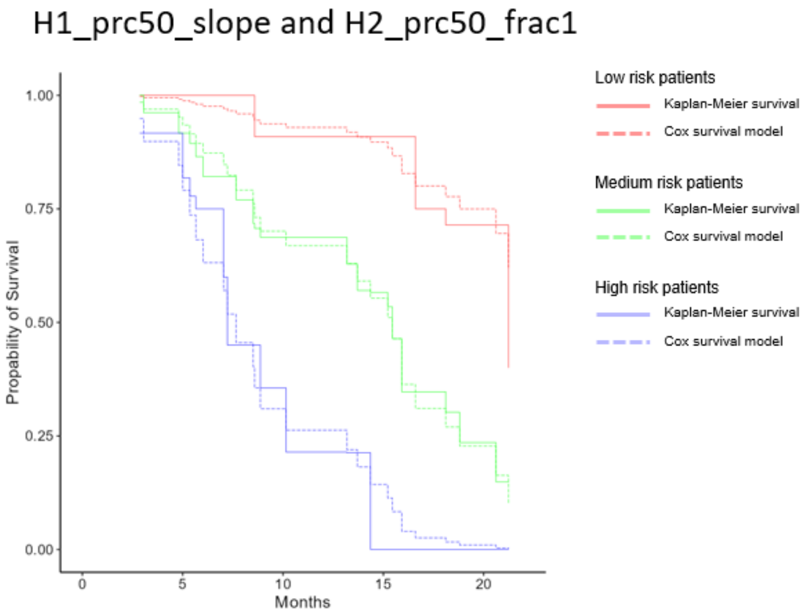


Figure A3.5. Comparison of the multivariable Cox model and the Kaplan Meier estimator for "H1_prc50_slope" and "H2_prc50_frac1". Patients were split into high, medium and low-risk groups based on the 25% and 75% percentiles of the calculated linear predictors, i.e. the high and low-risk groups each contained 25% of the patients, and the medium-risk group contained 50% of the patients. The division-coefficients were -0.68 and 0.82.

Appendix IIIb: Image decomposition using msNMF

To determine $\boldsymbol{W}$ and $\boldsymbol{H}$, the Frobenius norm of the residual, $||\boldsymbol{X} - \boldsymbol{WH}||_F^2$, was minimized under the constraints that $\boldsymbol{H}$ and $\boldsymbol{W}$ are non-negative, and that the components and the slopes of the components are monotonous. Minimization was performed using an alternating non-negative least squares (ANLS) algorithm, where $\boldsymbol{W}$ was optimized while $\boldsymbol{H}$ was kept constant and vice versa.

A stochastic gradient descent framework was used to solve the non-convex optimization problem. $\boldsymbol{X}$ was divided into bathes of 1000 voxels each, and for each batch, $\boldsymbol{W}$ was initialized as the $\boldsymbol{W}$ resulting from the previous batch while a random initialization was used for $\boldsymbol{H}$. The optimization continued until convergence. Data was shuffled between each epoch, i.e. each batch did not contain the same voxels for each epoch. Based on the resulting components, $\boldsymbol{W}$, a final optimization of $\boldsymbol{H}$ was performed for each patient while keeping $\boldsymbol{W}$ constant, i.e. by solving $\min_{\boldsymbol{H} \geq 0} ||\boldsymbol{X} - \boldsymbol{WH}||_F^2$. This was done both for patients scanned with DWI sequence 1 and 2. For patients with DWI sequence 2, $\boldsymbol{W}$ was interpolated/extrapolated to match the set of b-values in DWI sequence 2 ranging from 0-800 s/mm^2.

Appendix IIIc: Statistical analysis plan

Statistical analysis plan

Title:
Prediction of response in pancreas tumours using parameters derived from longitudinal diffusion-weighted MRI

SAP number with dates:
Version 2, date: 27.06.2023

Protocol version:
Project description of the Ph.D. project
"DW-MRI as a decision making tool in the in-room MRI guided radiotherapy pipeline " updated on the 10.01.2022 by Anne Bisgaard. See "Study 3".

SAP revision: This is the second version of the SAP. This SAP was completed on the 27.06.2023, before any analysis was performed on the outcome variables.

Roles and contributions:

Author: Anne Bisgaard, Ph.D.-student. Laboratory of Radiophysics, Department of Oncology, Odense University Hospital.

Main supervisor: Faisal Mahmood, Ph.D. Laboratory of Radiophysics, Department of Oncology, Odense University Hospital.

Co-supervisor: Carsten Brink, Ph.D. Laboratory of Radiophysics, Department of Oncology, Odense University Hospital.

Co-supervisor: Tine Schytte, Ph.D. Department of Oncology, Odense University Hospital.

This plan is based on the TRIPOD Checklist for Prediction Model Development, item 1-3 and item 6a through 11. (http://www.tripod-statement.org/TRIPOD/TRIPOD-Checklists/TRIPOD-Checklist-Prediction-Model-Development)

Title and abstract

Item 1. Title

Prediction of response in pancreas tumours using parameters derived from longitudinal diffusion-weighted MRI

Item 2. Abstract

Introduction

Item3.a. Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.

Pancreas cancer is the fourth most common cause of cancer-related death in the western world [2]. The 5-year survival rate is less than 10% [3]. The treatment include surgery, chemotherapy and radiotherapy (RT). Some patients have surgery after RT, however not all patients are suited for surgery. In any case, down-staging RT is used to achieve local control.

Prediction of response to neo-adjuvant treatment would allow for treatment adaptation, which could potentially improve the outcome or reduce toxicity for the individual patient. The rationale for developing the prediction model is to allow treatment adaptation based on information from quantitative MRI. The model is prognostic.

DWI is a potential biomarker for response to neo-adjuvant treatment for patients with pancreas cancer [4]. Typically, the “apparent diffusion coefficient” (ADC) is derived from DWI images using a mono-exponential model [4]. However, using models may lead to biased parameters, as they do not take into account partial volume effects. To overcome this problem, data driven approaches may be preferred. An example is the newly developed monotonous slope non-negative matrix factorization (ms-NMF), which uses a decomposition of the DWI signal to extract parameters [5]. The aim of this study is to test whether these parameters can be used to predict outcome in terms of time to local progression based on the RECIST criteria [6] and overall survival in patients with pancreas cancer treated with RT.

Item 3b. Specify the objectives, including whether the study describes the development or validation of the model or both.

Hypothesis:

- Parameters derived from longitudinal DWI acquired on the MR-linac can be used to predict time to local progression based on the RECIST criteria and overall survival in patients with pancreas cancer

Methods

Item 6a. Clearly define the outcome that is predicted by the prediction model, including how and when assessed.

The outcome is time to local progression based on the RECIST criteria and overall survival. According to the RECIST criteria, the patients are categorized in the following way, based on the longest tumour diameter (LD) at baseline and at follow up [6]:

- Complete response (CR): no target volume left at follow up
- Partial response (PR): at least 30% decrease in LD compared to baseline
- Progressive disease (PD): at least 20% increase in LD compared to baseline
- No change: no change in LD compared to baseline
- Stable disease (SD): the patient does not qualify for any of the above statements

The above categories are chosen based on changes in both the tumour and lymph nodes. Hence, a patient may be categorized as a PR even if the tumour diameter does not change between baseline and follow up. Therefore, in addition to the above categories (CR, PR, PD or SD), we also collect information about whether the progression is local or distant, based on the tumour diameter to be able to evaluate the response of the tumour alone using the RECIST criteria.

The following information is collected:

- The time for diagnosis (primary tumour or recurrence)
- The time for the start of the RT treatment.
- The time of progression (PD), either local *or* distant, based on the RECIST criteria (based on both tumour and lymph nodes).
- Local or distant progression? This is determined based on the RECIST criteria (based on tumour alone).
- The time of resection of the tumour.
- The time of death.

The endpoint is:

- Time to local progression, defined as the time between start of RT to the time of local progression.
- Overall survival, defined as the time from start RT to death.

The censoring criteria, when looking at time to local progression is:

- Distant progression (if the patient has distant progression, it is not possible to know if the patient also has local progression).
- Tumour resection
- Death
- Patient moved to another hospital
- None of the above and no progression at the cut-off date (30/03 2023)

The censoring criteria, when looking at overall survival is:

- Patient is alive at the cut-off date (30/03 2023)
- Patient moved to another hospital

Item 6b. Report any actions to blind assessment of the outcome to be predicted.

No actions have been taken to blind assessment of the predictors for the outcome and other predictors. This is not expected to influence the data, as none of the predictors or endpoinds depend on subjective decisions.

Item 7a. Clearly define all predictors used in developing the multivariable prediction model, including how and when they were measured.

Predictor	Measured how and when
Tumour GTV at baseline (cm^3)	Measured using MRI from Ingenia at baseline.
Primary tumour or local relapse	Whether the tumour being treated with RT is a local relapse or primary tumour.
DWI derived parameters	Parameters derived using the msNMF method developed by Sofie Rahbek.

Sex	M/F
Age at time of inclusion	
Time from diagnosis to RT start	
Performance status before RT start	Measured 0-4 weeks before RT

Note: All patients had a T-stage>=T3, and all patients received chemotherapy before RT (some patients also received chemotherapy after RT, and this might lead to differences in the outcome, however, it is not possible to obtain these data). These are important factors to predict outcome, however, since they are identical for all patients, they are not included in the model.

Item 7b. Report any actions to blind assessment of predictors for the outcome and other predictors.

No actions have been taken to blind assessment of the predictors for the outcome and other predictors. This is not expected to influence the data, as none of the predictors or endpoints depend on subjective decisions.

Item 8. Explain how the study size was arrived at.

All patient data meeting the eligibility criteria available from our institution by the beginning of end of March 2023 was evaluated to reach as high statistical power as possible. The eligibility criteria are:

- Primary tumour or relapse in pancreas, adenocarcinoma
- Treated with 5 fractions radiotherapy on MRL
- All five fractions were delivered
- DWI scans were acquired from at least one fraction
- The image quality of DWI scans is acceptable based on visual inspection.
- Follow up data is available for at least the 3 months follow up scan. I.e. patients

who started RT before the 1st of January 2023 are included.

Item 9. Describe how missing data were handled (for example, complete-case analysis, single imputation, multiple imputation), with details of any imputation method.

Predictor	**n missing**	**Reason for missingness – missing completely at random (MCAR)?**	**Handling of missing values**
Tumour GTV volume at baseline (cm^3)	0		
DWI derived parameters	5 scans	Missing for some fractions if DWI was not acquired.	If at least two fractions are available for a patient, the slope and intercept from a linear fit will be used.
Primary tumour or local relapse	*0*		
Sex	0		
Age	0		
Time from diagnosis to RT start	0		
Performance status before RT start	0		

Item 10a. Describe how predictors were handled in the analyses.

Predictor	**Scale/unit**	**Coding**	**Test of coding**
Tumour GTV volume	cm^3	Continous	
DWI derived parameters Chosen based on best subset selection using CV	mm^2/s	Continous	
Primary tumour or local relapse	Primary tumour or local relapse	Primary tumour: 0 Local relapse: 1	
Sex	M/F	Factor: Male: 0 Female: 1	

Age	Years	Continous.	
Time from diagnosis to RT start	Months	Continous	
Performance status before RT start	0, 1, 2	0, 1, 2	

Item 10b. Specify type of model, all modelbuilding procedures (including any predictor selection), and methods for internal validation.

The prediction framework is based on a decomposition of the DWI signal into components. This method produces “mixture maps”, showing the weights of each component for each voxel. The 10th and 90th percentiles of each component within the GTV (plus a 5 mm margin) are extracted. A linear fit is performed for each percentile as a function of fraction number. The slope of this fit as well as the value at fraction 1 are used as predictors. Likewise, the slope and value at fraction one of the median ADC are included.

Apart from the above-mentioned DWI predictors, 6 clinical predictors were included: tumour GTV volume at baseline, relapse (yes/no), sex, age, time from diagnosis to RT and performance status are included. Predictor selection is the best subset method utilizing bootstraping for cross-validation

- All subsets of variables are tested.
- For each subset, a Cox proportional Hazard model is build and tested using bootstrapping.
- The best-performing model in terms of the cross-validated likelihood is chosen
- Seedpoint used for bootstrapping: 42
- Number of bootstraps: 50

Item 10d. Specify all measures used to assess model performance and, if relevant, to compare multiple models.

One model will be chosen for each endpoint based on maximal likelihood.

Appendix IV: Variation reduction strategies

Detailed instructions on VTV delineation:

The VTV is delineated on the high b-value DWI image (b=500) following these steps:

1. Localize the target region (target region is described in the provided case descriptions)
2. Delineate hyper-intense regions, avoiding obvious artifacts or signal from a different location than the tumour site.
3. Exclude any regions that are hyper-intense on the T2W image. For this, overlaying of the co-registered images may be convenient. (T2W hyper-intense regions often reflect necrotic/cystic parts)
4. Ensure delineated regions respect anatomical boundaries, which are better visualized in the T2W images.)

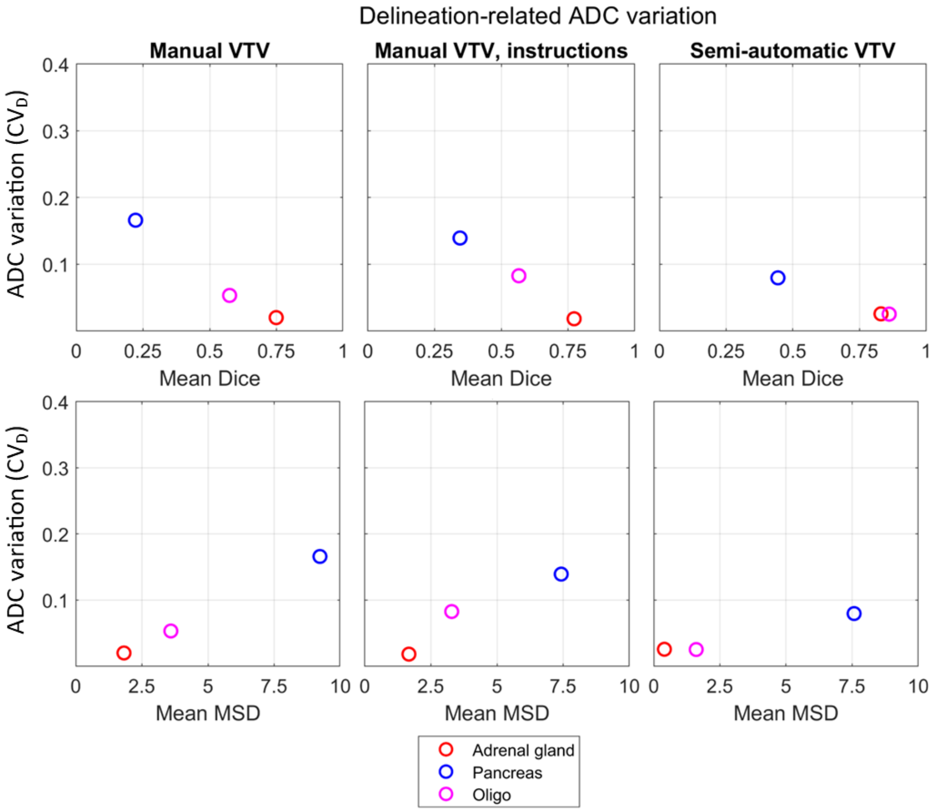


Figure A4.1. Variation reduction strategies. Delineation-related ADC variation (CV_D) as a function of Dice similarity coefficient and mean surface distance (MSD) for VTV delineations performed manually without instructions, manually with instructions, and semi-automatically using a tool. The ADC values used for comparison are median ADC values within the VTVs based on ADC maps calculated using b≥150 s/mm^2 using the scanner software.

Appendix references

[1] Stejskal EO, Tanner JE. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys. 1965;42:288–292.

[2] McGuigan A, Kelly P, Turkington RC, et al. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. World J Gastroenterol. 2018;24:4846–4861.

[3] Kirkegård J, Bojesen AB, Nielsen MF, et al. Trends in pancreatic cancer incidence, characteristics, and outcomes in Denmark 1980–2019: A nationwide cohort study. Cancer Epidemiol. 2022;80:102230.

[4] Beaton L, Bandula S, Gaze MN, et al. How rapid advances in imaging are defining the future of precision radiation oncology. Br J Cancer. 2019;120:779–790.

[5] Rahbek S, Madsen KH, Lundell H, et al. Data-driven separation of MRI signal components for tissue characterization. J Magn Reson. 2021;333:107103.

[6] Villaruz LC, Socinski MA. The clinical viewpoint: Definitions, limitations of RECIST, practical considerations of measurement. Clin Cancer Res. 2013;19:2629–2636.