Towards Diffusion Weighted MR for Evaluation of Rectal Tumours on an MR-Linac

Thomas Weststrate

In partial fulfilment of the requirements for the degree of
Master of Science
at Delft University of Technology,
to be defended publicly on 31 October 2022

Faculty: Mechanical, Maritime and Materials Engineering
Department: Biomedical Engineering
Programme: Medical Physics

Mentors / Supervisors: Casper Beijst, UMC Utrecht
                          Tim Schakel, UMC Utrecht
                          Marco van Vulpen, TU Delft
Graduation committee: Marco van Vulpen, TU Delft
                          Marlies Goorden, TU Delft
                          Sebastian Weingärtner, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl

Abstract—**Purpose:** Pre- and post-treatment MRI based Diffusion Weighted Imaging (DWI) has shown to be an effective predictor and indicator of treatment response in cancer. However, clinical applications of the Apparent Diffusion Coefficient (ADC) requires sufficient precision; this can be assessed with repeatability studies. Therefore, the purpose of this study was to assess the repeatability of ADC measurements on an MR-Linac for patients with locally advanced rectal cancer (LARC). Additionally, the relative change in ADC during treatment was compared against the repeatability.

Methods: For 17 patients, DWI was performed once on 3T MRI during pre-treatment, and twice on an 1.5T MR-Linac during each treatment fraction. Manual delineations of the Gross Tumour Volume (GTV) were created by the author. In addition two semi-automatic delineation methods were implemented, which used; a registration pipeline to propagate T2 delineations; a geometric distortion correction algorithm to correct for DWI susceptibility artefacts. Based on visual inspection the most accurate and consistent delineations were used to calculate the ADC; Bland-Altman repeatability coefficient (RC), within subject coefficient of variation (wCV), and the intraclass correlation coefficient (ICC). Lastly, the relative change in mean and median tumour ADC was compared against the RC.

Results: Manual delineations were determined to have the highest agreement with ADC tumour location and were used in subsequent calculations. The mean and median tumour ADC wCV was 7.6% (CI$_{95}$:5.5-9.2%) and 8.2% (CI$_{95}$:5.9-9.9%) respectively, which corresponds with a RC of 21.0% (CI$_{95}$:15.1-25.6%) and 22.7% (CI$_{95}$:16.3-27.4%). The reliability of the measured data was good, with an ICC of 0.81 (CI$_{95}$:0.72-0.87). In 6 out of 17 patients the relative change in ADC exceeded the RC at some point during treatment, which suggest a potential for future clinical applications.

Conclusion: Despite the low precision of rectal cancer ADC measurements, daily DWI imaging on MR-Linac has been shown to be viable for clinical applications in a subset of patients whom show significant ADC changes during treatment. Further studies are required to determine whether the measured ADC repeatability allows for clinically relevant observations.

Keywords—Diffusion weighted imaging, repeatability, rectum, MR-Linac

---◆---

1 INTRODUCTION

COLORECTAL cancer (CRC) is the colloquial term for the combination of colon, rectum and anal cancers. It is the third most commonly diagnosed form of cancer globally with an estimated 1.9 million cases in 2020, and is expected to grow to 3.2 million by 2040 [1]. CRC cases are divided into approximately 62% Colon, 38% rectum and <1% anal cancers [2]. CRC is the second leading cause of cancer related deaths in 2020 with 935.000 deaths [3] and has an overall 5-year survival rate of 65%. [4] Rectal cancer specifically is the 10th cause of cancer deaths [1] and has a overall 5-year survival rate of 67% [4].

Current standard treatment of locally advanced rectal cancer (LARC) consists of neoadjuvant chemoradiotherapy (NCRT), performed using cone-beam CT (CBCT) image guidance, followed by total mesorectal excision surgery [5]. In order to reduce the mortality rate and/or improve the quality of life for rectal cancer patients organ-sparing adaptive treatment is becoming increasingly common. The current NCRT adaptive treatment method typically uses CBCT image-guided RT prior to treatment combined with a library of plans strategy to reduce PTV margins by on average 15% [6].

In order to increase the survival rate and/or the quality of life of LARC patients, development of new treatment techniques has been gaining increasing interest in recent years. Particularly the integration of adaptive treatment plans, in which the treatment will be adapted to the specific patient and their response to the received treatment.

One promising method of improving the treatment of LARC patients is to adapt, or postpone, the surgery depending on the individual responses of the patients to the NCRT. Previous research by Sanghera et al. has shown that NCRT treatment response differs per patients, ranging from no response to a clinically complete response in 10 to 20 % of patients [7]. Adapting the treatment based on the NCRT treatment response could result in an organ sparing treatment for patients who respond well, e.g. Maas et al. showed that adapting a wait-and-see strategy with strict selection criteria for patients with a complete response resulted in equal results compared to patients with complete pathological response following total mesorectal excision surgery [8].

In order to use this method, first an accurate, and ideally non-invasive, method of determining patient response must be created. One method which seems to be able to fit these criteria is through the use of Diffusion Weighted Imaging (DWI) on MRI scanners. DWI is a functional MRI sequence which is used to measure the diffusion of hydrogen atoms. DWI allows for the acquisition of high-contrast images between the dense cellular cancerous tissue and surrounding healthy tissue, which are beneficial in tumour delineation [9]. In addition, by combining multiple DWI scans taken at different levels of diffusion sensitivity, a semi-quantized diffusion parameter can be calculated, which allows for more accurate assessment of diffusion properties. The most commonly used DWI diffusion parameter is the apparent diffusion coefficient (ADC).

Studies have found that pre-operative ADC tumour values and their change as a result of NCRT can be used as an early predictor of NCRT treatment response [10], and are a more accurate indicator of pathological response as compared to volumetric changes. [11] With the use of these

early predictions, not just the post-NCRT treatment can be adapted, but similarly the dosage during NCRT can be either decreased in order to spare tissue in bad responders or increased in moderate responders to increase the chance of a clinically complete response [12].

Another method which is used to improve cancer treatment makes use of the recently developed MR-Linac systems to perform online MRI-guided radiotherapy (MRgRT). The MR-Linac is a combined imaging and treatment system which consists of a MRI and a linear accelerator. With MR-gRT, the anatomy can be visualised daily with high quality MR images directly prior to treatment. These can be used to adapt the treatment plan based on anatomical changes such as bowel and bladder filling. Because of the higher soft tissue contrast of MRI compared to CBCT, MRgRT allows for better visualisation of the tumour and lymph nodes, thus improving dose margins, lowering radiation toxicity in the surrounding healthy tissue [13]. Furthermore, with MRgRT on MR-Linac allows for imaging to be performed while the dose is being delivered by the linear accelerator. As a result the system can be used to account for motion during treatment and similarly allows for smaller target volumes, again sparing more of the surrounding healthy tissues.

With the MR component of the MR-Linac it is possible to perform DWI scanning during treatment. With this functionality, treatment on MR-Linac would enable the acquisition of daily DWI scans, which could potentially provide information on treatment response on a daily basis. Yet, because of the structure of the MR-Linac, it is not known if the DWI measurement will provide proficiently accurate readings in order to use them for said objectives.

Any measurements taken on a MRI will include an amount of error, for example from noise, which causes the measured value to differ from the true value. An umbrella term which can be used to quantify the effects of measurement errors is the repeatability. Specifically Bland-Altman repeatability which is defined as the maximum error from the mean value in 95% of measurements [14]. In order for the DWI derived values to be used as a quantitative imaging biomarker for evaluation and prediction, it is imperative for the repeatability to be less than the variance of the values on which said conclusions are made [15]. If this would not be the case, any conclusion drawn from the data could be caused by the measurement error and not the actual patient response.

In order for the MR-Linac to integrate the linear accelerator with the MR scanner, the design of both systems had to be altered. For the MR scanner this means creating a gap for the radiation beam generated by the Linac to pass through. This results in a split gradient design, which impacts the gradient performance and can negatively impact image quality [16]. As the quality of DWI acquisitions benefit from state of the art hardware, e.g. high fidelity magnetic field gradients, the quality of DWI on MR-Linac systems is worse compared to clinical MRI systems. As a result the repeatability error on the MR-Linac is expected to exceed those measured on clinical MRI systems.

The aim of this paper is therefore to determine whether DWI repeatability performance on the MR-Linac during short-term NCRT of LARC patients is sufficient to use ADC values as a quantitative imaging biomarker for treatment evaluation and prediction. To achieve this aim, first two semi-automatic delineation methods, which rely on a registration and geometric correction workflow respectively, will be implemented in addition to standard manual delineations. The goal of these being, to determine whether these could improve the repeatability by removing human errors while also reducing treatment workload. Next, the ADC repeatability of rectal tumours on the MR-Linac will be calculated and compared against the change in ADC as a result of treatment in order to determine the effectiveness of the ADC values as a quantitative imaging biomarker. In addition a literature review about the effects of scan parameters on DWI repeatability will be performed. Based on the findings of the review several recommended alterations to the scanning sequence, which have been shown to improve repeatability performance, will be suggested.

2 THEORY

2.1 Diffusion Weighted Imaging

MRI sequences are typically divided into two different categories. These are, anatomical sequences which create an image of the structures in the measured subject, and functional sequences which sensitize the measured signal to a physiological process. Diffusion Weighted Imaging (DWI) is a functional sequence which is used to sensitize the scan towards the diffusion of hydrogen atoms in the subject. DWI performs well in the detection and evaluation of tumorous tissue. This is because the rapid cell division typical in tumorous tissues lead to an increased level of diffusion restriction in the inter and intra cellular spaces, an anisotropic diffusion, and is often associated with increased vasculature, each of which increases the contrast with the surrounding tissues in DWI.

DWI sensitizes the MRI signal to the diffusion in the subject through the use of a pair of identical gradients, symmetrically placed around a $180°$ refocusing pulse, a schematic illustration of and DWI sequence is shown in figure 1. Considering an example in which no diffusion took place the DWI sequence would first cause the atoms within a voxel to precess depending on the strength of the gradient in the voxel. Next, the atoms would flip $180°$ from the RF pulse, and finally experience a second gradient field, identical to the first, which would precess the atoms depending on the gradient strength in the voxel. In this example, without diffusion, the second gradient cancels out the phase drift created during the first gradient. As such the phase of the atoms will be re-aligned, subsequently increasing the signal strength. By introducing diffusion into the system the precession of the atoms during the second gradient will no longer be identical to that in the first, as a result the phase alignment between the atoms in not maintained, resulting in an attenuated signal strength [17], [18].

The DWI sequence can be adapted to be more or less sensitive to the diffusion by changing the gradient strength G, the interval between the gradients Δ and the duration of the gradients δ. The diffusion sensitivity is given by the b-value and is calculated with the previously mentioned variables, and the gyro-magnetic ratio γ (for hydrogen γ = 42.58 MHz/T), the equation for a square gradient pulse is:
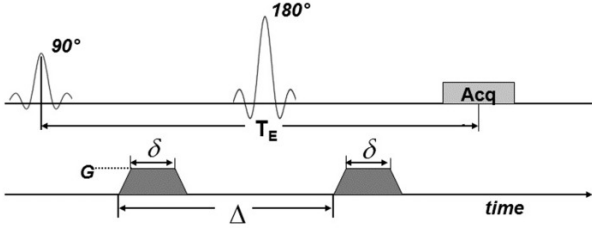
Fig. 1: Schematic representation of a DWI sequence. The sequence start with a 90° pulse, then is followed by diffusion gradients shown on the bottom line applied on around a 180° pulse, and end with a signal acquisition block (Acq). In the figure TE is the echo time, G is the amplitude of the diffusion gradients, δ is their duration and Δ is the delay time between them [19].

$$b = \gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3} \right) \qquad (1)$$

2.2 Apparent Diffusion Coefficient

The signal which is obtained from a DWI scan is in principle that of an adapted T2 scan in which the signal strength has been attenuated as a result of diffusion. For this reason the DWI signal is not solely dependent on the diffusion, and can therefor not be used in comparing diffusion across multiple scans. By combining multiple DWI scans at different levels of diffusion sensitivity it is possible to quantify the diffusion. The most common method of quantifying the DWI signal is through the use of the Apparent Diffusion Coefficient (ADC) which is calculated by fitting the data from measurements taken at multiple b-values to the following equation:

$$S(b) = S_0 \cdot e^{-ADC \cdot b} \qquad (2)$$

It should be noted that the processes of diffusion which take place in the human body are far more complex than those expressed in the function which is used to calculate the ADC. As such the ADC which is calculated with equation 2 does not reflect the true diffusivity, but is also influenced by other factors, e.g. the perfusion of fluids in the blood vessels.

2.3 Geometric Correction Algorithms

The most commonly used readout methods used for DWI is single-shot echo planar imaging (SS-EPI) as it allows for fast acquisition and low sensitivity to motion artefacts. [21] However, one large drawback of this method is that it is highly sensitive to geometric distortions in the phase encoding direction, caused by magnetic susceptibility artefacts mainly found on tissue and air boundaries. [22]

In order to correct for these errors Jezzard et al. proposed a geometric correction algorithm which measures the field imhomogeneities from a B_0 field map and uses it to unwrap the distorted MRI images [20].

In short, the method uses a double-echo gradient recalled echo phase map, to calculate the field inhomogeneities using equation 3, in which ΔB_0 denotes the field homogeneity, γ the gyromagnetic ratio, $\Delta\phi$ the phase evolution, and ΔTE the echo time difference.

$$\Delta B_0(x,y,z) = \frac{1}{2\pi\gamma \cdot \Delta T_e} \cdot \phi(x,y,z) \qquad (3)$$

Next, a third order 2D polynomial fit is applied onto the field inhomogeneities ΔB_0, with the fitted field map set to approach zero outside of the subject. Finally this deformation field map is converted into a pixel shift map which can be applied to correct for the geometric distortions. The pixel shift map is calculated using the equation below, with N pixels in the phase encoding direction, DW the dwell time, and τ_{ramp} the ramp time of the switched gradients [20].

$$\Delta r = \gamma \Delta B_0(x,y,z) \cdot N(2\tau_{ramp} + N \cdot DW) \qquad (4)$$

In figure 2 an example of the geometric distortion correction algorithm applied on a EPI imaged phantom is shown.

2.4 Registration

In the medical field images taken at either different moments and or systems are often used together. One problem that arises when combining these images is that often the subject on the images are not aligned, making direct comparison difficult. A popular solution to this problem is the use of image registration algorithms to realign the images, an example of registration between an CT and MRI scan is shown in figure 3.

Image registration algorithms consist of three main components which together allow them to determine an optimal transformation which aligns a moving image towards a fixed target image. The first part of these algorithms are the transformation models which are used to deform the moving image, depending on the model used the deformation
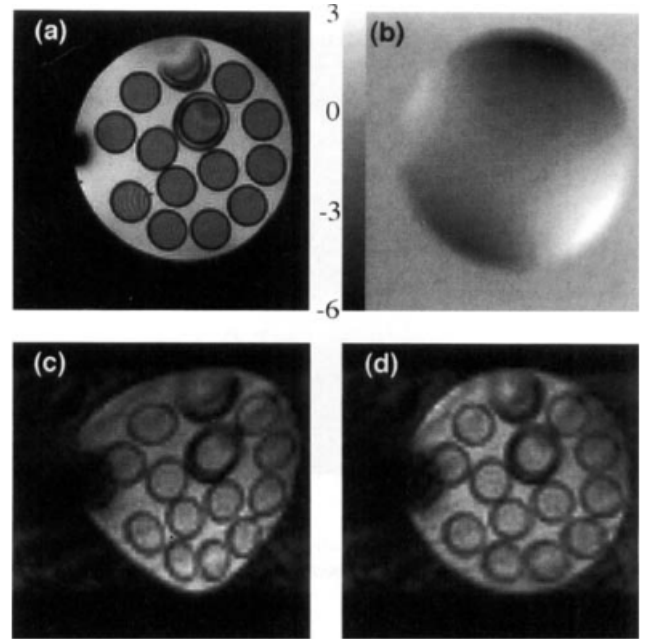

Fig. 2: The correct structure of the subject (a) is distorted because of EPI imaging. By applying the correct pixel shift map, which is based on the calculated field inhomogeneity (c), to the distorted image the original structure can be recovered (d) [20].
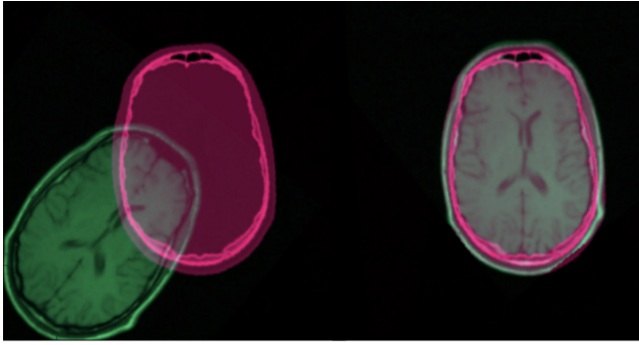
Fig. 3: Image registration example between a CT (red) and MRI scan (green) shown on the left prior to registration, and on the right post registration. After registration the two scans are aligned [23].

can range from using a combination of translation, scaling and rotation matrix's to full non-rigid deformations which can locally rerasterize the pixels. The second component of the algorithm is the evaluation function which gives a score to the current alignment. These functions are commonly known as loss functions. Two of the most popular loss functions are the least squares, which calculates the squared difference between all pixel values in the image, and the mutual information function which calculates the amount of joint information between the two images, i.e. how much information does knowing the fixed image give about the moving image. The final component of the registration algorithm is the optimisation function, which is responsible for speeding up the calculation by iteratively moving towards a better solution instead of attempting all possible deformations and selecting the best result.

3 LITERATURE STUDY

With the increased interest in the application of DWI derived diffusion parameters as quantitative imaging biomarkers for the prediction and evaluation of treatment response of cancer patients, more studies have been performed on the repeatability of said parameters so as to determine their validity as a biomarker. This research has been focused on both measuring the repeatability within different structures across the human body, as well as gaining a better understanding of how the repeatability changes depending on different scanning parameters.

In order to achieve a better understanding of how the DWI repeatability is influenced an overview of repeatability metrics across different structures was created, as well as an analysis of how several MRI procedures affect the repeatability.

3.1 Paper selection

The papers used in this study were acquired by scanning the MEDLINE database. Two different queries were performed with the search terms, *"DWI, ADC, repeatability"* and *"DWI, ADC, repeatability, rectal"*, these resulted in 129 and 8 papers respectively. Of these papers the title and abstract were examined to confirm they measured the DWI repeatability, those which did not were excluded from the study. After the

papers were filtered for relevancy 48 and 4 papers remained from the two queries for a total of 52 papers.

3.2 Overview

In recent years research on DWI has become increasingly popular, and has been shown to be beneficial in applications such as the early prediction and evaluation of treatment response of cancers [12], [24], [25], [26]. These methods rely on using the derived diffusion parameters from the DWI measurement as a quantitative imaging biomarker. However, before DWI can be used as a biomarker for clinical applications the repeatability and reproducibility of DWI measurements need to be validated [27]. The repeatability and reproducibility represent the short and long term error from repeated measurement respectively, and are vital to determine whether a change in the measured value should be considered as significant. Research on these metric is limited with only a small portion of publications on DWI focus on the measurement and improvement on repeatability [27].

In table 1 an overview of the measured repeatability of the studies included in this review are given. The statistics which are used to represent the repeatability across these studies vary, but can be separated into two different categories. The first group of statistics which are used are from Bland-Altman repeatability analysis and consist of, the Coefficient of Variation (CV), the within-subject Coefficient of Variation (wCV), the Repeatability Coefficient (RC) and the Limits of Agreement (LoA). The CV and wCV are used to express the relative variation in the measurements on the mean of all measurements, the difference being that the wCV is based on the assumed variation within a single set of measurements as opposed to the CV which uses the variation of the whole data set. When two measurements sets are used to calculate the CV and wCV they are related as such, $CV = \sqrt{2} \cdot wCV$. The RC and LoA denote the relative and absolute deviation from the mean measurement within which 95%[1] of all measurements are expected to lie. These values are calculated on the assumption that the errors are normally distributed around the true value. The RC is calculated with $RC = CV \cdot 1.96$. The second group of statistics are reliability metrics, and are used to ascertain the reliability of a dataset, i.e. in how much is the variance found in the dataset likely caused by real data. The metrics which have been used in the included papers are the Cohen's kappa (κ), the Agreement Index (AI), the Concordance Correlation Coefficient (CCC), and most commonly the Intraclass Correlation Coefficient (ICC). The ICC can be used to calculate both the intra-rater and inter-rater reliability, which represent the reliability from a single rater across exams and the reliability of multiple raters within the same exam. The ICC which was reported in table 1 is the inter-rater reliability. Following the guidelines of Koo et al. The ICC results can be interpreted as poor, moderate, good and excellent for the values 0-0.5, 0.5-0.75, 0.75-0.9, and >0.9 respectively [74].

In order to better compare the results across the papers the repeatability metrics were, where possible, converted to the wCV and summarised in table 2, showing the minimum

---

1. 95% is the most commonly used threshold for the repeatability, however other threshold can similarly be used

and maximum wCV within different structures of the human body.

From the results shown in table 2 it is clear that both across as well as within the different structures of the human body there exist large variances of reported repeatability levels, e.g. the repeatability for ADC measurements in the brain varied between 1.1% to 4.2%, while those in the pancreas were between 6.0% and 16.6%.

Similarly of note is the large variance in measured repeatability not just across but also within certain studied. In said studies the repeatability metric showed large variations based on certain changes applied on the DWI sequence, suggesting a high sensitivity of DWI performance towards certain operating conditions.

3.3 Factors affecting repeatability

From the selected papers nine different influences on DWI repeatability were explored. These can be divided into two categories, namely those associated with the measurement process and those related to the post-processing. The influences related to the measurement process which were found are; Selection of b-values used to determine quantized diffusion value, type of readout sequence used, magnetic field strength, inclusion of forms of (cardio)respiratory triggering, fat suppression methods and lastly DWI on an MR-Linac. The influences related to the post-processing are; Region of Interest type and size, DWI Diffusion parameters, and finally diffusion fitting methods.

3.3.1 b-values

To calculate the diffusion parameters which are used to quantify the diffusion in the body a combination of measurements, taken at different b-values need to be fitted to the equation associated to the diffusion parameter, such as the mono-exponentially descending equation used for ADC calculations, given in equation 2.

TABLE 1: DWI repeatability studies

	Author	Date	patients	Repeatability metrics
Brain	Bisdas et al. [28]	2013	22	ADC: CV=5.9%, D: CV=7.5% $\kappa = 0.70$
	Zakariak et al. [29]	2014	12	ADC: LoA$\approx \pm$0.09-0.12**
	Jerome et al. [30]	2017	17	ADC: CV=2.4-4.4%, D: CV=2.0-3.0%, DDC: CV=3.1-5.0%, D_k: CV=4.9-7.0%
	Lawrence et al. [31]	2021	59	ADC: wCV=0.9-1.1%
	Michoux et al. [32]	2021	24*	ADC: RC=7.5%
Breast	Aliu et al. [33]	2014	19	ADC: CV=11%
	Mürtz et al. [34]	2014	25	ADC: CV=2.2-9.9%
	Spick et al. [35]	2016	40	ADC: CV=3.2-8.3%, ICC = 0.92-0.99
	Newitt et al. [36]	2019	89	ADC: wCV=4.8%, ICC=0.97, AI=0.83
	Jerome et al. [37]	2021	20	ADC: CV=9.4%, D: CV=4.7-7.3%, DDC: CV=9.5%
	Granzier et al. [38]	2021	11*	ADC: CCC > 0.9 for 80% of radiomic features
Eyes	Lecler et al. [39]	2017	11*	ADC: CV=12-33%, ICC=0.54-0.96, D: CV=14-29%, ICC=0.59-0.94
	Zhou et al. [40]	2022	11*	ADC: RC=21.0-48.7%, ICC=0.38-0.92
Head & Neck	Hoang et al. [41]	2014	16	ADC: CV=7.4%, ICC=0.86
	Yonggang et al. [42]	2015	10*	ADC: wCV=4.9-17.3%, ICC=0.52-0.98
	Song et al. [43]	2020	43*	ADC: LoA=[-0.60-0.40]-[-0.26-0.26]
	Koopman et al. [44]	2021	10*	ADC: wCV$\approx$10%**, D: wCV$\approx$10-25%**
Intestines	Alyami et al. [45]	2021	10*	ADC: CV=5%, ICC=0.76, D: CV=10%, ICC=0.86
Kidney	M.Y.K. Bilgili [46]	2012	11*	ADC: CV=10.4-14.7%
	Michoux et al. [32]	2021	24*	ADC: RC=14-17%
liver	M.Y.K. Bilgili [46]	2012	11*	ADC: CV=7.3-10.6%
	Larsen et al. [47]	2013	12+10*	ADC: LoA=$\pm$0.1 $- \pm$0.25
	Lee et al. [48]	2015	12*	ADC: wCV=2.3-7.0%, D: wCV=3.2-16.2%
	Li et al. [49]	2017	12*	ADC: CV=7.2-14.9%, D: CV=6.5-13.7%
	Xiang et al. [50]	2028	18*	ADC: RC=10.1-10.3% D: RC=16.5-19.8%
	Pieper et al. [51]	2019	24	ADC: CV=6.9-8.1%, CCC=0.85-0.91, D: CV=7.5-9.5%, CCC=0.77-0.92
	Pathak et al. [52]	2019	8	ADC: CV=3.2-9.8%
	Pei et al. [53]	2020	23*	ADC: LoA=$\pm$0.010 $- \pm$0.035, $ICC = 0.74 - 0.91$
	Xie et al. [54]	2020	22*	ADC: LoA=$\pm$17.3 $- \pm$40.3, $ICC = 0.74 - 0.92$
	Michoux et al. [32]	2021	24*	ADC: RC=24%
	Sedlaczek et al. [55]	2022	21	ADC: RC=17%

TABLE 1: DWI repeatability studies (cont.)

	Author	Date	patients	Repeatability metrics
Lung	Weller et al. [56]	2017	23	ADC: wCV=3.9-9.6%, ICC=0.90-0.98
	Jiang et al. [57]	2017	18	ADC: RC=15.2-20.2%, ICC=0.72-0.99
	Swerkersson et al. [58]	2018	50	ADC: wCV=8.3-16.5%, D: wCV=7.8-14.8
	Wan et al. [59]	2019	15	ADC: CV=4.8-16.6%, ICC=0.95-0.98, D: CV=6.1-19.5%, ICC=0.92-0.98
Pancreas	Ma et al. [60]	2017	64	ADC: CV=10.1-15.9%, ICC=0.59-0.84
	Klaassen et al. [61]	2018	14	ADC: wCV=7.7%, D: wCV=6.7-10.0%, DDC: wCV=8.0%
	Chen et al. [62]	2021	26*	ADC: CV=6-9%, ICC=0.52-0.81
Phantom	Malyarenko et al. [63]	2013	-	ADC: 2-STD=2.8-11.3%
	Yonggang et al. [42]	2015	-	ADC: wCV=0.51-0.54%, ICC=0.89-0.99
	Belli et al. [64]	2016	-	ADC: max deviation=<5%
	Weiss et al. [65]	2017	-	ADC: CV=0.4-6.6%
	Michoux et al. [32]	2021	-	ADC: CV=0.2-1.2%
Prostate	Intven et al. [66]	2013	18	ADC: RC=5.4%
	Weiss et al. [65]	2015	16+10*	ADC: CV=11.7-14.0%
	Tamada et al. [67]	2016	56	ADC: RC=17.1-22.2%
	Michoux et al. [32]	2017	24	ADC: RC=17%
	Tsurata et al. [68]	2017	44	ADC: SD=11-21.6%, ICC=0.75-0.87
	Boss et al. [69]	2021	29	ADC: wCV=3.3-14.8%, ICC=0.76-0.94
Rectum	Intven et al. [66]	2013	18	ADC: RC=9.8%
	Chen et al. [62]	2018	32	ADC: CV=1.5-5.6%, ICC=0.92-0.99
	Sun et al. [70]	2018	66	ADC: LoA=$\pm0.12-\pm0.19$, $ICC=0.66-0.83$, D_k: LoA=$\pm0.24-\pm0.43$, $ICC=0.62-0.86$
	Bisgaard et al. [71]	2022	30	ADC: LoA (15.9-84.2th percentile)=$\pm0.04-\pm0.06$
Spleen	M.Y.K. Bilgili [46]	2012	11*	ADC: CV=8.2-13.4%
	Michoux et al. [32]	2021	24*	ADC: RC=30%
Uterus	Onodera et al. [72]	2018	33	ADC: wCV=3.3-8.9%
	Zhang et al. [73]	2020	84	ADC: CCC=0.91, D: CCC=0.89, DDC: CCC=0.92%

* Healthy volunteers

** Values were approximated from available figures

A previous study found that higher numbers of b-values used for fitting the diffusion parameter resulted in a decreased RC. However, these results are highly dependent on the range and distribution of b-values included in these calculations [62].

As the b-value changes so too does the level of diffusion sensitivity, as a result the main contributor to the signal can similarly change. For low b-value measurements ($0-200$ s/mm^2) the signal consists mainly on the effects of capillary perfusion, at higher sensitivity (>500 s/mm^2) the effects of perfusion are mostly attenuated and signal now originates mainly from lesion related diffusion, finally at very high sensitivity (>1000 s/mm^2) the signal from extracellular fluid diffusion is attenuated, resulting in signal mostly origination from intracellular diffusion [62].

In multiple studies the greatest influence on repeatability was correlated to the highest b-value measurement used in the diffusion parameter fitting. Inclusion of high b-value measurements (>1500 s/mm^2) were not found to increase repeatability [62], and was shown to decrease repeatability when measured on lower field strength scanner as the effect of Rician noise on the signal increases [64]. Similarly a decrease in the maximum b-value (<1000 s/mm^2) was also observed to have a markedly detrimental effect on repeatability, with Larsen et al. reporting an approximate factor two increase in RC when using (0, 500) instead of (0, 1000) as b-values for ADC calculations, suspectedly caused by the higher sensitivity in perfusion changes [47]. This is further supported by research from Bilgili which showed a significant increase in coefficient of variation for scans which used a lower maximum b-value [46].

3.3.2 Readout

Part of the DWI sequence is dedicated to the readout of the diffusion sensitised signal. Currently the default readout method for DWI is Single Slice Echo Planar Imaging [65]. Across the papers four alternative readout suggestions were compared to SS-EPI to determine their benefits and disadvantages.

Simultaneous multi-slice EPI (SMS-EPI) operates similarly to SS-EPI, however during each repetition time, multiple slices are obtained instead of just one, consequently decreasing scan duration, while preserving image quality. Weiss et al. compared the repeatability of ADC in prostate using both SS-EPI and SMS-EPI and found a comparable CV in patients using both readout sequences, while Simul-

taneous multi-slice EPI allowed for a 50% reduction in scan time [65]. Pei et al. compared the impact of the two different readout methods in combination with several breathing schemes on the ADC repeatability in the liver. According to their findings, the scans obtained using a breath-hold scheme similarly resulted in comparable repeatability performance between SS-EPI and SMS-EPI. However, when the breathing scheme was set to either respiratory triggering or free breathing, SS-EPI gave improved results compared to SMS-EPI [53]. One explanation of these findings could be that SMS-EPI for DWI is more sensitive to motion artefacts. Under these conditions the repeatability is expected to perform better in low motion sensitive organs such as the prostate, or under a breath-hold breathing scheme in which respiratory motion is limited, particularly in healthy volunteers with excellent breath hold capability such as in the study from Pei et al..

Readout-segmented EPI (RS-EPI) is again similar to SS-EPI, and can be considered the counter part to SMS-EPI,

as in RS-EPI each slice is separated into smaller segments along the readout direction. As a result the RS-EPI sequence requires a much shorter echo-spacing which results in a significant reduction of susceptibility and T_2^* blurring artefacts, especially on higher field strength scanners [75]. When comparing SS-EPI repeatability to readout-segmented EPI, Zhou et al. found a decrease in the RC likely caused by reduction of artefacts and blurring [40]. Contradictory, a similar study of the liver by Xie et al. found an increase in the RC when using readout-segmented EPI, which is likely caused by the increase in motion sensitivity of readout-segmented EPI caused by the increased scan time [54].

Reduced Field of View (rFOV) EPI, reduces the FOV from the conventional full FOV containing the whole anatomy to a reduced FOV containing only the region of interest. This is achieved by using two-dimensional spatially selective excitations. By reducing the FOV, only part of the subject needs to be spatially encoded in the phase encoding direction. Correspondingly the readout time of the scan

TABLE 2: Overview of minimum and maximum repeatability measured in different structures of the body

	Author	Date	patients	Repeatability**	

Brain	Bisdas et al. [28]
	Zakariak et al. [29]
	Jerome et al. [30]
	Lawrence et al. [31]
	Michoux et al. [32]
		2013
		2014
		2017
		2021
		2021
			22
			12
			17
			59
			24*
				ADC:
				  wCV: 1.1% - 4.2%
				D:
				  wCV: 1.4% - 5.3%
					DDC:
					  wCV: 3.5%
					D_k:
					  wCV: 5.4%

Breast	Aliu et al. [33]
	Mürtz et al. [34]
	Spick et al. [35]
	Newitt et al. [35]
	Jerome et al. [37]
	Granzier et al. [38]
		2014
		2014
		2016
		2019
		2021
		2021
			19
			25
			40
			89
			20
			11*
				ADC:
				  wCV: 2.6% - 11%
				D:
				  wCV: 4.7% - 7.3%
					DDC:
					  wCV: 9.45%
					D_k:
					  wCV: -

eyes	Lecler et al. [39]
	Zhou et al. [40]
		2017
		2022
			11*
			11*
				ADC:
				  wCV: 7.6% - 17.6%
				D:
				  wCV: 14%
					DDC:
					  wCV: -
					D_k:
					  wCV: -

Head & Neck	Hoang et al. [41]
	Yonggang et al. [42]
	Song et al. [43]
	Koopman et al. [44]
		2014
		2015
		2020
		2021
			16
			10*
			43
			10*
				ADC:
				  wCV: 4.9% - 17.3%
				D:
				  wCV: 9.4% - 23.4%
					DDC:
					  wCV: -
					D_k:
					  wCV: -

intestines	Alyami et al. [45]	2021	10*	ADC:
				  wCV: 5%
				D:
				  wCV: 10%
					DDC:
					  wCV: -
					D_k:
					  wCV: -

Kidney	M.Y.K. Bilgili [46]
	Michoux et al. [32]
		2012
		2021
			11*
			24*
				ADC:
				  wCV: 5.1% - 14.7%
				D:
				  wCV: -
					DDC:
					  wCV: -
					D_k:
					  wCV: -

Liver	M.Y.K. Bilgili [46]
	Larsen et al. [47]
	Lee et al. [48]
	Li et al. [49]
	Xiang et al. [50]
	Pieper et al. [51]
	Pathak et al. [52]
	Pei et al. [53]
	Xie et al. [54]
	Michoux et al. [32]
	Sedlaczek et al. [55]
		2012
		2013
		2015
		2017
		2018
		2019
		2019
		2020
		2020
		2021
		2022
			11*
			12 + 10*
			12*
			12*
			18*
			24
			8
			23*
			22*
			24*
			21
				ADC:
				  wCV: 2.3% - 14.5%
				D:
				  wCV: 4.5% - 9.8%
					DDC:
					  wCV: -
					D_k:
					  wCV: -

TABLE 2: Overview of minimum and maximum repeatability measured in different structures of the body (cont.)

	Author	Date	patients	Repeatability**	
Lung	Weller et al. [56]	2017	23	ADC:	DDC:
	Swerkersson et al. [58]	2018	18	wCV: 4.8% - 16.6%	wCV: -
	Jiang et al. [57]	2017	50	D:	D_k:
	Wan et al. [59]	2019	15	wCV: 6.1% - 19.5%	wCV: -
Pancreas	Ma et al. [60]	2017	64	ADC:	DDC:
	Klaassen et al. [61]	2018	14	wCV: 6.0% - 15.9%	wCV: 8.0%
	Chen et al. [62]	2021	26*	D:	D_k:
				wCV: 6.7% - 10.0%	wCV: -
Phantom	Malyarenko et al. [63]	2013	-	ADC:	DDC:
	yonggang et al. [42]	2015	10*	wCV: 0.5% - 11.3%	wCV: -
	Belli et al. [64]	2016	-	D:	D_k:
	Michoux et al. [32]	2021	24*	wCV: -	wCV: -
	Weiss et al. [65]	2017	16 + 10*		
Prostate	Intven et al. [66]	2013	18	ADC:	DDC:
	Weiss et al. [65]	2017	16 + 10*	wCV: 1.4% - 15.0%	wCV:
	Tamada et al. [67]	2018	56	D:	D_k:
	Michoux et al. [32]	2021	24*	wCV: -	wCV: -
	Tsurata et al. [68]	2022	44		
	Boss et al. [69]	2022	29		
Rectum	Intven et al.	2013	18	ADC:	DDC:
	Chen et al. [62]	2018	32	wCV: 1.4% - 5.6%	wCV: -
	Sun et al. [70]	2018	66	D:	D_k:
	Bisgaard et al. [71]	2021	30	wCV: -	wCV: -
Spleen	M.Y.K. Bilgili [46]	2012	11*	ADC:	DDC:
	Michoux et al. [32]	2021	24*	wCV: 8.2% - 13.4%	wCV: -
				D:	D_k:
				wCV: -	wCV: -
Uterus	Onodera et al. [72]	2018	33	ADC:	DDC:
	Zhang et al. [73]	2020	84	wCV: 4.4% - 8.1%	wCV: -
				D:	D_k:
				wCV: -	wCV: -

Healthy volunteers

** *minimum and maximum repeatability in structure, which could be converted to wCV%*

shortens which makes the scan less susceptible to motion and susceptibility artefacts [42]. In measurements by both Zhou et al. and lu et al. the implementation of reduced FOV EPI allowed for a significant decrease in the RC, for measurements of optic nerve and thyroid gland respectively [40], [42].

Finally, single shot Turbo spin echo (SS-TSE) is an alternative readout method which uses multiple 180° refocusing pulses per excitation to accelerate the readout process compared to conventional spin echo [76]. The benefit of SS-TSE over SS-EPI is its lack of sensitivity to susceptibility artefacts and deformations. For DWI measurements of the lung a clear decrease in the RC was observed when comparing the two methods, with a 55-65% reduction in CV for TSE measurements [59].

3.3.3 Scanner field strength

The field strength of an MRI scanner can have a large influence on the quality of the images generated. Current clinical scanners are often either 1.5T or 3T, with higher field strength 3T scanner becoming increasingly common in recent years [75]. In general higher field strength scanners can leverage the increased signal to create higher quality images by for example, increasing the image resolution or

increasing the SNR. However, higher field strengths suffer from increased levels of susceptibility artefacts [75].

Several multi-centre studies have been performed which use 1.5T as well as 3T scanner, in both phantoms and patients. In three of the four selected studies the measured repeatability was not affected by the field strength [36], [63], [72]. The remaining paper from Belli et al. conducted a multi-center study using a cylindrical doped water phantom and calculated the ADC using two different combinations of b-values with ranges (ADC_L: 0-1000 s/mm^2) and (ADC_H: 0-3000 s/mm^2). The repeatability of the ADC_L calculated with the low range was not effected by the scanner field strength, similar to the other studies. However, for the ADC_H scans the repeatability was Superior on the 3T scanner, this is likely a result of high Rician noise at high b-value on the 1.5T scanners [64]. Although field strength itself does not appear to influence the repeatability, higher field strength scanners are able to influence the maximum b-value scan that can reliably be measured, which in turn could affect the repeatability.

3.3.4 Triggered scanning

During measurements the structure of interest might be subject to motion as a result of respiratory or cardiac movement.

As a result of this motion the image is blurred and/or shows motion related artefacts, reducing its clinical applicability. In order to minimise the effect of motion on the image quality, in motion sensitive structures, different triggering strategies have been developed. These triggering strategies ensure that image acquisition is only performed while the triggering condition is met. As a result these methods are able to limit the effect of motion by, e.g. only scanning during breath-hold instead of during the whole respiratory cycle. Across the papers included in this study, all but one were related to DWI of the liver, which is highly sensitive to motion because of its proximity to the diaphragm. The studies compared five different triggering strategies, as well as free-breathing scans which can be considered as the baseline. The last remaining paper compared the use of respiratory triggering with free breathing for lung patients. The repeatability of DWI measurements of the liver was highest overall in measurements involving cardiac based triggering, these methods include respiratory-cardiac double triggering [49] and electrocardiography triggering [48], [50]. From the remaining strategies navigator triggering, which monitors the diaphragmatic movement as result of breathing, performed superior to regular respiratory triggering [54]. The effectiveness of respiratory triggering when compared to free breathing is inconclusive with studies finding either similar [47], [48], or improved performance [47], [53], [54]. Of note both studies from Pei et al. and Xie et al. found good performance for breath hold scanning, superior to respiratory triggering, however these studies were performed with healthy volunteers who were selected for excellent breath hold capability [53], [54]. For the lung Swerkersson et al. found similar repeatability performance in both respiratory triggering and free breathing [58].

3.3.5   Fat suppression

In order to improve the quality and clinical usefulness of MR imaging fat suppression is commonly applied. With fat suppression the strong signal from normal adipose tissue can be reduced to obtain better contrast with lesion tissue while also reducing chemical shift artefacts [77]. One paper by Mürtz et al. compared the influence of using either spectral presaturation by inversion recovery (SPIR) and short T1 inversion recovery (STIR) on ADC repeatability performance. According to their findings use of short T1 inversion recovery resulted in a significant decrease in the RC by over 50% in breast lesions [34].

3.3.6   MR-Linac

Because of the alternative design of the MR-Linac there are multiple differences as compared to a diagnostic 1.5T scanner that are detrimental to the imaging performance of the MR-Linac. Only one paper has performed a direct comparison between the repeatability performance of an MR-Linac and a MRI. In order to compare the performance of DW imaging of the brain Lawrence et al. measured the repeatability between a 1.5T MR-Linac and a 1.5T clinical MRI unit. Their results indicated similar repeatability for CSF, NAWM and NAGM ROIs on the MR-Linac as found on a clinical MRI system [31].

3.3.7   Region of Interest

There are multiple strategies that can be used to create the ROI such as, single slice, three slices and whole tumour. The single slice method uses delineations created in a single slice, typically the largest. For the three slice method the slices adjacent to that from the single slice method are also included. Lastly with the whole tumour method all slices which include the ROI are used [43], [67], [70]. Typically the slice delineations are drawn onto the whole tumour, however other methods such as partial ROIs inside the tumour are also used [60], [70]. In general repeatability performs best with whole lesion ROIs [43], [70], likely caused by the increased consistency on ROI delineations, and increases for larger tumours, likely caused by the increased number of pixels [57], [60]. However, this is not conclusive as research by Tamada et al. found a decrease in interobserver repeatability for the pancreas in whole tumour delineation as opposed to single slice ROIs [67].

3.3.8   Diffusion models

Most commonly the diffusion for DWI measurements is quantized with the ADC parameter which uses a mono-exponential model, however other methods do exist. The most common alternative method of quantisation found in the literature overview was the intravoxel incoherent motion (IVIM) followed by stretched exponential models (SEM) and Kurtosis.

The IVIM makes use of an bi-exponential function to separate the effects of diffusion into two components, the first component is the the molecular diffusion caused by the Gaussian diffusion of water molecules, and the second is the diffusion as a result of effects of perfusion. In the IVIM method the measured data across multiple b-values is fitted to equation 5, which includes the perfusion fraction f, the pure molecular diffusion parameter D, and D^* the pseudo-diffusion from incoherent micro-circulation [59].

$$\frac{S(b)}{S_0} = f \cdot e^{-bD^*} + (1-f) \cdot e^{-bD} \tag{5}$$

The SEM functions similarly to the MEM which is used to calculate the ADC,

The SEM functions similarly to the MEM in that the effects of diffusion are fitted with an mono-exponentially descending function. In SEM however, the exponents contains an additional factor α which reflects the intravoxel diffusion character spreading. The formula which is used for SEM modelling is given in equation 6, in which DDC (distributed diffusion coefficient) is the diffusion parameter [37].

$$S(b) = S_0 \cdot e^{-(b \cdot DDC)^{\alpha}} \tag{6}$$

Lastly Kurtosis modelling introduces a factor K which is used to model the effects of non-Gaussian diffusion at high b values. In the literature overview two papers used a Kurtosis model using different functions, Jerome et al. [30] used equation 7 and Sun et al. [70] used equation 8.

$$S(b) = S_0 \cdot e^{-\left(b \cdot D_k + K^2 \cdot b^2 / 6\right)} \tag{7}$$

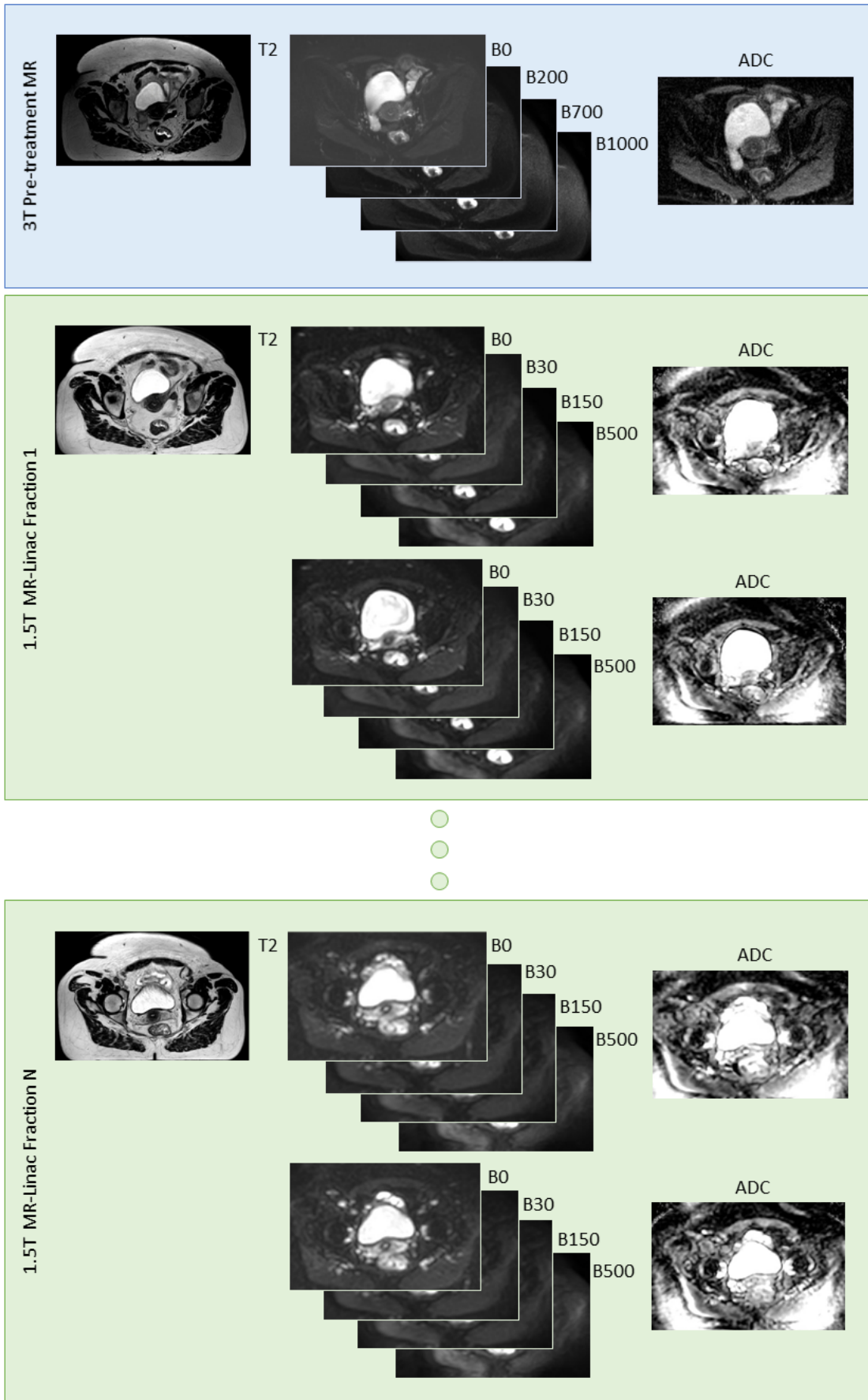$$S(b) = S_0 \cdot e^{-\left(b \cdot D_k + D_k^2 \cdot b^2 \cdot K / 6\right)} \tag{8}$$

Fig. 4: Schematic overview of measured data

From these three methods the diffusion parameters are ADC (MEM), D (IVIM), DDC (SEM), and D_k (Kurtosis). Across our findings no significant difference in repeatability between the four diffusion parameters was observed [28], [30], [37], [39], [45], [48], [49], [51], [57], [59], [61], [70]. The high correlation between the diffusion parameters implies that the signal is representative of the same diffusion source, further corroborating previous findings [30], [61].

Besides the diffusion parameters the other parameters which are used in the IVIM SEM and Kurtosis models (D^*, f, α, K) also are able to be used as biomarkers. The repeatability of these parameters, especially D^* and K, are often higher than those of the diffusion parameters, as such the resulting change needs to be significant for them to be useful.

3.3.9 Model fitting

To calculate the semi-quantized parameters from the diffusion models, the measured signal intensity across multiple b value scans are fitted. In order to determine the best fit for the provided data a loss function is introduced which is to be minimised. The most common loss function which is used for data fitting is the least-squared method, in this method the difference between the measured data points and the value which results from the fitted function are squared and summed. however other methods of determining the best fit could also be considered.

One study compared the use of nonlinear least-squares, Bayesian probability and a novel neural network for IVIM fitting. Koopman et al. found the worst performance with the simple nonlinear least-squares model, and significantly reduced wCV% for the other two methods. The neural network performed on average similarly to the Bayesian probability methods, however lacks consistency which limits it current applicability [44].

4 METHODS

To achieve to goal of the project the tasks were divided into three sub-groups. First, three different methods of creating tumour delineations were compared, secondly, the most accurate and consistent delineations from the previous task were used to calculate the ADC repeatability metrics, and finally the longitudinal ADC changes during treatment were compared against the repeatability to determine whether they can be used as a potential NCRT treatment response evaluation bio-marker.

4.1 Patient Cohort

Seventeen patients with locally advanced rectal cancer, which showed indication of treatment on the MR-Linac, were enrolled in the study and underwent DWI imaging on a 3T MRI scanner (Ingenia, Philips Healthcare, Best, The Netherlands) as part of their pre-treatment workflow. The patients received five daily fractions of 5 Gy of which six received three additional booster fractions of 5 Gy each. Treatment was performed using a 1.5T MR-Linac unit (Elekta AB, Stockholm, Sweden). Prior to the delivery of dose during each fraction two DWI sequences were performed, see figure 4 for a schematic overview.

4.2 MR Protocol

Two different systems were used to perform imaging in the study; a 3T MRI unit was used for pre-treatments scans, and a 1.5T MR-Linac during NCRT treatment fractions. A schematic overview of the data is shown in figure 4 and a table 3 provides an overview of DWI sequence parameters.

4.2.1 MR protocol

The MR protocol consisted of a transverse T2-, transverse B0- and a transverse DWI SPIR scan in addition to the default scans used for treatment planning. The Diffusion weighted scan was performed using a single shot spin-EPI readout scheme, SPIR fat suppression and free breathing. The DWI sequence consisted of 4 different diffusion scans with b values of 0, 200, 700, and 1000 s/mm^2. The resolution of the scans was 1.46 x 1.46 mm^2 and a slice thickness of 4 mm. The ADC map was calculated using the Philips Ingenia software, which uses a mono-descending exponential fit and a manual threshold to remove background pixels.

4.2.2 MR-Linac protocol

The MRL protocol consisted of two transverse T2-, one transverse B0-, and two DWI sequences. The Diffusion weighted scan was performed using DWI with a single shot spin EPI readout scheme and SPIR fat suppression. The DWI sequence consisted of 4 different diffusion scans with b values of 0, 30, 150, and 500 s/mm^2. The resolution of the scans was 1.92 x 1.92 mm^2 and a slice thickness of 5 mm. The ADC map on MRL was calculated with the same Philips algorithm as that on the MR protocol.

4.3 Delineation methods comparison

To calculate the ADC repeatability in the tumorous tissue the correct data must first be obtained from the measurement scans. To achieve this goal, delineations are created of the structure of interest. In order to correctly calculate the repeatability metrics it is imperative that the same structures accurately obtained from both the test and retest scans.

The goal of this experiment is to compare three different methods of acquiring gross tumour volume (GTV) delineations and determining which method is best suited for repeatability calculations. Three factors which are taken into consideration are the time spent on creating the delineations, their accuracy, and the consistency of the method.

TABLE 3: MRI and MR-Linac DWI sequence parameters

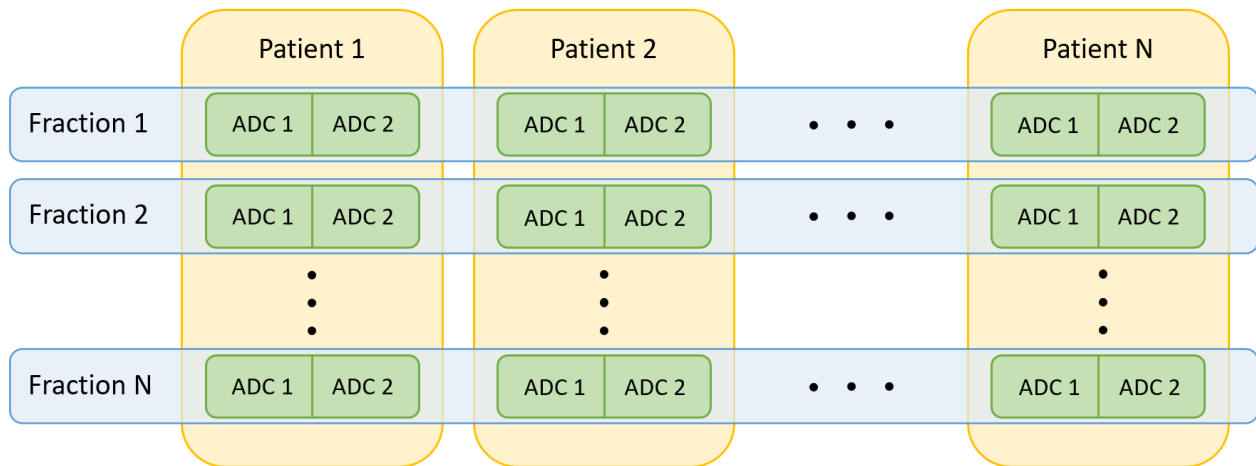| Parameters	MR DWI sequence	MR-Linac DWI sequence
Reading sequence	SS-EPI	SS-EPI
EPI-factor	97	53
FOV (mm)	420 x 420 x 160	430 x 430 x 100
Voxel size (mm)	1.46 x 1.46 x 4.0	1.92 x 1.92 x 5.0
TR (ms)	3366	3521
TE (ms)	83	66
Fat suppression	SPIR	SPIR
b-values (averages)	0(1), 200(2), 700(4), 1000(5) s/mm^2	0(2), 30(2), 150(4), 500(16) s/mm^2
Δ/δ (ms)	41.6 / 18.7	33.0 / 23.4
Bandwidth (Hz/px)	1421	2532
Scan time	4:02	4:03

Fig. 5: Data pair overview, patients groups are shown in yellow, fraction groups in blue, and data pairs in green

Of these the last two factors are the main consideration, as these relate to the quality of the delineations and the intra-reader repeatability.

During standard clinical workflow of LARC, delineations of the tumour are manually drawn by a trained physician, based on pre-treatment scans, including ADC maps. These delineations are necessary as input for the dose calculations of the treatment plan. As these delineations are already available their production time will be excluded from consideration.

In the first method an attempt was made to propagate the pre-treatment GTV delineations to the ADC maps acquired during the treatment fraction on the MR-Linac. In order to achieve this a rigid registration pipeline was applied between the ADC map of the first fraction (1F) and the pre-treatment scan. To propagate the delineations multiple steps are required; first the 1F and pre-treatment T2 scans are registered, next the pre-treatment and 1F ADC scans were resampled to their corresponding T2 scan, and finally the registration transformation between the two T2 scans was applied on the ADC scans. The registration process can be guided by supplying masks on which to evaluate the result. This process was performed using; No mask, GTV, GTV + femur, CTV, and CTV + femur masks. The registration between the two T2 scans was calculated using elastix [78], [79]. The registration parameters which were used are; the Euler transform with an adaptive stochastic gradient descend optimiser, Mattes mutual information metric, random sample size of 2048 pixels, and 4 resolution levels with each 1000 iterations.

For the second method, manual T2 based GTV delineations were created by a trained physician for each scan sequence, and the ADC maps were resampled to their corresponding T2 scan. Next, in order to correct for geometric distortions caused by susceptibility artefacts an in-house developed geometric distortion correction algorithm was applied on the ADC maps. The T2 based delineations were then subsequently added onto the corrected ADC maps.

In the last method GTV delineations were drawn manually by the author for each DWI scan sequence. The delin-

eations were created based on the highest b-value scan in the sequence, while having T2 based delineations drawn by the experienced physician available for reference.

4.4 Repeatability analysis

Repeatability is a key component in determining whether DWI can be used in treatment response evaluation as it determines the expected maximum variation in a measurement from its true value.

First, manually delineated ADC volumes were used to determine the parameters for which the repeatability will be calculated. These include the mean ADC value and the ADC histogram percentiles in steps of 2%. These values were subsequently paired up, such that each pair corresponds to the first and second DWI scan in a given measurement. Finally, the data pairs were separated into eight different groups, one for each MRL fraction, containing 17, 17, 17, 16, 15, 6, 6, & 5 data pairs respectively. A schematic overview of the data pairs and the groups is shown in figure 5.

From this data the within-subject coefficient of variation (wCV) and bias was calculated using all data points. The wCV for the mean and the percentile with the lowest wCV was subsequently converted to the repeatability coefficient (RC), which uses a 95% agreement limit as proposed by Bland-Altman. With this limit the RC gives the expected maximal variation which can be expected in 95% of all measurements. Finally, the 95% confidence interval (CI_{95}) of the RC was calculated. The resulting upper and lower bound denote the 2.5% and 97.5% confidence limits within which 95% of the population lies [80]. The results of each of these calculations were shown using Bland-Altman plots. As the measurements are performed on the same subject under similar conditions the expected bias should be zero. This was verified with a Wilcoxon signed-rank test. To verify the integrity of the calculations the reliability of the measurement method was assessed by calculating the Intraclass Correlation Coefficient (ICC) type (2,1).
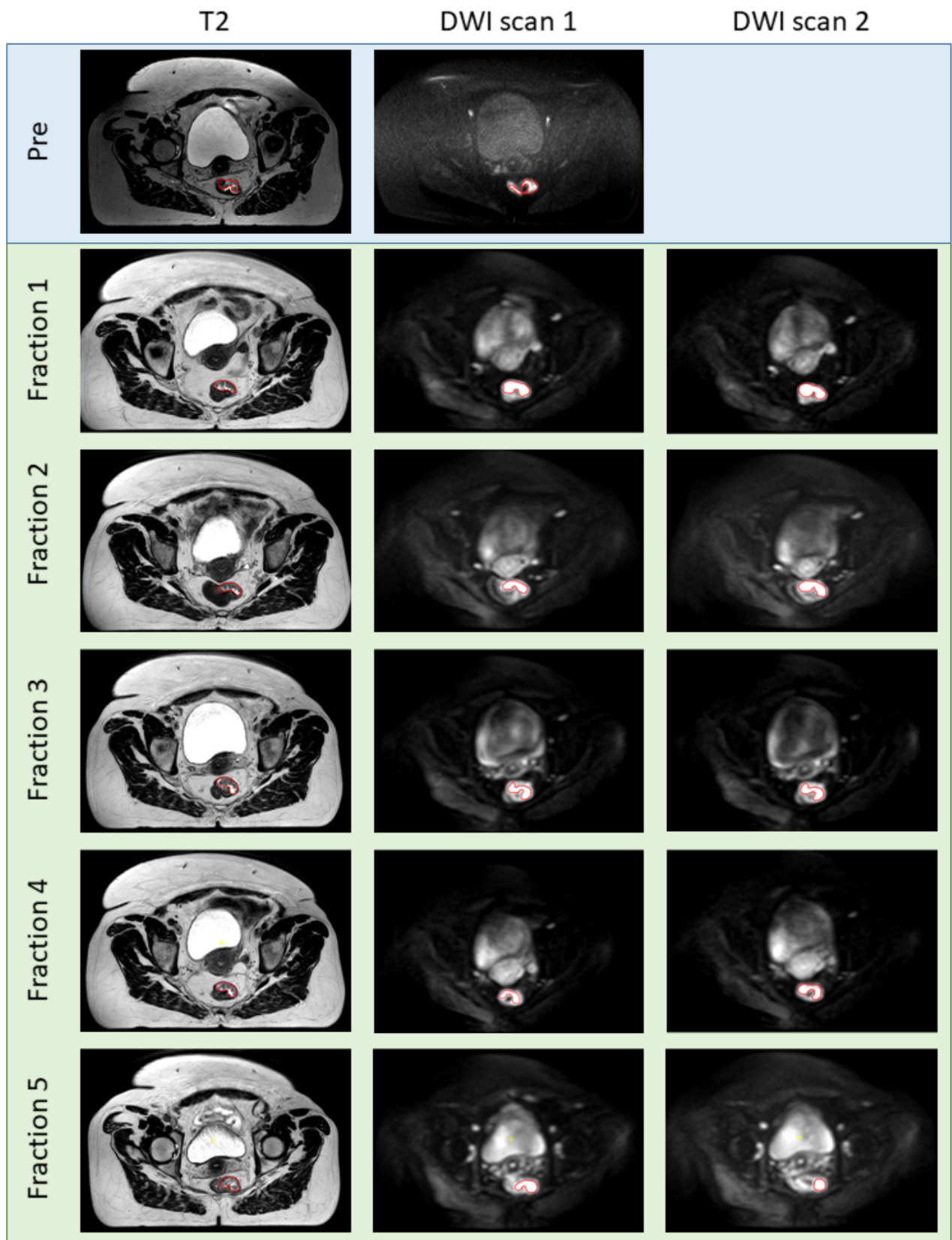
Fig. 6: Left column shows manual delineations (red) created on T2 scan, middle and right column show delineations created based on highest b-value scan (b1000 for MR (Blue) and b500 for MR-Linac (Green)).

4.5 Treatment response analysis

To determine whether daily ADC changes can be used to evaluate the response to treatment the measured change must exceed the repeatability error margin.

To verify whether this is the case, for each patient the mean and median ADC from the first DWI scan in each fraction was calculated using the manual delineations.

Next, for each patient the relative difference of the mean and median ADC (in %) from the first fraction DWI measurement, which is used as the base measurement, was plotted. In the same graph the previously calculated repeatability error threshold was plotted to determine which measurements showed sufficient change from baseline.

5 RESULTS

5.1 Delineation methods comparison

In the current paragraph, we investigate three methods for obtaining delineations of the target volume over the course of the treatment fractions; A) use registration to propagate pre-treatment GTV delineations, B) apply a geometric correction algorithm on the distorted ADC maps, C) create manual delineations.

5.1.1 Registration assisted delineations

To calculate the repeatability, accurate delineations of the structure of interest (the GTV) are of great importance. The first methods which was performed was by the use of registration. The registrations which were performed with a mask showed large visual discrepancies between the two registered images, as can be seen in the example in figure 9 which used a GTV mask to evaluate registration performance. The registrations created without the use of a mask showed far superior performance, with an average mutual information score of 0.70 (±0.13). An example of said registration is shown in figure 10. As can be seen in the figure, the registration shows an overall good performance following the contours for the majority of the major structures. However, near the edges of the rectal cavity, which is most important for our objective, signs of misalignment can be observed.

The resulting registration transforms were applied onto the resampled ADC maps to transfer the GTV delineation. The resulting propagated delineations for three patients are shown in figure 7.
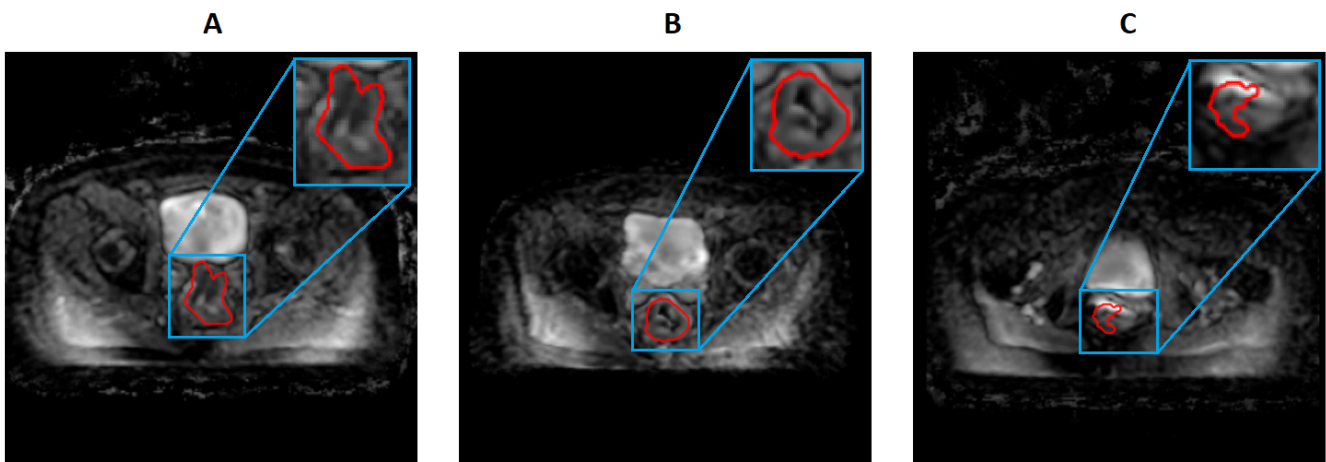

Fig. 7: Registered 1F ADC maps with delineation contours in red
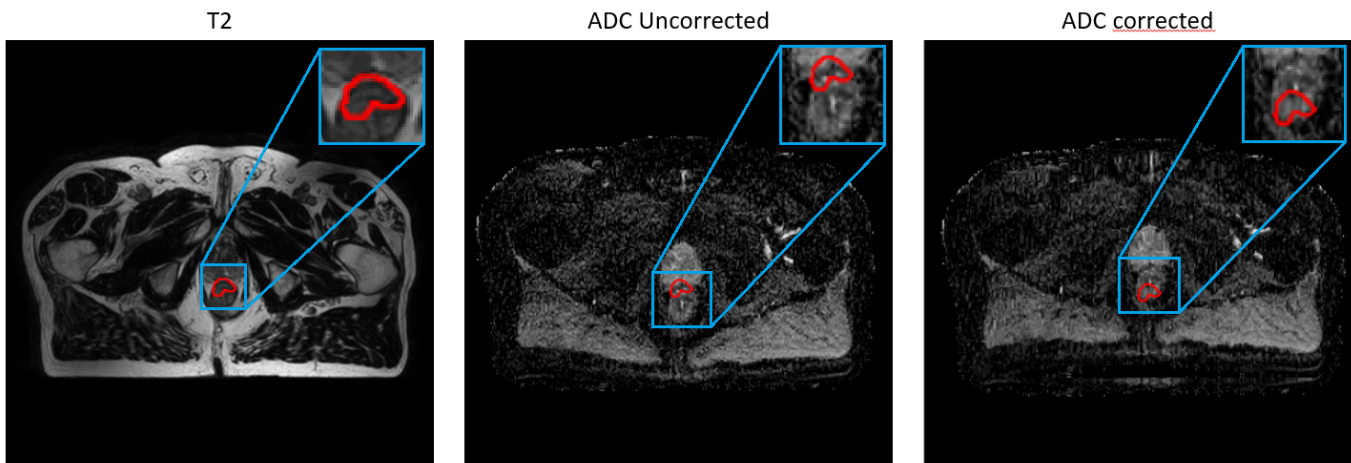

Fig. 8: Delineation contours (red) on the T2 scan (left, the original ADC scan (middle), and on the corrected ADC map (right)
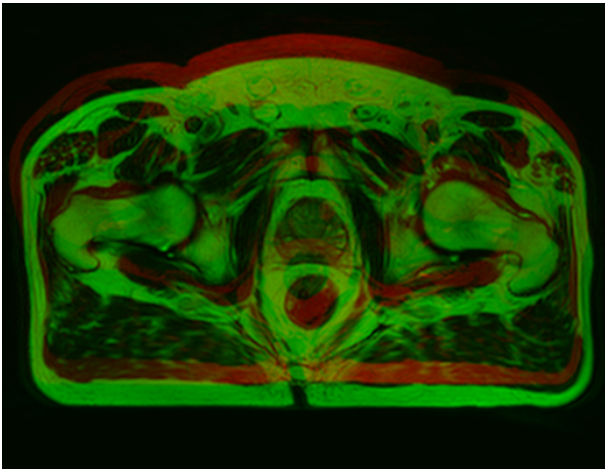
Fig. 9: Registration between pre-treatment (green) and 1F (Red) T2 scans using a GTV mask for evaluation. Mutual information metric = 0.06.
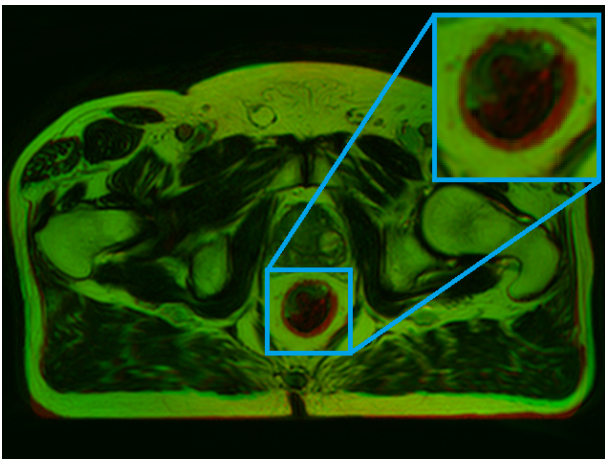

Fig. 10: Registration between pre-treatment (green) and 1F (Red) T2 scans without mask. Mutual information metric = 0.80. At the boundary of the rectum the registered images do not match, this is clearly visible in the enlarged cutout.
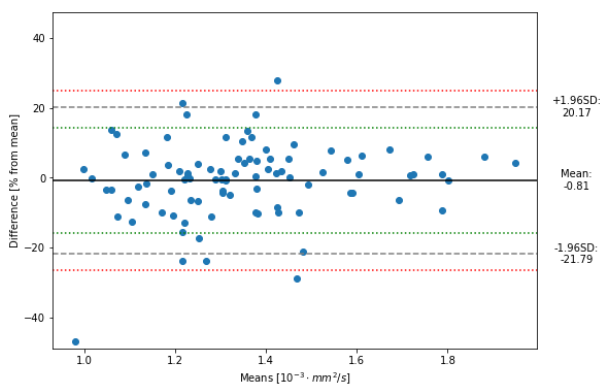

Fig. 11: Bland-Altman style plot for mean ADC repeatability. Mean value indicated by solid black line, repeatability coefficient by black dashed line, and upper and lower 95% CI shown in red and green respectively

These results show that registration has insufficient ac-

curacy and consistency for repeatability calculations. Although the method can achieve good delineations in certain patients, as seen in figure 7.A, the same method also results in misaligned delineations in other patients as can be seen on figure 7.C.

5.1.2 Geometric correction assisted delineations

In the second method new GTV delineations were created by a trained physician for each T2 scan. Next, the ADC maps were resampled to their respective T2 scan, and were subsequently subjected to an geometric distortion correction algorithm. In figure 8 an example of the results is shown. The centre image illustrates the distortion of the DWI scan by the mismatch between the delineation propagated from the T2 scan and the tumour. The right image shows the same, but with the corrected DWI scan. From this image it is clear that the algorithm has over-corrected for geometric distortions.

5.1.3 Manual delineations

In the final method manual GTV delineations were created by the author on both the test and retest scans. The delineations were drawn onto the highest b-value scan (b-1000 for MRI and b-500 for MRL) by adapting on existing T2 delineations. In figure 6 the delineations for one patient is shown. Visual assessment of the delineations shows that of the three methods that were evaluated, the delineations based on the highest b-value image showed the best agreement with the ADC maps.

5.2 Repeatability analysis

First the wCV repeatability metric was calculated for the mean ADC value using all available data. The wCV was calculated to be 7.59% with a CI_{95} of 5.46 – 9.24%, this corresponds to a RC of 21.0% and CI_{95} of 15.1 – 25.6%. The bias for the mean ADC data pairs was -0.81%, a Wilcoxon signed-rank test was performed between the data pairs and indicated insufficient proof to reject the null hypothesis of identical means between distributions (p = 0.95), therefore the bias can be assumed as null. The Results of the calculation can be found in figure 11. This means that for 95% of all measurements the value lies within 21% of the true value.

Next, the wCV was calculated for the ADC histogram percentiles in steps of 2%. The results, shown in figure 12, show that performance is lowest between the 50th (median) and the 90th percentile.

In Further calculations the 50th percentile is used. The wCV of the median ADC was 8.2% with CI_{95} of 5.9 – 9.9%. This corresponds to a RC of 22.7% with CI_{95} of 16.4 – 27.4%. The resulting bland-altman plot is shown in figure 13. Similar to the mean ADC repeatability calculation the Wilcoxon test showed insufficient proof to reject the null hypothesis (p = 0.59), as such the bias is considered null.

5.3 Treatment response analysis

For each patient the percent change in ADC mean and median at each fraction compared to start of treatment was plotted, the results are shown in figure 14 and figure 15 respectively.
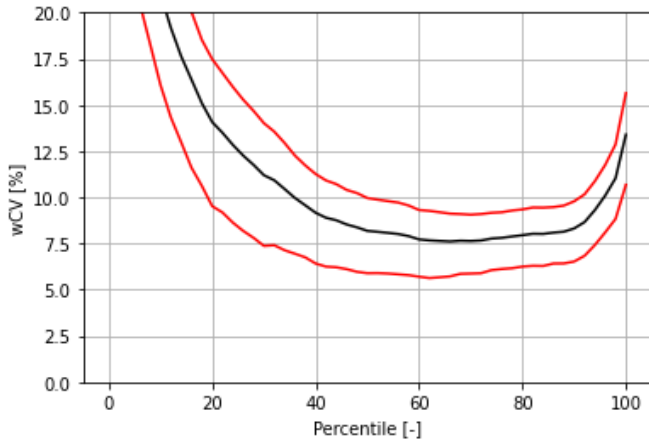
Fig. 12: The wCV as function of histogram percentile (Black) and the 95% confidence interval (red)
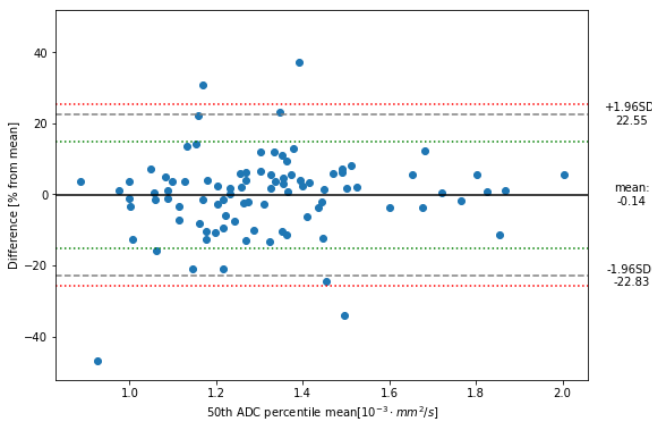

Fig. 13: Bland-Altman style plot for median ADC repeatability. Mean value indicated by solid black line, repeatability coefficient by black dashed line, and upper and lower 95% CI shown in red and green respectively

The ADC changes which occur during treatment are mainly within the repeatability error of the first base measurements. In five patients the mean ADC change exceeded the repeatability at some point during the treatment, all but the final fraction of patient 14 being at an increased ADC value as would be expected. For the median data the results were mostly similar, with the main differences being a decrease in ADC exceeding the repeatability in patient 17 during fraction 3, and a lower change in ADC for patient 13 fraction 7, and patient 14 fraction 3 which bring these point below the repeatability threshold.

6 DISCUSSION

In this paper the ADC repeatability of rectal tumours on an MR-Linac was measured and compared against the change in ADC as a result of treatment. In addition two semi-automatic delineation methods, that use registration and a geometric correction algorithm, were tested and compared against standard manual delineation procedures.

6.1 Delineation method comparison

Based on visual inspection the registration and geometric correction algorithms had inferior accuracy and consistency as compared to the manual delineations.

In the registration algorithm method the T2 based registration showed good performance overall with a mutual information score of $0.70(\pm 0.13)$ without use of a mask, yet based on visual observation the performance of the registration in the rectum specifically was not sufficient with misalignment's in the rectal walls which could have negatively impacted the resulting delineations. The biggest limitation for creating delineations with rigid registration methods were, however, not the small misalignment's previously mentioned. Instead, as the rectum contains air, the SS-EPI sequence results in large susceptibility artefacts in the rectum, which causes the ADC map to be locally distorted at the air-tissue boundaries. I.e. the tumour on the ADC map is distorted comparatively to the location found on the T2 scan. As these distortion artefacts are non-linear they can not be corrected for using rigid registration. A possible solution would be to implement non-rigid registration between the T2 and ADC map which could compensate for the artefacts, however these methods also run the risk of over-fitting to the desired output which could corrupt the data, in addition
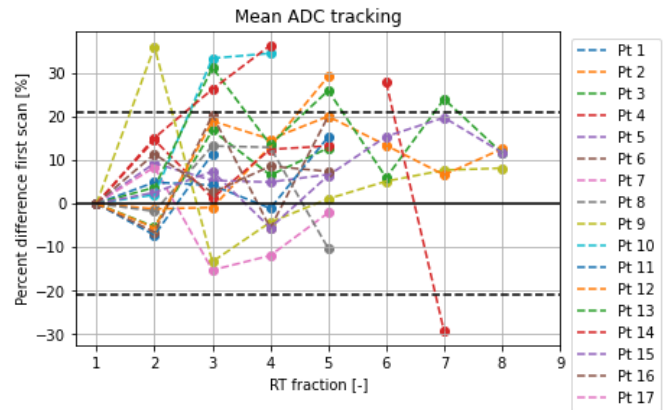

Fig. 14: The relative change in ADC mean during treatment, repeatability limits are given by the black dashed lines
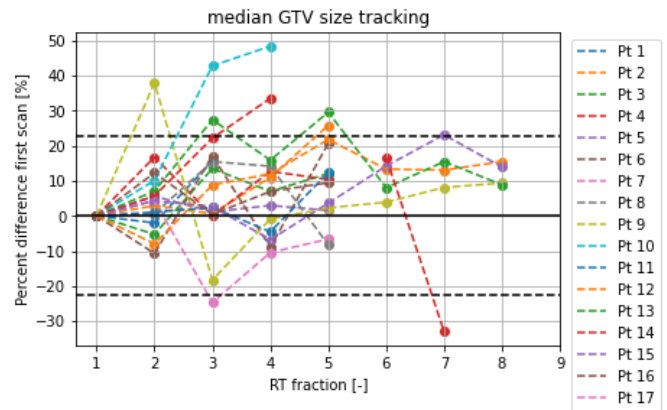

Fig. 15: The relative change in ADC median during treatment, repeatability limits are given by the black dashed lines

these methods need regularisation which is often hand-tuned by the user [81].

Another method used to compensate for the effects of susceptibility artefacts caused by the EPI readout is to implement a geometric distortion correction algorithms which through the use of a B_0 map calculates the deformation caused by said artefacts and corrects for them. In our results the correction algorithm over-corrected the susceptibility artefacts found in the ADC map. The geometric distortion correction algorithm requires information on the local magnetic field as input. This is provided from a additional B0 field map acquisition. A likely reason for the over-correction is a mismatch between the applied magnetic field shimming settings during acquisition between the DWI and B0 field map scan. This results in the magnetic field measured with the B0 field map acquisition not representing the magnetic field during the DWI acquisition. In retrospect the current B0 field map was not suited for this purpose. However, when acquired properly, this approach could still potentially provide a means to use the clinical GTV delineation.

Although both semi-automatic delineations propagation methods attempted in this paper allowed for the fast acquisition of ADC delineations with minimal workload, the resulting delineations were of insufficient accuracy and consistency for use in repeatability calculations. As such manual delineations were created by the author. The created delineations were drawn by adapting existing T2 delineations to the bright area's in the highest b-value scan which correlate to the regions with high diffusion restriction. The created delineations were validated by a trained physician to be of sufficient quality for repeatability calculations.

6.2 Repeatability analysis

In our calculations a RC of 21% (CI$_{95}$: 15.1 - 25.6%) was measured for mean rectum tumour ADC values, as measured with an MR-Linac.

Because of time limitations a second set of delineations could not be created, as such the intra-rater repeatability could not be calculated. Therefore, the effect of the non-experienced rater on the repeatability could not be determined.

In our dataset one entry was identified as being a possible outlier as the difference between the first and second ADC scan significantly exceeded those of all previous entries. However, as no concrete evidence, such as patient movement in-between b-scans in the sequence, was found it was not removed from the dataset. The RC without the suspected outlier would have been 18.8% (CI$_{95}$: 15.1 - 21.9%), 2.2% lower.

There have been four previous studies which have measured the ADC repeatability in rectal tumours. Their repeatability results were; a RC of 9.8% by Intven et al. [66]; A CV from 1.47 to 5.57% (converted to RC=2.8-10.9%) by Chen et al. [62]; a LoA of ±0.12 with a mean ADC of 0.687 (Converted to RC=17.5%) by Sun et al. [70]; and a 68.3% LoA from ±0.04 to ±0.06 by Bisgaard et al. [71]. Except for the results from Bisgaard et al. which can not be converted to RC, the result of al other studies which measured with a MRI outperform the results found with the MR-Linac in this study.

The two most likely sources for the increased repeatability with the MR-Linac as compared to the previous studies with MRI scanners are the lower SNR and the limitation on the maximum b-value which can realistically be used. The limited gradient field strength of the MR-Linac results in a lower maximum b-value used on the MR-Linac of B = 500 s/mm^2, which in previous studies on MRI systems has been found to lead to decidedly reduced repeatability as compared to sequences which use a maximum B value between 1000 and 1500 s/mm^2. [46], [47], [62]

The maximum b-value which was used in this study (b = 500 s/mm^2), is already the highest recommended b-value by the Elekta Unity MR-Linac consortium. This threshold was determined because of limitations on the decreasing SNR at higher b-values, higher sensitivity scans above the threshold were determined as unreliable. [82] However, a previous study by Lawrence et al. which compared the ADC repeatability of central nervous system tumours on a MRI and MR-Linac found similar results on both systems while using a maximum b-value of B = 800 s/mm^2, exceeding previously mentioned recommendations. [31] Therefore, further study on inclusion of a higher maximum b-value could also be considered on the MR-Linac for DWI of LARC, as it might lead to a noticeable improvement in repeatability. Besides increasing the maximum b-value, there are also multiple other methods identified in the literature study which could improve DWI performance.

The first adaptation which could be considered based on findings by Chen et al. is to increase the amount of b-values, between b=0-500 s/mm^2, in the sequence. In their study they found a significant decrease in CV for sequences which used more b-values. [62] Another modification which might lead to an improved repeatability is to switch the readout sequence to either a readout-segmented EPI, a Turbo Spin Echo sequence or a reduced field of view. Especially the use of a turbo spin echo sequence in lung cancer patients [59], and a reduced field of view sequence for the optic nerve [40] and thyroid glands [42] have been shown to significantly improve repeatability metrics as compared to conventional SS-EPI sequences. For readout-segmented EPI the results are contradictory, with the method seeming to improve the repeatability but only in the absence of motion [40], [54]. The improvement in repeatability by TSE is further corroborated by a recent study which found an increased repeatability for TSE over conventional EPI imaging [83]. One last modification which could have positive effects on the repeatability is to change the fat suppression from a SPIR to a STIR sequence. According to research by Mürtz et al. which compared these two methods for breast lesions, the adaptation of STIR resulted in a significantly improved repeatability [34].

One limitation in these improvements is that both the increase in the number of b-scans and the adaptation of the turbo spin echo and readout-segmented readout sequence will result in a noticeably longer scan duration, which might not to wanted and or possible. Under these circumstances a possible alternative could be to adapt the readout sequence to a simultaneous multi-slice EPI sequence. For organs that have a low motion sensitivity this sequence has been found to result in a significant decrease in scan duration while having minimal impact on repeatability [53], [65]. The time

which has been saved by the SMS-EPI sequence could then be invested into adaptations which could improve the repeatability, e.g. including more b-value scans.

Using these methods might result in a significant improvement in the repeatability of DWI measurements on an MR-Linac, further studies are required to determine the actual impact of said changes for rectal cancer patients.

6.3 Treatment response analysis

Numerous previous studies have shown that treatment prediction based on analysis of pre and post NCRT ADC values is viable with MRI systems, a summary of previous studies is given in a paper by Schurink et al. [84].

One major difference between these previous studies and this one is the time span in which the change in ADC is measured. To our knowledge this is the first study which tracked daily ADC changes in a low fractionated treatment plan. As a result the duration between the first and last DWI measurement on the MR-Linac never exceeded two weeks. In comparison, e.g. a previous study by Intven et al. which assessed the predictive potential of DWI for selecting good responders with LARC used MRI scans 1-2 weeks before NCRT, which lasted five weeks, and 1-2 weeks before surgery, which was 6-10 weeks post NCRT, totalling between 12-19 weeks between the two measurements.

The finding in this study indicated that the majority of measurements taken with an MR-Linac did not show a change in mean or median ADC, from start of treatment, exceeding the repeatability error. In total only 6 out of the 17 patients included in this study showed a change exceeding the repeatability error at some point during treatment, in 4 patients exceeding the upper bound, in one patients the lower bound, and in one patient both the upper and lower bound. Further study on the significance of these results are needed to determine their clinical relevancy.

Besides improving the repeatability the number of significant measurement could also possibly be increased by performing an additional scan post-treatment. Based on previous studies the parameter which is mainly used for treatment evaluation is the relative change in ADC between pre-treatment ADC scans and scans taken 6-8 weeks post treatment [85]. As the results in this study only include data measured during short-term NCRT, lasting at most 2 weeks, it is possible for the effects of treatment on ADC to manifest post treatment. Further studies are required to determine whether this is indeed the case.

7 CONCLUSION

The repeatability study of rectal cancer ADC measurements on an MR-Linac indicated a significant measurement error, exceeding previous findings on MRI systems. As a result only a limited amount of ADC measurements resulted in significant differences. However, despite the low precision of ADC measurements, the inclusion of significant measurements indicates that daily DWI imaging on MR-Linac could conceivably be used for clinical applications such as the prediction and evaluation of treatment response.

Further studies are required to determine whether the measured ADC repeatability allows for clinically relevant observations.

REFERENCES

[1] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors," *Gastroenterology Review*, vol. 14, no. 2, p. 89–103, 2019.
[2] Y. Xi and P. Xu, "Global colorectal cancer burden in 2020 and projections to 2040," *Translational Oncology*, vol. 14, no. 10, p. 101174, 2021.
[3] "Rectum - international agency for research on cancer."
[4] "Colorectal cancer - statistics," May 2022.
[5] C. J. van de Velde, P. G. Boelens, J. M. Borras, J.-W. Coebergh, A. Cervantes, L. Blomqvist, R. G. Beets-Tan, C. B. van den Broek, G. Brown, E. Van Cutsem, and et al., "Eurecca colorectal: Multidisciplinary management: European consensus conference colon amp; rectum," *European Journal of Cancer*, vol. 50, no. 1, 2014.
[6] C. Beekman, B. van Triest, S. van Beek, J.-J. Sonke, and P. Remeijer, "Margin and ptv volume reduction using a population based library of plans strategy for rectal cancer radiotherapy," *Medical Physics*, vol. 45, p. 4345–4354, Sep 2018.
[7] P. Sanghera, D. Wong, C. McConkey, J. Geh, and A. Hartley, "Chemoradiotherapy for rectal cancer: An updated analysis of factors affecting pathological response," *Clinical Oncology*, vol. 20, no. 2, p. 176–183, 2008.
[8] M. Maas, R. G. Beets-Tan, D. M. Lambregts, G. Lammering, P. J. Nelemans, S. M. Engelen, R. M. van Dam, R. L. Jansen, M. Sosef, J. W. Leijtens, and et al., "Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer," *Journal of Clinical Oncology*, vol. 29, no. 35, p. 4633–4640, 2011.
[9] H. Ligtenberg, T. Schakel, J. W. Dankbaar, L. N. Ruiter, B. Peltenburg, S. M. Willems, N. Kasperts, C. H. Terhaard, C. P. Raaijmakers, M. E. Philippens, and et al., "Target volume delineation using diffusion-weighted imaging for mr-guided radiotherapy: A case series of laryngeal cancer validated by pathology," *Cureus*, 2018.
[10] M. Intven, O. Reerink, and M. Philippens, "Diffusion-weighted mri in locally advanced rectal cancer," *Strahlentherapie und Onkologie*, vol. 189, no. 2, p. 117–122, 2012.
[11] M. Lambrecht, V. Vandecaveye, F. De Keyzer, S. Roels, F. Penninckx, E. Van Cutsem, F. Claus, and K. Haustermans, "Value of diffusion-weighted magnetic resonance imaging for prediction and early assessment of response to neoadjuvant radiochemotherapy in rectal cancer: Preliminary results," *International Journal of Radiation Oncology*Biology*Physics*, vol. 82, no. 2, p. 863–870, 2012.
[12] L. Jacobs, M. Intven, N. van Lelyveld, M. Philippens, M. Burbach, K. Seldenrijk, M. Los, and O. Reerink, "Diffusion-weighted mri for early prediction of treatment response on preoperative chemoradiotherapy for patients with locally advanced rectal cancer," *Annals of Surgery*, vol. 263, no. 3, p. 522–528, 2016.
[13] L. Boldrini, M. Intven, M. Bassetti, V. Valentini, and C. Gani, "Mr-guided radiotherapy for rectal cancer: Current perspective on organ preservation," *Frontiers in Oncology*, vol. 11, Mar 2021.
[14] J. Martin Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, p. 307–310, 1986.
[15] A. Shukla-Dave, N. A. Obuchowski, T. L. Chenevert, S. Jambawalikar, L. H. Schwartz, D. Malyarenko, W. Huang, S. M. Noworolski, R. J. Young, M. S. Shiroishi, and et al., "Quantitative imaging biomarkers alliance (qiba) recommendations for improved precision of dwi and dce-mri derived biomarkers in multicenter oncology trials," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, 2018.
[16] J. Habrich, S. Boeke, M. Nachbar, K. Nikolaou, F. Schick, C. Gani, D. Zips, and D. Thorwarth, "Repeatability of diffusion-weighted magnetic resonance imaging in head and neck cancer at a 1.5 t mr-linac," *Radiotherapy and Oncology*, 2022.
[17] D.-M. Koh and D. J. Collins, "Diffusion-weighted mri in the body: Applications and challenges in oncology," *American Journal of Roentgenology*, vol. 188, no. 6, p. 1622–1635, 2007.
[18] T. C. Kwee, T. Takahara, R. Ochiai, R. A. Nievelstein, and P. R. Luijten, "Diffusion-weighted whole-body imaging with background body signal suppression (dwibs): Features and potential applications in oncology," *European Radiology*, vol. 18, no. 9, p. 1937–1952, 2008.
[19] E. Vaghefi and P. J. Donaldson, "An exploration into diffusion tensor imaging in the bovine ocular lens," *Frontiers in Physiology*, vol. 4, Mar 2013.

[20] P. Jezzard and R. S. Balaban, "Correction for geometric distortion in echo planar images from b0 field variations," *Magnetic Resonance in Medicine*, vol. 34, no. 1, p. 65–73, 1995.

[21] G. Nketiah, K. M. Selnæs, E. Sandsmark, J. R. Teruel, B. Krüger-Stokke, H. Bertilsson, T. F. Bathen, and M. Elschot, "Geometric distortion correction in prostate diffusion-weighted mri and its effect on quantitative apparent diffusion coefficient analysis," *Magnetic Resonance in Medicine*, vol. 79, no. 5, p. 2524–2532, 2017.

[22] A. Tong, G. Lemberskiy, C. Huang, K. Shanbhogue, T. Feiweier, and A. B. Rosenkrantz, "Exploratory study of geometric distortion correction of prostate diffusion-weighted imaging using b0 map acquisition," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 5, p. 1614–1619, 2019.

[23] E. Kamoun, "Image registration: From sift to deep learning," Jan 2020.

[24] S. Marzi, A. Farneti, A. Vidiri, F. Di Giuliano, L. Marucci, F. Spasiano, I. Terrenato, and G. Sanguineti, "Radiation-induced parotid changes in oropharyngeal cancer patients: The role of early functional imaging and patient/treatment-related factors," *Radiation Oncology*, vol. 13, no. 1, 2018.

[25] Y. Gao, V. Ghodrati, A. Kalbasi, J. Fu, D. Ruan, M. Cao, C. Wang, F. C. Eilber, N. Bernthal, S. Bukata, and et al., "Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion mri and a deep neural network with generative adversarial network-based data augmentation," *Medical Physics*, vol. 48, no. 6, p. 3262–3372, 2021.

[26] X. Geng, D. Zhang, S. Suo, J. Chen, F. Cheng, K. Zhang, Q. Zhang, L. Li, Y. Lu, J. Hua, and et al., "Using the apparent diffusion coefficient histogram analysis to predict response to neoadjuvant chemotherapy in patients with breast cancer: Comparison among three region of interest selection methods," *Annals of Translational Medicine*, vol. 10, no. 6, p. 323–323, 2022.

[27] A. Shukla-Dave, N. A. Obuchowski, T. L. Chenevert, S. Jambawalikar, L. H. Schwartz, D. Malyarenko, W. Huang, S. M. Noworolski, R. J. Young, M. S. Shiroishi, and et al., "Quantitative imaging biomarkers alliance (qiba) recommendations for improved precision of dwi and dce-mri derived biomarkers in multicenter oncology trials," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, 2018.

[28] S. Bisdas, T. S. Koh, C. Roder, C. Braun, J. Schittenhelm, U. Ernemann, and U. Klose, "Intravoxel incoherent motion diffusion-weighted mr imaging of gliomas: Feasibility of the method and initial results," *Neuroradiology*, vol. 55, no. 10, p. 1189–1196, 2013.

[29] R. Zakaria, K. Das, M. Bhojak, M. Radon, V. Sluming, C. Walker, and M. Jenkinson, "The reliability of routine clinical post-processing software in assessing potential diffusion-weighted mri "biomarkers" in brain metastases," *Magnetic Resonance Imaging*, vol. 32, no. 3, p. 291–296, 2014.

[30] N. P. Jerome, K. Miyazaki, D. J. Collins, M. R. Orton, J. A. d'Arcy, T. Wallace, L. Moreno, A. D. Pearson, L. V. Marshall, F. Carceller, and et al., "Repeatability of derived parameters from histograms following non-gaussian diffusion modelling of diffusion-weighted imaging in a paediatric oncological cohort," *European Radiology*, vol. 27, no. 1, p. 345–353, 2016.

[31] L. S. Lawrence, R. W. Chan, H. Chen, B. Keller, J. Stewart, M. Ruschin, B. Chugh, M. Campbell, A. Theriault, G. J. Stanisz, and et al., "Accuracy and precision of apparent diffusion coefficient measurements on a 1.5 t mr-linac in central nervous system tumour patients," *Radiotherapy and Oncology*, vol. 164, p. 155–162, 2021.

[32] N. F. Michoux, J. W. Ceranka, J. Vandemeulebroucke, F. Peeters, P. Lu, J. Absil, P. Triqueneaux, Y. Liu, L. Collette, I. Willekens, and et al., "Repeatability and reproducibility of adc measurements: A prospective multicenter whole-body-mri study," *European Radiology*, vol. 31, no. 7, p. 4514–4527, 2021.

[33] S. O. Aliu, E. F. Jones, A. Azziz, J. Kornak, L. J. Wilmes, D. C. Newitt, S. A. Suzuki, C. Klifa, J. Gibbs, E. C. Proctor, and et al., "Repeatability of quantitative mri measurements in normal breast tissue," *Translational Oncology*, vol. 7, no. 1, p. 130–137, 2014.

[34] P. Mürtz, M. Tsesarskiy, A. Kowal, F. Träber, J. Gieseke, W. A. Willinek, C. C. Leutner, A. Schmiedel, and H. H. Schild, "Diffusion-weighted magnetic resonance imaging of breast lesions: The influence of different fat-suppression techniques on quantitative measurements and their reproducibility," *European Radiology*, vol. 24, no. 10, p. 2540–2551, 2014.

[35] C. Spick, H. Bickel, K. Pinker, M. Bernathova, P. Kapetas, R. Woitek, P. Clauser, S. H. Polanec, M. Rudas, R. Bartsch, and et al., "Diffusion-weighted mri of breast lesions: A prospective clinical investigation of the quantitative imaging biomarker characteristics of reproducibility, repeatability, and diagnostic accuracy," *NMR in Biomedicine*, vol. 29, no. 10, p. 1445–1453, 2016.

[36] D. C. Newitt, Z. Zhang, J. E. Gibbs, S. C. Partridge, T. L. Chenevert, M. A. Rosen, P. J. Bolan, H. S. Marques, S. Aliu, W. Li, and et al., "Test–retest repeatability and reproducibility of adc measures by breast dwi: Results from the acrin 6698 trial," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 6, p. 1617–1628, 2018.

[37] N. P. Jerome, I. Vidić, L. Egnell, T. E. Sjøbakk, A. Østlie, H. E. Fjøsne, P. E. Goa, and T. F. Bathen, "Understanding diffusion-weighted mri analysis: Repeatability and performance of diffusion models in a benign breast lesion cohort," *NMR in Biomedicine*, 2021.

[38] R. Granzier, A. Ibrahim, S. Primakov, S. Keek, I. Halilaj, A. Zwanenburg, S. Engelen, M. Lobbes, P. Lambin, H. Woodruff, and et al., "Test–retest data for the assessment of breast mri radiomic feature repeatability," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 2, p. 592–604, 2021.

[39] A. Lecler, J. Savatovsky, D. Balvay, M. Zmuda, J.-C. Sadik, O. Galatoire, F. Charbonneau, O. Bergès, H. Picard, L. Fournier, and et al., "Repeatability of apparent diffusion coefficient and intravoxel incoherent motion parameters at 3.0 tesla in orbital lesions," *European Radiology*, vol. 27, no. 12, p. 5094–5103, 2017.

[40] F. Zhou, Q. Li, X. Zhang, H. Ma, G. Zhang, S. Du, L. Zhang, T. Benkert, and Z. Zhang, "Reproducibility and feasibility of optic nerve diffusion mri techniques: Single-shot echo-planar imaging (epi), readout-segmented epi, and reduced field-of-view diffusion-weighted imaging," *BMC Medical Imaging*, vol. 22, no. 1, 2022.

[41] J. K. Hoang, K. R. Choudhury, J. Chang, O. I. Craciunescu, D. S. Yoo, and D. M. Brizel, "Diffusion-weighted imaging for head and neck squamous cell carcinoma: Quantifying repeatability to understand early treatment-induced change," *American Journal of Roentgenology*, vol. 203, no. 5, p. 1104–1108, 2014.

[42] Y. Lu, V. Hatzoglou, S. Banerjee, H. E. Stambuk, M. Gonen, A. Shankaranarayanan, Y. Mazaheri, J. O. Deasy, A. R. Shaha, R. M. Tuttle, and et al., "Repeatability investigation of reduced field-of-view diffusion-weighted magnetic resonance imaging on thyroid glands," *Journal of Computer Assisted Tomography*, p. 1, 2015.

[43] M. Song, Y. Yue, Y. Jin, J. Guo, L. Zuo, H. Peng, and Q. Chan, "Intravoxel incoherent motion and adc measurements for differentiating benign from malignant thyroid nodules: Utilizing the most repeatable region of interest delineation at 3.0t," *Cancer Imaging*, vol. 20, no. 1, 2020.

[44] T. Koopman, R. Martens, O. J. Gurney-Champion, M. Yaqub, C. Lavini, P. Graaf, J. Castelijns, R. Boellaard, and J. T. Marcus, "Repeatability of ivim biomarkers from diffusion-weighted mri in head and neck: Bayesian probability versus neural network," *Magnetic Resonance in Medicine*, vol. 85, no. 6, p. 3394–3402, 2021.

[45] A. S. Alyami, H. G. Williams, K. Argyriou, D. Gunn, V. Wilkinson-Smith, J. R. White, J. Alyami, P. A. Gowland, G. W. Moran, C. L. Hoad, and et al., "Test–retest assessment of non-contrast mri sequences to characterise and quantify the small bowel wall in healthy participants," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 34, no. 6, p. 791–804, 2021.

[46] M. Y. Bilgili, "Reproductibility of apparent diffusion coefficients measurements in diffusion-weighted mri of the abdomen with different b values," *European Journal of Radiology*, vol. 81, no. 9, p. 2066–2068, 2012.

[47] N. E. Larsen, S. Haack, L. P. Larsen, and E. M. Pedersen, "Quantitative liver adc measurements using diffusion-weighted mri at 3 tesla: Evaluation of reproducibility and perfusion dependence using different techniques for respiratory compensation," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 26, no. 5, p. 431–442, 2013.

[48] Y. Lee, S. S. Lee, N. Kim, E. Kim, Y. J. Kim, S.-C. Yun, B. Kühn, I. S. Kim, S. H. Park, S. Y. Kim, and et al., "Intravoxel incoherent motion diffusion-weighted mr imaging of the liver: Effect of triggering methods on regional variability and measurement repeatability of quantitative parameters," *Radiology*, vol. 274, p. 405–415, Feb 2015.

[49] J. Li, C. Zhang, Y. Cui, H. Liu, W. Chen, G. Wang, and D. Wang, "Intravoxel incoherent motion diffusion-weighted mr imaging of the liver using respiratory-cardiac double triggering," *Oncotarget*, vol. 8, p. 94959–94968, Oct 2017.

[50] Z. Xiang, Z. Ai, J. Liang, G. Li, X. Zhu, and X. Yan, "Evaluation of regional variability and measurement reproducibility of intravoxel

incoherent motion diffusion weighted imaging using a cardiac stationary phase based ecg trigger method," *BioMed Research International*, vol. 2018, p. 1–11, Apr 2018.

[51] C. Pieper, A. Sprinkart, G. Kukuk, and P. Mürtz, "Short-term measurement repeatability of a simplified intravoxel incoherent motion (ivim) analysis for routine clinical diffusion-weighted imaging in malignant liver lesions and liver parenchyma at 1.5t," *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 191, no. 03, p. 199–208, 2018.

[52] R. Pathak, J. Tian, N. A. Thacker, D. M. Morris, H. Ragheb, C. Saunders, M. Saunders, and A. Jackson, "Considering tumour volume for motion corrected dwi of colorectal liver metastases increases sensitivity of adc to detect treatment-induced changes," *Scientific Reports*, vol. 9, no. 1, 2019.

[53] Y. Pei, S. Xie, W. Li, X. Peng, Q. Qin, Q. Ye, M. Li, J. Hu, J. Hou, G. Li, and et al., "Evaluation of simultaneous-multislice diffusion-weighted imaging of liver at 3.0nbsp;t with different breathing schemes," *Abdominal Radiology*, vol. 45, no. 11, p. 3716–3729, 2020.

[54] S. Xie, I. B. Masokano, W. Liu, X. Long, G. Li, Y. Pei, and W. Li, "Comparing the clinical utility of single-shot echo-planar imaging and readout-segmented echo-planar imaging in diffusion-weighted imaging of the liver at 3 tesla," *European Journal of Radiology*, vol. 135, p. 109472, 2021.

[55] O. L. Sedlaczek, J. Kleesiek, F. A. Gallagher, J. Murray, S. Prinz, R. Perez-Lopez, E. Sala, C. Caramella, S. Diffetock, N. Lassau, and et al., "Quantification and reduction of cross-vendor variation in multicenter dwi mr imaging: Results of the cancer core europe imaging task force," *European Radiology*, 2022.

[56] A. Weller, M. V. Papoutsaki, J. C. Waterton, A. Chiti, S. Stroobants, J. Kuijer, M. Blackledge, V. Morgan, and N. M. deSouza, "Diffusion-weighted (dw) mri in lung cancers: Adc test-retest repeatability," *European Radiology*, vol. 27, no. 11, p. 4552–4562, 2017.

[57] J. Jiang, J. Yin, L. Cui, X. Gu, R. Cai, S. Gong, Y. Xu, H. Ma, and J. Mao, "Lung cancer: Short-term reproducibility of intravoxel incoherent motion parameters and apparent diffusion coefficient at 3t," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 4, p. 1003–1012, 2017.

[58] S. Swerkersson, O. Grundberg, K. Kölbeck, A. Carlberg, S. Nyrén, and M. Skorpil, "Optimizing diffusion-weighted magnetic resonance imaging for evaluation of lung tumors: A comparison of respiratory triggered and free breathing techniques," *European Journal of Radiology Open*, vol. 5, p. 189–193, 2018.

[59] Q. Wan, Q. Lei, P. Wang, J. Hu, T. Zhang, D. Yu, X. Li, and C. Liang, "Intravoxel incoherent motion diffusion-weighted imaging of lung cancer," *Journal of Computer Assisted Tomography*, vol. 44, no. 3, p. 334–340, 2020.

[60] C. Ma, X. Guo, L. Liu, Q. Zhan, J. Li, C. Zhu, L. Wang, J. Zhang, X. Fang, J. Qu, and et al., "Effect of region of interest size on adc measurements in pancreatic adenocarcinoma," *Cancer Imaging*, vol. 17, no. 1, 2017.

[61] R. Klaassen, O. J. Gurney-Champion, M. R. Engelbrecht, J. Stoker, J. W. Wilmink, M. G. Besselink, A. Bel, G. van Tienhoven, H. W. van Laarhoven, A. J. Nederveen, and et al., "Evaluation of six diffusion-weighted mri models for assessing effects of neoadjuvant chemoradiation in pancreatic cancer patients," *International Journal of Radiation Oncology*Biology*Physics*, vol. 102, no. 4, p. 1052–1062, 2018.

[62] L. Chen, F. Shen, Z. Li, H. Lu, Y. Chen, Z. Wang, and J. Lu, "Diffusion-weighted imaging of rectal cancer on repeatability and cancer characterization: An effect of b-value distribution study - cancer imaging," Nov 2018.

[63] D. Malyarenko, C. J. Galbán, F. J. Londy, C. R. Meyer, T. D. Johnson, A. Rehemtulla, B. D. Ross, and T. L. Chenevert, "Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom," *Journal of Magnetic Resonance Imaging*, vol. 37, no. 5, p. 1238–1246, 2012.

[64] G. Belli, S. Busoni, A. Ciccarone, A. Coniglio, M. Esposito, M. Giannelli, L. N. Mazzoni, L. Nocetti, R. Sghedoni, R. Tarducci, and et al., "Quality assurance multicenter comparison of different mr scanners for quantitative diffusion-weighted imaging," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, p. 213–219, 2015.

[65] J. Weiss, P. Martirosian, J. Taron, A. E. Othman, T. Kuestner, M. Erb, J. Bedke, F. Bamberg, K. Nikolaou, M. Notohamiprodjo, and et al., "Feasibility of accelerated simultaneous multislice diffusion-weighted mri of the prostate," *Journal of Magnetic Resonance Imaging*, vol. 46, no. 5, p. 1507–1515, 2017.

[66] M. Intven, O. Reerink, and M. E. Philippens, "Repeatability of diffusion-weighted imaging in rectal cancer," *Journal of Magnetic Resonance Imaging*, vol. 40, no. 1, p. 146–150, 2013.

[67] T. Tamada, C. Huang, J. M. Ream, M. Taffel, S. S. Taneja, and A. B. Rosenkrantz, "Apparent diffusion coefficient values of prostate cancer: Comparison of 2d and 3d rois," *American Journal of Roentgenology*, vol. 210, no. 1, p. 113–117, 2018.

[68] C. Tsuruta, K. Hirata, K. Kudo, N. Masumori, and M. Hatakenaka, "Dwi-related texture analysis for prostate cancer: Differences in correlation with histological aggressiveness and data repeatability between peripheral and transition zones," *European Radiology Experimental*, vol. 6, no. 1, 2022.

[69] M. A. Boss, B. S. Snyder, E. Kim, D. Flamini, S. Englander, K. M. Sundaram, N. Gumpeni, S. L. Palmer, H. Choi, A. T. Froemming, and et al., "Repeatability and reproducibility assessment of the apparent diffusion coefficient in the prostate: A trial of the ecog-acrin research group (acrin 6701)," *Journal of Magnetic Resonance Imaging*, vol. 56, no. 3, p. 668–679, 2022.

[70] Y. Sun, Q. Xiao, F. Hu, C. Fu, H. Jia, X. Yan, C. Xin, S. Cai, W. Peng, X. Wang, and et al., "Diffusion kurtosis imaging in the characterisation of rectal cancer: Utilizing the most repeatable region-of-interest strategy for diffusion parameters on a 3t scanner," *European Radiology*, vol. 28, no. 12, p. 5211–5220, 2018.

[71] A. L. Bisgaard, C. Brink, M. L. Fransen, T. Schytte, C. P. Behrens, I. Vogelius, H. D. Nissen, and F. Mahmood, "Robust extraction of biological information from diffusion-weighted magnetic resonance imaging during radiotherapy using semi-automatic delineation," *Physics and Imaging in Radiation Oncology*, vol. 21, p. 146–152, 2022.

[72] K. Onodera, M. Hatakenaka, N. Yama, M. Onodera, T. Saito, T. C. Kwee, and T. Takahara, "Repeatability analysis of adc histogram metrics of the uterus," *Acta Radiologica*, vol. 60, no. 4, p. 526–534, 2018.

[73] H. Zhang, Y. Zhou, J. Li, P. Zhang, Z. Li, and J. Guo, "The value of dwi in predicting the response to synchronous radiochemotherapy for advanced cervical carcinoma: Comparison among three mathematical models," *Cancer Imaging*, vol. 20, no. 1, 2020.

[74] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, p. 155–163, 2016.

[75] D. A. Porter and R. M. Heidemann, "High resolution diffusion-weighted imaging using readout-segmented echo-planar imaging, parallel imaging and a two-dimensional navigator-based reacquisition," *Magnetic Resonance in Medicine*, vol. 62, no. 2, p. 468–475, 2009.

[76] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic Resonance Imaging*. Wiley, 2014.

[77] E. M. Delfaut, J. Beltran, G. Johnson, J. Rousseau, X. Marchandise, and A. Cotten, "Fat suppression in mr imaging: Techniques and pitfalls," *RadioGraphics*, vol. 19, no. 2, p. 373–382, 1999.

[78] S. Klein*, M. Staring*, K. Murphy, M. A. Viergever, and J. P. Pluim, "elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 196 – 205, January 2010.

[79] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, M. Smits, S. Klein, and M. Staring, "Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer's disease," *Frontiers in Neuroinformatics*, vol. 7, pp. 1–15, January 2014.

[80] A. Carkeet, "A review of the use of confidence intervals for bland-altman limits of agreement in optometry and vision science," *Optometry and Vision Science*, vol. 97, p. 3–8, Jan 2020.

[81] I. J. Simpson, J. A. Schnabel, A. R. Groves, J. L. Andersson, and M. W. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, no. 3, p. 2438–2451, 2012.

[82] E. S. Kooreman, P. J. van Houdt, R. Keesman, F. J. Pos, V. W. van Pelt, M. E. Nowee, A. Wetscherek, R. H. Tijssen, M. E. Philippens, D. Thorwarth, and et al., "Adc measurements on the unity mr-linac – a recommendation on behalf of the elekta unity mr-linac consortium," *Radiotherapy and Oncology*, vol. 153, p. 106–113, Dec 2020.

[83] B. A. McDonald, T. Salzillo, S. Mulder, S. Ahmed, A. Dresner, K. Preston, R. He, J. Christodouleas, A. S. Mohamed, M. Philippens, and et al., "in vivo and phantom repeatability of diffusion-weighted mri sequences on 1.5t mri-linear accelerator (mr-linac) and mr simulator devices for head and neck cancers: Results from a prospective r-ideal stage 2a evaluation of tumor and nor-

mal tissue apparent diffusion coefficients as quantitative imaging biomarkers," May 2022.

[84] N. W. Schurink, D. M. Lambregts, and R. G. Beets-Tan, "Diffusion-weighted imaging in rectal cancer: Current applications and future perspectives," *The British Journal of Radiology*, vol. 92, no. 1096, p. 20180655, 2019.

[85] M. Ingle, M. Blackledge, I. White, A. Wetscherek, S. Lalondrelle, S. Hafeez, and S. Bhide, "Quantitative analysis of diffusion weighted imaging in rectal cancer during radiotherapy using a magnetic resonance imaging integrated linear accelerator," *Physics and Imaging in Radiation Oncology*, vol. 23, p. 32–37, Jul 2022.