# MR-only and PET-MR Radiotherapy for Pelvic Cancers

**Jonathan J Wyatt**

A thesis presented for the degree of
Doctor of Philosophy

Translational and Clinical Research Institute
Newcastle University
United Kingdom

April 2022

# Abstract

$42,500$ people are treated each year for cancers in the pelvis in the UK. Radiotherapy is a key treatment technique for pelvic cancers, with approximately 40% of patients receiving it as the primary or adjuvant treatment. Critical to targeting the radiation accurately is the quality of the imaging used to plan radiotherapy treatments. Conventional Computed Tomography (CT) imaging is geometrically robust and provides mass density information for accurate radiotherapy dose calculations. However in the pelvis the poor soft-tissue contrast of CT makes delineation of the tumour and nearby healthy Organs At Risk (OARs) difficult and CT is unable to provide functional or metabolic information about the tumour. In contrast, Magnetic Resonance (MR) has superb soft-tissue contrast, improving the accuracy of tumour and OAR delineation, and is also able to provide additional functional information to further characterise the tumour such as Diffusion Weighted (DW)-MR. When combined with the metabolic information available from Positron Emission Tomography (PET) in a simultaneous PET-MR scanner, this has great potential to enable the identification of tumour sub-volumes for receiving boost radiation doses and to characterise the tumour for more stratified radiotherapy dose prescriptions.

However there are significant scientific and technical barriers to using MR-only and PET-MR imaging for radiotherapy planning in the pelvis. The aim of this thesis was to develop technical solutions to enable MR-only and PET-MR for radiotherapy planning of pelvic cancers and to evaluate these solutions for clinical radiotherapy treatments.

The primary barrier to MR-only radiotherapy is that MR images cannot be used directly for radiotherapy dose calculations. This dissertation describes the development of a synthetic CT (sCT) Deep Learning model based on a novel zero echo time MR sequence, in collaboration with GE Healthcare, and its comprehensive evaluation for a range of pelvic radiotherapy treatments. Additionally, a separate Deep Learning algorithm that automatically contoured OARs, also developed by GE Healthcare, was evaluated for prostate, anal and rectal cancer sites. Finally, clinical implementation of MR-only radiotherapy also requires a method for ongoing Quality Assurance (QA) of the sCT dose calculation accuracy. A method using Cone Beam (CB)CT was developed and analysed on a cohort of clinical MR-only patients.

A major barrier for the use of PET-MR imaging for pelvic radiotherapy is the impact on both PET and MR image quality when acquiring images in the pelvic radiotherapy position. This image quality loss was quantified using phantoms and methods of incorporating the radiotherapy hardware into the PET Attenuation Correction (AC) map were developed. The impact of using these AC maps on tumour delineation and metabolic characterisation was then investigated in anal and rectal radiotherapy patients.

Acquiring an accurate PET image also requires an AC map of the patient. This is challenging for PET-MR because MR, unlike CT, cannot directly be used for PET AC. This could be overcome by using the MR-generated sCT evaluated for MR-only radiotherapy to generate a patient AC map. The accuracy of PET images reconstructed using sCTAC compared to gold standard CTAC and also the current MR-based MRAC was evaluated.

Another essential component to enabling the use of PET-MR imaging for radiotherapy planning is a QA programme focused on radiotherapy requirements. The tests needed for such a programme were developed and their same-day repeatability and monthly stability over 12 months evaluated.

Finally, fully utilising MR and PET-MR imaging for radiotherapy requires imaging times of 20 minutes or more so that high quality anatomical, functional and metabolic information can be acquired. However, for radiotherapy treatment planning it is critical that the positions of the internal anatomy need to be the same during imaging as they are for treatment. During image acquisitions of $\geq$ 20 minutes, organ motion from changes in bladder filling with be substantial. This organ motion was assessed in healthy volunteers and a method of correcting for the organ motion and developed and evaluated.

In summary, this thesis has developed and evaluated methods that enable MR-only and PET-MR imaging to be used for pelvic radiotherapy. These have included automated OAR delineation, bladder filling management, quantitative PET imaging in the radiotherapy position and accurate radiotherapy dose calculation, all underpinned by patient specific and system QA tests. Together, this thesis enables MR-only and PET-MR to be implemented for patients, paving the way for evaluation of their clinical benefits.

# Acknowledgements

I am very grateful to the large number of people at Newcastle University and Newcastle upon Tyne Hospitals NHS Foundation Trust who have enabled me to do this project. Especial thanks goes to the patients and volunteers who gave up their time to participate in the DeepMR-only RT study.

I particularly want to thank my supervisors, Dr Hazel McCallum and Dr Ross Maxwell. Their guidance, support and encouragement has been invaluable. In particular, I am incredibably appreciative for how they handled the major challenges of the Covid-19 pandemic and Brexit. Their determination to keep this project going despite those challenges and the many many hours and countless meetings they put in to ensure that, were essential to me completing this thesis. Throughout they managed to shield me from those challenges and gave me the freedom to focus on the scientific aspects of the project, for which I am very grateful. I am also thankful for their focus on my own personal scientific and professional development and their willingness to give me opportunities to present my work.

Additionally I would like to thank the team who helped deliver the clinical study. Especial thanks to Dr Rachel Pearson who was the Principal Investigator for the Deep MR-only RT study and who put a huge amount of effort into setting up and running the study, as well as recruiting patients and encouraging patient recruitment from other clinicians. I am grateful to Dr George Petrides and Ms Elizabeth Howell for their assistance and adivce in setting up and running the PET patient sub-study and for George's reporting and contouring PET images. Thanks also to radiographers Ms Paula Hindmarch, Mr Tim Hodgson and Ms Dorothy Wallace for scanning the patients and volunteers. Especial thanks to radiographer Mrs Rachel Pearson who not only provided training in setting up patients for radiotherapy, but also ensured patients could be scanned when I was absent, even at incredibly short notice. I am also grateful to nuclear medicine technologists Ms Sarah Atkinson, Ms Kendal Green and Ms Inez van der Westhuizen who injected the PET patients. I would also like to thank Mr Stuart Hamilton, Ms Jessica Williams, Ms Irene Jobson and all of the Clinical Trials team for managing the day-to-day running of the DeepMR-only RT study. And finally all the consultant clinical oncologists and specialist registrars at the Northern Centre for Cancer Care who took the time in busy clinic sessions to discuss the DeepMR-only RT study with their patients.

Thanks to Professor Andrew Blamire and Dr George Petrides for their annual project reviews and very useful advice and suggestions. I would also like to thank my fellow collaborators in the Deep MR-only RT consortium, especially Dr Florian, Sandeep, Cristina, Laszlo, Timio, Steven and Marta. I would also like to acknowledge assistance with specific aspects of different projects. The Northern Medical Physics and Clinical Engineering mechanical workshop designed and produced several different phantom holders and patient immobilisation lock-bars which were essential to accurate phantom and patient scanning. Dr Maelene Lohezic provided important technical advice about the PET-MR scanner. Ms Elizabeth Howell advised on PET SUV measurements and filled numerous PET phantoms for me. Mr Stephen Hedley modifed the open-source MR image quality software which I used for chapters 4 and 7.

Finally I would like to thank my family: My parents, who have always encouraged me to think for myself and to explore the world, my two children for the joy and distraction they bring, and most of all my wife whose unwavering love and support I could not have done without.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

AC Attenuation Correction

ACR American College of Radiologists

ADC Apparent Diffusion Coefficient

CBCT Cone Beam Computed Tomography

CNN Convolutional Neural Network

CT Computed Tomography

CTV Clinical Target Volume

DCE-MR Dynamic Contrast Enhanced-Magnetic Resonance

DSC Dice Similarity Coefficient

DTA Distance To Agreement

DVH Dose-Volume Histogram

DW-MR Diffusion Weighted-Magnetic Resonance

EPI Echo Planar Imaging

FDG FluoroDeoxyGlucose

FOV Field Of View

GAN Generative Adversarial Network

GI Gastro-Intestinal

GTV Gross Tumour Volume

GU Genito-Urinary

HU Hounsfield Units

IEC International Electrotechnical Commission

IGRT Image Guided RadioTherapy

IMRT Intensity Modulated RadioTherapy

kV kiloVoltage

MR Magnetic Resonance

MRAC Magnetic Resonance Attenuation Correction

MTV Metabolic Target Volume

MUSE MUltiplexed Sensitivity Encoding

MV MegaVoltage

NEMA National Electrical Manufacturers Association

OAR Organs At Risk

PET Positron Emission Tomography

PSMA Prostate Specific Membrane Antigen

PTV Planning Target Volume

QA Quality Assurance

RCR Royal College of Radiologists

RF Radio Frequency

ROI Region Of Interest

SABR Stereotactic Ablative Body Radiotherapy

sCT Synthetic Computed Tomography

SNR Signal-to-Noise-Ratio

SUV Standard Uptake Value

TLG Total Lesion Glycolysis

TSE Turbo Spin Echo

UTE Ultrashort Echo Time

VMAT Volumetric Modulated Arc Therapy

ZTE Zero Echo Time

# Chapter 1

# Introduction and Review of the Literature

## 1.1 Setting the Scene

Approximately 42,500 patients are treated annually in the UK for cancers in the pelvis, including gastrointestinal, gynaecological and urological cancers [1]. Radiotherapy is a fundamental treatment technique for these patients with only surgery being used more frequently [2]. Around 40% of patients with pelvic cancers are treated with radiotherapy in the UK [1]. Radiotherapy uses high energy radiation to destroy cancer cells but can also damage healthy tissue, resulting in treatment side-effects such as bladder and bowel incontinence, cystitis, rectal bleeding, fistulae, bowel obstructions and pelvic fractures. These can severely impact a patient's quality of life, both during treatment and for the rest of their lives [2]. 50% of patients receiving pelvic radiotherapy report permanent side-effects that adversely affect their quality of life [3].

Modern radiotherapy techniques such as Intensity Modulated RadioTherapy (IMRT) are able to create highly complex treatment plans which conform the high dose radiation to the cancer target and minimise the radiation to healthy tissues, reducing treatment side-effects whilst maintaining treatment effectiveness [4]. A meta-analysis of 23 studies in prostate cancer showed that IMRT significantly decreased grade 2-4 acute Gastro-Intestinal (GI) side-effects, chronic GI toxicity and chronic rectal bleeding compared to 3D conformal radiotherapy [5]. Similarly a meta-analysis for cervical cancer showed significant improvements for IMRT in acute GI and Genito-Urinary (GU) side-effects and chronic GU side-effects. For rectal cancer, retrospective reviews have shown significant reductions in grade 2+ GI toxicity and grade 2+ diarrhoea from 62% to 32% and 48% to 23% respectively when compared to 3D conformal radiotherapy, the previous clinical standard [6].

However, the much steeper dose gradients outside of the tumour produced by IMRT techniques make accurate targeting of the tumour critical [7]. The standard radiotherapy workflow involves a number of different steps (see figure 1.1), and the overall treatment accuracy is determined by the uncertainties within each step, with the step with the largest uncertainty dominating [8]. In the current standard workflow, the step with the largest uncertainty is manual delineation of the target volumes and Organs At Risk (OAR) by a clinician on the planning images, which has significant intra- and inter-observer variability [9]. Radiotherapy uncertainties are accounted for by the expansion of tumour volumes by a safety margin to produce a Planning Target Volume (PTV) [10]. This reduces the chance of the tumour being under-dosed but at the expense of the increased irradiation of healthy tissue surrounding the tumour. Reducing the delineation uncertainty would enable the PTV margin to be reduced, sparing more healthy tissue and so reducing treatment side-effects but without impacting on the treatment effectiveness [11]. The key to reducing this delineation uncertainty is to improve the quality of the planning images [7].

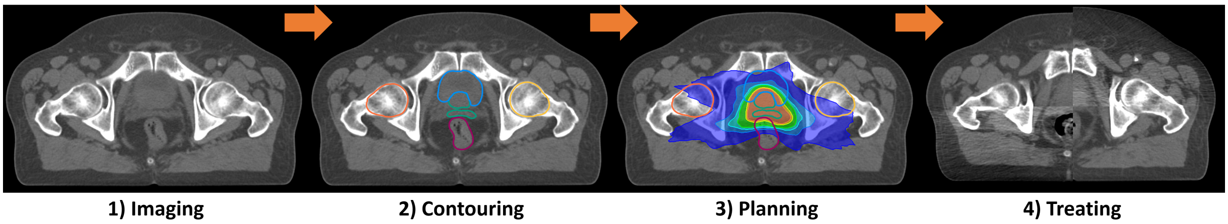


Figure 1.1: Radiotherapy workflow showing the main different steps for a prostate cancer treatment: 1) Imaging in the treatment position, 2) Manual contouring of target (green line) and OARs (bladder in blue, femoral heads in yellow and orange and rectum in purple), 3) IMRT treatment planning with the isodoses (regions receiving approximately the same radiation dose) shown as a colour wash from dark blue (33% of prescribed dose) to orange (95%) and 4) Treatment including onboard Cone Beam Computed Tomography (CBCT) verification imaging, CBCT registration to planning imaging, shifting the patient couch to the optimum position and delivering treatment beam(s).

Another aspect currently limiting radiotherapy treatments is targeting the entire tumour with a homogeneous dose [10], despite tumours being very heterogeneous [12]. For treatment to be successful this requires the entire PTV to receive a radiation dose sufficient to kill all clonogenic cells in the most active sub-volume, even though a lower radiation dose would have been sufficient for most of the volume [13]. Modern IMRT techniques enable a non-homogeneous dose distribution to be created which accurately targets the most active sub-volumes with 'boost' radiation doses [4]. This concept is called dose painting and can potentially improve tumour control probability without increase treatment side-effects [14]. However, dose painting depends critically on accurate delineation of the tumour sub-volume, which in turn depends on the images used to plan radiotherapy treatments [15].

The current standard imaging modality, Computed Tomography (CT), has poor soft-tissue contrast and is unable to accurately identify tumour sub-volumes for dose painting. The poor CT soft-tissue contrast is the primary cause of the significant inter-observer variability observed in tumour and OAR delineation [9]. This has motivated investigations of other imaging modalities to reduce delineation uncertainties and enable dose painting treatment strategies. Two of the most important are Magnetic Resonance (MR) and Positron Emission Tomography (PET) [16].

Anatomical MR imaging has demonstrated improvements in tumour and OAR delineation due to its superb soft-tissue contrast which makes organ boundaries much more visible (see figure 1.2 for an example of a prostate patient) [17]. This often also results in smaller overall target volumes [18] since in the presence of uncertainty oncologists tend to err on the side of larger volumes to minimise the risk of missing cancer cells. In addition, functional MR techniques such as Diffusion Weighted-Magnetic Resonance (DW-MR) can improve tumour delineation [19] and can also provide additional information about tumour heterogeneity, enabling the most active parts to be identified and treated using a higher 'boost' radiation dose [15].

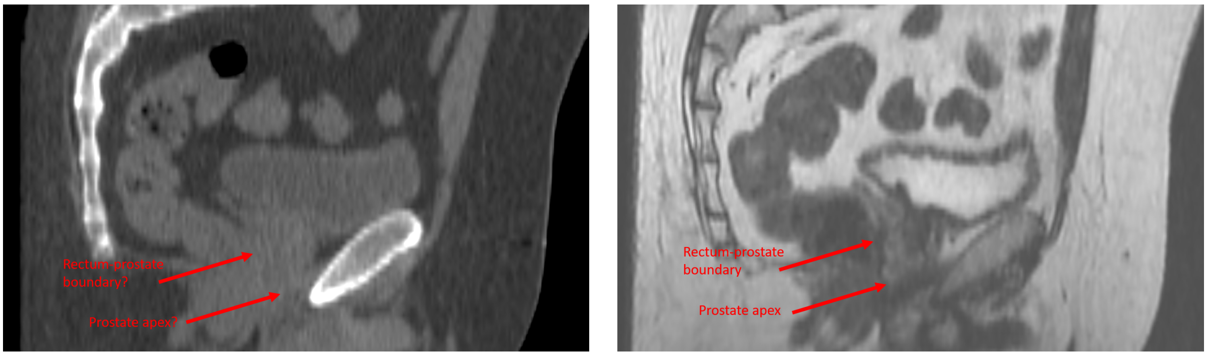


Figure 1.2: Example image of the improved soft-tissue contrast of MR (right) compared to CT (left) for a prostate radiotherapy patient. In particular the rectum-prostate boundary and prostate apex are visualised much more readily on MR than CT.

PET imaging is a crucial modality for diagnosis, staging and following up of most pelvic cancers [20]. PET can also improve tumour delineation for pelvic cancers [21, 22], and is able to identify metabolically active tumour sub-volumes [23]. This information is often complementary to that provided by functional MR techniques and the combination of PET and MR information outperforms either modality by itself [16, 23, 24]. Figure 1.3 shows a rectal cancer patient example using $^{18}$F-FluoroDeoxyGlucose (FDG) PET, anatomical and DW-MR images from a combined PET-MR scanner.

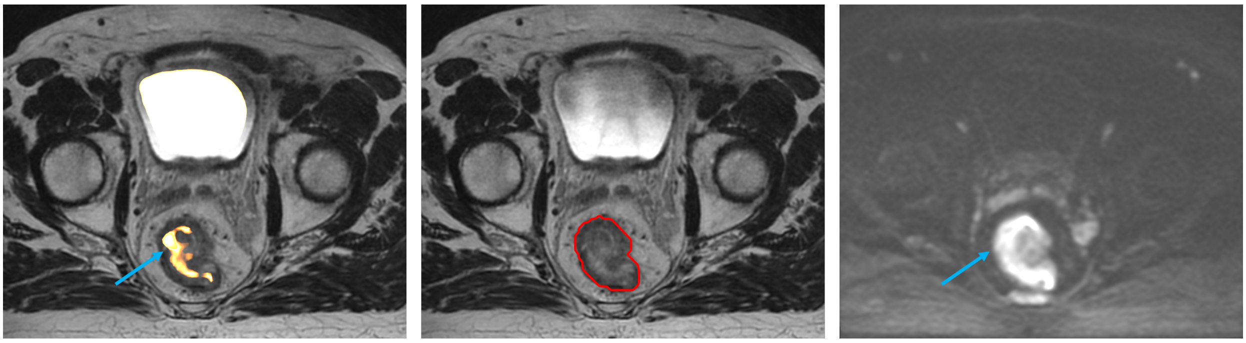


Figure 1.3: Example of $^{18}$F-FDG PET (left), T2-weighted anatomical MR (middle) and DW-MR(right) images. The primary tumour volume is shown on the anatomical MR (red contour), with the PET and DW-MR tumour sub-volumes show by the blue arrows.

PET and MR images can be incorporated into radiotherapy treatment planning by registering them to the planning CT. This enables the benefits of PET and MR for tumour and/or tumour sub-volume delineation to be combined with the high accuracy of radiotherapy dose calculations based on CT. However, accurate registration of images in the pelvis is difficult even when all images are acquired in the radiotherapy setup using flat couch tops, immobilisation devices and external lasers. This is due to the substantial variations in internal anatomy between imaging sessions that can occur from differences in bladder and rectal filling, as well as variations in patient posture and setup. The MR-CT registration brings a systematic uncertainty which is estimated to be 2 mm for prostate radiotherapy [25], and is likely to be larger for other pelvic cancers. In addition, the registration to CT often results in the PET or MR based contour being adjusted by the clinician to match the CT anatomy, since that is the image used to plan the radiotherapy treatment on [26]. This reduces the benefit from using MR and/or PET for delineation. Finally, the difficulties of registration mean that combining information from functional MR and PET imaging for tumour sub-volume delineation becomes difficult [27].

This has motivated MR-only radiotherapy where a CT is not acquired, removing the MR-CT registration uncertainty, and the MR alone is used for contouring and planning [28]. This has also motivated simultaneous PET-MR where PET and MR images are acquired with high spatial alignment [29]. These two approaches can also be combined to create a PET-MR-only radiotherapy pathway, where a single PET-MR imaging session provides all the information required for planning pelvic radiotherapy treatments [30].

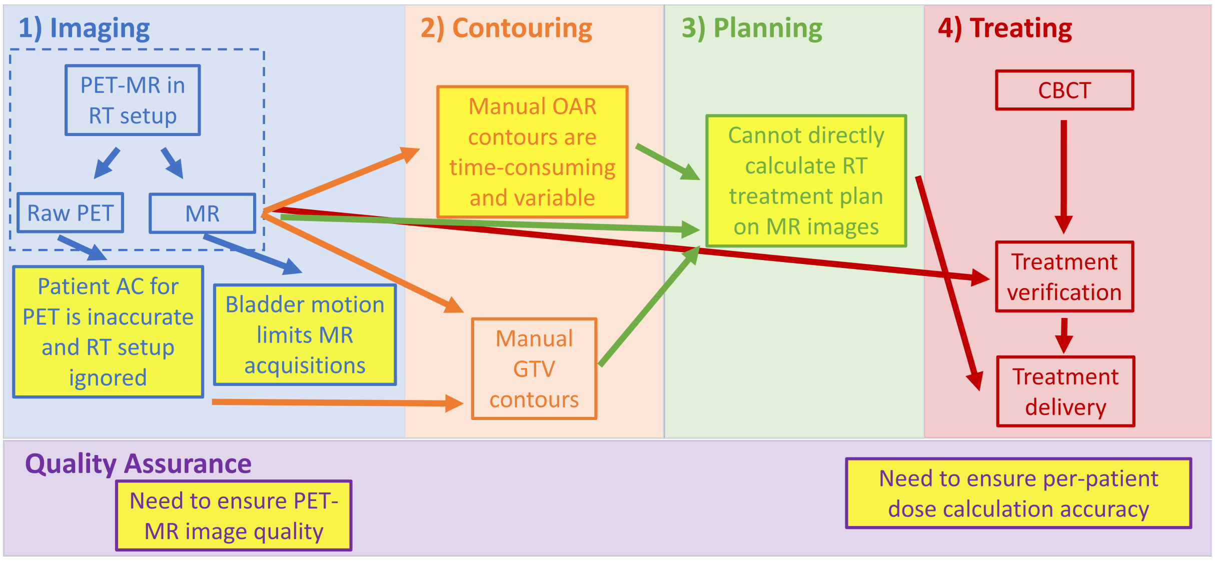

1) Imaging
2) Contouring
3) Planning
4) Treating
PET-MR in RT setup
Raw PET
MR
Patient AC for PET is inaccurate and RT setup ignored
Bladder motion limits MR acquisitions
Manual OAR contours are time-consuming and variable
Manual GTV contours
Cannot directly calculate RT treatment plan on MR images
CBCT
Treatment verification
Treatment delivery
Quality Assurance
Need to ensure PET-MR image quality
Need to ensure per-patient dose calculation accuracy

Figure 1.4: Diagram of the potential PET-MR-only radiotherapy pathway showing the current barriers to PET-MR-only radiotherapy implementation (boxes highlighted in yellow). These are 1) the current MR-based patient AC map for PET is inaccurate and the impact of the radiotherapy hardware is ignored, and the full potential of MR for radiotherapy treatment planning is limited by organ motion due to bladder filling, which results in maximum a MR acquisition time (< 20 minutes). 2) Manual OAR contouring is time-consuming and has significant inter-observer variability. 3) MR images cannot be used directly for radiotherapy dose calculations. Quality assurance methods to ensure PET-MR image quality for radiotherapy purposes and per-patient dose calculation accuracy of MR-only radiotherapy treatments need to be developed.

However there are significant scientific and technical barriers to using MR and PET-MR images to plan radiotherapy treatments. The primary issue is that, unlike CT images, MR images cannot by themselves be used for radiotherapy dose calculations [28]. This requires an algorithm to generate a Synthetic Computed Tomography (sCT) image from the MR, which can then be used for radiotherapy dose calculations. Doses calculated on sCT need to be very similar to dose calculated on gold standard CT. Clinical use of sCT algorithms will also require a method of on-going dose calculation accuracy Quality Assurance (QA). In addition, planning PET-MR images will need to be acquired in the radiotherapy treatment position utilising a flat couch and anterior coil bridge. This may impact on PET-MR image quality, which may in turn impact on the accuracy of target delineation. Therefore the effect on images needs to be quantified, and methods of compensating for it developed. Another issue for PET-MR is developing methods of patient attenuation correction. Similarly to radiotherapy dose calculations, MR images cannot by themselves be used for PET attenuation correction and require the generation of a CT-like image. The current commercially available solution for both manufacturers of PET-MR scanners, Magnetic Resonance Attenuation Correction (MRAC), does not reproduce bone in the image and so introduces PET quantification errors [31]. An attractive solution is to utilise the same sCT produced for MR-only radiotherapy, but this needs to evaluated. Also, the use of PET-MR for radiotherapy planning, which includes accurate functional information (PET and DW-MR), needs a QA programme tailored to radiotherapy requirements. These are additional to the QA tests needed for diagnostic purposes, such as high geometric accuracy over the full Field Of View (FOV) and high mechanical accuracy of couch and laser movements. Finally, fully utilising PET-MR imaging will require imaging times of $\geq 20$ minutes. However this is a significant issue for pelvic radiotherapy planning where the positions of the internal anatomy need to be the same during imaging as for treatment. During a $\geq 20$ minute imaging session organ motion such as bladder filling will change substantially, whereas on the treatment machine the time from starting CBCT verification imaging to completing the delivery of the radiation beam(s) is only a few minutes. Therefore fully utilising PET-MR imaging requires a method of tracking and compensating for such organ motion so that the planning images are representative of the patient anatomy on treatment. These barriers are illustrated in figure 1.4, which shows the potential PET-MR-only radiotherapy workflow and the current problems to be overcome.

The rest of this chapter reviews the literature in MR-Only radiotherapy, PET-MR for radiotherapy planning, PET-MR QA for radiotherapy and methods of tracking organ motion in the pelvis. The final section summarises the aims of this thesis and outlines the remaining chapters.

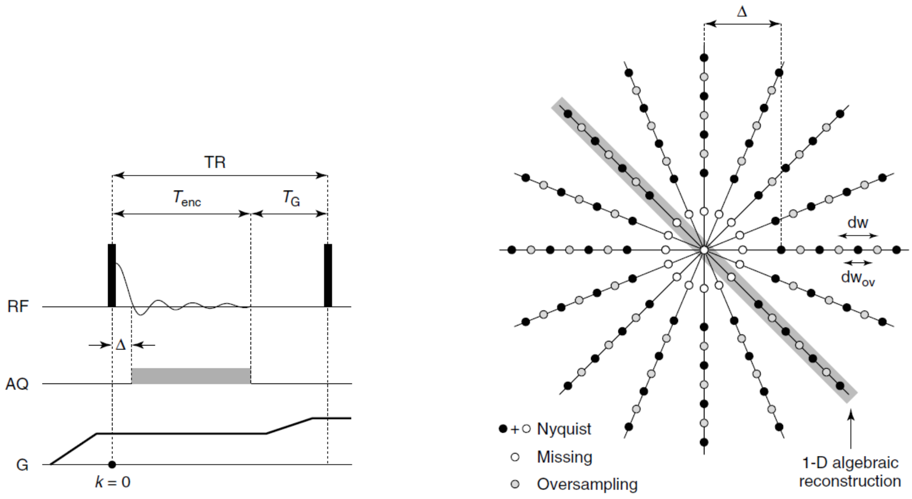
## 1.2 MR-Only Radiotherapy

MR-Only radiotherapy enables the superior soft-tissue contrast of MR [17] to be used within radiotherapy planning without the uncertainties caused by MR-CT registrations [25]. The primary challenge of MR-only radiotherapy is a method to generate a sCT from the MR that can be used for radiotherapy dose calculations [32]. A number of different methodologies have been proposed, using different MR sequences and computational techniques. The sCT methodology that is evaluated in chapter 2 of this thesis uses Zero Echo Time (ZTE) MR and Deep Learning algorithms [33]. Therefore brief background information to these topics, as well descriptions of the common dose evaluation methods used in the MR-only literature, are given first before turning to a review of the literature on sCT methodologies. Finally, another potential improvement in the consistency of contouring could be through the use of automatic contouring algorithms, which would also save time. Although CT-based automatic contouring solutions are commercially available, they have so far failed to be put into widespread clinical use due to limited accuracy. The superior soft-tissue contrast of MR potentially facilitates improved accuracy in automatic contouring as well as manual, and so MR-only radiotherapy also provides an opportunity for these automatic contouring methods to be safely implemented. A review of MR-based automatic contouring algorithms concludes this section.

### 1.2.1 Zero Echo Time MR

The primary attraction of ZTE MR is its ability to image tissues (principally bone) with extremely short transverse relaxation times which conventional MR sequences are unable to do [33]. The ability to image bone is very attractive from a MR-only radiotherapy perspective as it is distinguishing bone from air that is the primary challenge for sCT algorithms [32]. ZTE methods have been evaluated for PET-MR attenuation correction [33, 34] and for developing sCT for MR-only radiotherapy (see following section).

ZTE MR uses frequency encoding in a 3D radial centre-out k-space scheme with the encoding gradients applied prior to the initial Radio Frequency (RF) pulse [35]. This is effectively a free induction decay with a zero echo time [36], see figure 1.5. ZTE imaging requires the RF pulse to be uniform over the gradient bandwidth to avoid altering spin excitations, which is typically achieved through short, hard RF pulses leading to flip angles of $1 - 4^o$ [37]. Thus ZTE images are inherently proton-density weighted. By sampling k-space with sequential radial spokes the gradient direction is only changed incrementally, which leads to silent scanning and reduces eddy current effects [37]. A significant challenge for ZTE imaging is the dead-time caused by the scanner switching from transmit to receive mode, causing data at the centre of k-space to not be sampled [36]. Dead-times are typically in the range $6 - 9$ $\mu$s [35]. This can be solved through oversampling and algebraic reconstruction [38]. This uses data from two opposite sign radial gradients to reconstruct the 1D signal along that axis, which can then be inverse Fourier Transformed to give a complete k-space projection (see figure 1.5) [35].



(a) ZTE pulse sequence

(b) ZTE k-space

Figure 1.5: a) pulse sequence diagram for ZTE imaging. $\Delta$ indicates the dead-time as the scanner switches from transmit to receive, $T_{enc}$ the time the data for a particular gradient is acquired and $T_G$ the time taken for the gradient to adjust to the next projection. b) one plane in k-space showing the required points (white and black points) to adequately sample the data. Grey dots indicate the oversampled points and white dots the points that were missed due to the dead-time $\Delta$. A single 1D algebraic reconstruction is highlights by the grey bar. Diagrams taken from ref [35].

### 1.2.2 Deep Learning

Deep Learning is a very rapidly growing field for automated analysis of images [39], with a exponential increase in the number of papers being published on deep learning and deep learning methods becoming the dominant approach in medical image analysis [40]. There has also been a corresponding increase in interest in Deep Learning within radiotherapy, with over 500 papers published in 2017 [41]. Deep Learning refers to a computational model which has multiple non-linear data processing layers within it, with each layer using the output of the previous layer as input [39]. The aim is for each layer to extract different image features and therefore to develop a hierarchal representation of the data, combining sensitivity to small details with insensitivity to standard variations [41]. Critical to all Deep Learning approaches is the use of training data in sufficient quantity that each layer within the model can be iteratively optimised simultaneously using backpropagation algorithms [42]. Each layer has a number of parameters that define its output called weights, and the optimisation happens by adjusting these weights according to the error gradient [41]. This determines for each weight the amount the error defined by a prescribed loss function would increase or decrease for a given change in weight. The weight is then adjusted in the negative gradient direction. Typically a method called stochastic gradient descent is used where the gradient is calculated for a few samples, the weights updated and the process repeated for many iterations until the average of the loss function stops reducing [43]. The key to using the multiple layers of Deep Learning models is the fact that the error gradient can be backpropagated through each layer using the differentiation chain rule [44]. This means that the error gradient for one layer can be computed from the error gradient from the subsequent layer, starting with the final or output layer and backpropagated all the way through to the initial or input layer [43].

There are a number of different Deep Learning model architectures, but the dominant architecture within medical image analysis is a Convolutional Neural Network (CNN). [40]. This seeks to incorporate the spatial relationship contained in images through the use of kernels [41]. These are conventional image processing filters that extract particular image features from the input layer. A CNN contains an input layer and an output layer, with a number of convolution layers, activation layers and pooling layers in-between [42]. An example of the different layers in a CNN are shown in figure 1.6.

The convolutional layers consist of one or more kernels that are convolved with the input image to produce a feature map. It is this use of a convolution that ensures the spatial information of the image is preserved [41]. The kernel values are modified iteratively through the training phase, enabling the model to learn the optimum features to extract in each layer. The activation layer adds non-linearity to speed up the training, with the most commonly used activation layer being a rectified linear unit, which sets all negative inputs to zero and keeps all positive values the same [40]. Pooling layers downsample the feature map to reduce the computational load whilst preserving spatial extent [41]. A typical CNN model will have multiple sets of convolution layer, activation layer, pooling layer, with the output of one pooling layer being fed into the input of the next convolutional layer. The final output layer generates a probability score for classifying the image or each pixel within the image into the different available classes [40].

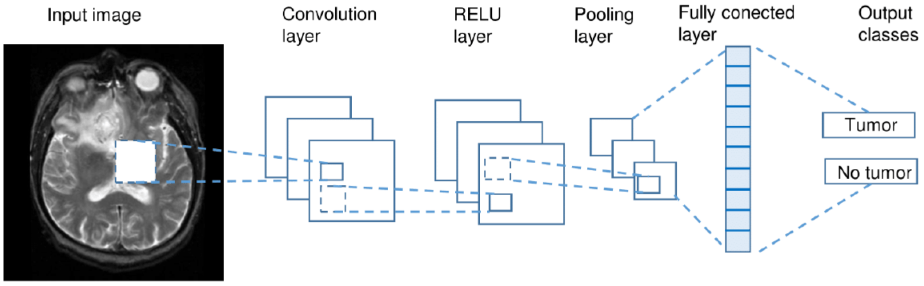


Figure 1.6: Example of a CNN to classify a MR image for the presence of tumour. The diagram shows a convolution layer, activation layer (labelled RELU layer for rectified linear unit), a pooling layer and the final output layer. Diagram taken from ref [40].

A variation on the CNN architecture that is commonly applied to medical images is called the U-NET [45]. This seeks to overcome the trade-off between sensitivity to coarse and fine resolutions [42] by adding to a conventional CNN architecture a series of up-sampling layers which mirror the down-sampling provided by the pooling layers combined with deconvolution layers which mirror the convolution layers [46]. In addition, at each resolution the layers on the down-sampling and up-sampling parts of the algorithm are connected. This crucially provides high resolution features to the deconvolution, enabling the algorithm to be sensitive to both coarse and fine resolutions [42]. A diagram of the original U-NET architecture is shown in figure 1.7.

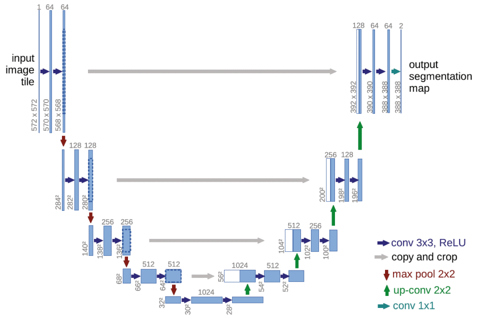


Figure 1.7: Diagram of the U-NET version of the CNN architecture. Diagram taken from ref [46].

Most medical images are inherently 3D, however utilising 3D CNNs are very computationally intensive [39]. Initially CNNs for medical image analysis were therefore 2D, treating each image slice completely independently. This is much less computationally intensive and also significantly reduces the amount of patient data required for training, since effectively the training data is the number of patient images multiplied by the number of slices in each image. However, it does mean that through-slice spatial information is lost [42]. One method for incorporating this information with a small increase in computational cost is a 2.5D CNN [45]. This trains the model on three 2D datasets which are on orthogonal planes (typically axial, sagittal and coronal). The output of each orientation model is then combined in the final layer, incorporating spatial information in all three planes [47]. More recently, a variant of the U-NET that is intrinsically 3D has been developed, called V-NET [48]. This uses volumetric kernels in the convolution, preserving the full 3D information at each layer of the CNN. The V-NET also incorporates a residual function at each convolutional stage, ensuring convergence which conventional U-NET architectures do not.

A critical component of a CNN is the loss function which is used to evaluate the model's performance during training. It is this loss that the model is looking to minimise, and so ensuring the metric used is relevant is very important for producing accurate models [44]. The most common loss function used in medical images is the cross-entropy loss which evaluates the model prediction for each voxel and averages the result over all voxels [39]. A variation of the cross-entropy loss is the weighted cross-entropy loss, which weights the foreground components of the image more highly in the loss function [49]. Foreground components in contouring algorithms refers to the organ being contoured and immediate surrounding area, with the background consisting of the rest of the image. An alternative loss function is the Dice loss, which is based on the widely used metric for assessing the overlap of one contour with another in medical images, the Dice Similarity Coefficient (DSC) [50]. The DSC is defined as twice the intersection volume of two contours divided by the sum of the two volume. This can also be extended to the generalised Dice loss, which is the DSC generalised to more than two contours [51].

There are a number of aspects to consider when developing and utilising CNNs. One common problem with medical image Deep Learning algorithms is over-fitting, where a model performs well on the training dataset but generalises poorly [39]. This is usually caused by too small amounts of training data, which is a widespread issue since high quality labelled medical data is scarce [44]. One method that is used to alleviate this problem is data augmentation which applies augmentations such as rotating, scaling, translating, flipping or distorting to some of the training data [39]. Another solution is to use transfer learning where models initially trained on different but related data are applied to the new training dataset rather than starting from a random initialisation of weights [40]. Finally training data can be increased through splitting up individual patients into separate patches, which are treated as independent [42]. This is in effect what 2D and 2.5D CNNs do with 3D medical images data. Nonetheless, over-fitting remains a problem in the application of Deep Learning to medical images [44].

Another common problem with Deep Learning algorithms for medical image analysis is class imbalance, where the foreground regions of interest are much smaller than the background, meaning the model is biased to the background [39]. Solutions to this problem include weighting the loss function to focus on the foreground regions [43] and splitting data up into separate patches and ensuring most of the patches used in training include foreground regions [42].

A third issue for Deep Learning models is robustness; how well the model copes with small deviations in the input data from the training data [44]. Finlayson et al. showed how the addition of small levels of noise to images changed the behaviour of skin cancer classification algorithm from malignant to benign or vice versa, with the model indicating high levels of confidence in the classification each time [52]. This demonstrates that Deep Learning models need to be evaluated on the full range of possible input images to be confident that it is robust. In addition, QA of the outputs of Deep Learning algorithms are also important to ensure robustness of output.

In conclusion, Deep Learning algorithms for medical image analysis are rapidly proliferating due to their significant improvement on previous methods. These Deep Learning algorithms are overwhelmingly designed using a CNN with a U-NET architecture. Important considerations for developing Deep Learning algorithms include the loss function, the amount and quality of training data and the size of the regions of interest relative to the background images.

### 1.2.3 Methods for Evaluating Synthetic CT Dose Accuracy

The standard approach to evaluating the dose accuracy of sCT images is to calculate the same treatment plan on sCT and CT and look at differences in the resulting dose distributions. Three different methods of evaluating those dose differences are commonly used in the MR-only literature [53, 54]: Dose-Volume Histogram (DVH) dose differences, radiological isocentre depth differences and gamma analyses. A DVH summarises a 3D dose distribution within a given organ by binning the dose in the organ for each dose bin plotting the volume of the organ receiving at least that dose [10]. A PTV DVH will typically be almost square since the treatment plan is optimised to produce a highly homogeneous dose (within −5% and +7% of the prescription dose) to the PTV. OAR DVHs will typically monotonically decrease from the top left (100% of the organ will receive ≥ 0 Gy) to the bottom right (0% of the organ will receive a dose greater than the maximum dose to the patient). Points on the DVH can be identified as either the dose received by a specified volume (eg D98% is the minimum dose received by the hottest 98% of the organ) or volume receiving a specified dose (eg V30 Gy is the volume receiving at least 30 Gy). An example DVH for a typical prostate radiotherapy treatment is shown in figure 1.8. PTVs are typically compared on near-minimum (D98%), median (D50%) and near-maximum (D2%) DVH points. OARs are typically compared on clinically relevant DVH constraints.

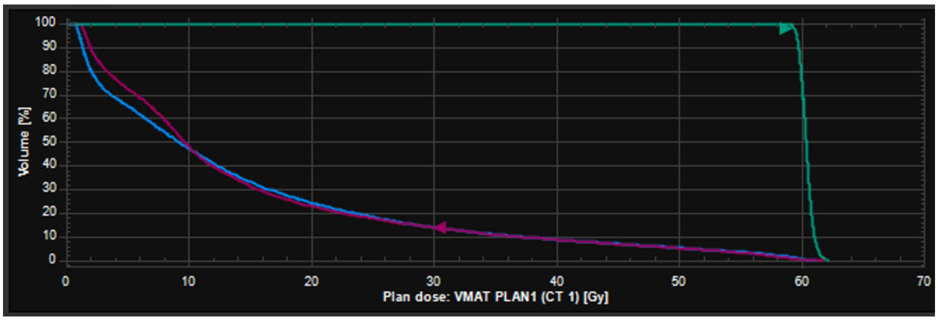


Figure 1.8: Example Dose-Volume Histogram for a prostate radiotherapy patient, showing PTV in green, bladder in blue and rectum in purple. The prostate D98% point and rectum V30 Gy are shown as arrows (green and purple respectively).

The second method commonly used is the radiological isocentre depth difference [55, 56]. The radiological depth is the depth of water that attenuates the beam by the same amount as the patient tissue in the beam path [57]. So for pure water the radiological depth is identical to the physical depth, whereas if the beam passes through more attenuating material such as bone the radiological depth will be larger than the physical depth and if it passes through less attenuating material (eg lung) it will be smaller. The radiological depth is calculated by multiplying the beam path length through each voxel by the electron density relative to water of that voxel, and then summing these scaled voxel path lengths for all the voxels along the beam axis. The relative electron density of each voxel is calculated from the voxel HU value multiplied by a CT scanner specific calibration curve that is measured using CT scans of a relative electron density phantom.

The final method used for comparing sCT and CT dose differences is a gamma analysis. This evaluates a comparison dose distribution relative to a reference dose distribution [58]. For each dose grid point in the comparison dose distribution the gamma index for all the nearest dose grid points (within a pre-defined distance) of the reference dose distribution is calculated using

$$\Gamma = \sqrt{\left(\frac{\Delta_D}{DD}\right)^2 + \left(\frac{\Delta_d}{dta}\right)^2}. \tag{1.1}$$

Here $\Delta_D$ is the difference in dose between comparison and reference point, and $\Delta_d$ the difference in distance. These differences are relative to a dose difference criterion ($DD$) and a distance to agreement criterion ($dta$), which are specified for a particular analysis (eg gamma criteria 2% of prescription dose and 2 mm distance to agreement). For each point in the comparison distribution the minimum gamma index calculated compared to all the points in the reference dose distribution is assigned as its gamma value. Gamma values $< 1.0$ are considered passes. Typically the percentage of points within a specified contour (eg PTV or external contour) passing the gamma analysis is reported. The gamma analysis provides a method of comparing two non-homogeneous dose distributions that is not too sensitive to either small dose differences in dose or small distance offsets. Typically gamma pass rates $\geq 95\%$ would be considered acceptable [58]. Figure 1.9 shows two example gamma maps, one with a high pass rate and one with a lower pass rate.

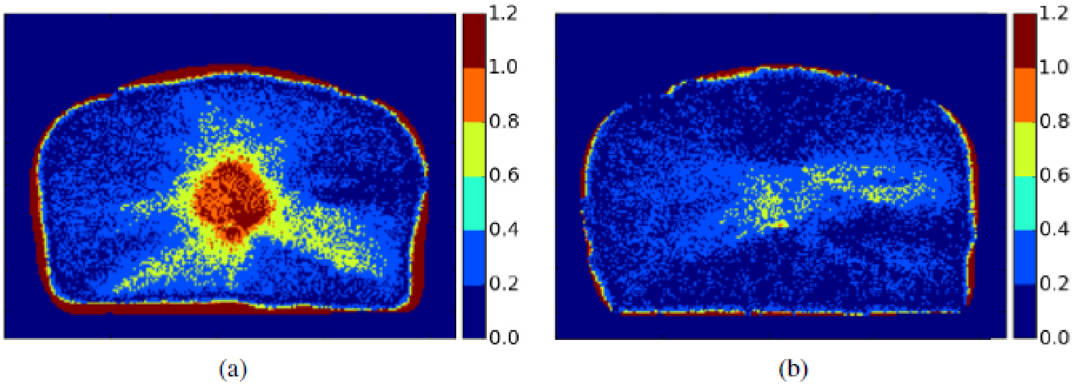


Figure 1.9: Example of gamma maps showing a low pass percentage (a) and a high pass percentage (b). Both maps calculated with gamma criteria 1%/1 mm. Images taken from ref. [59] (images used with permission).

### 1.2.4 Synthetic CT Algorithms

The crucial challenge for MR-only radiotherapy is to generate an image from the MR that can be used for radiotherapy dose calculations. A simple calibration of MR intensity values cannot be used since tissues with very different x-ray attenuation properties, such as bone and air, can have very similar intensity values in conventional MR images [32]. Therefore MR-only research has focused on developing algorithms that can produce a CT-like image, usually called a sCT, which can then be used in the same way as standard CT images for radiotherapy planning and dose calculations. sCT algorithms are often divided into three categories: bulk density override, atlas-based and voxel-based [28], with a fourth rapidly growing category being Deep Learning algorithms.

Bulk density override methods use automatic or manual segmentation techniques to identify difference tissue classes, which are then assigned population-based densities to generate a sCT [60]. The major challenge for bulk density methods is automatically segmenting the required tissue classes since manual segmentation is too time-consuming for clinical use [28]. Two tissue classes (soft-tissue and bone) are the minimum required [28], with additional tissue classes improving the dosimetric accuracy of the sCT, including distinguishing between cortical and spongy bone [61]. There are two commercial sCT algorithms for the pelvis currently available that use bulk-density methods: Magnetic Resonance for Calculating ATtenuation (MRCAT, Philips Healthcare) [62] and RT Dot Engine for pelvis (Siemens Healthineers) [63]. The two methods are very similar and are based on a DIXON sequence [64], which can separate water and fat soft-tissue classes using the chemical shift [65]. The RT Dot Engine uses four different tissue classes: air, soft-tissue, fat and bone [66]. The air is segmented as everything outside the MR patient external contour. The bone is segmented using an atlas-based technique where manually segmented MR images are deformably registered to the new MR and the bone contours copied across. The remaining tissue is segmented as soft-tissue and fat using thresholding based on the water and fat DIXON images [63]. The MRCAT algorithm uses a very similar methodology but with the bone divided into two tissue classes, spongy bone and cortical [62]. These are segmented through thresholding of the signal intensity from the in-phase DIXON image within the bone contour [67]. A schematic of the method is shown in figure 1.10. The MRCAT algorithm has been demonstrated to be highly dosimetrically accurate in clinical use, with mean dose differences $\leq 0.5\%$ between sCT and CT for 25 prostate patients [67] and $\leq 0.3\%$ for 20 rectum patients. The RT Dot Engine has been validated dosimetrically for 13 prostate patients, with mean gamma analysis agreement between sCT and CT with criteria 3% dose difference and 3 mm distance to agreement being $\Gamma_{3/3} = 98.7\%$ [66].

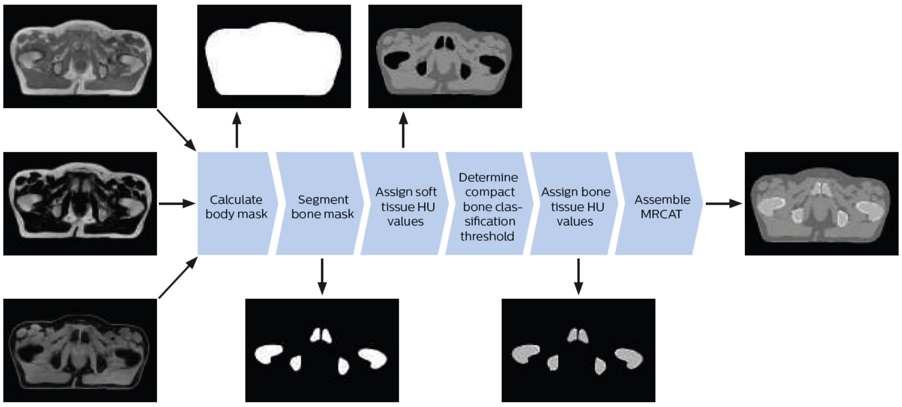

Calculate body mask
Segment bone mask
Assign soft tissue HU values
Determine compact bone classification threshold
Assign bone tissue HU values
Assemble MRCAT

Figure 1.10: Schematic of the MRCAT sCT algorithm showing the input images from the DIXON sequence, the air classification, bone model, soft-tissue and fat classification and the spongy and cortical bone classification. The RT Dot Engine for pelvis is very similar but with a single bone classification. Diagram taken from ref [62].

Atlas-based methods use an atlas of CT and MR pairs, which have been either deformably registered together [68] or registered using structure-guided deformable registration where the bones are held rigid but the soft-tissue is allowed to deform [55, 69]. Each atlas MR is deformably registered to the incoming MR, these deformations applied to the corresponding atlas CTs and the deformed CTs then combined to make a sCT using a local patch weighting process [28]. This process weights contributions from the atlas CTs to each voxel depending on the similarity of the voxels in the atlas MRs in the same place. A diagram of the atlas-based approach is shown in figure 1.11. Atlas-based methods have been shown to generalise to MR images acquired on different scanners and field strengths and with different acquisition parameters [53, 59]. There is an atlas-based commercially available sCT algorithm for the prostate, MriPlanner (Spectronic Medical AB), which has demonstrated high dosimetric accuracy with mean dose differences $\leq 0.3\%$ across 170 patients [53].

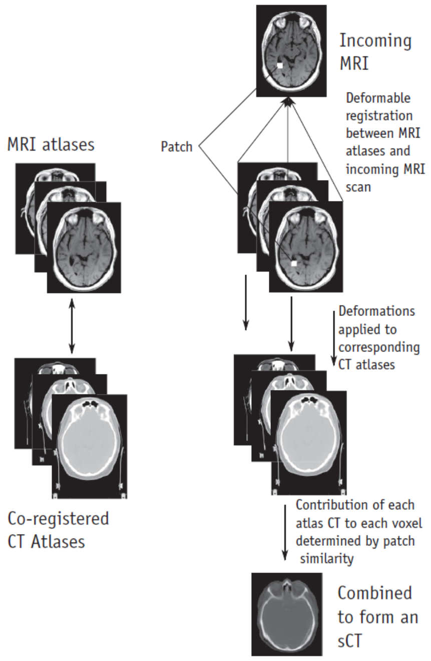

Incoming MRI
Deformable registration between MRI atlases and incoming MRI scan
MRI atlases
Patch
Deformations applied to corresponding CT atlases
Co-registered CT Atlases
Contribution of each atlas CT to each voxel determined by patch similarity
Combined to form an sCT

Figure 1.11: Diagram of an atlas-based sCT algorithm. Diagram taken from ref [28].

Voxel-based methods use multiple MR sequences and combine the image intensity values to generate a sCT [28]. There are a number of different approaches that have been investigated to determine the regression relationship between the multiple MR intensity values and the Hounsfield Units (HU) value in the sCT, including Gaussian approaches, discriminant analysis, principal component analysis and random forest [32]. All these approaches use an initial training set of MR-CT pairs that have been co-registered to determine the parameters of the regression relationship. There have been approaches reported in the literature using conventional T1- and T2-weighted MR sequences, but these have required additional manual bone segmentation for accurate results [70]. One approach that has been used to treat over 200 prostate patients uses a T1/T2*-weighted MR sequence with an atlas-based bone segmentation [71]. Two different regression models are applied to the normalised MR intensity values, one within the bone contour and one outside [72]. Within the bone an inverse relationship between MR signal intensity is used with a second order polynomial fit, with a patient-specific normalisation applied using a manually contoured region in the patient muscle [73]. The mean dose differences between sCT and CT were $\leq 1\%$ [72]. Most of the research literature has focused on incorporating more novel MR sequences that use ultrashort or zero echo times to generate signal from bone tissues [28]. A Gaussian regression approach with a T2-weighted Turbo Spin Echo (TSE) sequence and two Ultrashort Echo Time (UTE) sequences was found to accurately predict HU values with a mean average error in HU being 137 HU [74]. A simple linear relationship between normalised ZTE intensity values and CT intensity values in bone was found within the brain, enabling the creation of a sCT with mean dose differences across five patients of 0.2% [33]. A particular issue for voxel-based methods is correct classification at the air/soft-tissue and bone/soft-tissue boundaries due to partial volume effects [74]. So far these ultrashort or zero echo time approaches have focused on the brain anatomic site without any exploration of the pelvis [32].

Deep learning algorithms for sCT generation have only started appearing in the literature in the last few years, but are rapidly increasing [41]. A 2D CNN model with multiple embedding blocks included with the convolutional layers was developed by Xiang et al [75]. An embedding block reconstructs a tentative sCT from the feature map produced by the preceding convolutional layer, and then combines that result with the feature map extracted by the current convolutional layer. The final layer reconstructs the final sCT. This approach aims to speed up the MR-to-CT mapping being achieved through the CNN. The model was trained with 22 prostate patients with deformably registered MR and CT images. To increase the size of the training dataset the data was augmented through left-right flipping and the individual slices sub-divided into small patches. The model was evaluated using a leave-one-out approach with the mean absolute error being 42.5 HU. Fu et al. developed 2D and 3D CNN models with a U-NET architecture for prostate patients [76]. The models were trained with 20 T1-weighted MR and CT pairs deformably registered together. The loss function used was the mean absolute error between sCT and CT. The 3D model was identical to the 2D model except each convolutional layer was 3D. The mean absolute error for the 2D model was 40.5; HU with the 3D model slightly improving to 37.6; HU. Leynes et al. reported a similar deep 3D CNN model with U-NET architecture [77]. The model was trained using fat and water MR images from a DIXON sequence, ZTE images and paired CT images of 10 pelvic cancer patients. The CT and MR images were deformably registered. The model was trained with small volumetric patches containing each MR contrast and the corresponding deformed CT patch. A loss function combining mean absolute error, gradient difference loss and Laplacian difference loss was used, with the latter two functions improving image sharpness. The model was evaluated on a further 16 pelvic cancer patients, with a mean error in HU of −36; HU.

Maspero et al. used a different model architecture consisting of a Generative Adversarial

Network (GAN) [78]. This consists of two simultaneously-trained models, one generating sCTs and other seeking to discriminate between sCTs and the ground truth CTs [40]. The generator model consisted of a 2D CNN with U-NET architecture and the discriminator model worked on discriminating patches of the images, rather than each 2D slice as a whole [78]. The model was trained on in-phase, fat and water MR images from a DIXON sequence with rigidly registered CT images from 32 prostate cancer patients. The model was evaluated on 30 different pelvic cancer patients (10 prostate, 10 rectum and 10 cervix) with mean dose differences of 0.4%, 0.4% and 0.2% for the prostate, rectum and cervix patients respectively. The corresponding gamma pass rates at 2%/2 mm within the external contour were 95 ± 2%, 92 ± 3% and 93 ± 4%. Bird et al. also evaluated a GAN model trained on T2-weighted MR images and deformably registered CT images from 46 rectal radiotherapy patients [79]. Dose differences to the PTV D50% in 44 ano-rectal patients were small, 0.1 ± 0.2% (mean ± standard deviation, range $-0.6\%$, 0.6%) with very high 2%/2 mm gamma pass rates across the external contour of 99.5 ± 0.2% (99.0%,100.0%) when using deformably registered CTs as the comparison. The same analysis using rigidly registered CTs had similar mean dose differences of $-0.1\%$, but with a much wider range ($-3.5\%$,1.7%, estimated from graph). The gamma pass rates were also lower, 96% (range 81%,98%). Yoo et al evaluated sCTs generated from a GAN model and from two variants of the GAN, the cycle consistent GAN and the reference-guided GAN [80]. The cycle consistent GAN used two generators, forward (MR to sCT) and back (sCT to synthetic MR) and two corresponding discriminators (sCT compared to real CT and synthetic MR compared to real MR), with the model seeking to minimise both networks simultaneously. The reference-guided GAN further adjusted the output of the cycle consistent GAN with paired MR-CT data testing the sCT-CT differences voxel by voxel. The models were trained with 93 prostate patient images and evaluated on 20 patients. Mean dose differences to the PTV D50% were 0.4 ± 0.3%, 0.7 ± 0.6% and 0.6 ± 0.4% for the three models respectively. Corresponding gamma pass rates at 2%/2 mm were 93 ± 4%, 90 ± 5% and 94 ± 3%.

In conclusion there are a number of different approaches that have been used in sCT algorithms for pelvic cancers. Bulk density and atlas-based methods are mature approaches with commercially available solutions being used to clinically treat patients both in the context of research and as routine clinical practice, although as yet only for prostate cancers and not other pelvic cancers [81]. Voxel-based methods are less mature, but there is one in-house sCT algorithm that has been used clinically for several years. All these methods have been subject to rigorous evaluation with dosimetric analysis and not just image comparison metrics such as the mean absolute HU error. Deep Learning sCT algorithms are a much more recent but rapidly growing field, which are showing promising results. A significant advantage of Deep Learning algorithms is once the model is trained the implementation is very quick [76]. One issue for Deep Learning algorithms is the lack of dosimetric analysis of most models, with only a few papers investigating models in the pelvis with a dose evaluation. The remaining papers only report image similarity metrics, which bear little relation to the actual parameter of interest [78]. For Deep Learning algorithms to mature, it is important that dosimetric evaluations on significant patient numbers and on data acquired on different scanners to those used to train the model are carried out.

### 1.2.5 MR-Based Automatic Contouring

Modern radiotherapy techniques such as IMRT require all tumour targets and all nearby healthy OARs to be contoured for each patient [82]. This is essential for the treatment planning system optimizer to calculate the optimal dose distribution and to produce dose-volume statistics to enable clinical evaluation. The current clinical standard is for these contours to be manually delineated, which is both time-consuming and prone to inter- and intra-observer variability [83]. This has prompted extensive research into developing automatic contouring solutions, but to date these systems have not demonstrated sufficient accuracy and reliability for clinical use. This may be partly be due to the majority of these algorithms being based on CT images and so hampered by the same poor soft-tissue contrast that causes variability in manual contouring [84]. The superior soft-tissue contrast of MR may provide the additional imaging information required for accurate and robust automatic contouring, although MR also presents challenges for automatic contouring due to the large image intensity differences between patients [83]. For combined MR-CT pathways, typically the MR is only used for target delineation, whilst all OARs are contoured on CT. Therefore MR-only radiotherapy alone provides the opportunity for full MR-based automatic contouring techniques to be used in the clinical setting [85].

A relatively established approach to MR automatic contouring has been using an atlas of contoured MR images [83], although most applications have been to contouring for diagnostic purposes rather than radiotherapy [86]. Dowling et al. used such an multi-atlas technique based on T2-weighted prostate MR images to generate bone, bladder, prostate and rectum contours [69]. Comparisons to manual contours gave DSC agreement of $0.91 \pm 0.03$, $0.86 \pm 0.12$, $0.80 \pm 0.08$ and $0.84 \pm 0.06$ respectively (mean ± standard deviation). The results for prostate and rectum were equivalent to inter-observer results on the same images $0.84 \pm 0.11$ and $0.82 \pm 0.07$, whilst the bladder results were approaching the inter-observer variability of $0.95 \pm 0.01$. A limitation of atlas-based methods is increasing atlas size improves contour accuracy but at the cost of significant increase in computational time. Alvarez et al. developed an atlas-based approach that used a multiresolution similarity index to select a sub-atlas which was most similar to the image to be contoured, enabling a large atlas to be used without time penalty [87]. Comparisons with single-observer gold standard contours gave DSC score of $0.82 \pm 0.07$ for the prostate based on T2-weighted MR images.

Another approach to automatic contouring has been model-based, which uses a training data set to learn an average organ contour and characteristic shape variations, which are then applied to a new image using image features such as intensities, gradients and textures [84]. Pasquier et al. reported the results for a model-based algorithm for the prostate and seeded region-growing method for rectum and bladder on T2-weighted and T1-weighted images from 24 prostate patients [88]. The mean ± standard deviation DSC scores were $0.88 \pm 0.03$, $0.88 \pm 0.04$ and $0.94 \pm 0.02$ respectively (results converted from reported intersection over union (Jaccard similarity) results using $D = \frac{2J}{J+1}$, where $J$ was the Jaccard similarity and $D$ the DSC score [51]). Kuisma et al. investigated a commercial model-based algorithm on T2-weighted and T1-weighted DIXON images for on 65 MR-only prostate patients [86]. Compared to a single manual delineator, the automatic contours returned DSC scores of $0.84 \pm 0.04$, $0.92 \pm 0.04$ and $0.85 \pm 0.08$ for the prostate, bladder and rectum respectively. Automatic contours performed worse for the smaller and more variable organs: the seminal vesicles and penile bulb, with DSC scores of $0.56 \pm 0.17$ and $0.69 \pm 0.12$.

A more recent approach has been to use Deep Learning methods [83]. Wang et al. developed a 3D Deep Learning modified U-NET CNN, which included Deep Supervision [89]. Multiple loss functions were evaluated: cosine similarity which quantifies the similarity between two vectors in a certain space by the cosine of the angle between them, cross-entropy and the dice loss, as well as combinations of the three. The model was trained and evaluated on T2-weighted MR images from 40 prostate patients using a leave-five-out cross-validation method. Single-observer contours quality checked by an independent observer were used as the gold standard. The impact of the different loss functions were small, with all results agreeing within one standard deviation of the highest performing (DSC $0.85 \pm 0.04$). Elguindi et al. investigated two different 2D Deep Learning networks, both trained using the same 50 prostate patient T2-weighted MR images and evaluated on a further 50 using the clinically used contours as a gold standard [85]. One network was a very deep fully convolutional network with a large number of layers previously trained for segmenting natural images and initialised using those weights. Networks with large numbers of layers have not been used in medical image segmentation since training such very deep networks from scratch requires thousands of patient images [85]. The other was a conventional U-NET CNN with a few layers initialised with random weights. Both networks used data augmentation and a cross-entropy loss function. The previously trained network outperformed the U-Net on the bladder, prostate, penile bulb and rectum with median DSC scores respectively of 0.93, 0.83, 0.75 and 0.83 compared to the U-Net scores of 0.83, 0.75, 0.67 and 0.72 (inferred from graph). The two networks performed similarly for the urethra (median DSC 0.68 and 0.72).

Accurate and robust evaluation of automatic contouring methods is critical [90], with a number of different issues to consider: The first is to ensure that the algorithm is evaluated on the full range of patient images that may occur in a clinical pathway, including patients with unusual anatomies and images with significant artefacts [83].

The second consideration is what gold standard (or ground truth) will be used to evaluate the automatic contours against. This is a significant issue since manual contours are the current clinical standard but are known to have substantial intra- and inter-observer variability [86]. One method of accounting for this is to ensure comparisons are made to intra- and inter-observer variability on the same images [90]. If agreement with automatic contours to the gold standard is similar to or better than inter-observer variability in manual contouring, then the automatic contours should be considered clinically acceptable.

A third consideration is what metrics to compare the automatic and gold standard contours with. There are a wide range of different metrics reported in the literature [84]. Roach et al. investigated a number of different contouring metrics including volume, surface (eg distance to agreement), overlap (eg DSC) and centre-of-mass metrics, and correlated them with dosimetric differences between treatment plans created based on individual observer's contours and on gold standard consensus contours for prostate cancer patients [91]. Little to no correlation was found between metric values and significant differences in dosimetry, suggesting reporting contouring metrics on their own without dosimetric results is of little benefit in indicating clinical acceptability for radiotherapy planning. Developing new metrics which potentially are more relevant to radiotherapy is therefore important. Nikolov et al. developed an alternative metric called the surface DSC metric [92]. This seeks to measure agreement between the surfaces of the two structures relative to agreement within a clinically defined acceptable distance called $\tau$. This therefore ensures small clinically irrelevant boundary differences are ignored but larger ones are counted. Critical to the implementation of this metric is the value given to $\tau$ [85]. Another alternative metric is a subjective rating from an expert delineator(s) on the accuracy of the contour. Although this is not a quantitative result, it evaluates whether a volume would be clinically acceptable as it would be evaluated in clinical practice and so is the most relevant metric [83]. Song et al. graded contours from a CT-based Deep Learning model for rectal radiotherapy patients using five different yes/no criteria by two different observers [93]. Such metrics will be important in establishing automatic contouring solutions as part of routine clinical practice.

Fourthly, clinical implementation of automatic contouring will require per-patient QA [94]. Current Royal College of Radiologists (RCR) guidelines require all target contours for radically-treated patients to be peer-reviewed [95]. This process is directly applicable to automatic contours, but brings back the inter-observer variability and time-consuming problems that automatic contouring was designed to remove. Automatic methods of contour QA have been proposed using an atlas of previously contoured images as a knowledge base [94]. Metrics were evaluated such as size, shape and relative position of the new contours compared to the atlas and these metrics combined with experimentally determined weights to produce a pass/fail flag. The method detected 40/42 deliberate contouring errors introduced as well as flagging 9 false positives on head and neck cancer CT images. Ensuring the independence of automatic contouring and automatic checking methods will be important.

A final issue concerns quantifying the time-saving produced by automatic contours. This is a key motivation of automatic contouring and they may still reduce contouring time even if the automatic contours are not clinically acceptable in themselves. A number of studies have assessed this, however measuring the time-saving is quite a time-consuming task in and of itself. Vaassen et al. assessed the correlation between different metrics of automatically generated and manually adjusted volume and the amount of time spent by the delineator for 20 lung cancer patients [96]. They concluded that the surface DSC with $\tau = 0$ mm (ie with zero tolerance for differences) and the added path length required metrics did correlate with time savings, whereas standard DSC and distance metrics did not. This suggests that surface DSC is a clinically relevant metric for radiotherapy purposes.

An alternative approach is to assess dose differences due to differences in contours [91]. Variations in OAR contours may mean the true dose to the organ exceeds DVH constraints, potentially resulting in harm to the patient. This is a more clinically relevant measure but only a few papers have investigated it, with several methodologies used because contours are used for two purposes in treatment planning: firstly to optimise the treatment plan and secondly to evaluate it and report dose statistics. The first methodology reported in the literature is a plan optimisation assessment, evaluating the efficacy of automatic contours when used to optimise a treatment plan. This method requires two treatment plans, one based on manual contours ($\text{plan}_{\text{man}}$) and one on automatic contours ($\text{plan}_{\text{auto}}$). The doses to the manual contours ($\text{contour}_{\text{man}}$) from each plan are compared. The second method is an evaluation assessment, which investigates the accuracy of reported doses from automatic contours compared to manual contours from the same plan. A third combined option is to optimise two plans and then compare doses to manual contours from $\text{plan}_{\text{man}}$ and doses to automatic contours from $\text{plan}_{\text{auto}}$. Cao et al. investigated all three options using atlas-based automatic contours for five prostate and five head and neck patients [97]. For the prostate patients both the target and OAR volumes were assessed, meaning the PTV was different between the two plans, whereas for the head and neck patients only the OAR volumes were compared with both plans using the same (manual) PTV. There were no significant differences for any dose constraint for the prostate patients suggesting high contouring accuracy for both optimisation and evaluation. There were some larger dose differences for some of the organs for the head and neck patients, with issues for both optimisation and evaluation. For each patient, $\text{plan}_{\text{man}}$ and $\text{plan}_{\text{auto}}$ were manually optimised by the same planner. This potentially introduces a confounding difference between the plans caused by subtle differences in how the planner manually optimised the plan. Van Rooij et al. aimed to control for this by using an automatic knowledge-based planning algorithm to optimise plans based on Deep Learning automatic and manual contours [98]. 15 head and neck radiotherapy patients were evaluated with the same PTV used for both plans. Mean and maximum doses to manual OARs were calculated from both plans and dose differences determined (ie assessing optimisation only). Mean dose differences were small for all OARs ($\leq$ 2.2 Gy), and although statistically significant for two OARs were not deemed clinically significant. For a few OARs the maximum dose difference was substantially larger (up to 9.1 Gy), which would be considered clinically significant. Another study evaluated Deep Learning and atlas based automatic contouring compared to manual contours of the bladder and rectum for 15 prostate cancer patients, assessing accuracy for optimisation only [99]. Each contour was reviewed by a radiation oncologist and edited where necessary. The atlas contours required significantly more editing time (mean 10.2 minutes) than the manual contours (4.1 minutes) and the Deep Learning contours (4.7 minutes). Doses to manual contours derived from treatment plans optimised using edited Deep Learning contours and edited manual contours were compared. Standard optimisation weights were used for both plans without adjustment to ensure the plans were comparable. The mean dose differences were $\leq$ 0.2 Gy for all dose constraints evaluated, however the relevance of this is limited since the automatic contours were manually edited to be clinically acceptable. Vaassen et al. assessed both un-edited and edited automatic contours for dose differences in 20 lung cancer patients [100]. An atlas-based and a Deep Learning based automatic contour sets were evaluated compared to manual ones, plus edited versions of both automatic contouring methods. Plans were generated using automatic knowledge-based planning for all five contour sets using the same PTV for all plans. Optimisation, evaluation and combined assessments were performed and small differences were found for all with differences of the same magnitude as those due to intra-observer variability.

In summary, MR-based Deep Learning automatic contouring methods have significant potential for producing accurate contours, with significantly less computational time than more established atlas-based methods. However evaluations in the literature have primarily used standard overlap metrics such as the DSC, which have limited applicability to actual dosimetric differences. Further work is required to develop and validate metrics that are clinically relevant and to evaluate Deep Learning algorithms using these metrics on a variety of patients before widespread clinical adoption. In this context the surface DSC appears promising. Incorporating such metrics into the loss functions of Deep Learning algorithms may improve their performance further. Additionally quantifying the contouring time-saving is important for encouraging clinical uptake, which needs to include time for contour QA. Automatic methods may also have an important role here, but ensuring model and data independence from automatic contouring methods will be essential. The most clinically relevant assessment is of dose differences between automatic and manual contours. However this approach also has issues, depending on whether dose differences due to plan optimisation, plan evaluation or both are assessed. The literature to date has mostly compared contours for plan optimisation only (ie comparing doses to manual contours from a plan optimised on automatic contours and a plan optimised on manual contours). However plan evaluation (comparing doses to automatic and manual contours from the same plan) is more clinically relevant. Combined comparisons on their own do not provide sufficient information to assess contour accuracy since in principle the automatic and manual contours could be completely different shapes and locations but return identical DVH parameters. In addition, whether target and OAR volumes are compared or just OARs can have a substantial impact on the results. Finally the results presented are only relevant for the PTVs assessed, which will have limited impact for tumours with small variations in PTV between patients (eg prostate radiotherapy) but is significant for ones with larger variations (eg head and neck radiotherapy).

## 1.3 PET-MR Functional Information for Radiotherapy Planning

Conventional radiotherapy treatments aim for a homogeneous dose to the entire visible tumour, called the Gross Tumour Volume (GTV), plus a margin to cover microscopic disease, called a Clinical Target Volume (CTV) [10]. However, tumours are known to be very heterogeneous, with some regions far more active than others [12]. For conventional radiotherapy to be successful, this require the entire CTV to receive the radiation dose sufficient to ensure cell kill in the most active tumour sub-volume. The large volume receiving this high dose either causes significant side-effects [13] or due to unacceptable toxicities the dose to the whole CTV is reduced, reducing the tumour control probability [14]. The concept of dose painting aims at delivering 'boost' radiation doses to the active tumour sub-volumes, theoretically boosting tumour control probability without changing treatment side-effects [12]. This is supported by recurrence evidence, where the majority of local recurrences will occur at the location of the most active sub-volumes [101]. Modern radiotherapy techniques such as IMRT and Image Guided RadioTherapy (IGRT) are able to plan and accurately deliver complex dose distributions targeting different volumes with different doses whilst sparing OARs [4]. Therefore the key to clinical utilisation of dose painting depends on accurately identifying the active tumour sub-volumes [15]. It is important that functional imaging used for dose painting is robust and repeatable with high geometrical accuracy [27]. Two of the primary imaging modalities investigated for identifying tumour sub-volumes for radiotherapy dose painting are DW-MR and PET [16].

### 1.3.1 DW-MR for Radiotherapy Planning

DW-MR is a quantitative MR imaging technique which is sensitive to the diffusion of water molecules due to Brownian motion [102]. Due to the differences in restriction of Brownian motion in different biological tissues, DW-MR can generate image contrast due to differences in water diffusion. DW-MR is most commonly performed using pulsed gradient spin echo sequences, consisting of standard spin echo sequence with two equal diffusion gradients either side of the $180^o$ re-focusing pulse [103]. If there was no motion of water molecules then the two gradients would cancel each other out and there would be no signal loss [104]. However in the presence of molecular motion the proton spins experience a different phase shift from the second gradient to the first, resulting in signal loss from spin dephasing. For random Brownian motion this signal loss will be exponential and is given by [104]

$$S(TE, b) = S_0 e^{-TE/T_2} e^{-bADC}, \tag{1.2}$$

where $S_0$ is the signal at $b = 0$, $TE$ is the sequence echo time, $T_2$ is the spin-spin relaxation time constant of the tissue, $b$ describes the amount of diffusion weighting applied and is defined below and $ADC$ is the Apparent Diffusion Coefficient of the tissue. For the standard pulsed gradient spin echo sequence the b-value is given by [103]

$$b = \gamma^2 G^2 \delta^2 \left( \Delta - \frac{\delta}{3} \right), \tag{1.3}$$

where $\gamma$ is the gyromagnetic ratio, $G$ the magnitude of the diffusion-weighted gradient, $\delta$ is the pulse width and $\Delta$ the gradient spacing. By acquiring multiple images with different b-values and the same echo time, the Apparent Diffusion Coefficient (ADC) can be calculated by fitting the observed signal with (1.2) on a pixel-by-pixel basis. The ADC quantifies the diffusion of water molecules but does not measure the true water diffusion since in biological tissues the diffusion is restricted [105]. It is also sensitive to other sources of molecular motion, for example bulk flow [103].

DW-MR has been shown to be sensitive to the detection of tumours where diffusion is more restricted due to the tight packing of cancer cells, with tumours appearing dark on ADC maps [106]. A meta-analysis of 5892 lesions reported a sensitivity and specificity of DW-MR detection of prostate cancer as 0.69 and 0.89 respectively [107]. DW-MR is now the most important functional method for the detection and staging of prostate cancer [105] and is considered essential for rectal cancer diagnosis and staging [108]. This has in turn driven research into using DW-MR for boost volume identification in pelvic cancers [109].

Alexander et al. assessed the diagnostic accuracy of T2-weighted MR and DW-MR of identifying dominant intra-prostatic lesions which could be used for boost radiotherapy doses [110]. They concluded that T2-weighted MR and DW-MR, using b-values $b = 0 - 800\ \mathrm{s\,mm^{-2}}$, can robustly detect lesions with diameters $\geq 1.0$ cm to receive higher boost doses. A multi-centre randomised trial in prostate cancer investigated delivering a boost dose of 95 Gy to the dominant lesion identified using DW-MR and Dynamic Contrast Enhanced-Magnetic Resonance (DCE-MR) combined with 77 Gy in 35 fractions to the whole prostate compared to the standard arm of 77 Gy in 35 fractions to the whole prostate [111]. Early toxicity results show no significant differences between the two arms, suggesting dose escalation to a boost volume is safe and feasible.

For the detection of tumours within the prostate, DW-MR with b-values $b = 1500\ \mathrm{s\,mm^{-1}}$ and $b = 2000\ \mathrm{smm^{-1}}$ were reported to be better than lower b-values [112]. Rosenkrantz et al. investigated very high b-values up to $b = 5000\ \mathrm{s\,mm^{-1}}$ using computed b-values based on a mono-exponential ADC fit [113]. They reported that b-values in the range $b = 1500-2500\ \mathrm{s\,mm^{-1}}$ were optimal for tumour detection, although directly measured b-values of that magnitude are likely to result in images with very low Signal-to-Noise-Ratio (SNR).

Song et al. compared various MR techniques for tumour delineation in cervix cancer and concluded that the combination of T2-weighted and DW-MR was the most accurate [114]. Schernberg et al. investigated using DW-MR to delineate brachytherapy boost volumes for cervix cancer and reported that DW-MR volumes would have modified the T2-weighted boost volume in 73% of patients [115].

The conventional implementation of DW-MR is with a single-shot Echo Planar Imaging (EPI) sequence. This collects all the k-space data for an image slice with one RF pulse (or pulse combination) using gradient echoes [104]. A significant issue with single-shot EPI is its high sensitivity to magnetic field inhomogeneities causing significant geometric distortion [116]. In the pelvis this is mainly due to susceptibility effects from air in the rectum and main field inhomogeneity [117]. Distortion is mainly in the phase-encode direction due to the small bandwidth-per-pixel in this direction required by the gradient blipping in the single-shot EPI implementation [118]. This is a significant issue for radiotherapy since accurate treatment depends on geometrically accurate planning images [119]. Geometric distortion can be reduced through increasing the bandwidth per pixel in the phase encode direction [116]. This reduces the voxel shift caused by a given frequency shift. Another method is to use parallel imaging , which reduces the number of phase-encoding steps and so the addition of phase errors. However geometric distortion remains a significant issue for single-shot EPI based DW-MR. Another issue for single-shot EPI is its low resolution, which can limit accurate lesion delineation and ADC quantification due to partial volume effects [120].

These issues have led to developing alternative sequences for DW-MR. One methodology is reduced FOV single-shot EPI. This uses a 2D radiofrequency pulse to only excite spins within the phase-encode FOV as well as for a single slice, rather than just for a single slice in conventional single-shot EPI [121]. This increases the effective bandwidth-per-pixel in the phase-encode direction, reducing geometric distortion. The image resolution is also improved due to the smaller FOV. In addition it inherently supresses fat, which is essential for EPI due to the large chemical shift artifacts in the phase-encode direction [121]. Reduced FOV has been compared to conventional single-shot EPI DW-MR for imaging 44 prostate cancer patients prior to radical prostatectomy [122]. The reduced FOV images were graded by two radiologists as having significantly superior image quality and reduced geometric distortion. The image quality improvement was only significant when phase encoding was in the left/right direction and not the anterior/posterior direction. 52% of patients had poor or non-diagnostic scores improved to fully diagnostic using reduced FOV in the left/right direction. ADC values and correlations with Gleason score were equivalent between the two methods, suggesting reduced FOV can simply replace conventional EPI DW-MR in standardised reporting schemes and automatic segmentation algorithms. Finally, inter-observer agreement in delineation of lesions was significantly improved with reduced FOV DW-MR. A similar study compared reduced FOV with conventional single-shot EPI DW-MR for 15 prostate cancer patients [123]. They reported significantly better alignment of lesions between DW-MR and T2 anatomical images for the reduced FOV sequence (differences $3 \pm 4$ mm) compared to conventional EPI (differences $5 \pm 3$ mm), suggesting reduced geometric distortion, as well as improved subjective image quality. Similarly, Thierfelder et al reported reduced FOV DW-MR had improved image quality without significant change in ADC values [124].

Another alternative DW-MR sequence is multi-shot EPI. This divides k-space into a few segments which are each acquired by a separate EPI sequence [103]. This increases the effective bandwidth-per-pixel in the phase encode-direction, reducing geometric distortion at the cost of increased scan time, which increases sensitivity to motion artifacts [125]. DW-MR is highly sensitive to motion artifacts since the DW-MR signal depends on molecular motion. An improvement to multi-shot EPI is MUltiplexed Sensitivity Encoding (MUSE)). This uses the sensitivity profile of the receive coils to estimate motion-induced variations between the different interleaved EPI segments [126]. This enables the benefits of multi-shot EPI without shot-to-shot motion variations.

Finally, geometric distortion can be reduced through post-processing. Nketiah et al. reported using a conventional single-shot EPI acquisition with an additional $b = 0\,\mathrm{s\,mm^{-2}}$ image acquired with the same parameters but a reverse phase-encoding polarity, which was used to determine the distortion deformation field through hierarchical smoothing [118]. They reported that geometric distortion had a significant effect on ADC quantification, and that application of the correction reduced residual distortion significantly. Tong et al. investigated using a direct field map approach to quantifying $B_0$ inhomogeneity using a separate gradient echo sequence rather than an additional DW-MR sequence [127]. They reported significantly greater overlap of the prostate delineation on distortion corrected DW-MR images with T2 anatomical images compared to non-distortion corrected DW-MR images.

In summary, DW-MR is an essential component for the detection and staging of pelvic cancers, and is increasingly being investigated for radiotherapy boost target delineation. Conventional EPI DW-MR in the pelvis has low resolution and is prone to geometric distortion, which limits its use in radiotherapy treatment planning. However, radiotherapy boost treatment based on EPI DW-MR volumes is currently the focus of several ongoing multi-centre trials (such as the FLAME trial [111]). More advanced DW-MR techniques such as reduced FOV single-shot EPI and multi-shot EPI have shown promise of reduced geometric distortion, as have the application of post-processing techniques using reversed phase-encoding acquisitions. The geometric and ADC accuracy of radiotherapy DW-MR protocols need to be assessed prior to clinical implementation.

### 1.3.2 PET for Radiotherapy Planning

PET uses the injection of a radiotracer which decays via positron emission to form an image through the detection of pairs of photons emitted from positron-electron annihilation. Different radiotracers will be taken up differently, providing different metabolic information in the PET image. The most common radiotracer for cancer imaging is a glucose analogue, $^{18}$F-FDG. This uses the increased glucose metabolism of most types of cancer cells to detect and localise tumours [128]. It is widely used in the diagnosis, staging and evaluation of treatment response for a wide range of cancers. For pelvic cancers, $^{18}$F-FDG-PET-CT is used for staging, treatment strategy, patient prognosis and treatment response in cervix cancer [129], with PET critical for the identification of loco-regional involved nodes [130]. A recent meta-analysis concluded $^{18}$F-FDG-PET-CT adds value in staging locally advanced anal cancer [131]. A clinical trial is currently underway to investigate whether $^{18}$F-FDG-PET-CT or $^{18}$F-FDG-PET-MR change patient management in rectal cancer [132]. This in turn has lead to interest in using $^{18}$F-FDG-PET for the delineation of radiotherapy target volumes.

In a treatment planning study for cervix cancer patients, Esthappen et al. developed plans delivering a boost radiotherapy dose of 59.4 Gy to $^{18}$F-FDG-PET-CT defined boost volumes [133]. Lin et al. used $^{18}$F-FDG-PET-CT to guide brachytherapy boosts for cervix cancer, which improved dose coverage of the tumour [134]. Zhang et al. reported that $^{18}$F-FDG-PET-CT defined tumour volumes with a threshold of 40% of SUV$_{\mathrm{max}}$ showed good agreement with histopathology volumes for cervix cancer whereas anatomical MR and CT over-estimated the volume [135]. Krengli et al reported significant changes in $^{18}$F-FDG-PET-CT based radiotherapy tumour delineations compared to CT-based delineations in 15/27 anal cancer patients [21]. A study in rectum cancer patients showed reduced inter-observer variability for tumour delineations on $^{18}$F-FDG-PET-CT compared to CT alone [136].

PET is a quantitative imaging modality, although the limitations of the PET spatial resolution and heterogeneity of tumours means a semiquantitative measure is typically used, called the Standard Uptake Value (SUV) [137]. This holds the potential for automatic contouring and several studies have investigated different methods for pelvic cancers. Day et al. investigated three different automatic contouring methods for 18 rectal and anal cancer patients receiving radiotherapy compared to manual contours [138]. Using a threshold of 43% of the maximum SUV ($\text{SUV}_{\text{max}}$) resulted in volumes on average 56% smaller than the manual contours. Contours produced with a fixed threshold $SUV = 2.5\ \text{g ml}^{-1}$ were closer with a mean difference of 37%, whilst using a confidence connected region growing approach had contours that were only 9% different.

In addition, the quantitative nature of PET gives it potential as a prognostic tool, which can be used to drive dose escalation or de-escalation strategies [130]. The primary parameters investigated have been $\text{SUV}_{\text{max}}$, the Metabolic Target Volume (MTV) which is the size of an automatically contoured volume normally using a threshold of $\text{SUV}_{\text{max}}$, and the Total Lesion Glycolysis (TLG) which is the MTV multiplied by the mean SUV. In anal cancer, $\text{SUV}_{\text{max}}$ is significantly correlated with T-stage and histology, which in turn are significant prognostic factors for disease-free and overall survival [139]. In another study, larger MTV (delineated using a threshold of 50% of $\text{SUV}_{\text{max}}$) was significantly associated with poorer overal survival, even when adjusting for T classification, with an optimum cut-off of 26 $\text{cm}^3$ [140]. In rectal cancer, Ogawa et al. found TLG was an independent prognostic factor for disease-free and overall survival for patients treated with surgical resection without neoadjuvant treatment [141]. TLG was calculated using a volume from 30% of $\text{SUV}_{\text{max}}$ and receiver operator curve analysis was used to determine cut-offs for MTV and TLG of 25.23 $\text{cm}^3$ and 341 g respectively. Similarly Choi et al. found MTV using 50% of $\text{SUV}_{\text{max}}$ and a cut-off of 23.9 $\text{cm}^3$ and TLG with a cut-off of 125.84 g to be significant independent prognostic factors for disease-free survival and MTV to be significant for overall survival.

Unlike most types of cancer cells, prostate tumours tend to show low $^{18}$F-FDG uptake due to their low glycaemic activity [142]. This has motivated research into other PET tracers for detecting prostate cancer. One of these is $^{18}$F-Fluorocholine, a synthetic version of the choline molecule which is phosphorylated by the enzyme choline kinase [143]. This enzyme is typically over-expressed by cancer cells, including prostate cancer, resulting in increased uptake and therefore PET signal in prostatic cancer cells [144]. The Royal College of Radiologists evidence-based guidelines for the diagnostic use of PET recommend $^{11}$C-Choline or $^{18}$F-Fluorocholine for evaluation of high-risk patients, or patients with equivocal findings on other imaging [20]. $^{11}$C-Choline is preferred due to its reduced urinary excretion however its short half-life makes it only available for centres with a cyclotron onsite [20]. The use of $^{18}$F-Fluorocholine-PET for prostate cancer diagnosis has led to several investigations into the utility of $^{18}$F-Fluorocholine-PET for guiding radiotherapy planning.

Kwee et al. found the prostate sextant with the largest $\text{SUV}_{\text{max}}$ corresponded to the sextant with largest tumour volume in 13/15 patients who received $^{18}$F-Fluorocholine-PET-CT prior to prostatectomy and histopathological analysis [143]. PET imaging was acquired with activities of 3.3-4 $\text{MBq kg}^{-1}$ 10 minutes after injection. There was a significant difference between $\text{SUV}_{\text{max}}$ in malignant and benign sextants, with the benign sextants having a $\text{SUV}_{\text{max}}$ 63 % of the malignant ones. Pinkawa et al. used $^{18}$F-Fluorocholine-PET to automatically define dominant lesions(s) in 66 patients using a threshold of twice the $\text{SUV}_{\text{max}}$ in the lowest SUV 1 $\text{cm}^2$ region of the prostate [145]. Images were acquired 1 hour after an injection of 178-355 MBq of $^{18}$F-Fluorocholine. These patients were treated with a dose of 76 Gy in 38 fractions to the whole prostate and a boost dose of 80 Gy to the $^{18}$F-Fluorocholine-PET defined regions plus a 4 mm expansion. Equivalent uniform doses to the bladder and rectum were not significantly different between these patients and 18 patients treated in the same period without boost doses. A follow-up study reported no significant differences in long-term patient-reported quality of life between boosted and non-boosted patients [146]. Kuang et al. investigated prostate radiotherapy plans with and without a boost dose to the $^{18}$F-Fluorocholine-PET-CT defined dominant intraprostatic lesion on 30 patients [147]. Patients were imaged with activities of 2.6 $\text{MBq}\,\text{kg}^{-1}$ 30 minutes after injection. The boost region was defined as 60 % of $\text{SUV}_{\text{max}}$ within the prostate, with an inner region defined using 70 % of $\text{SUV}_{\text{max}}$, with a 6 mm expansion except for 3 mm anteriorly and posteriorly, excluded from overlapping with rectum or bladder. The boost regions were prescribed 100 Gy (60 %) and 105 Gy (70 %), with the rest of the prostate receiving 79 Gy in 39 fractions. The 60 % boost region covered all tumour-bearing sextants in 27/28 patients with histopathology, with a significant increase in theoretical tumour control probability and no significant difference in OAR doses between boost and no-boost plans.

In summary, $^{18}$F-FDG-PET-CT is used for staging and deciding treatment strategy for anal and cervix cancers. This suggests that there may be a benefit for radiotherapy boost target delineation. There have been some studies suggesting $^{18}$F-FDG-PET-CT to delineate target volumes for radiotherapy treatment planning could be beneficial. However the clinical evidence remains limited and further investigation is required. For prostate cancer, radiotherapy plans delivering boost doses to $^{18}$F-Fluorocholine-PET defined regions has been demonstrated to be technically feasible without significantly increasing OAR doses. Imaging with activities of 2.6-4 $\text{MBq}\,\text{kg}^{-1}$ have been used, with thresholds of 60% $\text{SUV}_{\text{max}}$ being shown to be histopathologically appropriate. However the uptake time between tracker injection and imaging varies widely between studies. One small pilot study reported no increase in long-term quality of life following treatment with 5% boost doses. Theoretical improvements in tumour control probability have been demonstrated though improvements in patient outcomes have not yet been reported. $^{18}$F-Fluorocholine-PET appears promising for the identification of boost doses for radiotherapy planning of prostate cancer.

### 1.3.3 Combined PET-MR for Radiotherapy Planning

Simultaneous PET-MR scanners enable acquiring both MR and PET functional information at the same time and so with high degrees of spatial alignment. This has the potential to improve boost volume delineation through utilising the complementary information from both modalities [29]. Kim et al found combined $^{18}$F-fluorocholine-PET-MR imaging better at identifying localised prostate cancer than either PET or DW-MR alone [148]. There is an ongoing clinical trial investigating using a combination of DW-MR, DCE-MR and $^{18}$F-flurocholine-PET-CT to delineate a radiotherapy boost volume to 68 Gy in 20 fractions, with two-year toxicity data showing the treatment is well tolerated [149]. Zamboglou et al. compared $^{68}$Ga-Prostate Specific Membrane Antigen (PSMA)-PET-CT with T2-weighted, DW-MR and DCE-MR for prostate boost volume delineation and reported only 42% overlap in volumes [150]. A further study found both imaging modality volumes had moderate overlap with the histopathology volume, which improved significantly when the union of each modality volume was used, suggesting potential benefit for combined PET-MR [151]. A recent review concluded that PSMA-PET and T2-weighted and DW-MR offer complementary information regarding prostate boost volume delineation and incorporating both significantly improves the potential tumour control probability [152]. However they reported there remains significant inter-observer variability in boost delineation using these modalities and the development of guidelines and/or automatic delineation methods to improve the reproducibility of delineation is necessary for routine clinical implementation.

Brandmaier et al. reported an inverse correlation between $^{18}$F-FDG-PET SUV and ADC values for cervical cancer tumours, but they did not report on differences between delineated volumes [153]. This was in contrast to previous studies using MR and PET-CT, which they hypothesised was due to the improved co-localisation of simultaneous PET-MR. Song et al. investigated various MR sequences and $^{18}$F-FDG-PET-CT for cervix cancer delineation. They reported that combined T2-weighted and DW-MR were the best sequence for target delineation, with PET volumes being significantly smaller and indicating potential for boosting [114]. Rusten et al. evaluated DW-MR and $^{18}$F-FDG-PET-CT in the delineation of anal cancer targets and found good agreement in the overall target volume but some variability in identifying the boost volume [22]. They also reported significant inter-observer variability within the same modality for the boost volume delineation.

In conclusion, the combination of DW-MR and PET acquired simultaneously has significant potential to improve boost volume delineation in pelvic cancers. However, there remains limited clinical evidence of its feasibility and effectiveness, implying further work is needed to evaluate the combination. In particular, most studies have reported using MR and PET-CT information acquired in different imaging sessions, and then registered together with consequent uncertainties in anatomical alignment. This highlights the potential benefits of combined PET-MR scanners. In addition several studies reported significant inter-observer variability when using these modalities, implying the development of robust automatic delineation methods may be required for routine clinical use.

### 1.3.4 PET-MR in the Radiotherapy Planning Position

It is essential that the images used to plan radiotherapy treatments are acquired with the patient in the same position to ensure accurate radiotherapy treatment [154]. This is because the treatment plan is created assuming the planning images accurately represent the patient as they will be on treatment. Therefore, any discrepancies between the patient position at imaging and on treatment could result in a delivered dose distribution that is different to the planned, potentially under-dosing the tumour and/or over-dosing OARs. Therefore acquiring PET-MR images for radiotherapy planning requires patients to be scanned on a radiotherapy flat couch-top which mimics the treatment machine couch, with patients in appropriate radiotherapy immobilisation devices and with the MR receive coils supported off the patient so that the patient external contour is not deformed [154]. The carbon fibre couches typically used for PET-CT imaging have low PET attenuation but produce significant MR artefacts, whereas the glass fibre MR couches do not interfere with the MR signal but are significantly PET attenuating [155]. This means dedicated PET-MR radiotherapy hardware needs to be developed that is MR-compatible and has low PET attenuation [156]. Acquiring PET-MR images in the radiotherapy position will have an impact on MR image quality [157] since the receive coils will be further from the patient anatomy, reducing the coil filling factor and therefore the SNR [158, 159]. The radiotherapy planning position will also impact on PET image quality since the flat couch-top and immobilisation devices will add additional and non-uniform PET attenuation, degrading the image quality [155]. Therefore it is important to assess the impact on PET-MR image quality of using the dedicated PET-MR radiotherapy hardware so that MR protocols can be modified to compensate for the MR signal loss [157] and software methods of correcting for the PET attenuation can be developed for accurate quantitative PET imaging [156].

Paulus et al. developed a radiotherapy flat couch-top for PET-MR imaging consisting of a foam core surrounded by a plastic outer layer [155]. A coil bridge for holding flexible anterior array coils was also developed for head and neck imaging. Attenuation correction maps were created using CT scans of the radiotherapy hardware. The impact on MR signal was assessed using a uniform nickel sulphate phantom, which found a 25% SNR drop in the radiotherapy setup compared to the standard diagnostic head and neck coil. Comparison of T1-weighted 3D volunteer scans showed no essential differences, with the radiotherapy setup providing slightly noisier images. The PET image quality was assessed using a cylindrical uniform phantom. Without correction the attenuation of the radiotherapy couch was 3.8% and of the coil bridge 13.8%. If the correction maps were applied this dropped to 0.6% and 0.8% respectively. Two patients were also assessed with PET images acquired with and without the head coil and coil bridge. The mean difference in $\text{SUV}_{\text{max}}$ across five ROIs was $-11.0\%$ and $-0.9\%$ without and with attenuation correction respectively. This approach has also been extended to pelvic radiotherapy, using the same radiotherapy couch with a pelvic coil bridge [156]. The coil bridges were adjustable so that the MR anterior array coil could be positioned as close to the patient as possible. CT scans were again used to generate attenuation correction maps of the pelvis coil bridge. The impact on PET image quality was assessed using the uniform component of the NEMA PET body phantom [160] with and without the coil bridge, with the coil bridge introducing a mean $-8.5\%$ difference, reducing to $-1.2\%$ when attenuation correction was used. PET-MR imaging with and without the coil bridge of three abdominal patients with active lesions showed mean underestimations of $\mathrm{SUV}_{\mathrm{max}}$ of $-11.1 \pm 2.0$ without attenuation correction, reduced to $-3.9 \pm 2.6\%$ with the correction map incorporated. Winter et al. investigated using the same couch and head and neck coil bridge to evaluate PET-MR image quality in 10 head and neck radiotherapy patients [161]. Each patient was scanned in a radiotherapy setup with flat couch and coil bridge and in a conventional diagnostic setup. They reported a median drop in the SNR of the T2-weighted MR images of 26%, with reductions of 38% and 31% for DW-MR with b-values of 150 and 800 $\mathrm{smm}^{-2}$ respectively. Contrast to Noise Ratios (CNR) of PET-defined lesions compared to surrounding tissues were also lower in the radiotherapy setup, with a reduction is 31% for the T2-weighted MR images. ADC measurements were very similar with median difference of $-1.7\%$ and a coefficient of repeatability between diagnostic and radiotherapy setups of 17.6%. High similarity between the same structures delineated on MR images from the two setups were also found with a DSC score of 0.85 (min 0.68, max 0.89), suggesting that MR image quality was sufficient for radiotherapy purposes. PET images were assessed using automatically segmented ROIs with thresholds of 50% of $\mathrm{SUV}_{\mathrm{max}}$ compared between diagnostic and radiotherapy setups. High similarity was found, with median DSC score of 0.88 (min 0.69, max 0.94).

An alternative radiotherapy couch and pelvis coil bridge was developed by Brynolfsson et al [162]. The couch consisted of 5 mm thick PMMA layer on top of a 35 mm thick foam base, cut to fit into the GE Healthcare Signa PET-MR patient couch. The pelvis coil bridge was constructed out of polyoxymethylene and polycarbonate. Attenuation correction maps were created using CT scans of the couch and coil bridge. Measured activity using a uniform PET phantom with couch was 5% lower compared to without, and 13% lower with the couch and coil bridge. Incorporating attenuation correction resulted in differences of 1.5% and 0.7% respectively. MR SNR measured using a uniform MR phantom was 74% and 67% with the couch and couch and coil bridge respectively compared to measurements without.

Finally Witoszynskyj et al. developed a plastic and fibre glass radiotherapy couch for PET-MR imaging [163]. Minimal differences in MR SNR with and without the couch were found using uniform MR phantoms. The PET NEMA phantom was imaged and the difference in imaged versus injected activity was $8.7 \pm 2.1\%$ with the couch, reducing to $1.2 \pm 3.9\%$ when CT-based attenuation map of the couch was included in the reconstruction. A $^{68}$Ge/$^{68}$Ga transmission source was also used to determine a couch attenuation map, but its performance was unchanged compared to the CT-based map.

In conclusion, radiotherapy flat couch tops and coil bridges have been developed for PET-MR imaging in the radiotherapy planning position. The impact on MR image quality varies, with some solutions having minimal impact, and others reductions in SNR of 38%. PET image quality is also impacted, with reductions in measured activity levels of up to 13%. This effect was shown to be largely reversible with attenuation correction maps, reducing activity discrepancies to $\leq$ 4%. Of note is that all phantom assessments of radiotherapy hardware have used uniform phantoms only and the impact of the reductions in PET and MR signal on image quality has not been assessed in the pelvis.

## 1.4 PET-MR Quality Assurance for Radiotherapy Planning

PET-MR for radiotherapy planning is still an emerging technique and to the best of my knowledge there is nothing in the literature directly on it. There are consensus guidelines on PET-MR QA for the diagnostic setting [164]. These include PET cross-calibration and image quality assessments, MR image quality assessments with different receive coils, and PET-MR alignment assessment. However radiotherapy imaging has different requirements to diagnostic imaging and this needs to be reflected in QA programmes [119].

MR QA for radiotherapy planning is more established [154] and PET-MR radiotherapy QA will need to include all the tests included within radiotherapy MR QA. These primarily focus on MR image quality and geometric distortion [119]. MR image quality for radiotherapy is assessed using the same methods as for diagnostic scanners (eg American College of Radiologists image quality phantom [165]). MR geometric distortion assessment is essential for radiotherapy planning [166]. In particular radiotherapy requires the geometric distortion to be measured over a large field of view that includes the patient external contour [167]. There have been a number of phantoms developed for MR large field of view geometric distortion measurements for radiotherapy [81], with at least one demonstrating repeatable measurements [167]. The distortion of a scanner over time may change due to degradations in the performance of the gradient coils but there has only been one study investigating the temporal stability of measurements [168]. They reported measurements were stable but recommended that large field of view distortion should be assessed as part of a regular QA programme.

DW-MR is more prone than conventional anatomic imaging to geometric distortion [116]. This means it is important to include assessments of geometric distortion with DW-MR sequences [168]. Conventional single-shot EPI implementations of DW-MR are particularly sensitive to distortions caused by susceptibility differences in air-tissue boundaries.

This has lead many assessments of distortion to be carried out using human subjects, with anatomic MR images being used as the gold standard. Winfield et al. reported on using a large uniform phantom filled with silicone to assess eddy current induced distortions [169]. Silicone was used since it has a very low diffusivity at room temperature. This meant differences in $b = 0$ smm$^{-2}$ and $b = 1000$ smm$^{-2}$ images were only due to distortion and not diffusion. They reported reductions in distortion with increased bandwidth and parallel imaging. However this phantom does not assess distortion due to susceptibility effects, the primary cause of distortion in DW-MR images in the pelvis. Lavdas et al developed a phantom with air and fat inhomogeneities, which could be used to assess geometric distortion due to these inhomogeneities [170]. To the best of the author's knowledge there are no other phantoms for assessing distortion in DW-MR images in the literature.

DW-MR also enables quantitative imaging with ADC maps, which has been suggested to use for dose painting and treatment response monitoring. This means it is essential to regularly assess the accuracy of ADC maps if DW-MR is going to be used quantitatively [171]. A common issue with ADC accuracy phantoms is the need to measure and/or control the temperature accurately since the diffusivity of materials is strongly dependent on temperature [172]. Chenevert et al. reported an ADC phantom consisting of distilled water held at $T = 0$ °C in an ice-water bath was repeatable to within ±5% [173]. The same group extended the phantom with four additional tubes of water at the four axial corners to assess the geometrical variation [172]. They found that there was a significant spatial variation in the superior-inferior direction, with a less marked variation right-to-left, due to gradient non-linearities. Day-to-day repeatability of the central tube was $\leq 3\%$ of literature value for 95% of sites. However, the ADC value of water at $T = 0$ °C is significantly above that found *in vivo*, which limits the relevance of ice-water phantoms [170]. Lavdas et al developed a phantom with three different compartments made from nickel-doped agragose and sucrose gels, imaged at $T = 21$ °C [170]. These gave ADC values representing a range of benign and malignant tissues. Fat and air sections were also added to assess the impact of non-homogeneities. Away from these non-homoegeneties the phantom showed good stability over an 8 week period, with coefficients of variation $< 1\%$. Winfield et al. used a ADC accuracy phantom consisting of a cylinder containing five plastic tubes containing sucrose solutions ranging from $0\% - 20\%$ sucrose in an ice-water bath at a temperature of $T = 0$ °C [169]. They reported high repeatability of repeated ADC measurements across three different MR scanners, with coefficients of variation $\leq 4\%$. The scanners showed good agreement, with all results $\leq 5\%$ to the mean ADC of all measurements. A phantom containing different concentrations of polyethelyne glycol and gadolinium-based contrast agent at $T = 20$ °C was developed [174]. This enabled the T2-relaxivity and ADC values of each solution to be controlled, ensuring SNR as well as ADC value mimicked *in vivo* values. No measurements of repeatability however were performed.

Free-diffusion phantoms as described above are only able to assess ADC values derived using a mono-exponential fit (equation (1.2)). However tumour imaging demonstrates significant variation from mono-exponential behaviour, which has lead to the development of several different fitting techniques (eg intra-voxel incoherent motion [175]). McHugh et al. developed a phantom which can be used to assess these more advanced DW-MR techniques through using cell-mimicking micron-scale hollow spheres grouped into a hollow cylinder immersed in distilled water at $T = 24 \pm 1$ °C [176]. Repeated measurements over a 10-month period had a coefficient of variation of $\leq 5\%$. However, currently this phantom is too small to be used for clinical systems.

PET images are routinely quantitatively analysed using SUVs, and the majority of proposed automatic segmentation algorithms for PET imaging are based on SUVs [177]. Therefore it is essential that SUVs measured by the scanner are accurate and consistent [178]. Most commonly this is performed using uniform flood phantoms [179].

A combined SUV and ADC accuracy phantom has been proposed for PET-MR scanners [180]. This used agarose, sucrose and sodium chloride solutions mixed in water with known activity of $^{18}$F-FDG, placed in 3D-printed hollow tumour models, and then in a uniform phantom filled with a lower amount of background $^{18}$F-FDG activity. Varying the activity and sucrose concentration enables different SUV and ADC values to be assessed. The phantom was demonstrated as a proof of concept for image reconstruction methods, but the same methodology could be applied to a QA phantom.

A crucial potential advantage of simultaneous PET-MR imaging is the high degree of anatomical alignment between the MR and PET images due to being acquired in the same scanning session. However, this depends on the accuracy of the alignment between the PET and MR components. In simultaneous PET-MR scanners this is dependent on the physical locations of the PET and MR gantries, which will not change except after major services. Valladares et al. recommended PET-MR alignment tests occur after mechanical manipulation of the PET-MR gantry and software updates [164]. However the PET-MR alignment tolerance for the GE PET-MR scanner is 5 mm, which is unacceptable for radiotherapy purposes. In addition, to the best of the author's knowledge, the repeatability and stability of PET-MR alignment measurements have not been assessed. This, along with a tighter tolerance, would be important to assess in order for PET-MR to be used for radiotherapy purposes.

Finally, assessment of the electromechanical performance of the scanner couch and external lasers is also important for radiotherapy imaging to ensure accurate and reproducible imaging [181]. Standard tests have been developed for CT imaging [182] and are directly applicable to PET-MR imaging for radiotherapy [154].

In summary images used for radiotherapy have different requirements to those used for diagnostic purposes, which need to be considered when developing QA programmes. Radiotherapy MR QA is being actively researched, with phantoms and methods developed for assessing anatomic image quality and geometric distortion. Phantoms and methods for assessing DW-MR ADC accuracy have also been reported in the literature. PET methods for assessing SUV accuracy are well established and even a combined ADC-SUV accuracy phantom has been reported. However to the best of the author's knowledge QA methods for DW-MR geometric distortion, PET geometric accuracy and PET-MR alignment have not been reported in the literature. There is a need for the development of a PET-MR radiotherapy programme that includes anatomic MR image quality and large field of view distortion, DW-MR ADC accuracy and geometric distortion, PET SUV accuracy and geometric accuracy and PET-MR alignment.

## 1.5 Tracking Organ Motion in the Pelvis

Radiotherapy treatment planning requires the images used for treatment planning are representative of the patient for every treatment fraction. Flat couch tops, external lasers and patient-specific immobilisation devices are used to ensure the external patient position is consistent. However for pelvic radiotherapy it is also vital for the internal anatomy position to be as similar as possible. This requires both high concordance between different planning images and between planning images and the patient position on treatment [183]. CT is a rapid imaging modality with a limited range of image contrasts available and each image requires delivering an ionising radiation dose. Therefore typically a single planning CT image is acquired, which results in a very similar length of time from patient setup to end of CT acquisition as to the equivalent time on treatment. MR on the other hand is a much slower imaging modality and can provide a number of different image contrasts, including both anatomic and functional information, without any ionising radiation. Therefore, it is possible to acquire multiple different MR images in a single session. However, to utilise the full capabilities of MR for pelvic radiotherapy planning is likely to require image acquisition durations $\sim 5 - 10$ times longer than treatment durations. Therefore it becomes critical to understand the organ motion in the pelvis over these timescales and to develop methods of tracking and compensating for it.

The two organ motions of primary relevance to pelvic radiotherapy are bladder and rectal filling, with bladder filling being the most significant [184]. As the bladder fills it impacts on other nearby organs such as the rectum, small bowel and uterus, changing their shape and position which can impact on the doses these organs receive [185]. In radiotherapy for gynaecological cancers changes in bladder filling produces significant variation in target and OAR volumes on treatment compared to at planning, requiring large PTV margins which cause treatment toxicities [186]. For patients receiving prostate radiotherapy, bladder volumes of a certain size are important for reducing bladder toxicity, as well as being consistent between planning and treatment. In one study those with bladder volumes $< 180$ cm$^3$ had significantly higher rates of acute grade 2+ GU toxicities than patients with bladder volumes greater than this threshold [187].

Several studies have investigated bladder filling and methods of managing it for radiotherapy. Lotz et al. scanned 18 healthy volunteers with 7 sequential T1-weighted MR scans at 10 minute intervals over one hour following a bladder preparation protocol [185]. This consisted of bladder voiding followed by drinking 300 $\mathrm{cm}^3$ of water, with the MR acquisition starting 15 minutes after drinking. 10/18 volunteers also received two further repeat sessions separated by at least a month. There was a wide range in participant bladder filling rates, ranging from 2.1 $\mathrm{cm}^3\,\mathrm{min}^{-1}$ to 15.0 $\mathrm{cm}^3\,\mathrm{min}^{-1}$, with a mean rate of $9 \pm 3$ $\mathrm{cm}^3\,\mathrm{min}^{-1}$ (± standard deviation). Intra-volunteer variation was much smaller with a mean standard deviation of bladder fill rate over the three separate sessions of 0.4 $\mathrm{cm}^3\,\mathrm{min}^{-1}$, although for pre-menopausal women there were larger differences relating to the phase of the female cycle. The bladder filling rate was negatively correlated with age ($\rho = -0.50$,$p = 0.038$).

McBain et al. assessed bladder motion using cine and volumetric MR imaging for 10 bladder cancer patients and 5 healthy controls following bladder voiding having fasted from fluids for the previous hour [188]. Volumetric MR imags were acquired at 0, 14 and 28 minutes post-void. Participants completed two identical imaging sessions. The bladder expanded primarily in the superior and secondarily in the anterior direction, with the largest recorded movement of the bladder wall being 58 mm. Significant inter-patient variation was observed. There was also significant differences between the behaviour of bladder cancer patients with healthy controls, with cancerous bladders expanding more variably. Bladder fillings rates were much lower than reported by Lotz et al., with mean filling rates of $1.5 \pm 1.7$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$ (± standard deviation, range $0.3, 6.1$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$) for patients (session 1) and $0.9 \pm 0.7$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$ (range $0.2, 2.0$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$) for controls. This indicates that prior hydration does have a significant impact on bladder filling rates. Intra-participant variability was similar to Lotz et al, with mean standard deviation of fill rate over the two sessions of $0.6 \pm 0.4$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$ (range $0.1, 1.2$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$) for patients and $0.3 \pm 0.5$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$ (range $0.1, 1.1$ $\%\mathrm{cm}^3\,\mathrm{min}^{-1}$) for controls.

Fransson et al. acquired 4D T2-weighted MR images of the pelvis of 9 healthy male volunteers using a radial blade acquisition [189]. These are 3D images acquired a multiple time points over 10-15 minutes with a temporal resolution of approximately 1 minute. The prostate, bladder and rectum for each time frame were manually contoured and the frames deformable registered to the first frame. Principal component analysis of the first five deformation vector fields was used to determine a motion model, which were then validated against the last four deformation fields. They reported that the deformable registration algorithm accurately accounted for organ motion and using one or two principal component vectors could predict organ motion to within 1 mm. This has the potential to enable motion tracking, however applications of the model to longer time points than a few minutes were not evaluated.

Hynds et al. used an ultrasound probe to measure bladder volumes immediately post-void and immediately prior to planning CT for 30 patients treated with prostate radiotherapy [183]. Patients drank 500 $\text{cm}^3$ of water in the first 15 minutes post-void. Patients followed the same bladder preparation protocol for each of 37 fractions of radiotherapy they received, with their bladder volume measured with an ultrasound probe prior to each one. For the last fraction bladder volume was also measured immediately post-void. There was significant intra-patient variation in bladder volume between fractions, with 53% of fractions varying by more than 100 $\text{cm}^3$ to the planning bladder volume (approximately a change of one third relative to the mean planning volume over all patients of $282 \pm 145$ $\text{cm}^3$), and 36% of fractions varying by more than 150 $\text{cm}^3$, despite following the same preparation protocol. Mean pre-treatment bladder filling rate was $4.6 \pm 2.9$ $\%\text{cm}^3\,\text{min}^{-1}$ (± standard deviation) and post-treatment it was $2.5 \pm 1.8$ $\%\text{cm}^3\,\text{min}^{-1}$. This is also lower than Lotz et al., despite following a similar preparation protocol. Intra-patient variability was ascribed to variations in time of day of scanning, patient's hydration status, concomitant medications and patient compliance with the protocol.

Despite the intra-patient variability, bladder preparation protocols are still the primary method used to conform bladder volume at treatment to that at planning. A systematic review of bladder volume reproducibility for prostate radiotherapy found a wide range of bladder preparation protocols evaluated in clinical trials, with water volumes to be drunk ranging from 300 $\text{cm}^3$ to 1080 $\text{cm}^3$ and drinking until 'comfortably full' [190]. They evaluated mean difference in bladder volume for each treatment fraction relative to volume at planning CT as a measure of intra-patient variability. Mean difference was lowest for $300 - 400$ $\text{cm}^3$ protocols ($-12$ $\text{cm}^3$, 95% confidence interval $[-52, 28]$ $\text{cm}^3$) and increased as the drinking volume increased, with the 'comfortably full' protocol having a mean difference of $-46$ $\text{cm}^3$ $[-79, -13]$ $\text{cm}^3$.

Grün et al. evaluated an additional biofeedback mechanism to improve the consistency of bladder volumes following preparation protocols [187]. Patients emptied their bladder, drank 500 $\text{cm}^3$ of water and waited 30-45 minutes for their planning scan. Immediately after the scan they were asked to void into a measuring container, and the volume of urine was recorded (aiming to be $200 - 300$ $\text{cm}^3$). Patients followed the same protocol for each treatment fraction, and were told to aim to reproduce the same volume they recorded at treatment, constituting a biofeedback mechanism. This did not produce a significant difference in bladder volume consistency, although there was a trend to more consistent volumes.

Rectal filling also impacts on pelvic radiotherapy, with significant variation between planning and treatment [184]. For prostate radiotherapy rectal filling causes a posterior displacement of the prostate and seminal vesicles, with larger volumes ($> 60$ $\text{cm}^3$) causing larger displacements ($> 3$ mm) [191]. However, unlike bladder filling, the timescales of rectal filling are such that differences over the course of a 45 minute MR scan are minimal [188]. A range of methods of ensuring consistency in rectal filling between planning and treatment have been investigated including evacuation techniques, dietary advice, laxatives and enemas, with no strong evidence to recommend any particular method [192]. McNair et al. found rectal filling consistency was not improved when using dietary advice and a bowel movements diary to produce consistent rectal volumes in 22 patients treated with prostate radiotherapy [193]. Yahya et al. evaluated a high-fibre diet leaflet, micro-enemas and no bowel preparation in 30 prostate radiotherapy patients [194]. Variation between fractions were assessed using on-treatment CBCT images and demonstrated significantly smaller rectal volumes and less prostate movement in the micro-enema cohort. However measurements of rectal volume were on a single slice and may not be representative of full rectal volume.

In summary, organ motion in the pelvis depends primarily on bladder and rectal filling. Bladder and bowel preparation protocols are commonly used to ensure consistency in organ size and position between planning and treatment. However evidence to their efficacy is limited and significant inter- and intra-patient variation has been reported. Bladder filling rates vary significantly between individuals, with significant dependence on prior hydration. Bladder filling happens over a timescale of minutes meaning longer planning acquisitions typical for MR and PET-MR imaging (30-45 minutes) will result in substantial differences in bladder volume during the course of the acquisition. Rectal filling happens over longer timescales and so changes should be minimal. Changes in bladder filling impact the position of other organs in the pelvis. Therefore developing methods of reliably tracking and correcting for bladder filling motion would enable longer planning sessions to be used, facilitating maximising the benefit of advanced MR and PET-MR methods for radiotherapy planning.

## 1.6 Aims and Outline of Thesis

This thesis aimed to overcome the scientific and technical barriers to using MR-only and PET-MR for radiotherapy planning of pelvic cancers (see figure 1.4). It has focused on three areas: MR-only radiotherapy (chapters 2 and 3), PET-MR for radiotherapy planning (chapters 4-7) and organ motion tracking (chapter 8).

Chapter 2 describes the development and evaluation of a novel Deep Learning sCT algorithm based on a novel ZTE sequence and Deep Learning automatic contouring algorithm based on conventional T2-weighted MR. The sequences and algorithms were developed by our commercial collaborators in the Deep MR-Only RT project, GE Healthcare. I scanned the 57 patients in the study cohort which were used in training and evaluating the Deep Learning model, provided dose evaluations and feedback on preliminary sCT and automatic contouring models and then carried out a comprehensive dose evaluation of the final sCT and automatic contouring version for a range of pelvic radiotherapy sites.

Ongoing QA of sCT dose calculation accuracy is an important clinical concern. Chapter 3 presents a QA method developed using a different clinical MR-only radiotherapy patient cohort (n=49) utilising a different atlas-based sCT algorithm at the Northern Centre for Cancer Care. The method uses the first-fraction CBCT as an independent reference geometry for assessing sCT dose calculation accuracy and is the first to assess this in clinical practice and derive tolerance levels for clinical use. This work has been published [195].

The next four chapters turn to consider PET-MR for radiotherapy planning. Chapter 4 starts with a focus on the impact on PET and MR image quality from imaging patients in the pelvic radiotherapy position using a flat couch and anterior coil bridge. The broader impact on imaging quality using standard image quality phantoms has not been evaluated in the literature previously. This chapter discusses methods of incorporating the radiotherapy hardware in attenuation correction maps and their impact on PET quantitation accuracy. This work has been published [196].

Chapter 5 continues this theme by applying the developed PET attenuation correction methods to PET-MR images of ano-rectal radiotherapy patients. The impact of applying or not applying these methods on GTV delineation and quantification of metabolic parameters was evaluated using a sub-set (n=17) of the Deep MR-Only RT patient cohort described in chapter 2.

Quantitative PET requires accurate attenuation correction of the patient as well as the radiotherapy hardware. Chapter 6 investigates using the radiotherapy sCT described in chapter 2 for PET attenuation correction compared to gold standard CT and to the current clinical MRAC method used on the scanner. This was evaluated using a sub-set (n=10) of the Deep MR-Only RT patient cohort (chapter 2). The impact on automatic GTV delineation and the quantification accuracy of metabolic parameters were evaluated.

Chapter 7 finishes the discussion of PET-MR for radiotherapy planning with a description of the development and year-long evaluation of a PET-MR QA programme for radiotherapy. Six tests were developed to assess: MR image quality, MR geometric accuracy, mechanical accuracy, PET-MR alignment accuracy, DW-MR ADC accuracy and PET SUV accuracy. The same-day repeatability and monthly stability over a year were evaluated for each test. This work has been published [197].

The final area this thesis has covered is tracking organ motion in the pelvis. Chapter 8 describes developing a MR sequence and evaluation algorithm developed by myself to do this. This method was evaluated in 9 healthy volunteers recruited and scanned by myself, utilising the automatic contouring algorithm evaluated in chapter 2.

The last chapter discusses the conclusions that can be drawn from the thesis as a whole and the next steps to be taken in developing the methods for MR-only and PET-MR radiotherapy to be applied in the clinic.

# Chapter 2

# Comprehensive Dose Evaluation of Pelvic Deep Learning sCT and Automatic Contouring for MR-only Radiotherapy

## 2.1 Introduction

MR-only radiotherapy enables the superior soft-tissue contrast of MR to be used for delineation without the uncertainty of a MR-CT registration, improving the geometric accuracy of treatments and potentially reducing patient side-effects [81]. MR cannot be used directly for radiotherapy dose calculations and so a method of generating a sCT from the MR needs to be developed [28]. A number of different methods have been proposed for pelvic radiotherapy including bulk-density and atlas-based methods (see chapter 1 section 1.2.4). Several commercial solutions have been developed based on these approaches and evaluated on patients with prostate cancer with very small mean dose differences to CT ($\leq 0.5\%$) [53, 67, 198]. More recently a number of different Deep Learning methods have been reported in the literature, although only a few have been evaluated for dose calculation accuracy in the pelvis [78–80]. Dose calculation accuracy is the only clinically relevant parameter [78].

Most Deep Learning sCT algorithms have used conventional diagnostic MR images (T1-weighted, T2-weighted or Dixon sequences) as the input image. Unlike conventional MR images, ZTE images generate signal from cortical bone, which potentially facilitates improved bone generation in the sCT. Leynes et al. evaluated a Deep Learning sCT algorithm using a combination of ZTE and Dixon images for PET-MR attenuation correction [77]. To the best of my knowledge, sCTs generated from Deep Learning algorithms using ZTE images as the single input have not been dosimetrically evaluated for pelvic

MR-only radiotherapy.

Most dose evaluations of algorithms in the pelvis have focused on using clinical prostate treatment plans. This has limited relevance to other treatment sites where PTVs may extend significantly superiorly and/or inferiorly of prostate PTVs. The dose calculation accuracy of the sCT is not robustly assessed in these areas because it is in the low dose region of the prostate treatment plan. A challenge for evaluating sCTs dosimetrically for pelvic radiotherapy sites apart from the prostate is that there can be significant variability between patients in the size and location of the PTV(s). This requires large patient evaluation cohorts to ensure the sCT is accurate for all potential patients, which can be difficult especially in rarer cancers. An alternative approach is to use a small 'dummy' PTV at multiple points in the inferior-superior direction to assess dose differences. This enables a smaller number of patients to be used to evaluate sCTs for all pelvic radiotherapy sites. This also ensures that doses are not averaged over large volume PTVs, potentially masking small regions within the PTV with large dose differences. This approach has not been reported in the literature previously, to the best of my knowledge.

MR-only radiotherapy also enables the MR soft-tissue contrast to be utilised for automatic contouring methods of OARs, which are typically contoured on CT for MR-CT fusion treatment pathways [83]. A very large number of OAR automatic contouring algorithms have been reported in the literature, with Deep Learning methods typically producing the highest accuracy results [199]. However, automatic contours are often evaluated using delineation metrics which have little correlation with clinically relevant dose differences [91] and there is no consensus in the literature on the best method [7]. Evaluating the dose difference from automatic OAR contours is much more clinically relevant, but only a few studies have investigated this [97–100], with only two evaluating OARs for prostate radiotherapy, and none of them for ano-rectal radiotherapy. Therefore, there is a need for a dosimetric evaluation of automatic OAR contours for pelvic radiotherapy treatments.

This chapter presents an evaluation of a novel sCT algorithm based on novel ZTE MR sequence combined with Deep Learning and a Deep Learning automatic contouring algorithm, both developed by GE Healthcare as part of the DeepMR-Only RT consortium which also included clinical partners Erasmus Medical Center, The Netherlands, Szeged University, Hungary and Newcastle University and Newcastle upon Tyne Hospitals NHS Foundation Trust, UK. The aim of this study was to comprehensively evaluate the dose accuracy of a Deep Learning sCT algorithm for all pelvic radiotherapy sites and to evaluate the dose impact of a Deep Learning MR-based automatic OAR contouring algorithm for anal, prostate and rectal radiotherapy.

## 2.2 Evaluation of sCT Dose Calculation Accuracy

### 2.2.1 Materials and Methods

#### Patient Data Collection

56 patients enrolled in the Deep MR-only RT study (research ethics committee reference 20/LO/0583) who were planned for radical/neoadjuvant (chemo)radiotherapy for anal, rectal and prostate cancers were included in this study. Exclusion criteria included contraindicated for MR scanning, medical implants in the pelvic area (eg hip prostheses) and external contour greater than the scanner field of view. Patients were divided into training and evaluation cohorts, with patient characteristics for each detailed in table 2.1.

Table 2.1: Study patient characteristics in the training and evaluation cohorts. Disease staging according to [200]. All patients were M0.

| Characteristic | Training Cohort | Evaluation Cohort |
|---|---|---|
| Total patients | 36 | 20 |
| Female/Male split | 7/29 | 6/14 |
| Median age (range) | 65 (48,82) years | 66 (49,79) years |
| Patients with anal cancer (stages) | 6 (T1N0-T4N3) | 6 (T1/2N0M0-T2N1) |
| Patients with prostate cancer (stages) | 23 (T2N0-T4N0) | 10 (T2N0-T4N0) |
| Patients with rectal cancer (stages) | 7 (T2N1-T4N2 | 4 (T2N1-T3b/4N0) |
| Patients with hydrogel rectal spacers | 5 | 0 |

All patients received an MR scan on a SIGNA PET/MR 3T scanner (version MP26 GE Healthcare, Waukesha, USA) after their radiotherapy planning CT scan and before their first treatment fraction. Patients were scanned in the radiotherapy treatment position on a flat couch-top with a coil bridge for the anterior MR coil (see figure 2.1). Patients were positioned to match their radiotherapy planning CT scan using a combined customisable foot and knee rest (Civco Medical Solutions, Coralville, Iowa, USA) and external lasers matched to patient tattoos. Immediately prior to entering the scan room patients emptied their bladder and drank 400 ml of water.

MR images were acquired using a novel 3D radial ZTE sequence [33] and a T2-weighted 3D fast spin echo sequence. MR imaging parameters are given in table 2.2. Both sequences used the vendor supplied 3D distortion correction and the maximum receive bandwidth to minimise geometric distortions. During the time period that patient images were acquired a monthly QA programme was carried out which included measurements of geometric distortion (see chapter 7). The ZTE images were reconstructed using a Deep Learning denoising algorithm [201] with a reconstructed field of view twice that of the acquired field of view.

All patients received a planning CT scan (Sensation Open, Siemens, Erlangen, Germany) in the same model of customisable foot and knee rest (Civco Medical solutions) with a voxel size of $1.1 \times 1.1 \times 3$ mm$^3$ and a tube voltage of $V = 120$ kVp. Patients being treated

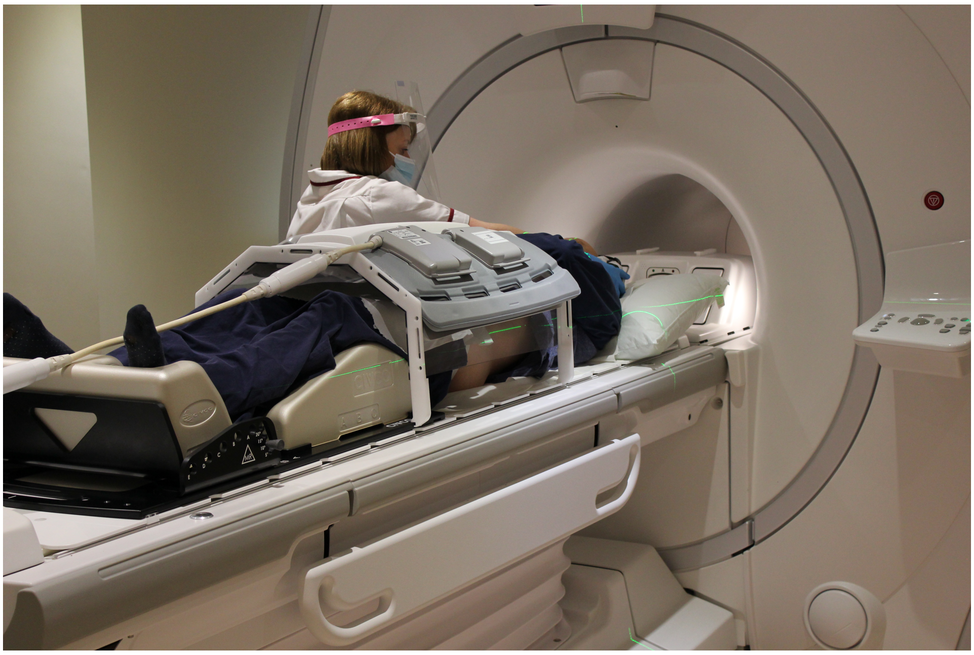


Figure 2.1: Example of patient setup.

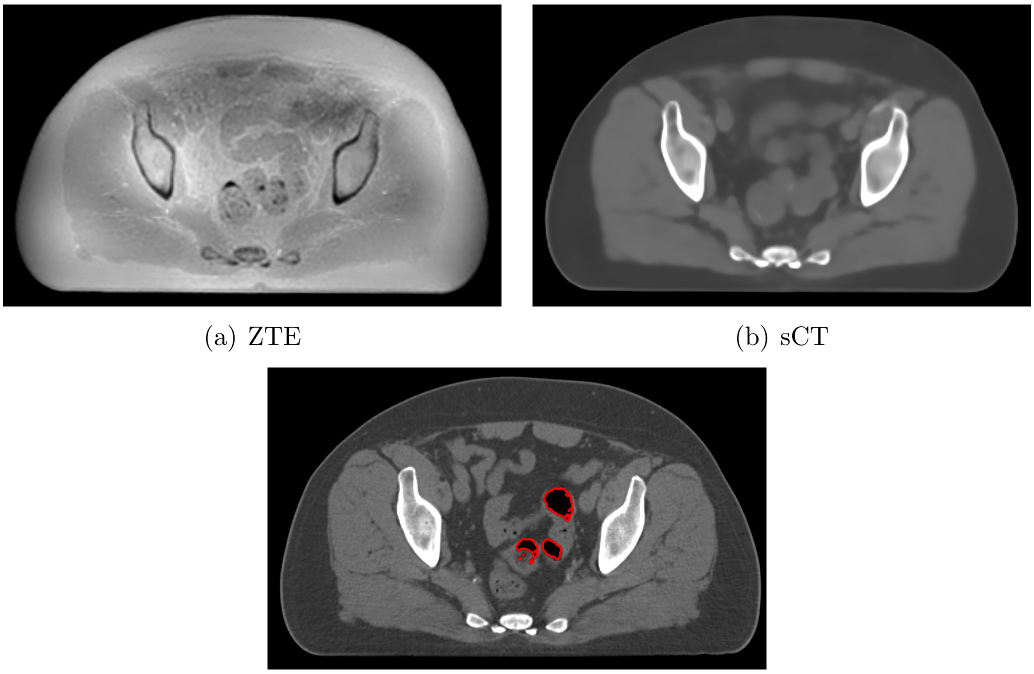
Table 2.2: MR imaging parameters for the ZTE and T2-weighted fast spin echo sequences.

| **Parameter** | **3D ZTE** | **3D FSE** |
|---|---|---|
| Voxel size | $2.0 \times 2.0 \times 2.0$ mm$^3$ | $1.0 \times 1.0 \times 2.0$ mm$^3$ |
| Field of view | $360 \times 360 \times 300$ mm$^3$ | $380 \times 304 \times 360$ mm$^3$ |
| Echo time | 0.02 ms | 148 ms |
| Repetition time | 543 ms | 2000 ms |
| Receive bandwidth | 694 Hz pixel$^{-1}$ | 658 Hz pixel$^{-1}$ |
| Acquisition time | 65 s | 372 s |

for anal and rectal cancers received a contrast-enhanced CT scan. Patients were imaged following routine bladder preparation consisting of an empty bladder 30 minutes prior to the scan, followed by drinking 400 ml of water, and bowel preparation consisting of the application of a micro-enema 60 minutes prior to the scan followed by bowel emptying. CT images were acquired within a median time of 6 days from the PET-MR acquisition (range 1-15 days) for the evaluation cohort.

### Synthetic CT Creation

The sCT was generated from the ZTE image using a 2D CNN U-NET model with a bone focused loss function, as described in [202]. The model had three tasks: whole image regression, bone segmentation and image value regression within the bone region. The logic is to separate segmentation and regression tasks and optimise the model to simultaneously reduce errors in both, each task implicitly reinforcing the other [203]. Each task was driven by an individualised loss function and generated an associated output image. The three output images were then combined using the voxel intensity values from the whole image regression output, except for voxels within the bone segmentation output, which were assigned values from the bone regression output. The model was trained with patients from the training cohort, with ZTE and CT images registered together using an affine transformation. The ZTE images were normalised with bias correction to ensure consistency across the dataset. Representative reconstructed ZTE and corresponding sCT and CT images are shown in figure 2.2.



(a) ZTE

(b) sCT

(c) CT

Figure 2.2: Example ZTE image (a) and corresponding sCT image (b) for a patient in the evaluation cohort. CT image for the same patient is shown in (c) with the air contour shown in red.

### Determining the Minimum Longitudinal CT Extent for Accurate Dose Calculations

Scattered radiation means dose deposition can occur outside of the beam penumbra. Therefore accurate radiotherapy dose calculations require the patient CT image to extend beyond the PTV, typically considered as a minimum of 3 cm. For some of the patients treated for prostate cancer, the longitudinal CT extent was substantially shorter than the sCT, potentially creating a confounding effect in the dose accuracy comparison. A potential solution in this case is to copy the most inferior CT slice enough times to provide sufficient longitudinal coverage. To determine what the minimum PTV-CT distance required was and the effectiveness of copying the last slice, a cylindrical dummy PTV with a diameter of 10 cm and a length of 5 cm was created centred on a midpoint between the femoral heads for one patient. A 6 MV single arc Volumetric Modulated Arc Therapy (VMAT) plan was optimised to deliver 50 Gy in 25 fractions to this PTV with a general dose fall-off function. This plan was recalculated on the CT with inferior slices removed and the doses compared to the plan calculated on the unchanged CT. The original reference CT had a PTV-to-end-of-CT extent of 9.4 cm. Shortened CTs were created with extents of [4.8,3.9,3.0,2.1,1.2,0.6,0.3,0.0] cm. The plan was then further recalculated on some of the CTs with slices removed, with the last slice copied ten times to add a further 3 cm of longitudinal CT extent beyond the PTV edge. These were done for the CTs with extents of [2.1,1.2,0.6,0.3,0.0] cm. Plans were compared to the unchanged reference CT on dose-volume histogram parameters to the PTV and in dose differences in a line profile extending ±5 cm in the superior-inferior direction.

### Comprehensive sCT Dose Evaluation

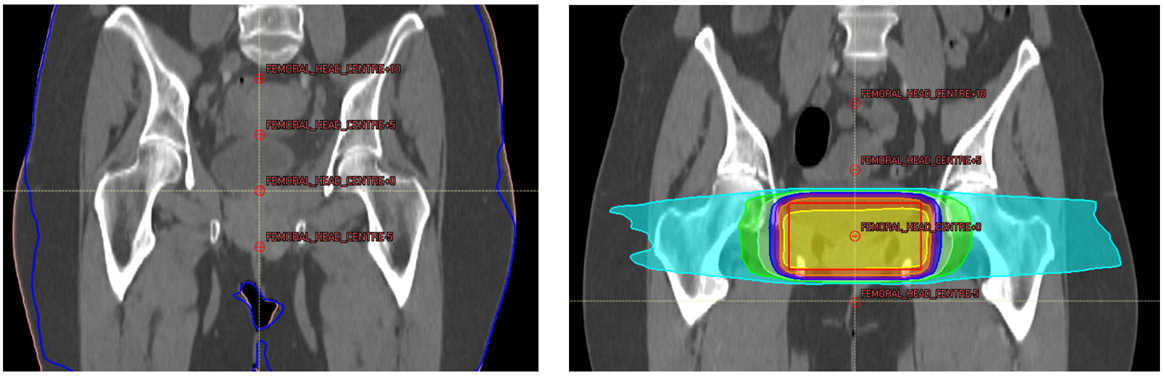
This aimed to evaluate the sCT at different longitudinal points using all patients, providing dose accuracy measurements that would be relevant to all pelvic radiotherapy treatments. The patient CT was rigidly registered to the sCT using the automatic mutual information algorithm within RayStation (v9, RaySearch Laboratories, Stockholm, Sweden). The sCT was calibrated using a Hounsfield Unit-mass density curve measured on the CT scanner used for the training CT cases. The CT was calibrated using data measured on the clinical CT scanner. Any air in the rectum in the CT or sCT was automatically contoured using a threshold method. Voxels falling between -1000 HU and -300 HU within the external body contour were thresholded, and the resulting structure set to water density. Patients who had received a contrast-enhanced CT had the contrast delineated and set to a mass density of 1.015 g $\text{cm}^{-3}$, the density of artery tissue. The contrast was contoured using a semi-automatic method within RayStation, using the 'Bone ROI' tool. This uses a thresholding and connected regions approach, with two thresholds: a definitely bone threshold where HU above this value are always included, and a definitely not bone threshold where HU below this value are always excluded. Voxels with intensity between these two thresholds are either included or excluded depending on whether they are connected to an included region. To delineate contrast, firstly a ROI was created using the 'Bone ROI' tool with the definitely bone threshold $> 150$ HU and the definitely not bone threshold $< 50$ HU. This included the contrast voxels, but also all the bone voxels. A second ROI was created with the 'Bone ROI' tool with thresholds $> 250$ HU and $< 150$ HU, which excluded the contrast voxels but included most of the bone. This second ROI was expanded by a uniform 1 cm margin and subtracted from the first ROI to leave just the contrast contoured. This was then visually inspected and manually modified to ensure accuracy.

The true external contour was automatically determined for both sCT and CT using a threshold of $-250$ HU. In addition, the intersection volume of the two true external contours was also calculated (the intersection external contour). This enabled dose differences to be compared without the confounding effect of small differences in external contour arising from the images being acquired in two imaging sessions. One patient had the most inferior CT slice copied 10 times, lengthening the CT by 3 cm to ensure sufficient lateral scatter for accurate dose calculation on the CT (see subsection 'Minimum Longitudinal

CT Extent' above). The inferior CT extent was not sufficient for one patient and the superior CT extent for 8 patients. A $2 \times 2 \times 2$ mm$^3$ dose grid was used which covered the external contours of both CT and sCT, with the dose grid voxel size and the dose grid position kept the same on both image sets.

Four separate isocentres were positioned on the CT, longitudinally separated by 5 cm. The first isocentre was positioned using manual contours of the femoral heads on the CT. The middle of each contour (left and right) was determined and the midpoint between them calculated. This isocentre was labelled FH. A further three points were determined with the same left-right and anterior-posterior position and 5 cm inferior (FH-5), 5 cm (FH+5) and 10 cm (FH+10) superior (see figure 2.3).

The 6 MV radiological water equivalent isocentre depth was calculated in the axial plane at $5^o$ angles for both sCT and CT at each of the four isocentres using $1 \times 1$ cm$^2$ beams (see section 1.2.3 for more background information on radiological isocentre depth analysis) [55, 56]. For each isocentre two sets of measurements were made, one using the true external contour for each image and one using the intersection external contour. The radiological isocentre depth using the intersection external contour gave a measure of the accuracy of the HU assignment in the sCT and the physical isocentre depth using the true external contour a measure of the differences in external contour due to differences in patient setup or residual geometric distortion at the patient periphery. The radiological isocentre depth using the true external contour gave an overall measure incorporating both HU assignment and geometric accuracy (the physical isocentre depth using the intersection contour would by definition have a difference of zero). The radiological isocentre depth was calculated on both CT and sCT using the density over-rides described above. The difference in radiological depth (sCT - CT) at each gantry angle was determined for each external contour method. The difference in physical isocentre depth was also determined at each gantry angle with both external contour methods.



(a) Isocentre Points

(b) Dummy Plan

Figure 2.3: Example CT showing the four isocentre points marking the FH-5, FH, FH+5 and FH+10 levels (a) and the cylindrical PTV (red line) and dummy plan dose at the FH level (b). The true external contour (pink line) and intersection contour (blue line) are shown on image a).

A cylindrical dummy PTV 5 cm long with a diameter of 10 cm was drawn on the CT centred on the isocentre in the femoral head plane (see figure 2.3). The 5 cm length was chosen because the FH isocentre points were spaced 5 cm apart so PTVs 5 cm long ensured continuous coverage in the superior-inferior direction. The diameter of 10 cm was selected as approximately representative of pelvic PTVs. A 6 MV single $360^o$ arc VMAT plan was optimised to deliver 50 Gy in 25 fractions to this PTV, using a general dose fall-off function to ensure conformality of the high dose region to the PTV. Dose was calculated using the RayStation collapsed cone algorithm (version 5.2), which calculates dose-to-water. This plan was recalculated on the sCT keeping the dose grid the same using the true external contour. The difference in dose to the PTV median dose (D50%), near-minimum (D98%) and near-maximum (D2%) were calculated as a percentage of the prescription dose (see section 1.2.3) [10]. The PTV and treatment plan were then reassigned to each of the other three isocentres in turn, with the plan recalculated (but not re-optimised) for each isocentre on CT and sCT and the doses compared using the same methodology. This analysis was then repeated within the intersection external contour. For one patient the cylinder PTV was outside the true external contour for the FH-5 point. For this patient, the PTV was cropped to the intersection external contour to ensure it was the same volume for CT and sCT.

### Site-Specific sCT Dose Evaluation

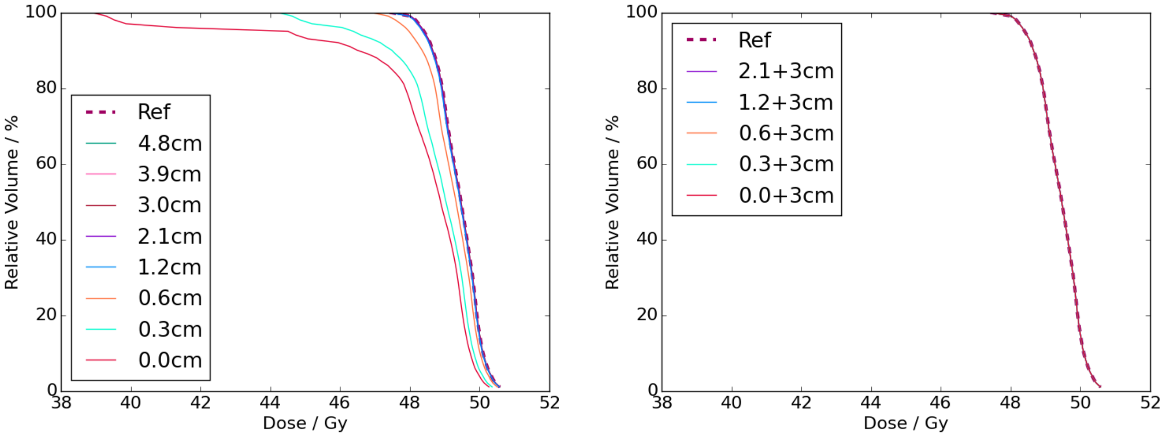
An additional dose evaluation was carried out using the clinical treatment plan to ensure clinical relevance. Unlike the FH point analysis though, the clinical plan analysis would only apply for the prostate, anal and rectal cancers contained within the evaluation cohort. So this analysis was more clinically relevant but less generalisable. The same CT registration, air and contrast density overwrites and additional CT slices (if required) were used as in the comprehensive dose evaluation. Only the true external contour was used for dose calculations, which may result in confounding patient setup differences but also ensures geometric accuracy of sCT is assessed and the results are comparable to the literature. The PTV(s) and relevant clinical OARs were copied from the CT to the sCT and the clinical treatment plan recalculated on the sCT keeping the monitor units and dose grid the same. The difference in dose to the PTV(s) DVH points D2, D50 and D98 and OAR DVH points were calculated. A gamma analysis (see section 1.2.3) was performed comparing the dose calculated on the sCT to the CT for the clinical treatment plan using the Medical Interactive Creative Environment Toolkit (version 2021.1.2, Umeå University, Sweden) [204]. A 3D global gamma analysis was carried out within the following structures: the external contour, the union of all patient PTVs (primary, nodal and elective if present) and the volume enclosed by the 50% isodose line of the prescription dose. The union of patient PTVs ensured all the high dose regions were included within the PTV evaluation. The gamma analysis criteria used were a dose difference of 1% of prescription dose and distance to agreement 1 mm and dose difference 2% and distance to agreement 2 mm. All points below 10% of the prescription dose were excluded from the analysis.

To enable comparisons of the site-specific dose differences to the FH point dose differences form the comprehensive dose evaluation, the superior-inferior length of the union of all patient PTVs was determined relative to the FH points for each patient.

## 2.2.2 Results

### Minimum Longitudinal CT Extent

DVH dose differences only occurred for CT extents $\leq$ 0.6 cm (see figure 2.4) without the last slice copied. Changes in the most sensitive parameter, the D99%, were still only $-0.6\%$ even at 0.6 cm extent, although this rapidly increased for shorter distances. Copying the last slice ten times resulted in zero differences in dose volume histograms, even for the CT with 0.0 cm extent (figure 2.4).



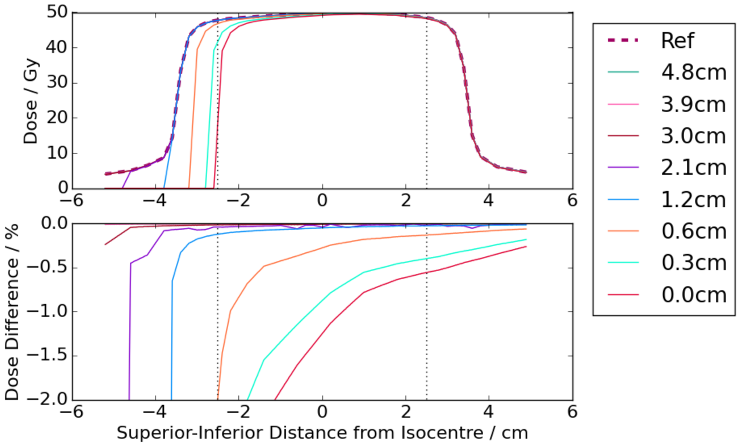
(a) Without Last Slice Copied

(b) Last Slice Copied

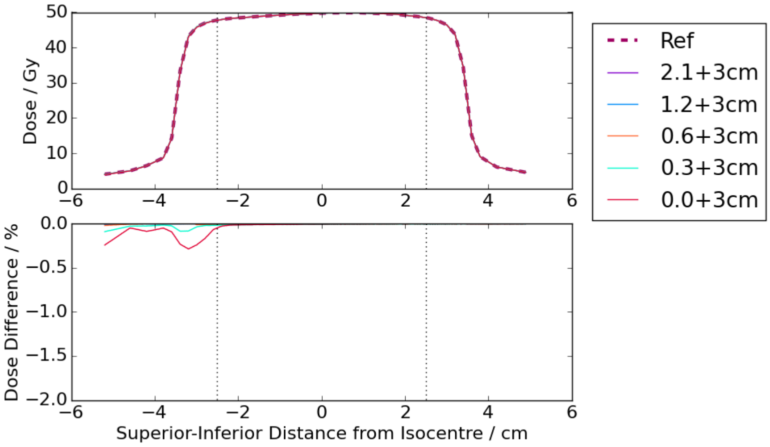
Figure 2.4: Dose volume histograms for CTs with different longitudinal distance extensions beyond the PTV edge (a). Panel (b) shows the equivalent figure but with CT scans where the last slice has been copied ten times.

The line profiles were more sensitive to the CT longitudinal extent, with differences within the PTV starting at the 1.2 cm distance (figure 2.5). These differences were still very small though, $< 0.2\%$. Copying the last slice ten times was very effective at reducing dose differences to zero except for extents $\leq$ 0.3 cm from the PTV. Even at distances $\leq$ 0.3 cm (the slice thickness) the differences were $\leq 0.3\%$ and 0.0% within the PTV.

In conclusion, CT's that extend longitudinally $\geq$ 1.5 cm beyond the PTV edge can be used for dose comparisons to sCT without modification. For CTs with extents $<$ 1.5 cm but $\geq$ 0.3 cm, the last CT slice should be copied and the resulting modified CT used for dose comparisons to sCT. The data suggests that the last slice only needs to be copied enough times to enure $\geq$ 1.5 cm beyond the PTV. However for simplicity a standard 10 times was used when applied in the sCT evaluation. For CTs with extents $<$ 0.3 cm, the most superior/inferior PTV (FH+10/FH-5) should not be used. This approach was



(a) Without Last Slice Copied



(b) With Last Slice Copied

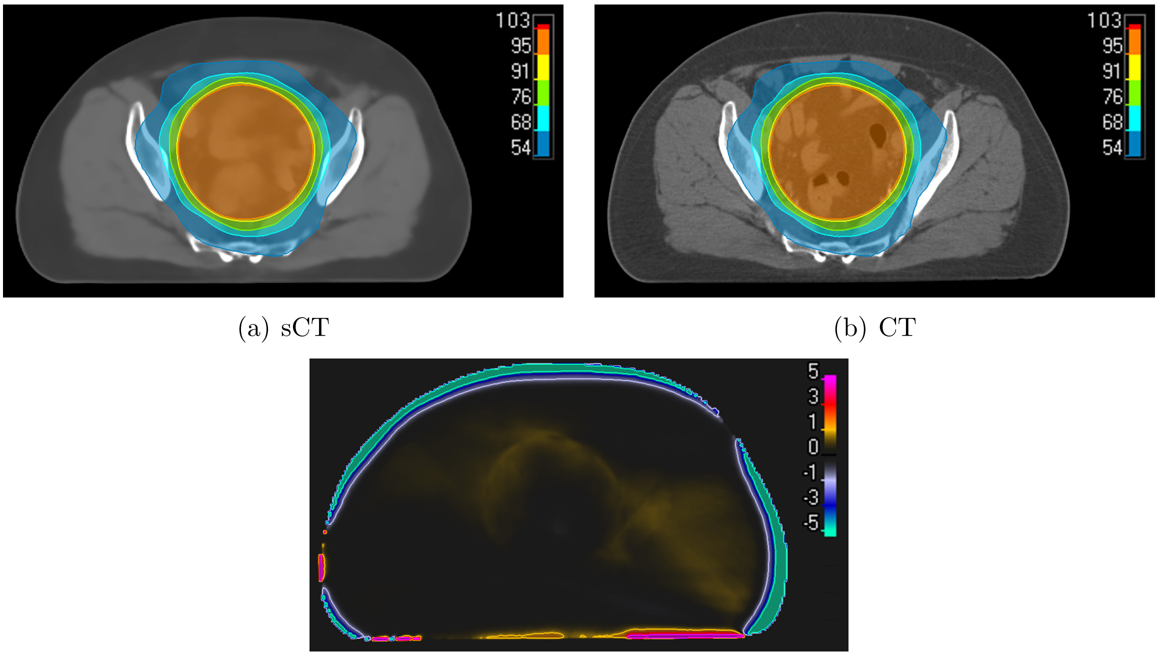
Figure 2.5: Profiles of doses along a line running longitudinally from the FH+5 to FH-5 points calculated on CTs with different longitudinal distances from PTV edge to edge of CT. a) shows results from the CT without modification and b) results with the last CT slice copied ten times. In each sub-figure the top panel shows the dose profile and the bottom panel the dose difference profile to the reference CT.

applied to the sCT evaluations for one patient (see next section). These results only apply to the beam energy evaluated here (6 MV flattened beam) since lateral scatter is beam energy dependent and higher beam energies are likely to require larger CT extents.

### Comprehensive sCT Dose Evaluation

sCTs were successfully generated for all patients. Example dose distributions and dose difference maps can be seen in figure 2.6. The mean radiological isocentre depth differences for the true external contour were $0.2 \pm 0.1$ mm (FH+10), $-0.3 \pm 0.1$ mm (FH+5), $-0.6 \pm 0.1$ mm (FH), and $-2.1 \pm 0.1$ mm (FH-5) (± standard error, see figure 2.7). The corresponding mean physical isocentre depth differences were $-0.5 \pm 0.1$ mm (FH+10), $-1.2 \pm 0.1$ mm (FH+5), $-1.5 \pm 0.1$ mm (FH), and $-2.1 \pm 0.1$ mm (FH-5). There did appear to be larger differences in both radiological and physical isocentre depths at the more inferior points, particular at approximate angles $60^o$, $120^o$, $240^o$ and $300^o$. The mean radiological isocentre depth differences for the intersection external contour were

$0.6 \pm 0.1$ mm (FH+10), $0.6 \pm 0.1$ mm (FH+5), $0.6 \pm 0.1$ mm (FH), and $-0.3 \pm 0.1$ mm (FH-5). There was no trend apparent in differences within the intersection contour with either angle or FH isocentre point.



(a) sCT

(b) CT

(c) Dose Difference Map

Figure 2.6: Example dose distribution on sCT (a), CT (b) and dose difference map (c) for the FH+5 point with the true external contour. Isodoses and dose differences shown as percentages of prescription dose (50 Gy). Patient was selected as closest D50% dose difference to mean value for the FH+5 point.

The mean dose differences within the true external contour at the different FH isocentre points were small, $\leq 1.1\%$ and even smaller for the intersection contour $\leq 0.3\%$ (see table 2.3). The distribution of dose differences were larger for the more inferior FH points within the true external contour, but similar across all points within the intersection contour (see figure 2.8).

### Site-Specific sCT Dose Evaluation

The superior-inferior length of the combined primary, nodal and elective PTVs for each patient is shown in figure 2.9. For prostate-only patients (n=9) only the dose differences at the FH-5 and FH points were relevant, as the PTV does not extend superiorly to the FH+5 point. For the prostate with nodal treatment (n=1), rectum and anus patients all four points were relevant to the dose evaluation, although only two patients extended beyond FH+10. Two anus patient's PTVs extended more than 4 cm inferiorly of the FH-5 point, suggesting an evaluation at the FH-10 point would have also been relevant for them.

Mean dose differences at clinically relevant DVH points for the PTV(s) and clinically relevant OARs are shown in table 2.4. Example gamma maps for a representative anus patient are shown in figure 2.10. Gamma pass rates at 2%/2 mm within the external

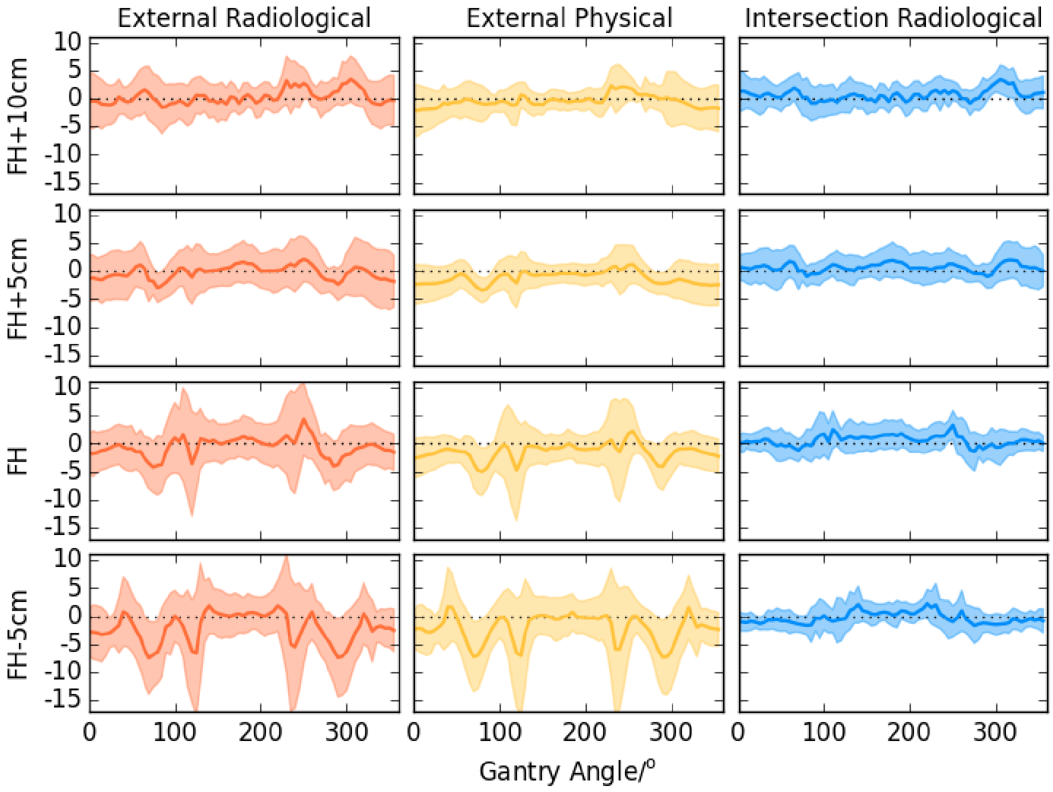


Figure 2.7: The mean isocentre depth differences (sCT - CT) using radiological depth (orange) and physical depth (yellow) within the true external contour and radiological depth (blue) for the intersection external contour. Differences are shown as a function of gantry angle, 0° indicating beams incident on the patient anterior and 90° on the patient left (all patients treated in head-first supine position). Dotted horizontal line indicates zero difference. The shaded area shows ± one standard error on the mean.

Table 2.3: Cylindrical PTV dose differences at the DVH constraints for the FH point plans. The number of patients included for the analysis at each FH point is indicated. Dose differences within the true external contour and intersection external contour reported as mean ± standard error (minimum, maximum).

| **Point** | **Patients** | **Constraint** | **Mean Dose Difference / %** | |
|---|---|---|---|---|
| | | | *True External* | *Intersection External* |
| FH+10 | 12 | D2% | −0.2 ± 0.2 (−1.2,0.7) | −0.3 ± 0.1 (−0.9,0.1) |
| | | D50% | −0.1 ± 0.1 (−1.0,0.4) | −0.3 ± 0.1 (−0.7,0.1) |
| | | D98% | 0.0 ± 0.1 (−0.7,0.7) | −0.2 ± 0.1 (−0.7,0.3) |
| FH+5 | 20 | D2% | 0.3 ± 0.2 (−1.1,2.6 | 0.0 ± 0.1 (−0.8,0.9) |
| | | D50% | 0.2 ± 0.2 (−1.1,2.3) | −0.1 ± 0.1 (−0.9,0.7) |
| | | D98% | 0.2 ± 0.2 (−1.3,2.5) | −0.1 ± 0.1 (−0.8,0.8) |
| FH | 20 | D2% | 0.6 ± 0.2 (−1.5,3.6) | 0.0 ± 0.1 (−1.1,0.6) |
| | | D50% | 0.4 ± 0.2 (−1.6,2.1) | −0.1 ± 0.1 (−1.2,0.6) |
| | | D98% | 0.2 ± 0.1 (−1.4,1.4) | −0.1 ± 0.1 (−0.9,0.5) |
| FH-5 | 19 | D2% | 1.1 ± 0.3 (−1.0,3.0) | 0.2 ± 0.1 (−0.6,0.7) |
| | | D50% | 0.8 ± 0.2 (−1.2,3.1) | 0.1 ± 0.1 (−0.7,0.6) |
| | | D98% | 0.5 ± 0.2 (−1.4,2.2) | 0.0 ± 0.1 (−0.7,0.5) |

contour were highest for the prostate patients, 98.9 ± 0.3% (96.9%,99.8%), followed by rectum, 98.4 ± 0.5% (97.3%,99.2%), and anus patients 97.8 ± 0.6% (96.0%,99.4%, see figure 2.11). Gamma pass rates at 1%/1 mm were lower: 97.7 ± 0.5% (94.4%,99.5%),

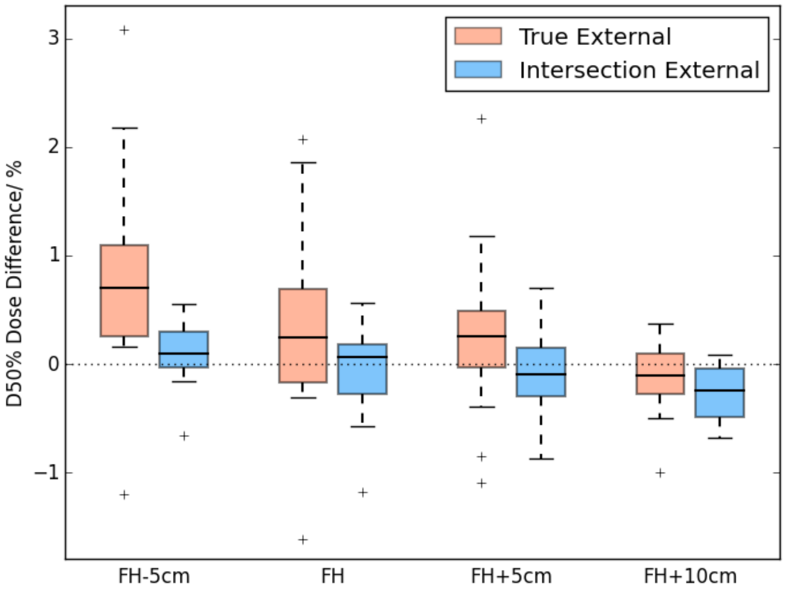


Figure 2.8: Boxplots of D50% dose differences within the true external contour (orange) and intersection external contour (blue) for each FH point.The dotted line indicates zero dose difference. The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points.

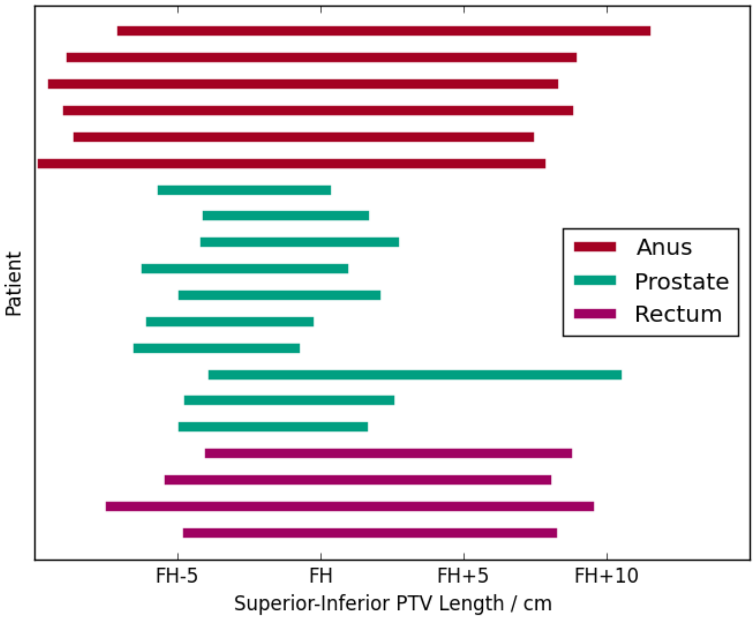
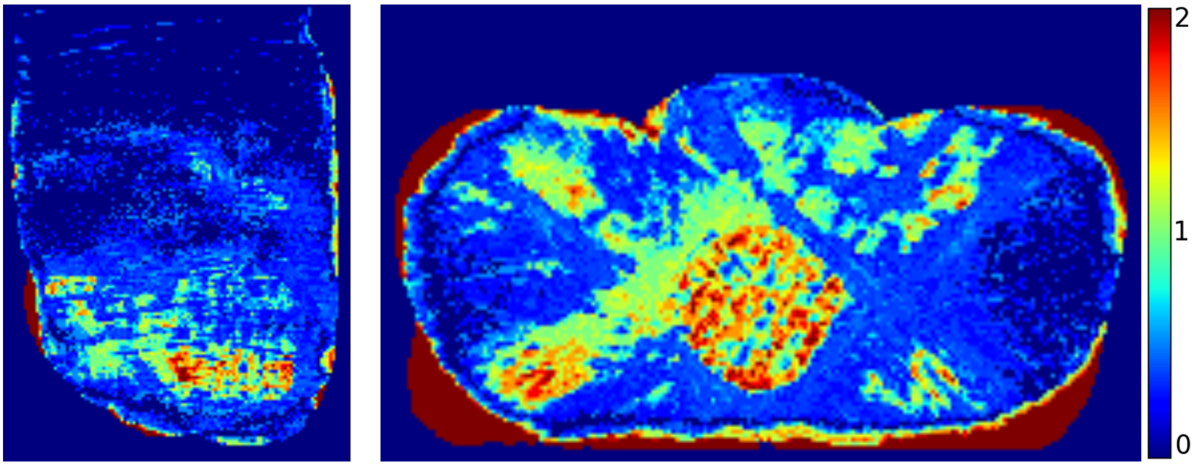


Figure 2.9: Plots of clinical PTV superior-inferior extents relative to the FH points used in the comprehensive evaluation. Each patient is shown as a separate line with treatment site shown with colours: anus (red), prostate (green) and rectum (purple). For patients with multiple PTVs with different dose levels the line indicates the combined superior-inferior extent.

$94.8 \pm 3.1\%$ (88.0%,100.0%) and $94.7 \pm 1.0\%$ (91.6%,98.2%) for the prostate, rectum and anus patients respectively.

Table 2.4: Dose differences at DVH constraints for the clinical plans for different regions of interest. The number of patients varies for the different PTVs for the ano-rectal patients as not all patients had elective or nodal volumes. All results given as mean ± standard error (minimum, maximum).

| Site | Region of Interest | Patients | Constraint | Dose Difference/% |
|---|---|---|---|---|
| Anus | PTV Primary | 6 | D2% | $1.0 \pm 0.5$ $(-0.7,2.6)$ |
| | | | D50% | $0.6 \pm 0.4$ $(-1.1,1.8)$ |
| | | | D98% | $0.2 \pm 0.3$ $(-1.2,1.0)$ |
| | PTV Nodal | 4 | D2% | $-0.1 \pm 0.4$ $(-0.7,0.7)$ |
| | | | D50% | $-0.2 \pm 0.3$ $(-0.8,0.7)$ |
| | | | D98% | $-0.1 \pm 0.3$ $(-0.7,0.6)$ |
| | PTV Elective | 6 | D2% | $0.2 \pm 0.2$ $(-0.8,0.7)$ |
| | | | D50% | $0.2 \pm 0.2$ $(-0.4,0.6)$ |
| | | | D98% | $-0.0 \pm 0.1$ $(-0.5,0.4)$ |
| | Bladder | 6 | D50% | $0.3 \pm 0.4$ $(-1.0,1.5)$ |
| | Small Bowel | 6 | D30% | $-0.1 \pm 0.2$ $(-0.6,0.4)$ |
| Prostate | PTV | 10 | D2% | $0.7 \pm 0.2$ $(-0.1,2.5)$ |
| | | | D50% | $0.7 \pm 0.2$ $(-0.0,1.9)$ |
| | | | D98% | $0.6 \pm 0.1$ $(0.0,1.2)$ |
| | Bladder | 10 | D50% | $0.0 \pm 0.1$ $(-0.3,0.2)$ |
| | Rectum | 10 | D30% | $0.1 \pm 0.1$ $(-0.3,1.1)$ |
| Rectum | PTV Primary | 4 | D2% | $0.5 \pm 0.2$ $(0.1,0.9)$ |
| | | | D50% | $0.3 \pm 0.1$ $(0.0,0.6)$ |
| | | | D98% | $0.3 \pm 0.1$ $(0.0,0.6)$ |
| | PTV Elective | 2 | D2% | $0.4 \pm 0.3$ $(0.1,0.7)$ |
| | | | D50% | $0.3 \pm 0.4$ $(-0.1,0.7)$ |
| | | | D98% | $0.1 \pm 0.2$ $(-0.1,0.3)$ |
| | Bladder | 4 | D50% | $0.3 \pm 0.2$ $(-0.1,0.6)$ |
| | Bowel Bag | 4 | D30% | $-0.1 \pm 0.1$ $(-0.2,0.1)$ |



(a) Sagittal

(b) Axial

Figure 2.10: Example sagittal (a) and axial (b) gamma pass maps for an anus patient with criteria 1%/1 mm. Patient was selected as the closest to the mean pass rate in the external contour.

### 2.2.3 Discussion

This study aimed to evaluate the dose calculation accuracy of a Deep Learning sCT algorithm for pelvic MR-only radiotherapy, both comprehensively for all pelvic sites and

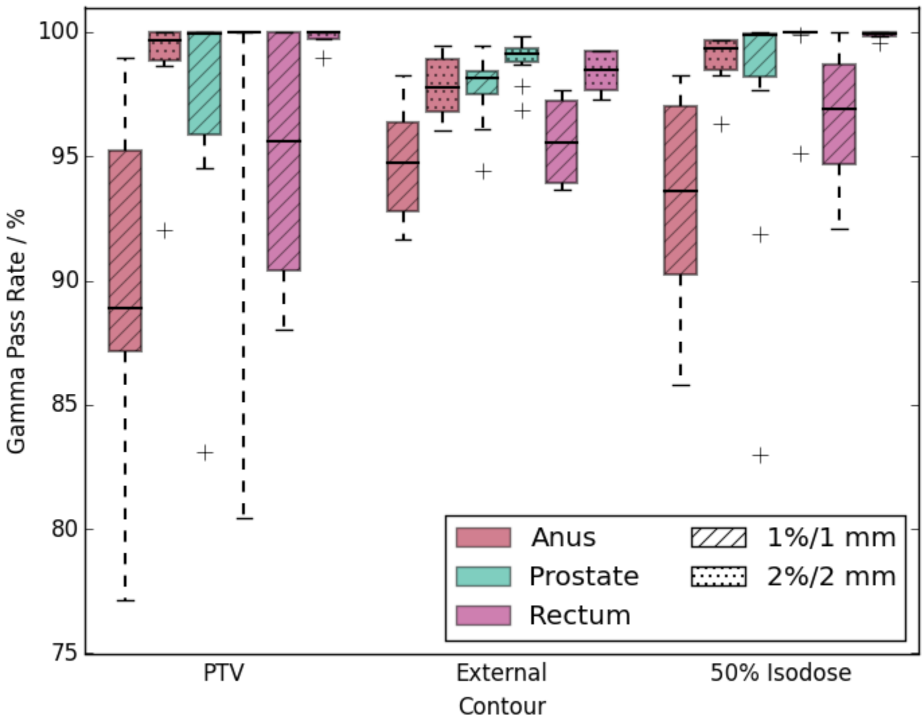


Figure 2.11: Boxplots of gamma pass rates for clinical plans within the PTV, external contour and 50% isodose contour for the anus (red bars, n=6), prostate (green bars, n=10) and rectum (purple bars, n=4). Gamma pass rates with criteria 1%/1 mm are shown with right diagonal hatching and with 2%/2 mm with dotted hatching. One outlier result for the prostate pass rate with 1%/1 mm criteria is not shown (53.8%). The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points.

for the clinical prostate and ano-rectal patient plans included in this study. The comprehensive dose evaluation used all patients and evaluated dose differences from a cylindrical PTV treatment plan centred on 4 longitudinally spaced isocentres (FH-5, FH, FH+5 ad FH+10). Dose differences were evaluated using two external contours: an intersection external contour which was identical for both sCT and CT to evaluate the dose differences caused by the sCT HU values, and a true external contour which evaluated the geometric fidelity of the sCT as well HU accuracy. The clinical plan evaluation used the clinical treatment plan for each patient (10 prostate, 6 anus and 4 rectum patients) within the true external contour only.

In the comprehensive dose evaluation, the absolute mean radiological isocentre depth differences were $\leq 0.6$ mm for all FH isocentre points except for FH-5 which was 2.2 mm. At the FH-5 point there were some angles with large differences, especially oblique angles approximately $\pm 30^o$ of the horizontal ($90^0$ and $270^o$) (figure 2.7). These differences were very similar to the physical differences in the true external contour, and completely absent in the intersection contour. This implies the differences were due to differences in the patient external contour and not incorrect HU assignment. These differences were likely due to differences in patient posture, even though they were setup using a radiotherapy pelvis board with matched settings to the CT image using external lasers. The oblique angles would be most affected by differences in external contour. The possibility of patient posture differences was much higher at the FH-5 point compared to superior points due to the greater impact of differences in patient posture such femur angle or buttock clenching.

The absolute mean dose differences to the FH plans within the true external contour were small, $\leq 1.1\%$, although there were five patients with D2% dose differences $> 2\%$ at at least one FH point isocentre. The dose differences within the intersection external contour were smaller, absolute mean differences $\leq 0.3\%$ and all patients within $\pm 1.2\%$. This again suggests that the primary reason for the larger dose differences for some patients was not incorrect HU assignment but differences in the external contour. The external contour dose differences were largest for the FH-5 point, followed by the FH point, with a systematic bias towards positive dose differences, implying the sCT over-estimates the dose. This fits with the under-determination of the radiological depth difference (figure 2.7) as the sCT has lower beam attenuation, resulting in a too high dose at the target.

The clinical plan dose differences were also small, with all mean dose differences $\leq 1.0\%$ (table 2.4). The dose differences in the OARs tended to be lower than the PTVs, probably due to having lower doses within them (all dose differences were relative to the prescription dose). Gamma pass rates with criteria 2%/2 mm were high, with the mean pass rate $\geq 97.8\%$ within all contours and for all treatment sites. At 1%/1 mm, pass rates were substantially lower, especially within the PTV(s). The anus patients had the lowest pass rate, probably due to having the most inferior PTVs where there were the largest discrepancies between CT and sCT external contours. Currently there are no published tolerances for acceptable gamma pass rates for the evaluation of sCT accuracy. However the results in this study do compare well to other results reported in the literature (see below).

There was good agreement between clinical PTV dose differences and the closest FH point to that clinical PTV (see figure 2.9). The prostate PTVs were located between FH-5 and FH points, and the mean D50% dose difference ($0.7 \pm 0.2\%$) was between the $0.4 \pm 0.2\%$ and $0.8 \pm 0.2\%$ for the FH and FH-5 points respectively. The rectum primary PTVs were close to the FH point and the elective PTVs close to the FH+5 point, and both had D50% dose differences that agreed with those FH dose differences within 0.1%. Similarly, the anus primary PTVs agreed within 0.2% with the FH-5 dose differences, the elective PTVs within 0.2% of the FH+5 differences and the nodal PTVS within 0.1% of the FH+10 differences. This implies that the FH plan evaluation was clinically relevant to the anus, prostate and rectum patients studied. This could be useful method going forward for evaluating dose differences with one patient cohort for a wide range of treatment sites.

A key question for MR-only radiotherapy is what magnitude of dose differences would be considered clinically significant. Dose differences $\geq 2\%$ are often suggested in the MR-only literature as the threshold above which differences are clinically significant [205]. The mean PTV dose differences at all FH points and all the clinical dose differences for all patients except one anus patient and one prostate patient were within ±2%. This implies the sCT was sufficiently accurate for clinical use. However, the justification for the ±2% value needs to be considered further. Over the whole treatment pathway the dosimetric uncertainty should be within 3% and the geometric uncertainty $2-4$ mm [8]. This requires the uncertainty of any individual component to be $< 1.0\%$ and $< 1.0$ mm to not significantly increase the overall uncertainty [8]. Therefore clinically significant dose differences would be $> 1.0\%$. Another consideration is that MR-only radiotherapy is essentially trading-off increased dosimetric uncertainty from the sCT for reduced geometric uncertainty through removing the MR-CT registration. This latter uncertainty is estimated at 2 mm for the pelvis [25], and so this also suggests generating an overall improvement in radiotherapy accuracy would require sCT dose differences to be $\leq 1.0\%$. All the FH plans PTV mean dose differences were $\leq 1.0\%$ except for the FH-5 D2%, which was 1.1%. And all the mean dose differences for the clinical plans were also $\leq 1.0\%$. This suggests that the sCT is sufficiently accurate for clinical use for most patients, although the larger dose differences for some patients ($\geq \pm 2\%$) implies some form of patient-specific QA of sCT dose calculation accuracy would be warranted (see chapter 3). Future work could investigate these setup differences further by examining the on-treatment CBCT images of these patients. If the variation in leg position on between different fraction CBCTs was similar to the magnitude of differences between CT and sCT then this suggests that these dose differences are not clinically relevant.

The results from this study compare well with other Deep Learning sCT approaches in the pelvis. Maspero et al. evaluated a Deep Learning sCT model for the pelvis and found dose differences of 0.4% and 0.4% to the PTV D50% for prostate and rectum patients respectively [78], compared to the $0.7 \pm 0.2\%$ and $0.3 \pm 0.1\%$ reported here. Gamma pass rates at 2%/2 mm were $95 \pm 2\%$ and $92 \pm 3\%$ for prostate and rectum patients, less than the $98.9 \pm 0.3\%$ and $98.4 \pm 0.5\%$ found in this study. Similarly, Yoo et al. found PTV D50% dose differences for prostate patients of $0.4 \pm 0.3\%$ and gamma pass rate (2%/2 mm) of $93 \pm 4\%$ [80]. A study looking at ano-rectal patients reported mean PTV dose differences of $-0.1\%$, with a 1%/1 mm gamma pass rate of 96% [79]. These agreed well with the $0.3 \pm 0.1\%$ and $0.2 \pm 0.3\%$ dose differences and $95 \pm 3\%$ and $95 \pm 1\%$ gamma pass rates for the rectum and anus patients respectively reported here. This suggests that the sCT algorithm performs equivalently well as those presented in the literature as sufficiently accurate for clinical use.

A potential limitation of this study was that rigid sCT to CT registration was used and not deformable registration. Both approaches are used in the literature, with deformable registration having the potential advantage of removing confounding differences in patient set-up from the different imaging sessions [55]. However, this does mean the sCT is not evaluated as it would be used for a clinical MR-only pathway, because it is deformed to match the CT [79]. This potentially masks geometric issues in the MR such as residual geometric distortion, which could have an impact on the dose calculation accuracy of sCT images. This study has used a rigid registration to ensure the sCT is evaluated in a clinically representative manner, but with an additional analysis within the intersection external contour, removing any differences in external contour between sCT and CT. This enables the scale of dose differences due to patient alignment discrepancies compared to HU assignment inaccuracies to be assessed. The results showed mean differences within the true external and intersection external were similar (within 0.7%), but with the range of differences substantially reduced.

A limitation of this study is that only data from a single centre and single manufacturer scanner, which was the same scanner used for the training cohort, has been used in the evaluation. This is particularly an issue for Deep Learning algorithms where over-fitting is a recognised problem, leading to poor generalisation of the algorithm. Future work could extend this evaluation to patient images acquired at different centres on different scanners to validate the generalisability of the results reported here. A related issue would be to investigate how sensitive the algorithm was to variations in the input images, such as different image resolutions, fields of view and noise levels [44].

## 2.3 Evaluation of MR Automatic Contours

### 2.3.1 Materials and Methods

#### MR Automatic Contour Creation

The automatic contours were generated by a Deep Learning algorithm produced by GE Healthcare, consisting of CNNs trained using manually contoured MR images (see section 1.2.2 for further background on the different Deep Learning models described here). The training dataset was drawn from several sources: the training patient cohort described in table 2.1, healthy volunteer images from the same 3T PET-MR scanner and prostate patient images from a different manufacturer 1.5 T scanner. All images were acquired with T2-weighted 3D TSE sequences, but with some variability in scan parameters. All images were manually contoured following organ-specific guidelines by medical students who had been trained by clinical oncologists. The number of training patients for each OAR model varied (see table 2.5). The prostate & seminal vesicles organ was a combined contour whereas the prostate contour only contained the prostate itself.

Individual CNN models were used for each OAR except for the femoral heads which were contoured by a single model and the penile bulb and urethra, which were also contoured by single model. Different approaches were used for different OAR types. For the large organs (bowel bag and pelvis body) where images may not completely cover the organ extent, 2D axial U-NETs were used for contouring each slice separately. The U-NETs used three contiguous slices as an input, with contours returned for the central slice.

Table 2.5: The OARs generated by the automatic contouring algorithm for male and female patient images, along with the number of patients used in training the model.

| Organ | Training Patients | Female Patients | Male Patients |
|---|---|---|---|
| Bladder | 92 | ✓ | ✓ |
| Bowel Bag | 39 | ✓ | ✓ |
| Femoral Head L | 95 | ✓ | ✓ |
| Femoral Head R | 95 | ✓ | ✓ |
| Pelvis Body | 81 | ✓ | ✓ |
| Penile Bulb | 77 | - | ✓ |
| Prostate | 80 | - | ✓ |
| Prostate & Seminal Vesicles | 80 | - | ✓ |
| Rectum | 93 | ✓ | ✓ |
| Urethra | 77 | - | ✓ |

The intermediate organs (bladder, prostate, prostate & seminal vesicles and rectum) were localised using three 2D U-NETS which contoured the OAR in low-resolution axial, coronal and sagittal planes. These were combined to produce a bounding box for the organ. The original image was then cropped to this bounding box, and a 3D V-NET model used on the cropped image to contour the OAR. The femoral heads were contoured using the same methodology but on both images inverted in the right-left axis and with the original orientation, thus enabling left and right femoral heads to be localised and contoured by the same model. Finally the penile bulb and urethra were localised with the prostate and contoured using a single joint 3D V-NET model.

All models included post-processing steps to smooth the contours, remove holes and resample to the input image resolution and the final contours were then exported as a DICOM radiotherapy structure set objects.

### Clinical Evaluation of MR Automatic Contours

The automatic contour algorithm was evaluated using the 20 evaluation patients described in table 2.1. Manual contours for comparison were produced by the same trained medical students who contoured the training images. The automatic contours were compared to the manual contours using three comparison metrics calculated in RayStation: the DSC, the mean Distance To Agreement (DTA) and the maximum DTA. The DSC is defined as twice the intersection volume of the two contours divided by the sum of the volumes of the two contours [91]. The DTA calculates the minimum distance from each point on the manual contour to a point on the automatic contour. The mean DTA is the mean of all distances and the maximum is the maximum distance. There is no currently no consensus in the literature on the best delineation metrics to use and so multiple metrics were evaluated to enable comparisons to the literature and correlations with clinical rating.

A fourth novel methodology was employed called the margin expansion method. This gave the average margin needed to expand the automatic contour so that it encompassed a certain percentage volume of the manual contour. This method aims to give an estimate of the PTV margin needed to encompass delineation uncertainties from the automatic contour, so providing something that is more clinically relevant than conventional delineation metrics. This method took the automatic contour and created an expansion around this using an increasing margin size with a step of 0.5 mm. At each new margin size the percentage volume of the manual contour that was contained within the automatic contour plus margin was calculated. This was then repeated for each patient. The mean margins required for 99% coverage of the manual contour were determined ($m_{99\%}$).

In addition, the automatic OAR contours for each patient were rated for clinical acceptability by an experienced consultant clinical oncologist on a five-point scale:

1. Delete: The contour is not acceptable for clinical use, complete recontouring needed
2. Major: The contour is not acceptable for clinical use, significant correction needed
3. Intermediate: The contour is not acceptable for clinical use, some correction needed
4. Minor: The contour is acceptable for clinical use, but minor corrections could be done
5. Accept: The contour is acceptable for clinical use at it is

Distance metrics have been suggested to be the most clinically relevant assessments of radiotherapy contours [91], so the clinical ratings were compared to plots of maximum against mean DTA metrics to visually assess correlation. The novel $m_{99\%}$ metric was also plotted as a function of clinical rating to visually assess correlation.

### Dosimetric Evaluation of MR Automatic Contours

The dose impact of contouring differences was evaluated using treatment plans produced for the relevant treatment site and calculated on CT. The planning CT was rigidly registered to the T2w-MR using the same parameters used in the sCT dose evaluation and both the automatic and manual OARs were copied onto the CT without modification. An automatic external contour was created on the CT using a threshold of −250 HU, which was used for all dose calculations. Any air within the patient contour was delineated and over-ridden to water density. For all sites, each patient had two VMAT plans generated: one using the automatic OAR contours in the optimisation and one using the manual OAR contours, labelled $\text{plan}_{\text{auto}}$ and $\text{plan}_{\text{man}}$ respectively. Both plans were optimised to deliver the same prescribed dose to the same PTV, using the same number of VMAT arcs with the same isocentre. The same optimisation parameters were used for both plans to remove operator bias and ensure differences in plans were due to differences in contours only.

For the prostate patients (n=10), the manually drawn prostate and seminal vesicles volume were used as the CTVs and expanded using PTV margins of 4 mm for the prostate-only PTV and 8 mm for the prostate and seminal vesicles PTV, in accordance with the PACE trial protocol [206]. The prostate-only PTV was prescribed 60 Gy and the prostate and seminal vesicles PTV 47 Gy in 20 fractions [206]. Single $360^o$ arc 6 MV VMAT plans were created to deliver a homogeneous dose to the PTV and minimise doses to the OARs. The local clinical standard set of optimisation weights were used (class solution) for both $\text{plan}_{\text{auto}}$ and $\text{plan}_{\text{man}}$ and no optimisation weights were changed in the planning process. All plans met mandatory dose constraints.

For the rectal cancer patients (n=4) the clinical CT-based PTV was used as no manual GTV and CTV contours had been delineated on the T2w-MR. Patients were prescribed doses of 50.4 Gy in 28 fractions (n=2), 45 Gy in 25 fractions (n=1) and 25 Gy in 5 fractions (n=1). Dual $360^o$ arc 6 MV VMAT plans were optimised using the optimisation objectives and weights and beam geometries from the clinical CT-based plan. All plans met mandatory dose constraints except those which had not been achieved in the clinical plan.

Similarly, for the anal cancer patients (n=6) the clinical CT-based PTV was used. Patients were prescribed 53.2 Gy (n=4) or 50.4 Gy (n=2) in 28 fractions. The optimisation objectives and weights and beam geometries from the clinical CT-based quadruple $360^o$ arc 6 MV VMAT plan were used to optimise $\text{plan}_{\text{auto}}$ and $\text{plan}_{\text{man}}$. Three patients had clinical treatment plans that were not accessible to the research team. For those patients, a quadruple $360^o$ arc VMAT plan was optimised using the CT-based structure set by an experienced planner and those optimisation parameters and beam geometries used for $\text{plan}_{\text{auto}}$ and $\text{plan}_{\text{man}}$. All plans met mandatory dose constraints except those which had not been achieved in the clinical plan.

DVH parameters were extracted for the manual contours from both plans ($\text{plan}_{\text{auto}}\text{contour}_{\text{man}}$ and $\text{plan}_{\text{man}}\text{contour}_{\text{man}}$ and for the automatic contours from the plan based on automatic contours ($\text{plan}_{\text{auto}}\text{contour}_{\text{auto}}$). This enabled two comparisons to be made: 1) the accuracy of the automatic contours for treatment plan optimisation, $\text{plan}_{\text{auto}}\text{contour}_{\text{man}}$ vs gold standard $\text{plan}_{\text{man}}\text{contour}_{\text{man}}$ and 2) the accuracy of the automatic contours for treatment plan evaluation, $\text{plan}_{\text{auto}}\text{contour}_{\text{auto}}$ vs gold standard $\text{plan}_{\text{auto}}\text{contour}_{\text{man}}$. The first comparison kept the evaluation contours the same but varied the planning contours, whereas the second comparison kept the planning contours the same but varied the evaluation contours.

Both comparisons were made on clinical relevant dose-volume constraints for the appropriate treatment site. The prostate dose-volume constraints were taken from the recommendations in a review paper of hypofractionated radiotherapy for prostate cancer [207]. Additional dose constraints for the bowel bag were taken from the RCR rectum cancer radiotherapy guidance document [208] and for the urethra from a trial protocol evaluating urethra-sparing prostate Stereotactic Ablative Body Radiotherapy (SABR) [209]. The femoral head dose constraints were replaced by those taken from the RCR rectum cancer radiotherapy guidance document [208] because only 4/10 prostate patients had doses above the 30 Gy prostate dose constraint. The prostate and prostate & seminal vesicles contours were excluded from the analysis as they were target contours and not OARs for this patient group. The pelvis body contour was excluded from the analysis as there was the same (CT-based) external contour used for all plans.

The anal and rectal cancer dose constraints were taken from the UK national IMRT guidance documents published by the RCR [208,210]. Additional dose constraints for the penile bulb were added from those used for prostate cancer [207] and for the urethra from the same source as above [209]. The prostate and seminal vesicles are not considered OARs in most other pelvic radiotherapy and so there were no dose constraints for them in the literature. Instead the dose differences at the D5%, D50% and D95% were assessed. The constraints are summarised in results tables 2.7 and 2.8. For each constraint the absolute difference in cumulative dose or absolute/relative volume was used as the measure of difference to avoid biasing the results towards constraints with very low values. The rectum contours were excluded from the analysis as they were target contours and not OARs for this patient group, and the pelvis body contour excluded for the same reason as for the prostate patients.

The differences in dose-volume constraints for the evaluation assessment were also correlated with the clinical rating. A single DVH constraint was selected for each OAR which showed the maximum dependence of the evaluation dose difference with clinical rating, determined by visual inspection of plots for each OAR. Summary plots of all OARs were then produced for the prostate and ano-rectal patients separately.

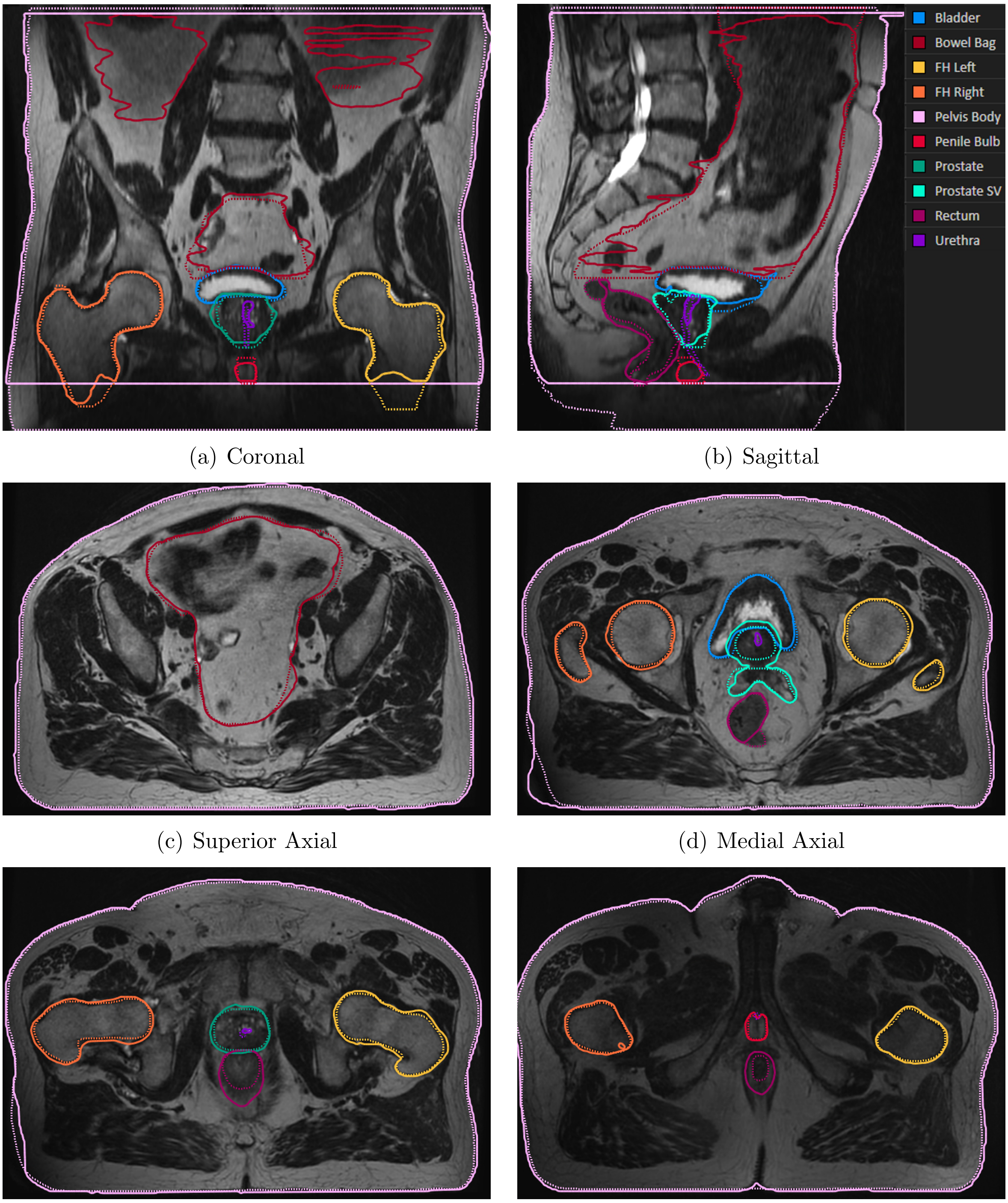
## 2.3.2 Results

### Clinical Evaluation of MR Automatic Contours

Automatic contours were successfully generated for all OARs and all patients, although the penile bulb contour on one patient and the urethra contour on two patients were completely misplaced (inside the bladder), suggesting the localisation model failed for those patients. Example contours for a representative patient can be seen in figure 2.12.

The delineation metrics are summarised for each OAR in table 2.6, with the different OAR models having quite different levels of performance. The volume that the automatic contours overlapped the manual contours as a function of expansion margin showed a similar variability between OARs (figure 2.13).

The clinical ratings also showed varying performance levels (figure 2.14), with only the femoral heads and pelvis body contours having all patients clinically acceptable, although only the bowel bag and prostate contours had median values below clinically acceptable. There was limited correlation between the clinical ratings and the margin expansion metric of the $\mathrm{DTA}_{\mathrm{max}}$ vs $\mathrm{DTA}_{\mathrm{mean}}$ plots (figures 2.13 and 2.15).



(a) Coronal

(b) Sagittal

(c) Superior Axial

(d) Medial Axial

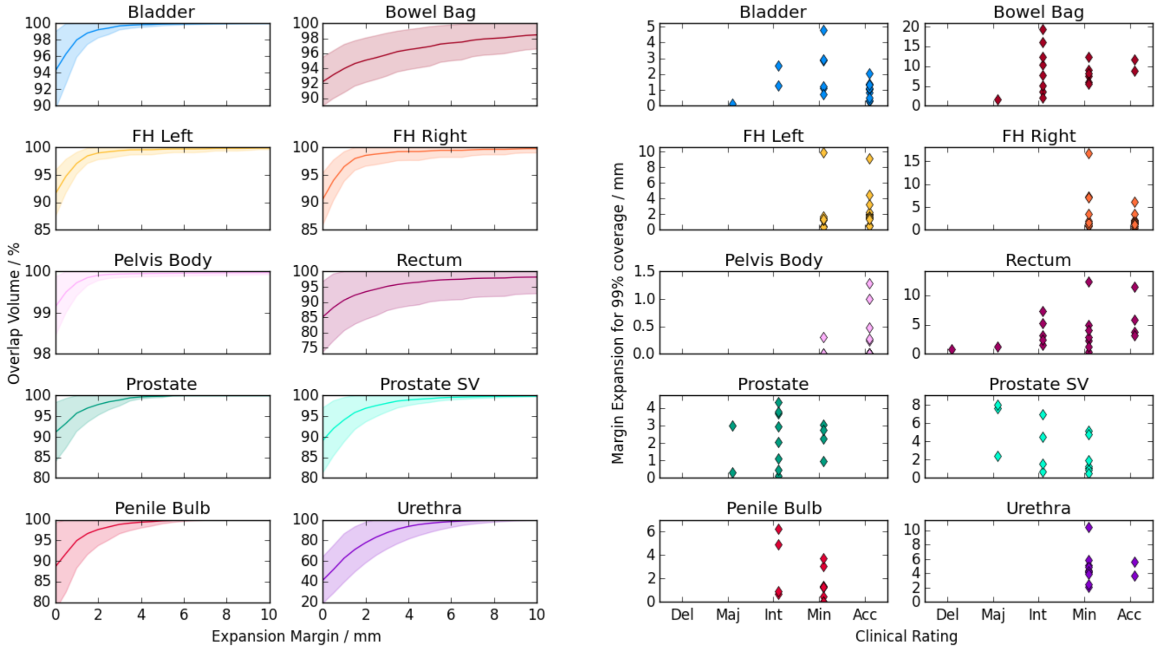
(e) Medial Axial

(f) Inferior Axial

Figure 2.12: Automatic (solid lines) and manual (dotted lines) contours for a representative patient. The patient was selected as having the most organs with the closest $\text{DTA}_{\text{mean}}$ results to the mean of all patients.

Table 2.6: Summary of delineation metrics for automatic contours (n=20 except for the organs only contoured on male patients, indicated by §, where n=14).$m_{99\%}$ indicates the expansion margin required by the automatic contour to cover 99% of the manual contour. All results given as mean ± standard error (minimum, maximum). *One (**two) patient(s) excluded because the automatic contour was completely misplaced.

| **Organ** | **DSC** | **$\mathbf{DTA_{mean}}$/mm** | **$\mathbf{m_{99\%}}$/mm** |
|---|---|---|---|
| Bladder | $0.92 \pm 0.01$ (0.81,0.97) | $1.1 \pm 0.1$ (0.6,1.8) | $1.5 \pm 0.3$ (0.1,4.8) |
| Bowel Bag | $0.89 \pm 0.01$ (0.82,0.94) | $4.3 \pm 0.3$ (2.6,8.4) | $8.7 \pm 1.2$ (1.7,$>$ 20.0) |
| Femoral Head L | $0.93 \pm 0.01$ (0.87,0.96) | $1.1 \pm 0.1$ (0.5,1.7) | $2.4 \pm 0.6$ (0.4,9.9) |
| Femoral Head R | $0.92 \pm 0.01$ (0.87,0.96) | $1.2 \pm 0.1$ (0.5,2.6) | $3.2 \pm 0.9$ (0.5,16.7) |
| Pelvis Body | $0.95 \pm 0.01$ (0.92,0.99) | $3.9 \pm 0.5$ (0.6,7.1) | $0.2 \pm 0.1$ (0.0,1.3) |
| Rectum | $0.80 \pm 0.01$ (0.69,0.93) | $2.1 \pm 0.4$ (0.7,8.8) | $4.1 \pm 0.8$ (0.4,$>$ 20.0) |
| Penile Bulb§* | $0.67 \pm 0.04$ (0.36,0.88) | $1.6 \pm 0.2$ (0.6,2.3) | $1.9 \pm 0.5$ (0.0,6.2) |
| Prostate§ | $0.84 \pm 0.01$ (0.76,0.91) | $1.9 \pm 0.2$ (1.0,3.1) | $2.2 \pm 0.4$ (0.1,4.3) |
| Prostate & SV§ | $0.81 \pm 0.01$ (0.71,0.90) | $1.8 \pm 0.2$ (0.8,3.0) | $3.4 \pm 0.7$ (0.5,8.0) |
| Urethra§** | $0.36 \pm 0.04$ (0.14,0.66) | $1.9 \pm 0.4$ (0.7,5.3) | $4.8 \pm 0.6$ (2.1,10.5) |



(a) Margin Expansion

(b) $m_{99\%}$ with Clinical Rating

Figure 2.13: a) Plots of percentage volume of manual contour included within automatic contour plus a margin as a function of that margin for each organ. Solid lines indicate mean over all patients and shaded areas the ± one standard error. Top six plots show results for combined organs (n=20) and bottom four for male-only organs (n=14). The y-axis scale is different for each plot. Subfigure (b) shows the 99% expansion metric $m_{99\%}$ for each patient as a function of clinical rating, where 'Del', 'Maj', 'Int', 'Min' and 'Acc' stand for 'Delete', 'Major', 'Intermediate', 'Minor' and 'Accept' respectively. Again the y-axis scale is different for each plot.

### Dosimetric Evaluation of MR Automatic Contours

For the prostate patients, the mean evaluation dose differences for the femoral heads and urethra contours were $\leq \pm 0.1$ Gy (see table 2.7). The penile bulb mean dose difference was larger and with a wider range of differences. For one patient the penile bulb and urethra automatic contours were completely misplaced, yielding dose differences to the D50% and

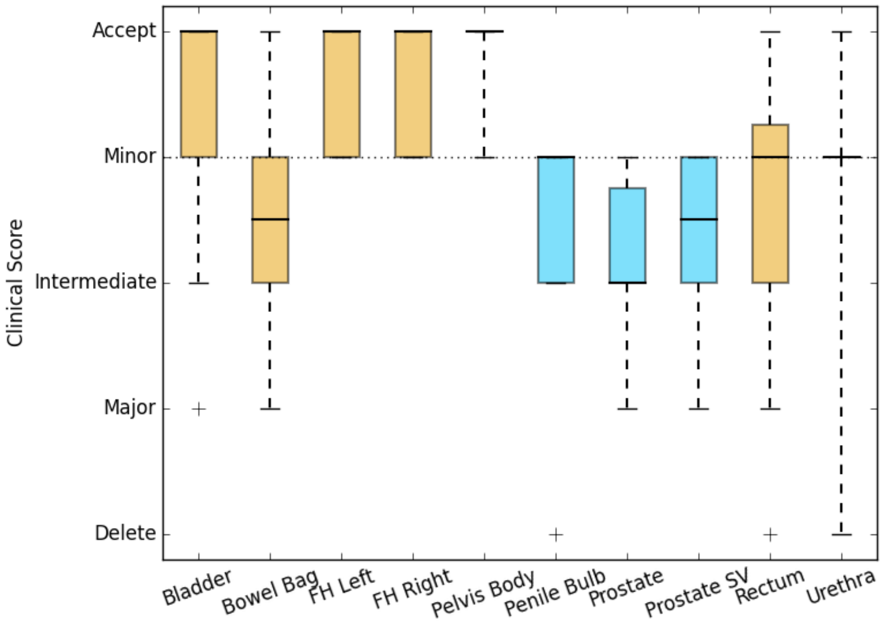


Figure 2.14: Boxplots of the clinical ratings of each organ. Results for the combined organs (n=20 patients) are shown in yellow and for the male-only organs (n=14 patients) in blue. The dotted horizontal line indicates ratings considered clinically acceptable ($\geq$ 4). The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points. The 'Delete' scores for the penile bulb and urethra contours were from the totally misplaced contours which were excluded from the metric and dosimetric evaluation.

D2% respectively of $-14.5$ Gy and 45.6 Gy. These contours have been excluded from the analysis. The relative volume differences for the bladder and rectum were also larger, but still small ($< \pm 1.0$ percentage point), with a trend to larger differences for the volumes at lower doses. The bowel bag mean dose differences were larger again (absolute differences 1.6 to 1.9 Gy), with the largest differences at the smallest D0.5 cm$^3$ constraint. The optimisation differences were smaller than the evaluation differences for all OARs except the femoral heads.

The mean evaluation dose differences for the ano-rectal patients were similarly small ($\leq \pm 0.7$ Gy) for all OARs, except for the bowel bag D180 cm$^3$ which was 1.4 Gy. For one rectum patient the urethra automatic contour was completely misplaced (D2% dose difference of $-1.0$ Gy) and the contour excluded from analysis. Similar to the prostate patients the optimisation dose differences were smaller than evaluation differences except for the femoral heads.

The correlation between the evaluation dose differences at selected dose constraints and the clinical rating of the quality of automatic contours are shown in figures 2.16.

### 2.3.3 Discussion

This sub-study has evaluated MR based automatic contours for pelvic OARs in male and female patients. Contours were evaluated compared to manual contours using delineation

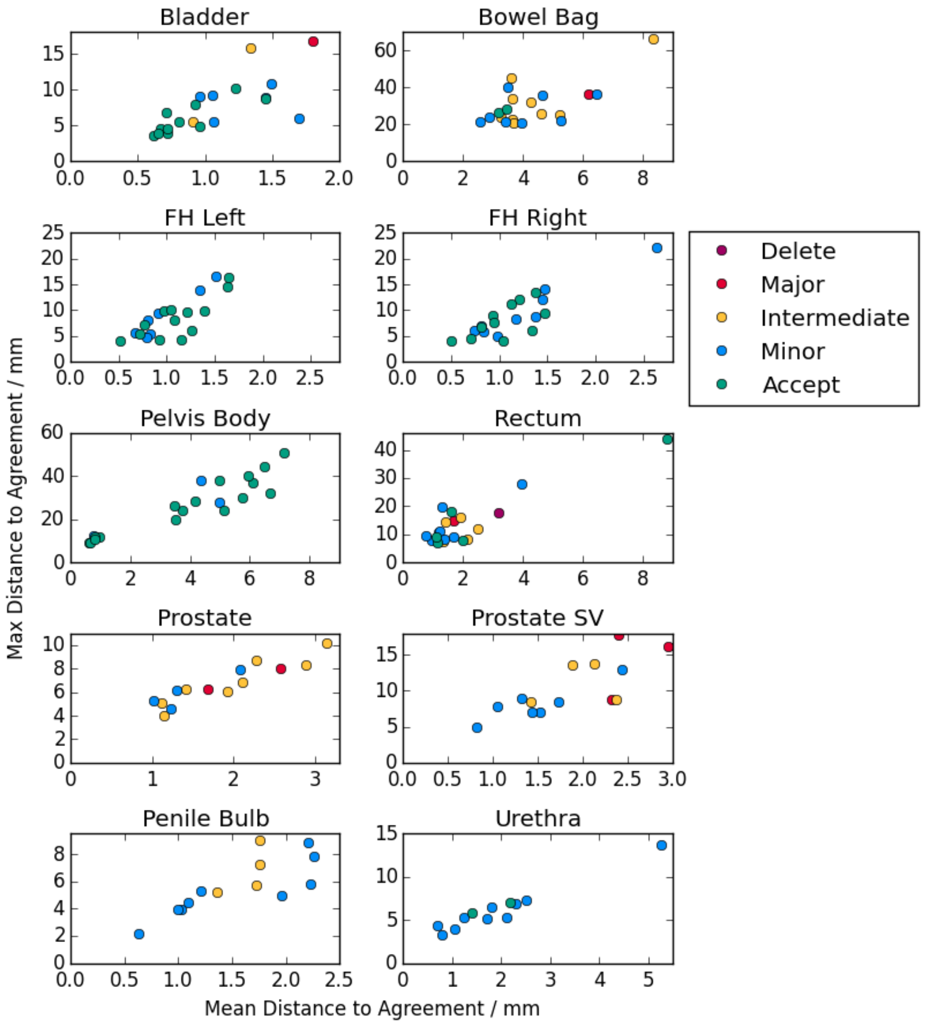


Figure 2.15: Plots of DTA$_{\text{max}}$ against DTA$_{\text{mean}}$ for automatic contour compared to manual contour for each organ. The colour of each point indicates the clinical rating of the automatic contour. Top six plots show results for combined organs (n=20) and bottom four for male-only organs (n=14). The x- and y-axis scales are different for each plot.

metrics, including a margin expansion method and were rated by a clinician. The dosimetric impact was assessed by comparing dose distributions from the same plan on the manual and automatic contours (evaluation assessment) and doses to manual contours from plans optimised using manual and automatic contours (optimisation assessment).

### Clinical Evaluation

The clinical rating demonstrated good performance for several OARs, with median values considered clinically acceptable for bladder, femoral heads, body, penile bulb, rectum and urethra contours (figure 2.14). However only the femoral heads and body contours had all patients clinically acceptable, and even for the bladder with a median rating of accept (no modification possible), three patients were rated as unacceptable requiring intermediate or major modification. This highlights the importance of patient-specific QA of automatic contouring solutions, even with very good average performance over a population. The urethra showed very good results except for the two patients were it was

Table 2.7: Differences at clinically relevant dose constraints for the prostate cancer patients (n=10). The square brackets in the second column indicate the units of the differences in the third and fourth column (% for differences in relative volume, Gy for differences in dose). The third column lists the results for the optimisation assessment, $\text{plan}_{\text{auto}}\text{contour}_{\text{man}}$ - $\text{plan}_{\text{man}}\text{contour}_{\text{man}}$. The fourth column gives the equivalent results for the evaluation assessment $\text{plan}_{\text{auto}}\text{contour}_{\text{auto}}$ - $\text{plan}_{\text{auto}}\text{contour}_{\text{man}}$. All results given as mean ± standard error (minimum, maximum).[*]One patient excluded because the automatic contour was completely misplaced. [†]One patient excluded because the volume receiving 60 Gy was zero.

| **Organ** | **Dose Constraint** | **Mean Differences** | |
|---|---|---|---|
| | | *Optimisation* | *Evaluation* |
| Bladder | V60 Gy[%] | 0.3 ± 0.2 (−0.2,1.4) | 0.0 ± 0.1 (−1.0,0.3) |
| | V50 Gy [%] | 0.1 ± 0.1 (−0.6,0.7) | −0.4 ± 0.5 (−3.7,1.5) |
| | V40 Gy [%] | 0.1 ± 0.1 (−0.7,0.5) | −0.4 ± 0.5 (−4.2,1.2) |
| | V30 Gy [%] | 0.0 ± 0.2 (−0.6,0.9) | −0.5 ± 0.6 (−5.1,0.8) |
| Bowel Bag | D180 $\text{cm}^3$ [Gy] | 0.1 ± 0.2 (−0.3,1.3) | 1.8 ± 2.1 (−5.7,1.4) |
| | D100 $\text{cm}^3$ [Gy] | 0.3 ± 0.2 (−0.3,1.8) | 1.8 ± 2.2 (−9.2,15.7) |
| | D65 $\text{cm}^3$ [Gy] | 0.4 ± 0.3 (−0.6,2.2) | 1.6 ± 2.5 (−12.7,14.2) |
| | D0.5 $\text{cm}^3$ [Gy] | 0.1 ± 0.5 (−2.8,2.1) | −1.9 ± 4.2 (−32.4,15.4) |
| Femoral Head L | D50% [Gy] | 0.4 ± 0.5 (−1.7,3.3) | 0.1 ± 0.2 (−0.5,1.0) |
| | D35% [Gy] | −0.3 ± 0.6 (−3.3,2.1) | 0.1 ± 0.1 (−0.3,0.6) |
| | D5% [Gy] | −0.6 ± 0.7 (−3.3,3.8) | 0.1 ± 0.04 (−0.1,0.2) |
| Femoral Head R | D50% [Gy] | −0.7 ± 0.7 (−6.5,1.2) | −0.1 ± 0.2 (−1.0,0.6) |
| | D35% [Gy] | −0.7 ± 0.6 (−3.7,1.4) | 0.0 ± 0.1 (−0.8,0.4) |
| | D5% [Gy] | −0.4 ± 0.6 (−4.7,2.5) | 0.1 ± 0.04 (−0.1,0.3) |
| Penile Bulb | D50% [Gy] | 0.0 ± 0.08 (−0.2,0.6) | 0.9 ± 0.6 (−0.8,4.8)[*] |
| Rectum | V60 Gy [%] | −0.1 ± 0.1 (−0.5,0.1) | 0.1 ± 0.03 (−0.1,0.2)[†] |
| | V57 Gy [%] | 0.0 ± 0.1 (−0.3,0.3) | 0.6 ± 0.3 (−1.8,1.5) |
| | V30 Gy [%] | 0.5 ± 0.4 (−2.6,1.8) | 0.9 ± 1.1 (−5.0,5.0) |
| Urethra | D2% [Gy] | 0.0 ± 0.1 (−0.6,0.7) | −0.1 ± 0.2 (−1.3,0.4)[*] |

completely misplaced. This suggests that the contouring part of the model is performing well but the localisation part is sensitive to errors. The bowel bag and rectum contours were more variable, with two patients rating accept but also two patients rating major or delete. The penile bulb, prostate and prostate & seminal vesicles contours performed the worst, with no patient contours not requiring modification. These models were trained with fewer patients (77-80) compared to the bladder, femoral heads and rectum (92-95), possibly indicating larger training cohort would improve the results, although the bowel bag was trained on substantially fewer patients (39) and returned better results.

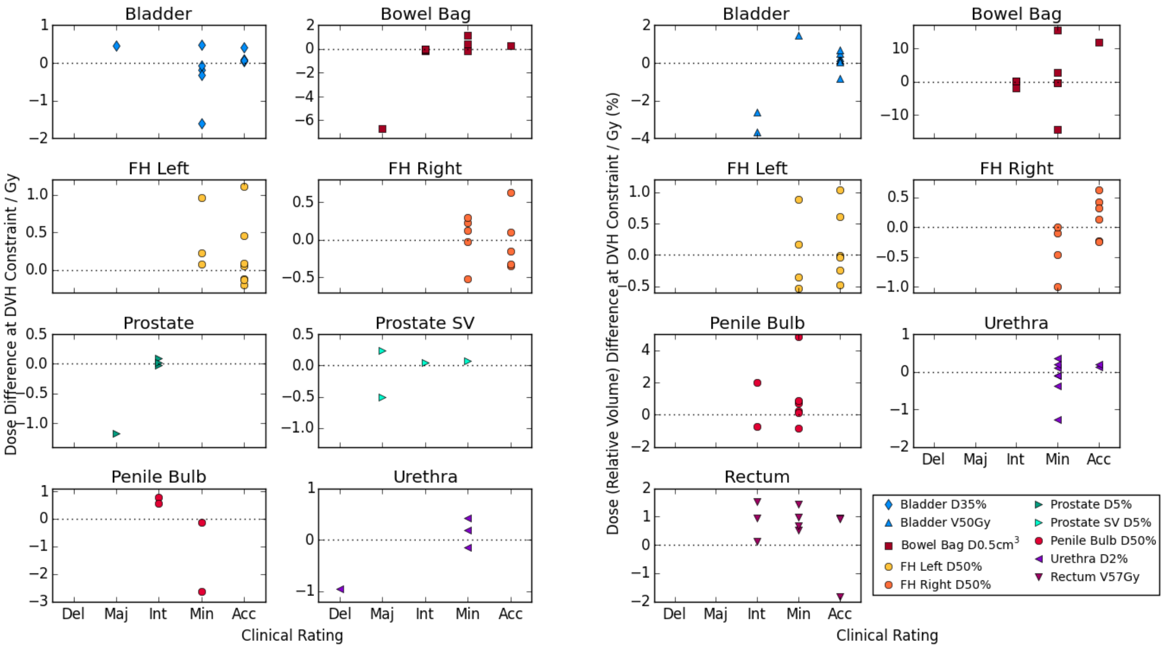
The mean DSC results were $\geq 0.9$ for the bladder, femoral heads and body contour and $\geq 0.8$ for the bowel bag, rectum, prostate and prostate & SV contours (table 2.6). This is similar to DSC results reported for the inter-observer variability in manual MR-based contouring of the bladder (0.95), rectum (0.82) and prostate (0.84) [69]. However DSC has no correlation with dosimetric differences [91] or clinical acceptability [96] and so is likely of limited value. This can be seen in the very low urethra DSC, even though all contour patients were rated clinically acceptable (except for two misplaced contours which were excluded from the DSC analysis).

Table 2.8: Differences at clinically relevant dose constraints for the anal and rectal cancer patients (n=10 except for the organs only contoured on male patients, indicated by §, where n=4). The square brackets in the second column indicate the units of the differences in the third and fourth column (% for differences in relative volume, Gy for differences in dose). The third column lists the results for the optimisation assessment, plan$_{\mathrm{auto}}$contour$_{\mathrm{man}}$ - plan$_{\mathrm{man}}$contour$_{\mathrm{man}}$. The fourth column gives the equivalent results for the evaluation assessment plan$_{\mathrm{auto}}$contour$_{\mathrm{auto}}$ - plan$_{\mathrm{auto}}$contour$_{\mathrm{man}}$. All results given as mean ± standard error (minimum, maximum). *One patient excluded because the automatic contour was completely misplaced.

| **Organ** | **Dose Constraint** | **Mean Differences** | |
|---|---|---|---|
| | | *Optimisation* | *Evaluation* |
| Bladder | D50% [Gy] | $-0.2 \pm 0.2$ ($-1.0$,0.6) | $-0.2 \pm 0.1$ ($-0.9$,0.5) |
| | D35% [Gy] | $-0.2 \pm 0.2$ ($-1.2$,0.5) | $-0.1 \pm 0.2$ ($-1.6$,0.5) |
| | D5% [Gy] | $-0.1 \pm 0.03$ ($-0.2$,0.1) | $0.0 \pm 0.1$ ($-0.4$,0.7) |
| Bowel Bag | D180cm$^3$ [Gy] | $0.0 \pm 0.1$ ($-0.7$,0.4) | $1.4 \pm 1.5$ ($-2.5$,12.5) |
| | D100cm$^3$ [Gy] | $0.0 \pm 0.1$ ($-0.2$,0.3) | $0.0 \pm 0.4$ ($-1.4$,3.4) |
| | D65cm$^3$ [Gy] | $0.1 \pm 0.1$ ($-0.2$,0.6) | $-0.1 \pm 0.4$ ($-1.6$,3.2) |
| | D0.5cm$^3$ [Gy] | $0.1 \pm 0.2$ ($-0.9$,0.9) | $-0.5 \pm 0.7$ ($-6.7$,1.1) |
| Femoral Heads L | D50% [Gy] | $0.0 \pm 0.3$ ($-1.2$,1.8) | $0.3 \pm 0.1$ ($-0.2$,1.1) |
| | D35% [Gy] | $-0.7 \pm 0.3$ ($-3.0$,0.1) | $0.3 \pm 0.2$ ($-0.1$,1.4) |
| | D5% [Gy] | $0.0 \pm 0.2$ ($-1.3$,1.0) | $0.2 \pm 0.1$ ($-0.4$,0.8) |
| Femoral Heads R | D50% | $0.3 \pm 0.2$ ($-0.4$,2.0) | $0.0 \pm 0.1$ ($-0.5$,0.6) |
| | D35% [Gy] | $0.2 \pm 0.2$ ($-0.9$,1.1) | $0.0 \pm 0.1$ ($-0.6$,0.7) |
| | D5% [Gy] | $0.0 \pm 0.2$ ($-0.8$,0.7) | $0.2 \pm 0.1$ ($-0.2$,0.7) |
| Penile Bulb§ | D50% [Gy] | $-0.1 \pm 0.1$ ($-0.3$,0.0) | $-0.4 \pm 0.8$ ($-2.7$,0.8) |
| Prostate§ | D95% [Gy] | $0.2 \pm 0.2$ (0.0,0.8) | $0.7 \pm 0.4$ ($-0.2$,1.5) |
| | D50% [Gy] | $0.1 \pm 0.1$ (0.0,0.2) | $0.0 \pm 0.1$ ($-0.4$,0.2) |
| | D5% [Gy] | $-0.1 \pm 0.1$ ($-0.2$,0.0) | $-0.3 \pm 0.3$ ($-1.2$,0.1) |
| Prostate & SV§ | D95% [Gy] | $0.2 \pm 0.2$ (0.0,0.6) | $0.7 \pm 0.4$ ($-0.5$,1.4) |
| | D50% [Gy] | $0.1 \pm 0.1$ (0.0,0.3) | $0.1 \pm 0.1$ ($-0.1$,0.2) |
| | D5% [Gy] | $0.0 \pm 0.02$ ($-0.1$,0.0) | $0.0 \pm 0.2$ ($-0.5$,0.2) |
| Urethra§ | D2% [Gy] | $-0.1 \pm 0.1$ ($-0.4$,0.0) | $0.2 \pm 0.2$ ($-0.2$,0.4)* |

The novel margin expansion metric ($m_{99\%}$) showed similar results to the DTA$_{\mathrm{mean}}$ metric for most OARs, although it was substantially larger for the bowel bag and rectum. This was due to discrepancies in longitudinal extent impacting $m_{99\%}$ more than DTA$_{\mathrm{mean}}$ (see figure 2.13). This is also likely the reason for the poorer agreement for the combined prostate & seminal vesicles contour compared to the prostate only, even though the DTA$_{\mathrm{mean}}$ results were similar. There was also a volume effect, with very large volumes with small $m_{99\%}$ (body contour) and small volumes larger ones (urethra and penile bulb). This suggests that the novel metric does not substantially improve on standard distance-to-agreement metrics.

The delineation metrics showed limited correlation with clinical rating. Figure 2.15 shows plots of DTA$_{\mathrm{max}}$ against DTA$_{\mathrm{mean}}$, colour-coded by clinical rating. Using both metrics attempted to capture both contours which were mostly accurate except on a few slices with large discrepancies (small DTA$_{\mathrm{mean}}$, large DTA$_{\mathrm{max}}$) and contours which were systematically different but with no large discrepancies (large DTA$_{\mathrm{mean}}$, small DTA$_{\mathrm{max}}$). However,



(a) Ano-Rectal Patients

(b) Prostate Patients

Figure 2.16: Dose or relative volume differences at selected DVH constraints between automatic and manual contours for the automatic plans as a function of clinical rating of the quality of the automatic contours. Clinical rating was recorded on a 5-point scale: ‘Del’, ‘Maj’, ‘Int’, ‘Min’ and ‘Acc’, which stood for ‘Delete’, ‘Major’, ‘Intermediate’, ‘Minor’ and ‘Accept’ respectively. Differences are shown for the ano-rectal patients (a) and prostate patients (b). The legend for both subfigures is shown on subfigure (b). For the prostate patients results for one penile bulb and one urethra contour are not shown (dose differences 45.6 Gy and −14.5 Gy respectively, clinical rating for both contours ‘delete’ ).

there was a general correlation between mean and maximum results and no correlation with clinical rating except for the prostate & seminal vesicles, although even then results for a patient rated ‘minor’ were larger than for a patient rated ‘major’. The rectum contours had a patient rated ‘accept’ with the highest mean and max DTA results due to large but clinically insignificant variations in superior extent, and a contour rated ‘delete’ with much lower DTA values. This suggests that even combined mean and maximum DTA metrics are of very limited value in assessing contouring quality. The $\mathrm{m}_{99\%}$ also did not show any correlation with clinical rating (figure 2.13), again implying it does not add value.

## Dosimetric Evaluation

For the prostate patients, the mean evaluation dose/relative volume differences (same plan, different contours) suggested good contouring accuracy for the bladder, femoral heads, penile bulb, rectum and urethra, with mean dose differences $< 1.0$ Gy or mean relative volume differences $< 1$ percentage point (table 2.7). This was similar to the mean dose differences $< 1.0$ Gy from manual contour inter-observer variability reported for lung cancer OARs [100]. This suggests that the mean dose differences reported here for all OARs except the bowel bag are likely to be similar to those arising from inter-observer variability in manual contouring. Future work could validate this by investigating the dosimetric impact of OAR contouring variability in prostate patients. This suggests that these contours would be clinically acceptable, however only the femoral heads had all patients with differences $< 1.0$ Gy or $< 1$ percentage point, suggesting that manual clinical review would be required for the other contours.

There was no correlation with clinical rating in dose or volume differences for any of the OARs for the prostate patients except for the bladder, where there were larger volume differences for two patients ($< -2\%$), which did correlate with the two contours rated 'intermediate' (see figure 2.16). This is an interesting result as both clinical rating and dose differences have been reported as clinically relevant parameters to evaluate automatic contours on, and so the lack of correlation between them is surprising. A partial explanation for this might be that the dosimetric analysis is sensitive to differences that are not clinically relevant. For example the bladder and rectum differences were largest for the low dose volumes (V30 Gy), suggesting the contours were accurate in the high dose regions near the prostate target. This could explain the largest relative volume difference for the rectum occurring in a patient rated 'accept'. In contrast, the bowel bag highest differences were in the dose to the hottest 0.5 cm$^3$ volume. This was likely due to discrepancies between the manual and automatic contours in how inferiorly the contour extended. Because the inferior edge of the bowel bag was close to the prostate PTV, differences in inferior extension of a few mm could result in large dose differences due to moving down a steep dose gradient. This highlights that some large dose differences may be caused by small geometric differences, and that these large dose differences may still not be clinically relevant. This may partially explain why there was a lack of correlation between dose differences and clinical rating.

The mean optimisation dose and relative volume differences were $\leq 0.7$ Gy and $\leq 0.5$ percentage points for all OARs, suggesting the automatic contours were accurate enough to optimise treatment plans with. The optimisation differences compared doses to the manual contours from the plan based on automatic contours (plan$_{\text{auto}}$contour$_{\text{man}}$) and the plan based on manual contours (plan$_{\text{man}}$contour$_{\text{man}}$). So the contour is the same but the plan is different. The mean optimisation dose differences were smaller than the corresponding mean evaluation dose differences for all OARs except the femoral heads. This suggests that all OARs were sufficiently accurate to optimise a treatment plan with but not all were accurate enough to evaluate it. Automatic contours that are sufficient for optimisation but not evaluation have limited clinical relevance because a treatment plan optimised using automatic contours would still need to be evaluated using manual contours, and therefore there would be no time-saving from using automatic contours. The increase in the dose differences to the femoral heads on optimisation compared to evaluation is likely due to inter-plan variation, since the femoral head contours were all similar. Inter-plan variation was attempted to be minimised by using the same set of optimisation parameters used for both automatic and manual contour plans, but there is intrinsic variability in the optimisation process with multiple optimal treatment plans possible [211].

For the ano-rectal patients, the mean evaluation dose differences were similarly small, $< 1.0$ Gy for all except the bowel bag D180 cm$^3$ (table 2.8). Again the optimisation dose differences were $\leq$ to the evaluation dose differences for all OARs, except the left femoral head D35%. Dose differences tended to be larger for the larger volume constraints, especially in the bowel bag, prostate and prostate & seminal vesicles. There were some correlations with clinical rating, with dose differences for contours rated 'delete' or 'major' being larger than contours rated $\geq$ 'intermediate' for the bowel bag, prostate and urethra (figure 2.9). However, dose differences to the bladder and prostate & seminal vesicles patients rated 'major' were similar to dose differences for patients rated as clinically acceptable.

The bowel bag had much smaller dose differences for the ano-rectal patients than the prostate patients, despite the distribution of clinical ratings being similar (figure 2.16). This is likely due to the ano-rectal PTVs extending considerably superior to the prostate PTVs (figure 2.9) leading to larger volumes of the bowel bag receiving higher doses and less steep dose gradients, which reduces the impact of small geometric discrepancies in contours. This highlights the influence of the particular treatment site being used when assessing the dosimetric impact of automatic contours. It also potentially explains the lack of correlation between DVH differences and clinical rating, since the rating was of the suitability of the OAR contour for all possible pelvic radiotherapy treatments, whereas the DVH differences were just for the prostate/ano-rectal plans investigated.

Comparing these mean dose differences to those arising from inter-observer variability in lung cancer OARs suggests that the bladder femoral heads, penile bulb, prostate, prostate & seminal vesicles and urethra would all be clinically acceptable for ano-rectal radiotherapy. Future work could validate this by calculating the dose impact of inter-observer variability in manual OAR contours for ano-rectal radiotherapy. Similar to the prostate patients, all OARs except femoral heads had some patients with larger differences, implying manual review of contours on a patient-by-patient basis would be warranted.

Only a few studies have investigated the dose impact of automatic OAR contouring. Cao et al. reported evaluation DVH differences in 15 prostate patients using automatic compared to manual contours [97]. Mean differences in relative volume for the bladder V40 Gy were 1.5% (estimated from bar chart) compared to $-0.4\%$ in this study. Mean differences in rectum V32 Gy were 7% compared to 0.9% for the V30 Gy in this study, both showing significant improvements for the automatic contouring algorithm evaluated here. They also reported no correlation in delineation metrics (including distance metrics) with dose differences, as found in this study. Zabel et al. reported mean optimisation dose differences in DVH parameters for rectum and bladder automatic contours $\leq 0.2$ Gy for 15 prostate cancer patients. This is smaller than differences reported here ($\leq 0.5$ percentage points in relative volume), however, the dose differences were assessed after manual correction of the automatic contours which would reduce dose differences to a minimum.

Evaluations of automatic contouring in other sites also showed comparable or larger dose differences to those reported in this study. Vaassen et al. assessed the dosimetric impact of automatic versus manual OAR contouring and of inter-observer variations in manual contours in non-small cell lung cancer patients [100]. They reported mean dose differences $< 1.0$ Gy for automatic compared to manual contours, which was comparable to differences due to inter-observer variability in manual contours. This is very similar to the results reported here for all OARs except the bowel bag. Another study reported evaluated head and neck OAR automatic contours using the same clinical PTV for both manual and automatic contour plans in 15 patients [98]. Mean optimisation dose differences were $\leq 1.3$ Gy for all OARs except two, which were statistically significantly different 1.4 Gy and 2.2 Gy. These optimisation dose differences were slightly larger than those reported in this study ($\leq 0.7$ Gy).

Determining whether the automatic contours are sufficiently accurate for clinical use is difficult, since there are no established criteria in the literature. Using a limit of 1 Gy from Vaasseen et al., the bladder, femoral heads, penile bulb, prostate, prostate & seminal vesicles and urethra would all be suitable for clinical implementation, subject to manual review and correction. However, the use of manual review and modification by clinicians suggests that the clinical rating of each organ is the most important. On this basis, the prostate and prostate & seminal vesicle contours would not be considered clinically acceptable, since these had median scores below the clinically acceptable threshold.

This study had some limitations. These included only using one rater for the clinical rating, who was different to the clinicians who provided the manual contours used in the quantitative contouring evaluations. This may have introduced discrepancies in the correlations of clinical rating with dose differences. Future work could investigate using several raters to reduce the subjectivity of the result and evaluate the inter-observer variability in rating. The clinicians could also modify the automatic contour if required, and dose differences between modified and unmodified contours evaluated to investigate correlations with ratings. A limitation of the dosimetric evaluation was that the optimisation parameters were fixed for both automatic and manual contour plans. This was done to minimise operator bias, but did mean that each plan was not optimised as tightly to the contour set as it could have been, which could have under-estimated the impact of contouring differences in the optimisation evaluation. A solution to this would be to use automatic planning algorithms, which are becoming more widely available and clinically implemented [212]. Finally, evaluations using larger patient numbers would be able to assess the correlations between dose differences and clinical ratings more fully.

## 2.4 Conclusions

A ZTE-based Deep Learning algorithm successfully generated sCTs for all patients. Mean dose differences to the PTV D98% in the comprehensive dose analysis were $\leq 0.5\%$ for all FH points within the true external contour. For the clinical plans PTV D98% dose differences were similarly small, $\leq \pm 0.6\%$, with mean gamma pass rates at the stringent criteria of 1%/1 mm being $97.7 \pm 0.5\%$, $94.8 \pm 3.1\%$ and $94.7 \pm 1.0\%$ for the prostate, rectum and anus patients respectively. These mean dose differences are $< 1\%$, ensuring there is not a significant increase in overall dosimetric uncertainty in the radiotherapy pathway, and agree well with other results presented in the literature as appropriate for clinical use.

Automatic MR-based contours were successfully generated for all patients for all OARs except penile bulb (one failed patient) and urethra (two failed patients). Median clinical ratings for all OARs except bowel bag, prostate and prostate & seminal vesicles were clinically acceptable, although only for the femoral heads and body contours were all patients considered acceptable. Mean DVH differences were $< 1.0$ Gy or 1 percentage point for bladder, femoral heads, penile bulb, rectum and urethra for both prostate and ano-rectal treatment plans. These differences are similar to or less than those reported in the literature. This suggests the algorithm is sufficiently accurate for clinical use for these OARs for prostate and ano-rectal radiotherapy plans, subject to manual review and modification prior to treatment planning.

A combination of accurate sCT generation and MR-based automatic contours enables a streamlined MR-only radiotherapy pathway to be implemented, using a single patient imaging session to generate high quality images for manual tumour delineation, automatic OAR contours and sCT for treatment plan optimisation. This reduces geometric treatment uncertainties without significantly increasing dosimetric uncertainties, improving the accuracy and efficiency of the pelvic radiotherapy pathway.

# Chapter 3

# CBCT for Dose Calculation QA for MR-only Radiotherapy

## 3.1 Introduction

The implementation of MR-only radiotherapy into the clinic depends not only on a sCT that provides accurate radiotherapy dose calculations, but also on a method that provides assurance of the sCT dose accuracy on a per-patient basis. A number of sCT algorithms have been evaluated in the literature (see chapter 1 section 1.2.4 and chapter 2), with commercially available solutions currently being used clinically in the treatment of prostate cancer [81, 213].

These sCT algorithms have demonstrated high dose calculation accuracy [53,67], yet there may be situations where they fail to generate accurate tissue densities. These situations include artefacts in the MR image, particularly those that affect the patient external contour such as motion and phase wrap artefacts, and for atlas-based algorithms patients substantially larger than the atlas patients [69]. In addition, MR images can suffer from geometric distortion [166], which can vary depending on the patient and scanning parameters [214]. Dose accuracy depends on both the correct assignment of tissue densities and the geometric accuracy of the image. The magnitude of the error introduced would be very variable, but in principle could be clinically significant. For example, failure to apply vendor 3D geometric distortion can increase geometric distortions at the patient external contour from $< 2$ mm to nearly 8 mm [215], which could produce a dose difference of $\sim 4\%$ in a 6 MV beam. Dose uncertainties of this magnitude would be more problematic than the removal of the 2 mm MR-to-CT registration uncertainty [25] that MR-only enables.

This means ongoing QA of sCT dose calculation accuracy is important for MR-only pathways [32]. The standard method for evaluating the dose accuracy of sCTs has been recalculating radiotherapy treatment plans on a CT image of the same patient and comparing the dose distributions. But patients on clinical MR-only pathways will not have CTs and so an alternative methodology needs to be developed for ongoing dose accuracy QA of sCTs.

Independent monitor unit check calculations are a well-established part of the radiotherapy workflow. These aim to check the dose calculation using an independent dose algorithm modelled with independent beam data, increasing the likelihood of detecting errors in the treatment plan before the patient starts treatment [216]. However the aim of the check calculations is to check the dose calculation algorithm, not the accuracy of the image used for that dose calculation. So all independent monitor unit check calculations use the patient geometry as defined by the planning image and therefore will not detect geometric inaccuracies. In addition, algorithms that use the HU from the planning image will not detect inaccuracies in tissue density assignment in the sCT. Therefore conventional independent monitor unit check methods are not sufficient for ongoing dose accuracy QA of a MR-only pathway.

Edmund et al. proposed using the first-fraction CBCT as a QA tool for MR-only radiotherapy [217]. A CBCT needs to have a calibration curve to convert CBCT HUs into relative electron densities/mass densities needed for radiotherapy dose calculations, in a similar way to CT. However CBCT images have a much greater variation in photon scatter between patients, leading to a variable relationship between CBCT HU and tissue density [218]. This is a potential issue for using dose calculations on CBCT as a QA tool for sCT accuracy. However Edmund et al. demonstrated that using a population-based calibration curve gave good agreement between CT and CBCT relative electron densities for six brain patients, suggesting that CBCT could be used to evaluate the dose accuracy of sCT [217]. This methodology was then retrospectively evaluated on 10 prostate patients with sCT-CBCT dose differences agreeing within 1% of gold standard sCT-CT dose differences [219]. An alternative method of calculating radiotherapy doses on CBCT has been developed using patient-specific thresholds to segment the CBCT into six tissue classes, which are then applied population bulk densities [220]. The aim of this study was to extend the comparison of sCT-CBCT and sCT-CT dose evaluations using this dose calculation method and to prospectively evaluate dose accuracy QA using CBCT in a clinical MR-only radiotherapy pathway.

## 3.2 Materials and Methods

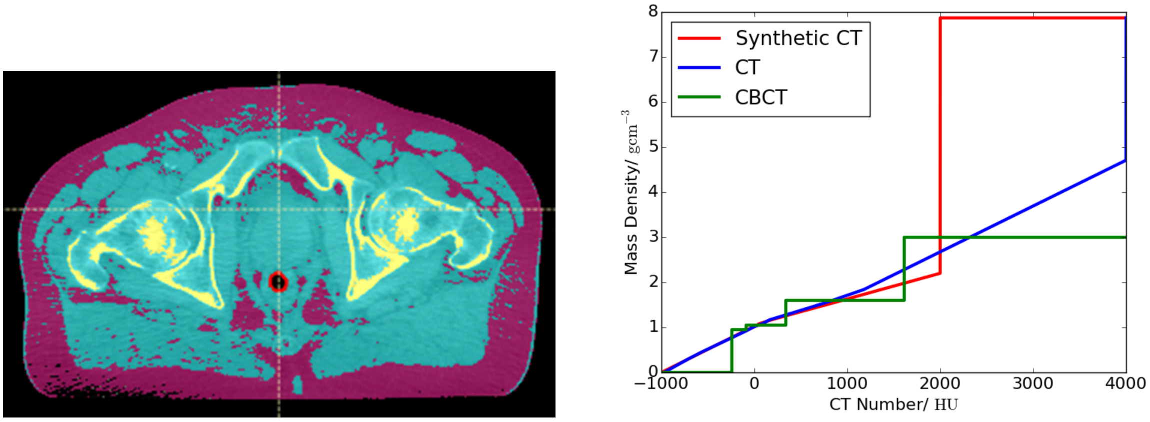
### 3.2.1 Patient Data Collection

A total of 49 patients treated with MR-only radiotherapy for prostate cancer at the Northern Centre for Cancer Care, Newcastle upon Tyne, UK were included in this study. The consent for radiotherapy treatment included consent for data to be used for research purposes. Patients were divided into two cohorts: the first 20 patients (Cohort 1) and the remaining 29 patients (Cohort 2). All patients were treated with prostate and seminal vesicle radiotherapy only, with no nodal irradiation.

All patients received a radiotherapy planning MR (1.5 T Magnetom Espree, Siemens, Erlangen, Germany) performed on a flat couch top with local standard prostate radiotherapy immobilisation. Patients in Cohort 1 also received a back-up CT (Sensation Open, Siemens) whereas Cohort 2 patients didn't. Prior to each scan and treatment fraction patients underwent routine bladder and bowel preparation, consisting of the application of a micro-enema 60 minutes prior to the scan, bladder and bowel emptying at 30 minutes prior and drinking 400 ml of water. The MR images were acquired using a 6 channel flexible receive coil (Siemens Body Matrix) supported over the patient by an in-house manufactured coil bridge and the 24 channel spine receive coil contained in the couch (Siemens Spine Matrix).

The MR images were acquired with a T2-weighted 3D turbo spin echo SPACE (Sampling Perfection with Application optimised Contrasts using different flip angle Evolution) sequence with a field of view of $450 \times 450 \times 180$ mm$^3$, covering the patient external contour. Geometric distortion was minimised through using a bandwidth of 601 Hz Pixel$^{-1}$ and applying the Siemens 3D distortion correction algorithm. Measurements with a GRADE phantom (Spectronic Medical, Helsingborg, Sweden) [167] found 99% of phantom markers within the sequence field of view with distortion $D < 2.0$ mm. The sCT images were generated from the MR images using Mriplanner (prostate model version 1.1.7, Spectronic Medical) [55]. The Cohort 1 CT images were acquired with a voxel size of $1.1 \times 1.1 \times 3$ mm$^3$ and a tube voltage of $V = 120$ kVp.

All patients were planned with a 6 MegaVoltage (MV) single 360$^o$ volumetric modulated arc therapy treatment plan optimised on the sCT delivering a prescription dose of 60 Gy in 20 fractions to 50% of the central PTV [221] in Raystation (version 7, RaySearch Laboratories, Stockholm, Sweden). A sCT-specific HU to mass density curve provided by Spectronic Medical was used for dose calculations (see figure 3.1). All dose calculations were made with the same beam model using the RayStation collapsed cone algorithm, which calculates dose-to-water. All patients received daily kiloVoltage (kV) CBCT imaging using a TrueBeam STx (version 2.7 MR3, Varian Medical Systems, Palo Alto, USA), with a voxel size of $0.9 \times 0.9 \times 2$ mm$^3$, a tube voltage of $V = 125$ kVp and a field of view of 46.5 cm. CBCT images were soft-tissue matched to the planning MR image by treatment radiographers for on-treatment image guidance [222]. This involved an automatic rigid registration between the CBCT and MR images, followed by a manual adjustment to ensure the prostate and seminal vesicles target as visually assessed from the CBCT were included within the PTV delineated on the MR. All patients were then shifted to the soft-tissue match position and treated.



(a) Threshold

(b) Calibration Curves

Figure 3.1: a) Automatic threshold of the CBCT into air (black), adipose (purple), tissue (blue) and cartilage/bone (yellow) (left) and the outlined air in the rectum (red), which was set to $\rho = 1.0$ gcm$^{-3}$, for a representative patient. b) Plot of CBCT voxel value to mass density curve for the same patient, as well as the Hounsfield Units to mass density curves for the CT and sCT.

### 3.2.2 Dose Calculations on CBCT and CT

The first-fraction CBCT was imported in RayStation and registered to the sCT using the registration matrix from the online treatment match. The treatment plan was recalculated on the CBCT using the patient-specific step-wise HU to mass density curve available in RayStation. This converted the CBCT image into six tissue classes using automatically determined patient-specific HU thresholds and assigned the following bulk mass densities: air - 0.00121 gcm$^{-3}$, lung - 0.26 gcm$^{-3}$, adipose - 0.95 gcm$^{-3}$, tissue - 1.05 gcm$^{-3}$, cartilage/bone - 1.6 gcm$^{-3}$, and other - 3.0 gcm$^{-3}$ (see example in figure 3.1) [220]. These thresholds were reviewed for each patient and the adipose - tissue threshold manually adjusted in $<$ 10 patients. Dose differences to CT with this method for the pelvis has been reported as 0.2 ± 1.6% (mean ± standard deviation) [220]. The body outline on the CBCT was automatically outlined using a threshold based automatic body contour in RayStation. Any air in the rectum was outlined and assigned unit density since this process was included in the sCT generation process. The treatment plan was recalculated on the CBCT keeping the monitor units, dose grid voxel size and dose grid position the same.

For the Cohort 1 patients the back-up CT was rigidly registered to the sCT using the automatic mutual information algorithm with six degrees of freedom focused on the PTV in RayStation. A HU to mass density curve derived from data measured on the CT scanner was applied (figure 3.1) and the treatment plan recalculated on the CT with the same monitor units, dose grid voxel size and dose grid position. Any air in the rectum was outlined and assigned unit density.

### 3.2.3 Dose Evaluation

For both patient cohorts the doses calculated on CBCT and sCT were compared using differences in isocentre dose and a 3D global gamma analysis, similar to those described in chapter 2 (see section 1.2.3 for a fuller background to these methods). The percentage difference in isocentre dose was calculated using

$$\Delta D_{CBCT} = 100\frac{D_{CBCT} - D_{sCT}}{D_{prescription}}, \tag{3.1}$$

where $D_{CBCT}$ was the dose at the isocentre for the CBCT, $D_{sCT}$ was the dose at the isocentre for the sCT and $D_{prescription}$ the prescription dose. In addition the PTV from the sCT was copied onto the CBCT without modification and the difference in dose to the PTV mean dose, near maximum (D2) and near minimum (D98) were calculated as a percentage of the prescription dose [10].

A gamma analysis was performed comparing the dose calculated on the sCT to the CBCT using the Medical Interactive Creative Environment Toolkit (version 1.0.8, Umeå University, Sweden) [204]. Separate gamma analyses were carried out within the external contour and the volume enclosed by the 50% isodose line, using 1% global dose difference of the prescription dose (60 Gy) and 1 mm distance-to-agreement, and 2%/2 mm criteria. All points below 10% of the prescription dose were excluded.

In addition, the 6 MV radiological water equivalent isocentre depth was calculated at $5^o$ angles for each image in the isocentre plane in the same way as described in chapter 2 [55, 56]. For the CBCT the radiological isocentre depth was calculated using density over-rides for any air in the rectum. For each patient the difference in radiological and physical isocentre depth (CBCT - sCT) at each gantry angle was measured and the mean difference over all gantry angles was calculated. The physical isocentre depth difference was a measure of external contour differences between the images.

For patient Cohort 1, the same dose evaluation methodology was also applied between the sCT and the CT.

### 3.2.4 Data Analysis

Firstly, Cohort 1 sCT-CT and sCT-CBCT dose differences were compared. Bland-Altman plots of isocentre dose differences and mean physical and radiological isocentre depth differences were generated and the 95% limits of agreement calculated [223].

Secondly, sCT-CBCT data from Cohort 1 was used to generate QA tolerance levels. Only Cohort 1 data was used since the dose accuracy of each patient's sCT had been demonstrated through dose differences with the CT. The 95% confidence interval of the sCT-CBCT isocentre dose difference was calculated and rounded to generate clinical tolerance levels.

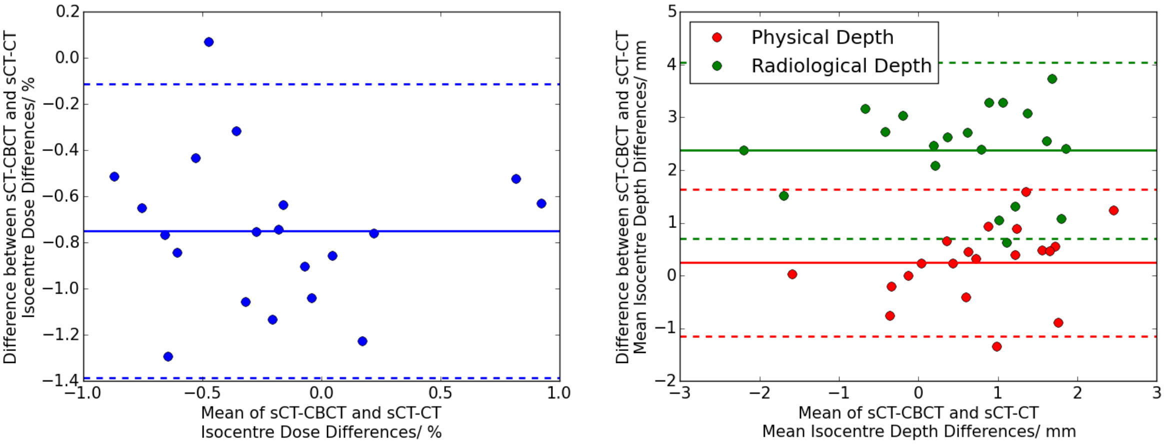
Thirdly, sCT-CBCT data from Cohort 2 was evaluated to determine if any patients were outside these tolerance levels, and the cause investigated.

Finally, sCT-CBCT data from cohorts 1 and 2 were evaluated to characterise two factors which might have impacted the results. Firstly, the CBCT dose calculation was carried out at the patient's treated position, a manual adjustment from the optimum MR-CBCT registration (soft-tissue match). To assess the impact of this, the magnitude of the vector shift between the automatic and soft-tissue match positions was calculated for each patient and correlated with the absolute isocentre dose difference. Secondly, the time from MR scan to first-fraction CBCT (3-5 weeks) may have increased the probability of the patient external contour changing (through weight gain or loss). Therefore this time was recorded and the correlation in absolute physical isocentre depth difference (a measure of patient contour change) and absolute isocentre dose difference with time calculated. There was also a time gap between CT and MR, however this gap was only 1-3 days and so any weight changes were considered negligible.

## 3.3 Results

### 3.3.1 Dose Comparisons sCT-CT vs sCT-CBCT

The sCT-CBCT isocentre dose differences in Cohort 1 were lower than the sCT-CT differences by $-0.7 \pm 0.6$ % (mean ±95% limits of agreement, see figure 3.2). There were minimal differences between the mean sCT-CT and sCT-CBCT isocentre physical depth differences, $0.2 \pm 1.4$ mm, suggesting that the CBCT was geometrically similar to the CT. However, there was a substantial difference in mean sCT-CT and sCT-CBCT isocentre radiological depth differences of $2.4 \pm 1.7$ mm.



(a) Isocentre Dose Difference

(b) Isocentre Depth Difference

Figure 3.2: a) Bland-Altman plot of the difference in sCT-CT and sCT-CBCT isocentre dose differences as a function of the mean of the sCT-CT and sCT-CBCT isocentre dose differences. b) Bland-Altman plot comparing mean physical and radiological isocentre depth differences for sCT-CT and sCT-CBCT (red and green respectively). In all plots the circles show the data points, the solid line the mean difference and the dashed line the 95% limits of agreement. Both plots show data from Cohort 1 only.

The sCT-CBCT gamma pass rates were lower than the sCT-CT pass rates, especially for the 1%/1 mm gamma criteria (see figure 3.4). The mean sCT-CBCT gamma pass rate within the body contour for Cohort 1 was $85 \pm 1\%$ (± standard error of the mean, range 75%, 94%), compared to $98.4 \pm 0.2\%$ (95.6%, 99.4%) for the sCT-CT.

### 3.3.2 Tolerance Levels for sCT-CBCT Isocentre Dose Difference

The mean sCT-CBCT dose difference for Cohort 1 was $\Delta D_{CBCT1} = -0.6 \pm 0.1\%$ , ($-1.3\%$, 0.6%). The 95% confidence interval on the mean was [$-1.5\%$,0.4%]. This was rounded to produce asymmetric tolerance levels of [$-2\%$,1%]. The mean differences in PTV mean dose, D2 and D98 were $-0.6 \pm 0.1\%$, $-0.6 \pm 0.1\%$ and $-0.8 \pm 0.1\%$ respectively.

The equivalent results for Cohort 2 were $\Delta D_{CBCT2} = -0.6 \pm 0.1\%$ ($-2.3\%$, 2.3%) for the isocentre dose difference, and $-0.7 \pm 0.1$, $-0.6 \pm 0.2$ and $-0.8 \pm 0.1$ for the mean differences in PTV mean dose, D2 and D98 respectively. Only 2/29 patients were outside the proposed tolerance levels.

### 3.3.3 Evaluation of sCT-CBCT Dose Differences

The mean sCT-CBCT dose differences across both cohorts was $\Delta D_{CBCT1\&2} = -0.6 \pm 0.1\%$ ($-2.3\%$, 2.3%). The 95% confidence interval across both cohorts was [$-1.9\%$,0.7%] which fitted well with the proposed tolerance levels. The CBCT appears to be systematically lower than the sCT dose, with negative dose differences in 43/49 patients (see figure 3.3).

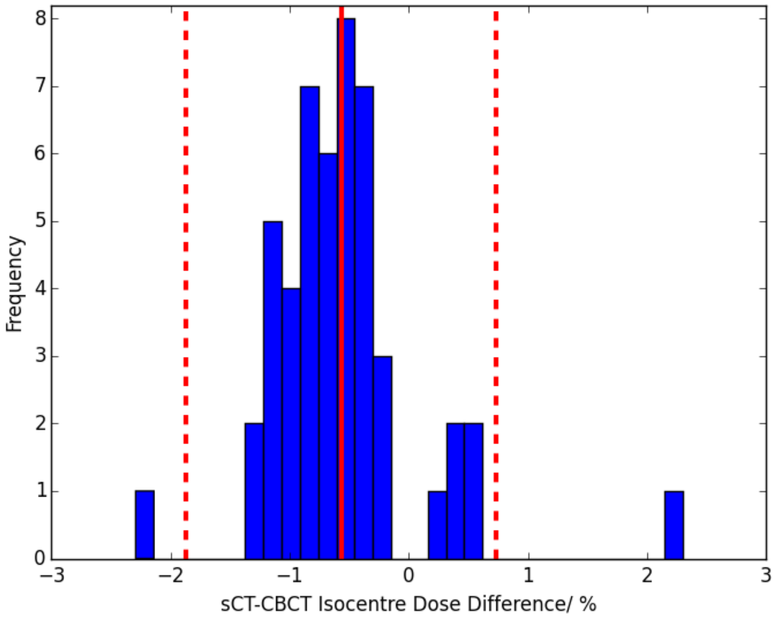


Figure 3.3: Histogram of sCT-CBCT isocentre dose differences (CBCT- sCT) for both patient cohorts. The red vertical line indicates the mean, and dashed lines the 95% confidence interval.

The gamma analysis showed reasonable agreement between sCT and CBCT with a mean gamma pass rate across both cohorts within the external contour with gamma criteria

1%/1 mm of $86.4 \pm 0.7\%$ (74.5%, 93.6%) and at 2%/2 mm of $96.1 \pm 0.4\%$ (85.4%, 99.7%). The interquartile range of the sCT-CBCT gamma passes overlapped substantially between cohorts 1 and 2 (figure 3.4).

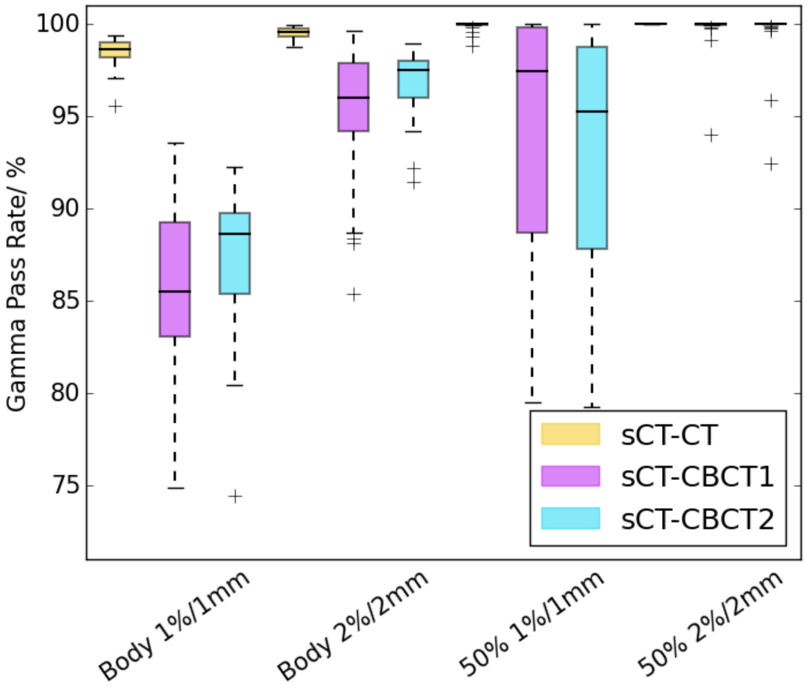


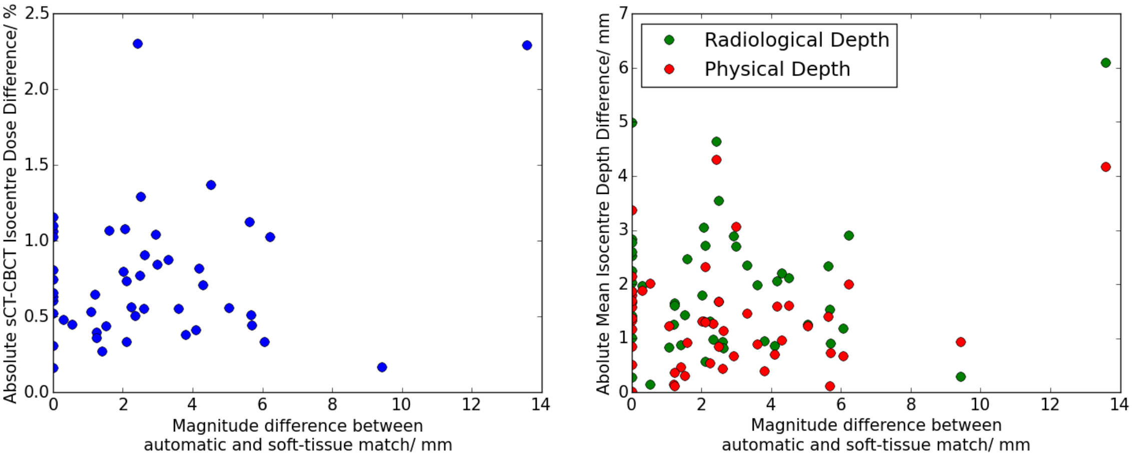
Figure 3.4: Boxplot showing gamma pass rates with criteria 1%/1 mm and 2%/2 mm within the external contour (Body) and the volume enclosed by the 50% isodose line (50%) for sCT-CT (yellow), Cohort 1 sCT-CBCT (purple) and Cohort 2 sCT-CBCT (blue) . The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points. One outlier for the 50% isodose line 1%/1 mm results for sCT-CBCT cohort two has been omitted (value 61.5%).

The mean difference in mean radiological isocentre depth was $\Delta d_{Rad} = 1.6 \pm 0.2$ mm (−4.6 mm, 6.1 mm). The mean difference in mean physical isocentre depth was similarly small, $\Delta d_{Phys} = 0.9 \pm 0.2$ mm (−4.3 mm, 4.2 mm).

There was a weak correlation between the magnitude of the vector shift between automatic and soft-tissue match and the absolute sCT-CBCT isocentre dose difference, with Pearson's $r = 0.3$ (see figure 3.5). This correlation was approaching statistical significance ($p = 0.06$). There was no statistically significant correlation with absolute mean physical or radiological isocentre depth difference (Pearson's $r = 0.2$, $p = 0.23$ and $r = 0.2$, $p = 0.18$ respectively). There was no correlation between time from MR to first-fraction and change in physical isocentre depth or isocentre dose difference (Pearson's $r = 0.1$, $p = 0.51$ and $r = -0.1$, $p = 0.40$ respectively).

## 3.4 Discussion

This study has evaluated using the first-fraction CBCT as a method of dose accuracy QA for a prostate MR-only pathway. Compared to gold standard CT, CBCT has a small



(a) Dose Difference

(b) Depth Difference

Figure 3.5: The correlation between the magnitude of the vector shift between automatic and soft-tissue match and absolute sCT-CBCT isocentre dose difference (left) and absolute mean sCT-CBCT isocentre depth difference (right) for both cohorts.

but systematic shift in isocentre dose difference of −0.7 %. This was also reflected in a mean sCT-CBCT isocentre dose difference of $\Delta D_{CBCT1\&2} = -0.6 \pm 0.1\%$. There was good agreement in sCT-CBCT gamma pass rates with gamma criteria 2%/2 mm and no correlation between sCT-CBCT isocentre dose difference and the magnitude of soft-tissue match shift or time between MR and first-fraction CBCT.

The sCT-CBCT isocentre dose difference appeared to be systematically lower than the sCT-CT difference, with the 95% limits of agreement not encompassing zero difference (see figure 3.2). This systematic under-estimation comes from the difference in radiological isocentre depth (mean difference $2.4 \pm 1.7$ mm) rather than physical (mean difference $0.2 \pm 1.4$ mm), suggesting it is the CBCT density assignment, rather than the geometrical accuracy, which caused the dose difference. This is consistent with the high geometric accuracy of CBCT, within 1 mm [224]. Dose calculations on CBCT images are known to be less accurate due to the variable photon scatter causing a variable relationship between HU and tissue density [218]. Different methods of dose calculation have been proposed, including applying standard CT, site-specific or patient-specific HU-density curves, bulk density over-rides or deformable registration of a planning CT [218]. Bulk density methods in prostate patients have reported CBCT dose differences to CT of $0.2 \pm 1.3\%$ (mean ± standard deviation) [220] and 0.4% (median, min 0.0%, max 0.4% [218]. The magnitude of the difference observed in this study was slightly larger but agreed within the 95% limits of agreement (±0.6%).

The lower CBCT dose resulted in worse gamma pass rates than sCT-CT (figure 3.4). The errors in density assignment for the CBCT may also have reduced the gamma pass rates. There was still very good agreement with gamma criteria 2%/2 mm ($96.1 \pm 0.4\%$ within the body contour), which is commonly used in the MR-only literature [53,67] and is likely sufficient for QA purposes. The key question is whether the systematic dose difference due to uncertainties in CBCT dose calculations is small enough that the CBCT is sufficiently accurate for QA of the sCT. This depends on both the magnitude of the mean difference and the spread of differences given by the 95% confidence interval.

The asymmetric 95% confidence interval on the mean sCT-CBCT isocentre dose difference ([$-1.5\%$,0.4%]) from Cohort 1 lead to the adoption of asymmetric dose tolerances of [$-2.0\%$,1.0%]. Applying these to Cohort 2 resulted in 27/29 patients with isocentre dose differences within these tolerances. The two outlier patients both had substantial changes in the patient external contour (mean physical isocentre depth difference $|\Delta d_{Phys}| >$ 4 mm) which were readily observable on the CBCT image, due to weight gain in one instance and different patient posture (clenched buttocks) in the other. This suggests that the CBCT method with tolerances of [$-2.0\%$,1.0%] can accurately detect geometric differences between sCT and CBCT, such as may be caused by geometric distortion in the MR used to generate the sCT. This indicates that though the sCT-CBCT dose differences were larger than sCT-CT dose differences, the mean difference and spread of differences were sufficiently small for CBCT to be used for QA of the sCT. Although in this case both results were false positives, in clinical practice doses differences outside of these tolerances at the start of treatment would be cause for concern and should be investigated, monitored over several fractions and if necessary replanned. This implies that the suggested asymmetric dose tolerance would be appropriate for clinical use.

Only one other paper to our knowledge has investigated sCT-CBCT dose differences, reporting mean sCT-CBCT differences in PTV mean dose of $-0.8 \pm 0.6\%$ (CBCT - sCT, ± standard deviation) [219]. This agrees with the results given here ($-0.64 \pm 0.09\%$ across both cohorts) with the CBCT dose being lower than the sCT dose. They also found the sCT-CBCT dose difference to be larger than the sCT-CT dose difference, although only by 0.3% rather than 0.7% reported here. This may be due to the fact that both CT and CBCT images were calibrated using the same HU to relative electron density curve. This has been reported to improve the mean agreement in dose calculations between CT and CBCT but also increase the standard deviation of differences and the number of outliers [218, 225].

A potential confounding factor in this study was the use of the soft-tissue match registration between sCT and CBCT which means the sCT-CBCT registration used for treatment was not necessarily the optimal registration between the two image sets, but had been modified to ensure that the prostate and seminal vesicles were covered by the PTV. For example, if a patient has an identical posture but a larger rectum on CBCT compared to the planning MR, the automatic MR-CBCT registration would result in good alignment of the bone but the prostate will be shifted anteriorly by the rectum, and so will only be partially covered by the PTV. The soft-tissue match will apply a corresponding posterior couch shift to bring the prostate back into the planned PTV, however this will result in the bones now being partially misaligned, introducing a small dose difference between the planned (sCT) dose and that calculated on the CBCT. This is the optimal strategy for treatment because it ensures that the prostate and/or OARs are not under- or over-dosed respectively. However it does introduce a confounding factor for QA of the sCT since dose differences may be due to the soft-tissue match rather than errors in the sCT. The soft-tissue match can be up to 10 mm different in each direction compared to the automatic MR-CBCT registration, although for the vast majority of patients the differences are $\leq 5$ mm.

This confounding factor could have been avoided through using the original automatic MR-CBCT registration. However the only registration matrix stored on the linear accelerator was the treatment position (the soft-tissue match). The radiographers manually recorded the automatic registration position, but to use this position for the sCT-CBCT dose QA process would have required additional manual steps in the process and precluded full automation. The use of soft-tissue matching appears not to have had a substantial impact, with a small correlation (Pearson's $r = 0.3$) between absolute isocentre dose difference and magnitude of the vector shift difference between the automatic and soft-tissue matches. Although this correlation was approaching statistical significance ($p = 0.06$), this was due to the two outlier patients described above. If they are excluded, the correlation disappears (Pearson's $r = 0.0$, $p = 0.79$). This highlights the dosimetric robustness of MR-CBCT soft-tissue matching and suggests there is only an issue when the patient external contour is substantially different between CBCT and sCT.

Another potential method to avoid the confounding impact of the soft-tissue match would have been to use a deformable registration between CBCT and sCT to ensure the external contours were identical. However, a deformable registration would mean that the geometry of the CBCT would not have been maintained, removing one of the main rationales for using the CBCT as a QA tool, namely the potential geometric distortion in the MR used to generate the sCT.

Another potential clinical concern for using CBCT for dose accuracy QA for MR-only radiotherapy is that patients may gain or lose weight between planning MR and first-fraction CBCT. This would introduce a confounding dose change, which may be dependent on the length of time between MR and first fraction since longer times would increase the probability of weight change. However, there was no significant correlation in time between physical isocentre depth or isocentre dose difference, suggesting this is not an issue within the period evaluated in this study (maximum 34 days).

A practical consideration of using CBCT for dose accuracy QA for MR-only radiotherapy is the resource implications of recalculating the treatment plan. However, the recalculation process presented here was highly automated utilising scripts within the treatment planning system and took less than 10 minutes.

Future work will look to characterise the clinical effectiveness of the method further by determining false positive and false negative rates. This could be through generating deliberate errors such as not applying the 3D distortion correction post-processing to the MR image before generating a sCT and seeing the impact on sCT-CBCT dose differences. Alternatively, retrospective MR images with potential issues (eg significant image artefacts) could be used to generate sCTs and sCT-CT and sCT-CBCT dose differences compared. An investigation procedure for out-of-tolerance results should also be developed and evaluated. This could include visual inspection of the sCT and CBCT, examining the sCT-CBCT registration and the MR used to generate the sCT for potential issues (eg 3D distortion correction applied, appropriate receive bandwidth used, and presence and position of image artefacts). If the out-of-tolerance investigation identifies an issue with the sCT or MR then the patient should be rescanned and replanned.

## 3.5 Conclusions

The first-fraction CBCT appears a promising method for dose accuracy QA of sCT in a MR-only prostate radiotherapy pathway, with a high sCT-CBCT gamma pass rate with 2%/2 mm criteria. There was a small systematic difference in dose between sCT and CBCT, suggesting that asymmetric dose tolerances of [$-2.0\%$,1.0%] would be appropriate clinically. There was no correlation between sCT-CBCT isocentre dose difference and the magnitude of soft-tissue match shift or time between MR and first-fraction CBCT. sCT dose accuracy QA using the first-fraction CBCT would enable departments to safely implement MR-only radiotherapy without the need for back-up CTs.

# Chapter 4

# Evaluating the Image Quality of Combined PET-MR Images Acquired in the Pelvic Radiotherapy Position

## 4.1 Introduction

Combined PET-MR scanners have great potential for improving radiotherapy with molecular PET information obtained simultaneously with functional and anatomical MR information [29]. In particular PET-MR images may facilitate radiotherapy dose painting through identifying active tumour sub-volumes to receive radiotherapy 'boost' doses [23]. It is important that the delineation of active tumour sub-volumes is robust and repeatable which requires accurate, quantitative imaging [27].

Radiotherapy planning images need to be acquired in the radiotherapy position for accurate treatment and registration with other planning images [154]. Acquiring PET-MR images for pelvic radiotherapy planning therefore requires patients to be scanned on a radiotherapy flat couch-top which mimics the treatment machine couch, with patients in appropriate radiotherapy immobilisation devices and with the MR receive coils supported away from the patient so that the patient external contour is not deformed [154]. The carbon fibre couches typically used for PET-CT imaging have low PET attenuation but produce significant MR artefacts, whereas glass fibre MR couches do not interfere with the MR signal but significantly attenuate the 511 keV photons detected in PET [155]. This means dedicated PET-MR radiotherapy hardware needs to be developed that is MR-compatible and has low PET attenuation [156].

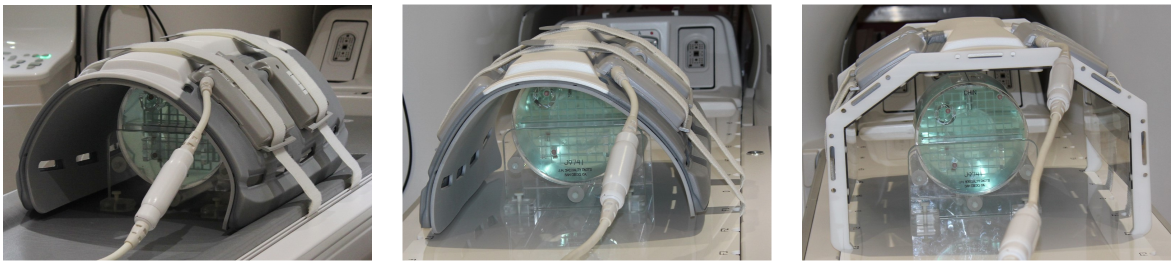
Acquiring PET-MR images in the radiotherapy position will have an impact on MR image quality [157] since the receive coils will be further from the patient anatomy, reducing the coil filling factor and therefore the SNR [158, 159]. The radiotherapy planning position will also impact on PET image quality since the flat couch-top and immobilisation devices will add additional and non-uniform PET attenuation, degrading the image quality [155]. Therefore it is important to assess the impact on PET-MR image quality of dedicated PET-MR radiotherapy hardware so that: (i) MR protocols can be modified to compensate for the MR signal loss [157], and (ii) software methods of correcting for the PET attenuation can be developed for accurate quantitative PET imaging [156]. Previous studies have investigated the impact of PET-MR imaging in the pelvic radiotherapy position using uniform MR and PET phantoms [156, 162, 163]. However, the broader impact on PET-MR image quality relative to diagnostic image quality using standard image quality phantoms has not been assessed. Further the PET attenuation from the anterior MR receive coil for pelvis imaging has not been considered for the GE Signa PET-MR. The aim of this study was to evaluate the impact of using a flat couch top and coil bridge on PET-MR image quality and PET quantification for radiotherapy pelvis imaging.

## 4.2 Methods

### 4.2.1 Imaging

All images were acquired on a SIGNA PET/MR software version MP26 3T scanner (GE Healthcare, Waukesha, USA). Three different experimental setups were used for both the MR and PET image quality assessments: diagnostic, couch and radiotherapy (figure 4.1). The diagnostic setup consisted of the image quality phantom (PET or MR) placed on the soft foam overlay on the PET-MR couch with the anterior array coil placed directly on phantom. The couch setup comprised the phantom placed on the radiotherapy flat couch-top with the anterior array coil directly on phantom. The radiotherapy setup had the phantom placed on the radiotherapy flat couch-top with the anterior array coil placed on a pelvis coil bridge. The radiotherapy flat couch-top and pelvis coil bridge were produced by Knightec (Stockholm, Sweden) and were similar to those evaluated by Brynolfsson et al [162]. The couch-top consisted of a 5 mm polymethyl methacrylate top sheet bonded to a 35 mm thick low density polyethylene foam base. This base had been cut to match the PET-MR patient table. The coil bridge was made from polyoxymethylene and polycarbonate and designed to just fit inside the patient bore, maximising the range of patient sizes who could be imaged. The coil bridge connected to indexing points along the sides of the couch top, facilitating reproducible set-up at a number of different positions along the length of the couch.

The MR image quality assessment was carried out using the American College of Radiologists (ACR) large image quality phantom [165]. The phantom was imaged in three different imaging sessions on separate days. Each imaging session included all three setups. The phantom was placed in an in-house manufactured holder (figure 4.1). The



(a) Diagnostic (b) Couch (c) Radiotherapy

Figure 4.1: Photographs of the diagnostic (a), couch (b) and radiotherapy (c) experimental setups with the ACR phantom and holder used for the MR image quality assessment. The same experimental setups with the NEMA phantom setup on foam blocks without the phantom holder were used for the PET image quality assessment.

Table 4.1: The MR parameters used for the MR image quality assessment.

| **Parameter** | **Sequence** | | |
|---|---|---|---|
| | Localiser | ACR T1 | ACR T2 |
| Field of view ($mm^2$) | 250 × 250 | 250 × 250 | 250 × 250 |
| Matrix | 256 × 256 | 256 × 256 | 256 × 256 |
| Slice and slice gap thickness (mm) | 20.0 | 5.0 | 5.0 |
| Slices/slice gaps | 1/0 | 11/10 | 11/10 |
| Echo time (ms) | 20 | 20 | 80 |
| Repetition time (ms) | 200 | 500 | 2000 |
| Bandwidth ($\mathrm{Hz\,pixel^{-1}}$) | 651 | 651 | 651 |

holder had three screws which enabled easy levelling of the phantom in two axes using a spirit level. For each setup the phantom was imaged using the recommended ACR sequences consisting of a sagittal localiser, an axial T1-weighted spin echo (ACR T1) and an axial double-echo T2-weighted spin echo (ACR T2) (table 4.1) [165]. The second echo images in the ACR T2 series were used for all image analyses.

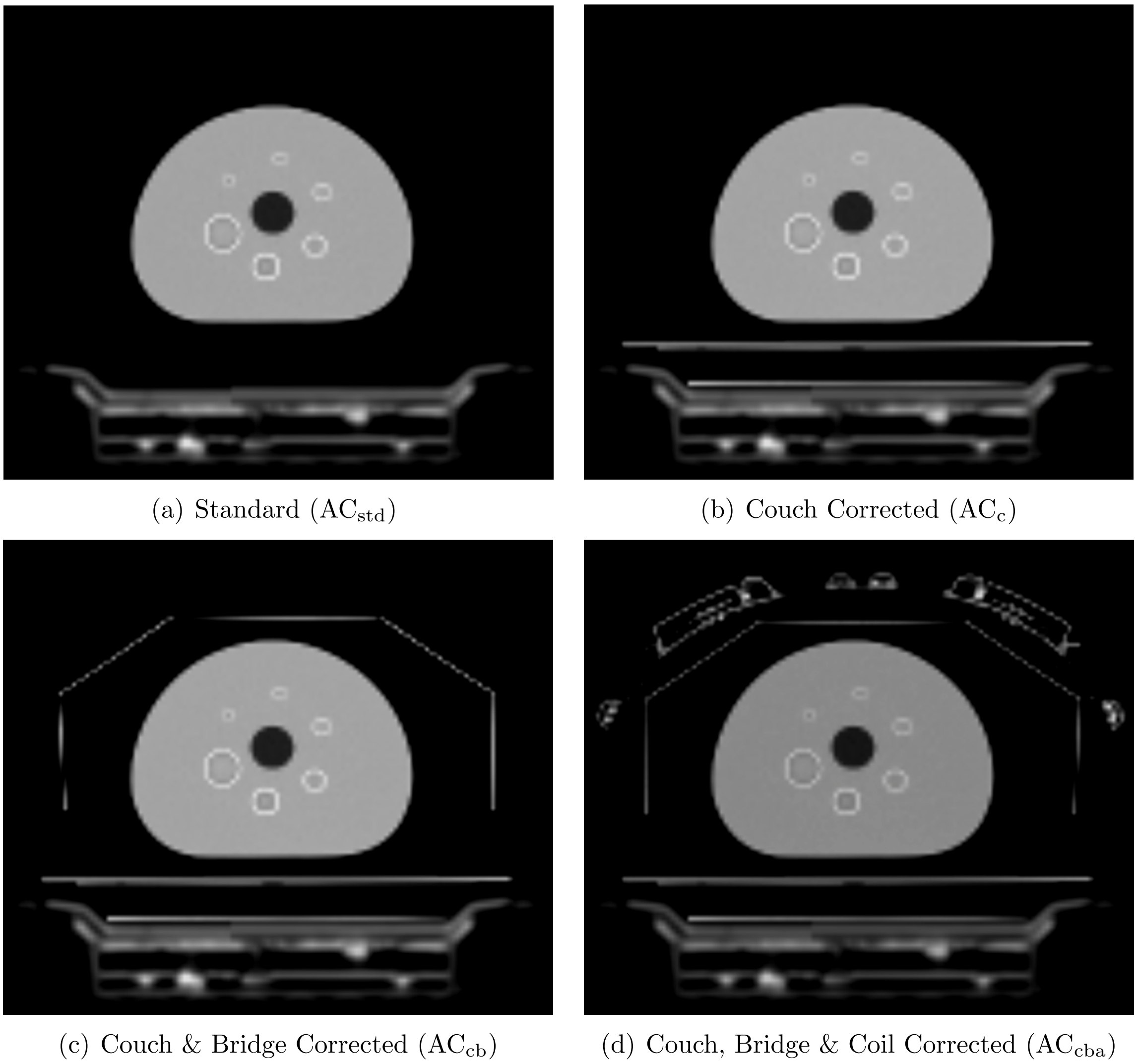
The PET image quality assessment was carried out using an International Electrotechnical Commission (IEC) 61675-1 emission phantom (PTW, Freiburg, Germany). The phantom was set up with the six spheres and the background filled with a mixture of $^{18}$F-FDG and water, with the activity concentration within the spheres being approximately four times more than in the background, as specified by the National Electrical Manufacturers Association (NEMA) NU 2-2007 standard [226]. Unlike the NEMA specification all six spheres were hot compared to the background as this is more representative of a radiotherapy planning context. The lung insert was used for all measurements. For the couch and radiotherapy setups the phantom was placed on two small foam blocks (height 20 mm) to approximately centre the phantom in the scanner bore. For the diagnostic setup foam blocks with twice the height were used, to compensate for the lack of the flat couch-top. The holder shown in figure 4.1 was not used for the PET acquisitions. The phantom was filled on two separate imaging sessions on separate days, with the positions of the spheres within the phantom kept the same for both sessions. This was so that the same phantom attenuation correction map could be used, however it did mean that the same size sphere was closest to the anterior coil in both scans. Each imaging session consisted of six sequential acquisitions. There were two acquisitions in each of the three experimental setups, one with the anterior array coil and one without. The position of the phantom relative to the anterior array was kept the same for each setup and between the two sessions. All acquisitions consisted of one bed position with the phantom centred in the PET field of view. The first acquisition in each session used a five minute bed position with an activity concentration of 5.5 $\mathrm{kBq\,ml^{-1}}$ and subsequent acquisitions used longer bed positions to allow for radioactive decay, giving approximately the same number of counts in each.

For each PET acquisition and attenuation map two reconstruction algorithms were used: an OSEM reconstruction with 16 subsets and 4 iterations and a 5.0 mm Gaussian filter, and a Bayesian penalized-likelihood iterative image reconstruction (Q.Clear) with a relative noise regularizing term factor of $\beta = 350$ [227]. Both reconstructions used point spread function correction and time of flight information.

### 4.2.2 Attenuation Correction

All PET reconstructions incorporated a standard attenuation map consisting of a model of the phantom and the coil components contained within the scanner bed for attenuation correction of the PET data ($\mathrm{AC_{std}}$). Additional images were reconstructed for the couch and radiotherapy setups with a modified attenuation correction map which included a kVCT scan of the radiotherapy couch ($\mathrm{AC_c}$). For the radiotherapy setup further images were also reconstructed with another modified attenuation correction map that included kVCT scans of both the radiotherapy couch and the radiotherapy coil bridge ($\mathrm{AC_{cb}}$). The kVCT scan of the coil bridge were positioned relative to the centre of the phantom using measurements of the distance from the end of the phantom to the end of the coil bridge for both superior and inferior ends. The kVCT scans of both the couch and the coil bridge were acquired using a Somatom Open scanner (Siemens, Erlangen, Germany) with a tube voltage of 140 kVp, a voxel size of $1.2 \times 1.2 \times 1.5$ $\mathrm{mm^3}$ and an axial field of view of $600 \times 600$ $\mathrm{mm^2}$. The CT scan was converted into a PET attenuation map by using the PET-MR vendor (GE Healthcare) supplied mapping from 140 kVp HU to 511 keV linear attenuation coefficients.

Finally a MVCT of the coil bridge with the anterior array coil on it was acquired using a TomoHD TomoTherapy helical linear accelerator (Accuray, Sunnyvale, California, USA). MVCT images were obtained due to the high atomic number elements in the array coil creating substantial streak artefacts on kVCT imaging [228]. Images were acquired with the detuned imaging beam energy ($\sim$ 1 MV) and 2 mm slice thickness. Three MVCT images were acquired: with the bridge centred, laterally displaced to the right, and to the left. The left and right images were registered to each other via registration to the central image, cropped to the midpoint of the bridge and merged to produce one MVCT image containing the whole coil and bridge. A relative electron density phantom was also imaged on both the MVCT and kVCT scanners to derive relative electron density as a function of MVCT HU and kVCT HU respectively. Combining these with the vendor-supplied PET linear attenuation coefficient as a function of kVCT HU enabled a PET linear attenuation coefficient as a function of MVCT HU to be calculated. This was applied to the MVCT image of the anterior coil on the coil bridge and combined with $AC_c$ to produce a couch, bridge and anterior coil corrected attenuation map ($AC_{cba}$) The four different attenuation maps can be seen in figure 4.2.



(a) Standard ($AC_{std}$)

(b) Couch Corrected ($AC_c$)

(c) Couch & Bridge Corrected ($AC_{cb}$)

(d) Couch, Bridge & Coil Corrected ($AC_{cba}$)

Figure 4.2: Example slices of the attenuation maps used. a) The standard attenuation map contained a model of the phantom and the scanner bed. b) The couch corrected map added a model of the radiotherapy couch. c) The couch and bridge corrected map added a map of the bridge and d) the couch, bridge and coil corrected map added in a map of the anterior coil.

### 4.2.3 MR Image Quality Assessment

MR images were analysed according to the ACR recommendations by evaluating geometric accuracy, high-contrast spatial resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity, percent-signal ghosting and low-contrast object detectability. In addition SNR was also assessed in the T1- and T2-weighted images.

The first six ACR tests were analysed using in-house developed Matlab software (version R2017a Mathworks, Natick, USA). This was based upon open source software [229], with modifications to reduce the influence of the partial volume effect when calculating the signal from the phantom, improve the accuracy of profiles and edge detection through up-sampling and make the analysis more robust to image artefacts such as air bubbles.

The software performed the six ACR tests evaluating geometric accuracy, high-contrast spatial resolution, slice thickness accuracy, slice position accuracy, image intensity uniformity and percent-signal ghosting. The geometric test compared measured and known lengths in the phantom. The resolution test used the smallest diameter holes that could be distinguished in a horizontal or vertical array. The slice thickness test used the measured profile of two angled ramps. The slice position test used crossed $45^o$ wedges at the inferior and superior edge of the phantom. The uniformity test used the near-maximum and near-minimum pixel values in the uniform section of the phantom. The ghosting took the ratio of mean pixel values of four Region Of Interest (ROI)s against the edges of the field of view (outside the phantom) to a $\sim 200$ cm$^2$ ROI in the uniform section of the phantom. The seventh ACR test, low-contrast object detectability, was performed manually using RayStation (version 7, RaySearch Laboratories, Stockholm, Sweden). The low-contrast detectability score was the total number of visible ‘spokes’ of disks of decreasing diameter (7.0 mm to 1.5 mm) and contrast (5.1% to 1.4%).

In addition the MR SNR was calculated using the methodology of McCann et al. with one $20 \times 20$ mm$^2$ ROI centred on the phantom centre and four more $20 \times 20$ mm$^2$ ROIs centred at $(\pm 40, \pm 40)$ mm from the phantom centre [230]. This method robustly calculated SNR from a single image by: (i) smoothing the image by convolution with a square boxcar filter, (ii) subtracting the smoothed image from the original image to create a noise image, and (iii) calculating the SNR for each ROI using

$$SNR_i^{MR} = \frac{S_i}{\sigma_i}. \tag{4.1}$$

Here $S_i$ was the mean pixel value within ROI $i$ in the original image and $\sigma_i$ was the standard deviation of pixel values within ROI $i$ in the noise image. The SNR for the whole image was calculated as the mean over all ROIs. The SNR analysis was carried out using MICE Toolkit [204]. All MR image quality measurements were presented as the mean over three repeats ± the standard error of the mean.

### 4.2.4 PET Image Quality Assessment

All PET images were analysed using four metrics: background activity deviation, PET SNR, contrast recovery and background variability. Spherical ROIs matching the known sphere volume were drawn on each sphere using RayStation. Twelve cylindrical ROIs with a diameter of 15 mm and a length of 5 mm were placed in the background in the central slice passing through the spheres. These background ROIs were repeated contiguously down the longitudinal length of the phantom. Within a phantom setup these ROIs were drawn on one image and copied onto all the others. The background activity deviation was calculated as the difference between known activity concentration and the reconstructed activity concentration averaged over the background ROIs as a function of longitudinal distance in the phantom. The relative difference in background deviation for all setups and reconstructions to the background deviation for the diagnostic setup without anterior coil was calculated since this was the gold standard for PET image quality (no additional radiotherapy hardware and no anterior MR coil). The PET SNR was determined using [231]

$$SNR_i^{PET} = \frac{c_i - c_{bg}}{\sigma_{bg}}, \tag{4.2}$$

where $c_i$, was the mean reconstructed activity concentration in sphere ROI $i$, $c_{bg}$ was the mean reconstructed activity concentration in the twelve background ROIs in the central slice only, and $\sigma_{bg}$ was the standard deviation of the reconstructed activity concentration in the same background ROIs. The injected activity concentration ratio between the spheres and background was compared to the measured ratio to derive the contrast recovery, defined as [160]

$$C = 100\frac{c_i/c_{bg} - 1}{a_{sp}/a_{bg} - 1}. \tag{4.3}$$

Here $a_{sp}$ was the injected activity concentration in the spheres, $a_{bg}$ the injected activity concentration in the background and $c_i$ and $c_{bg}$ as defined in equation (4.2). The background variability was calculated using [160]

$$N = \frac{\sigma_{bg}}{c_{bg}}, \tag{4.4}$$

where $\sigma_{bg}$ and $c_{bg}$ were as defined in equation (4.2). All PET image quality measurements were reported as the mean over two repeats ± the standard error of the mean.

## 4.3 Results

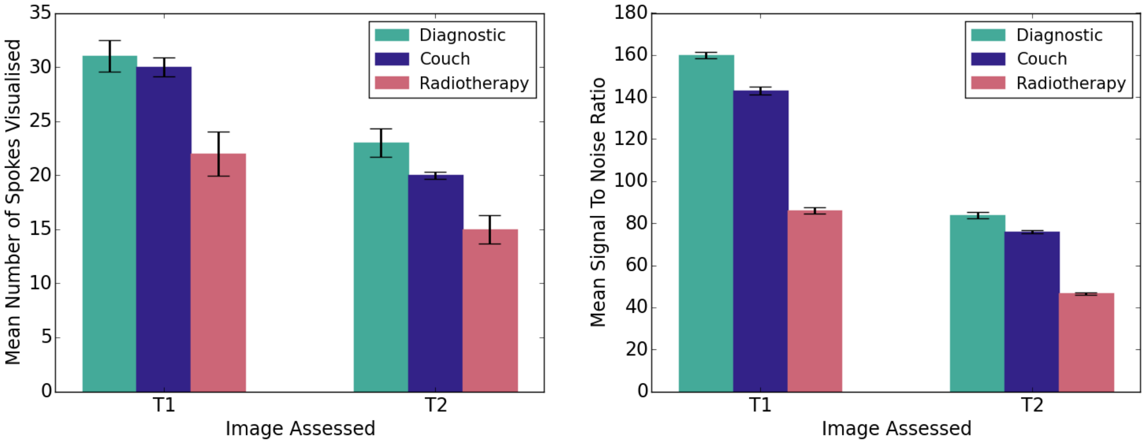
### 4.3.1 MR Image Quality Assessment

The results for all the ACR image quality metrics agreed within one standard error between the different setups (table 4.2) except for the low-contrast detectability test, where the couch setup was lower than the diagnostic, and the radiotherapy substantially lower (figure 4.3). The MR SNR for the couch and radiotherapy setups was lower than for diagnostic setup, being $89 \pm 2\%$ and $54 \pm 1\%$ of the diagnostic setup respectively for the ACR T1 images (figure 4.3). The ACR T2 image results were similar, with the couch

SNR being 91 ± 2% of the diagnostic setup and the radiotherapy SNR 56 ± 1%.

Table 4.2: MR image quality assessment: standard ACR tests. All results given as mean ± one standard error of the mean. For the geometric accuracy, spatial resolution and slice position tests with multiple measurements per image, the mean of those measurements is shown. The units for each test are displayed with the test name.

| **Test** | **Sequence** | **Reference** | **Setup** | | |
|---|---|---|---|---|---|
| | | | *Diagnostic* | *Couch* | *Radiotherapy* |
| Geometric Accuracy [mm] | Localiser | *148 ± 2* | $146.8 \pm 0.1$ | $146.73 \pm 0.03$ | $146.8 \pm 0.1$ |
| | T1 | *190 ± 2* | $188.9 \pm 0.2$ | $189.8 \pm 0.2$ | $189.7 \pm 0.2$ |
| | T2 | *190 ± 2* | $189.6 \pm 0.2$ | $190.0 \pm 0.2$ | $189.8 \pm 0.2$ |
| Spatial Resolution [mm] | T1 | $\leq$ *1.0* | 1.0 | 1.0 | 1.0 |
| | T2 | $\leq$ *1.0* | 1.0 | 1.0 | 1.0 |
| Slice Thickness [mm] | T1 | *5 ± 0.7* | $5.4 \pm 0.2$ | $5.1 \pm 0.1$ | $5.1 \pm 0.3$ |
| | T2 | *5 ± 0.7* | $5.0 \pm 0.3$ | $4.7 \pm 0.1$ | $5.0 \pm 0.3$ |
| Slice Position [mm] | T1 | $\leq$ *5* | $2.3 \pm 0.5$ | $3.4 \pm 0.3$ | $3.3 \pm 0.6$ |
| | T2 | $\leq$ *5* | $2.3 \pm 0.4$ | $3.4 \pm 0.3$ | $3.2 \pm 0.5$ |
| Image Uniformity [%] | T1 | $\geq$ *82* | $66.5 \pm 0.5$ | $60.3 \pm 0.9$ | $67.0 \pm 0.6$ |
| | T2 | $\geq$ *82* | $62 \pm 2$ | $54 \pm 1$ | $66.1 \pm 0.3$ |
| Ghosting [%] | T1 | $\leq$ *3* | $9 \pm 5 \times 10^{-3}$ | $12 \pm 8 \times 10^{-3}$ | $7 \pm 3 \times 10^{-3}$ |
| | T2 | $\leq$ *3* | $30 \pm 15 \times 10^{-3}$ | $28 \pm 8 \times 10^{-3}$ | $72 \pm 20 \times 10^{-3}$ |



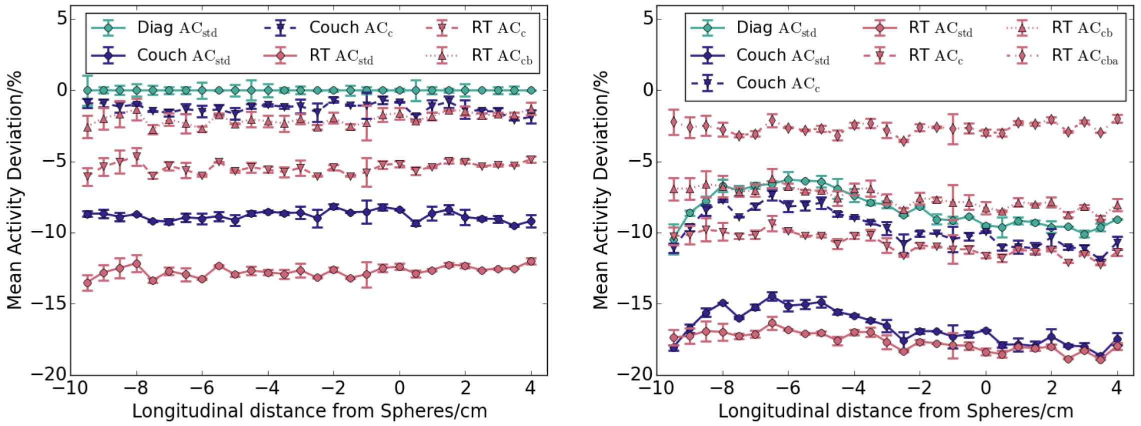
(a) Low Contrast Detectability

(b) MR Signal to Noise Ratio

Figure 4.3: MR image quality assessment. a) Low contrast detectability, measured in the number of spokes visible in the images and (b) MR SNR. The mean over three acquisitions performed on separate days is represented for each experimental setup for the T1 and T2 MR images. Error bars are one standard error of the mean.

## 4.3.2 PET Image Quality Assessment

The background activity deviation was approximately uniform (within 1%) along the length of the phantom for each setup acquired without the anterior array coil in place (figure 4.4). Using the $AC_{std}$ map, the mean background deviation of the couch setup relative to the diagnostic setup without coil was $-9.0 \pm 0.1\%$ and for the radiotherapy setup it was $-13.0 \pm 0.1\%$. Using $AC_c$ instead reduced this to $-1.0 \pm 0.1\%$ and $-5.0 \pm 0.1\%$ respectively. For the radiotherapy setup, using $AC_{cb}$ led to a $-2.0 \pm 0.1\%$ difference. The images acquired with the anterior array coil in place showed a non-uniform background activity deviation along the phantom length, with differences from the images acquired without the coil in place between −6% and −12% (figure 4.4). The mean difference to diagnostic setup without coil for the diagnostic setup with coil was $-8.3 \pm 0.2\%$. For the couch and radiotherapy setups using $\text{AC}_{\text{std}}$ the difference was $-16.7 \pm 0.2\%$ and $-17.7 \pm 0.1\%$ respectively. Correcting for the radiotherapy hardware improved the performance, with $\text{AC}_{\text{c}}$ included reducing the activity difference to $-9.7 \pm 0.2\%$ (couch setup) and $-10.8 \pm 0.1\%$ (radiotherapy setup). Using $\text{AC}_{\text{cb}}$ in the radiotherapy setup gave similar performance to the diagnostic setup (activity difference $-7.5 \pm 0.1\%$). With $\text{AC}_{\text{cba}}$, the radiotherapy setup outperformed the diagnostic setup and was only $-2.7 \pm 0.1\%$ different to the diagnostic setup without anterior coil. In all setups there was no difference between the OSEM and Q.Clear reconstructions.

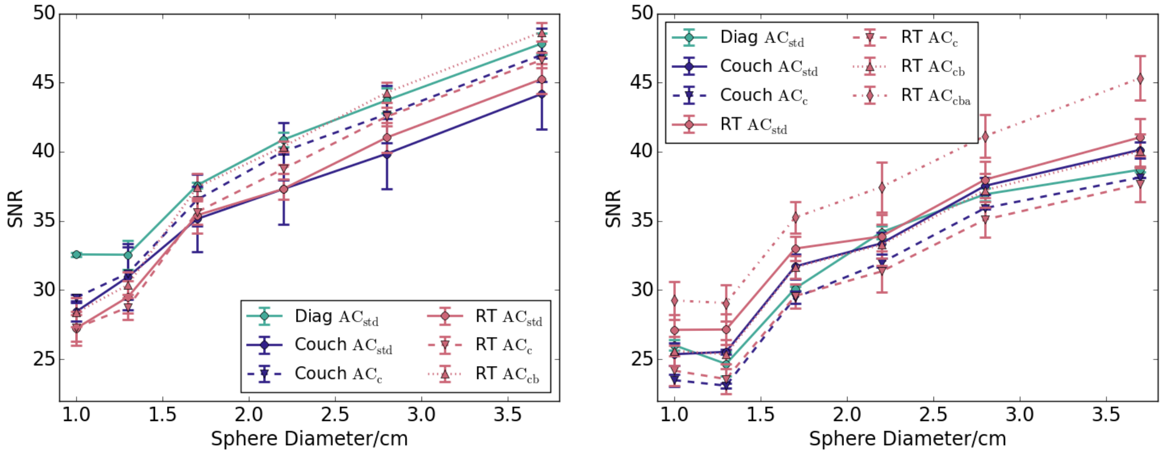


(a) Without Anterior Coil

(b) With Anterior Coil

Figure 4.4: PET image quality assessment: Percentage difference in background activity deviation for a given setup and attenuation map to the background activity deviation of the diagnostic setup without anterior coil with standard attenuation as a function of longitudinal distance from the largest sphere (negative indicates superior and positive inferior directions). Plots show data acquired without (a) and with (b) the anterior coil in place. Green, blue and purple lines indicate the diagnostic (diag), couch and radiotherapy (RT) setups respectively. Solid lines/circular markers show images reconstructed with the standard attenuation map ($\text{AC}_{\text{std}}$), dashed lines/downward triangular markers images incorporating the attenuation of the couch ($\text{AC}_{\text{c}}$), dotted lines/upward triangular markers the couch and coil bridge ($\text{AC}_{\text{cb}}$) and dash-dotted lines/diamond markers the couch, bridge and anterior coil ($\text{AC}_{\text{cba}}$). Data shown used the Q.Clear reconstruction.

The PET SNR as a function of sphere diameter is shown in figure 4.5. When the anterior array coil was not in place the diagnostic setup showed the best performance, with the $\text{AC}_{\text{std}}$ corrected couch and radiotherapy setups being worse (> one standard error), but similar to each other. Correcting for the couch and coil bridge improved the performances of both couch and radiotherapy setups to a similar quality to the diagnostic setup (within one standard error). Including the anterior array coil in the acquisition caused a general decrease in SNR of $\sim 17\%$ across the setups with the exception of couch, coil bridge and coil corrected radiotherapy setup. The differences between the setups and corrections, with the above exception, agreed within one standard error. The couch, coil bridge and coil correction of the radiotherapy setup approached the performance of the diagnostic setup without anterior coil. The Q.Clear reconstructions performed better than the OSEM for all different setups with and without the anterior coil by approximately 5%.



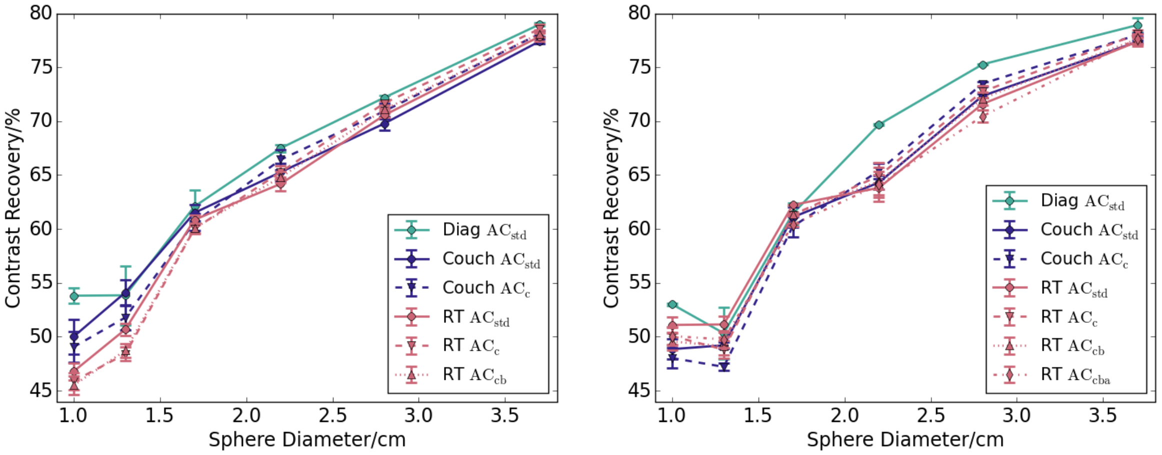
(a) Without Anterior Coil

(b) With Anterior Coil

Figure 4.5: PET image quality assessment: PET SNR for the different sphere diameters without (a) and with (b) the anterior coil for the Q.Clear reconstructions. Green, blue and red lines indicate the diagnostic, couch and radiotherapy setups respectively. Solid lines/circular markers show images reconstructed with the standard attenuation map ($\mathrm{AC_{std}}$), dashed lines/downward triangular markers images incorporating the attenuation of the couch ($\mathrm{AC_{c}}$), dotted lines/upward triangular markers the couch and coil bridge ($\mathrm{AC_{cb}}$) and dash-dotted lines/diamond markers the couch, bridge and anterior coil ($\mathrm{AC_{cba}}$). Error bars indicate one standard error of the mean.

The contrast recovery increased as a function of sphere size due to the partial volume effect of the relatively poor PET resolution. Figure 4.6 shows the results for the Q.Clear reconstruction. Without the anterior array coil there was not a large difference between the setups except for the smallest sphere diameter, where the diagnostic setup performed best. The couch and coil bridge corrections did not appear to significantly change the performance of the couch and radiotherapy setups. With the anterior array coil the diagnostic setup performed better for most sphere diameters, with small differences between the other setups and corrections. Similarly to the SNR results, the Q.Clear reconstruction outperformed the OSEM reconstruction by $\sim 5\%$ for all setups and corrections.

The background variability for all setups with the Q.Clear reconstructions is given in table 4.3. The OSEM reconstructions (see supplementary material) had a higher ($\geq 1.0\%$) background variability for all setups and attenuation corrections. The presence of the anterior coil increased the background variability for all setups and corrections. Without the coil the background variability was lowest in the diagnostic setup and decreased in the other setups once the appropriate attenuation corrections were applied ($\mathrm{AC_{c}}$ for the couch and $\mathrm{AC_{cb}}$ for the couch and coil bridge). With the coil the effects were more variable, with the couch setup using $\mathrm{AC_{std}}$ and the radiotherapy setup using $\mathrm{AC_{std}}$ and using $\mathrm{AC_{cba}}$) giving the lowest values, although the results were within 0.8% of each other so the differences were relatively small.



(a) Without Anterior Coil

(b) With Anterior Coil

Figure 4.6: PET image quality assessment: Contrast recovery curves without (a) and with (b) the anterior coil for the Q.Clear reconstructions. Green, blue and red lines indicate the diagnostic, couch and radiotherapy setups respectively. Solid lines/circular markers show images reconstructed with the standard attenuation map ($\mathrm{AC}_{\mathrm{std}}$), dashed lines/downward triangular markers images incorporating the attenuation of the couch ($\mathrm{AC}_{\mathrm{c}}$), dotted lines/upward triangular markers the couch and coil bridge ($\mathrm{AC}_{\mathrm{cb}}$) and dash-dotted lines/diamond markers the couch, bridge and anterior coil ($\mathrm{AC}_{\mathrm{cba}}$). Error bars indicate one standard error of the mean.

## 4.4 Discussion

PET-MR imaging has great potential for radiotherapy treatment planning and radiotherapy images need to be acquired in the planning position. This study has investigated the impact on both PET and MR image quality from acquiring PET-MR images in the radiotherapy planning position for treatment of pelvic cancers.

This is the first study to investigate the impact on MR image quality of the pelvic radiotherapy hardware using an MR image quality phantom in a PET-MR scanner. The impact was substantial with the couch SNR being $91 \pm 2\%$ of the diagnostic setup and the radiotherapy SNR $56 \pm 1\%$. This was likely due to the receive coils being at a greater distance from the phantom and so reducing the SNR [158]. This consequently gave a substantial reduction in low-contrast detectability (figure 4.3) but not on any other of the evaluated image quality metrics. This suggests that MR parameters in radiotherapy PET-MR protocols need to be modified, for instance by increasing signal averages, to take into account the reduction in SNR in order that MR images retain sufficient quality for accurate organ delineation. Alternatively, noise reduction reconstruction techniques, such as the recently proposed deep learning reconstruction could be used to improve image quality [201]. These techniques would not require compromises in acquisition time, voxel size or field of view as modifying the scan parameters would, but would require validation for the specific clinical task. Further work could assess the impact of this image quality reduction in patient images using radiotherapy sequences, including the impact of noise reduction reconstruction techniques.

Brynolfsson et al. evaluated the same couch and a similar coil bridge using a large uniform

Table 4.3: PET image quality assessment: The mean background variability for each PET image with Q.Clear reconstructions. Values reported as mean ± standard error on the mean. Coil present indicates images acquired with (yes) and without (no) the anterior coil. The standard attenuation map included the phantom and scanner table ($AC_{std}$), the couch attenuation map added the radiotherapy couch ($AC_{c}$), the couch & bridge map added the coil bridge ($AC_{cb}$) and the couch, bridge & coil map added the anterior array coil ($AC_{cba}$).

| SetUp | Coil Present | Attenuation | Background Variability |
|---|---|---|---|
| Diagnostic | No | Standard | $5.8 \pm 0.4\%$ |
| Couch | No | Standard | $6.2 \pm 0.1\%$ |
| Couch | No | Couch | $5.8 \pm 0.1\%$ |
| Radiotherapy | No | Standard | $6.0 \pm 0.1\%$ |
| Radiotherapy | No | Couch | $5.9 \pm 0.2\%$ |
| Radiotherapy | No | Couch & Bridge | $5.6 \pm 0.2\%$ |
| Diagnostic | Yes | Standard | $7.2 \pm 0.4\%$ |
| Couch | Yes | Standard | $6.8 \pm 0.2\%$ |
| Couch | Yes | Couch | $7.2 \pm 0.3\%$ |
| Radiotherapy | Yes | Standard | $6.6 \pm 0.1\%$ |
| Radiotherapy | Yes | Couch | $7.3 \pm 0.1\%$ |
| Radiotherapy | Yes | Couch & Bridge | $6.8 \pm 0.1\%$ |
| Radiotherapy | Yes | Couch, Bridge & Coil | $6.0 \pm 0.1\%$ |

MR phantom [162]. They reported SNRs of 74% when using the couch only and 67% when using the couch and coil bridge compared to the diagnostic setup. These results are slightly different to those reported here, which may be due to the use of a larger uniform phantom. This would increase the contribution of the spinal coil in the scanner bed relative to the anterior coil to the MR signal, thus increasing the SNR reduction due to the couch setup. It would also reduce the distance between the phantom and the anterior coil when the coil bridge was used, therefore reducing the SNR reduction from the coil bridge. Paulus et al. developed a radiotherapy flat couch-top for PET-MR imaging consisting of a foam core surrounded by a plastic outer layer [155] and an adjustable pelvic coil bridge [156]. Quantitative assessment of MR image quality using this setup was not carried out, but subjective image quality of three abdominal patients was reported to be similar to the diagnostic setup. This suggests the SNR reduction did not substantially reduce subjective image quality, which may be due to the adjustable coil bridge enabling the coil-patient distance to be minimised, although this is also likely to be dependant on the MR sequence and protocol used. Witoszynskyj et al. developed a plastic and fibre glass radiotherapy couch for PET-MR imaging and found minimal differences in MR SNR with and without the couch using a uniform MR phantom [163]. However, their images were acquired with the integrated body coil, and so the radiotherapy hardware did not change the distance of the receive coil from the phantom. The integrated body coil is not used clinically and so these results are not relevant to clinical practice.

This is also the first study to investigate the impact on standard PET image quality metrics using the NEMA phantom from PET-MR images acquired in the pelvic radiotherapy setup. The radiotherapy hardware reduced PET SNR for all spheres, but incorporating the couch and bridge attenuation correction recovered the PET SNR performance to within one standard error (figure 4.5). Including the coil attenuation correction resulted in the radiotherapy setup considerably outperforming the diagnostic setup with anterior coil and approached the performance of the diagnostic setup without anterior coil. However, the corrections appeared to make minimal difference to the contrast recovery curves (figure 4.6), with the diagnostic setup performing better but within 2% of the couch and radiotherapy setups. This suggests that qualitative image quality, of which contrast recovery is a surrogate measure, was not substantially changed by the presence of the radiotherapy hardware. Similarly, the presence of the anterior coil appears to make minimal difference to all setups, despite the significant attenuation it introduces, confirming that qualitative diagnostic imaging does not require correction for the coil [232]. Several authors have evaluated PET image quality on PET-MR scanners acquired in the diagnostic setup (with no anterior coil). Grant et al. reported contrast recovery values of $[35.2, 48.9, 59.9, 68.6]\%$ for the four smallest sphere sizes acquired on the same scanner model with similar acquisition settings, which shows reasonable agreement (within 6%) of the OSEM diagnostic setup values reported here, except for the smallest sphere size which was 11% lower [233]. Similarly measurements on a Siemens Biograph mMR PET-MR scanner using a different image quality phantom showed good agreement for the three smallest spheres [234]. For some of the setups with the anterior coil in place the smallest sphere has a larger PET SNR and contrast recovery than the second smallest sphere. This is a counter-intuitive result since the increasing sphere volume should reduce the impact of partial volume effects. Potentially this may be due to the impact of the coil since the second smallest sphere was the most anterior sphere and so would have the most lines of response passing through the coil. This would explain why the setups without the coil in place do not show the same effect. This is a potential confounding factor for this study. The impact of this could be investigated doing repeat images with the different sphere positions within the phantom, although this would require using different phantom attenuation correction maps.

The radiotherapy setup also significantly reduced PET background activity, with losses of $-12.7\pm0.5\%$ (without the anterior coil) and $-10.2\pm0.5\%$ (with anterior coil) compared to the diagnostic setup without and with anterior coil respectively. The smaller activity loss with the anterior coil in place was possibly due to the coil bridge raising the anterior coil away from the phantom, and so reducing the number of lines of response passing through the coil and therefore being attenuated. These results are similar to those reported by Brynolfsson et al. using the same couch and coil bridge [162]. They also found correcting for the attenuation of the radiotherapy hardware reduced these differences to $-1.5\%$ and $-0.7\%$ for the couch and coil bridge respectively, very similar to the $-1.5\%$ and $0.8\%$ reported in this study. These results were similar to other approaches in the literature. Paulus et al. reported $-3.8\%$ and $-8.5\%$ activity differences from their couch and pelvic coil bridge respectively, which reduced to $-0.6\%$ and $-1.2\%$ when attenuation correction was used [156]. Witoszynskyj et al. found their couch reduced PET activity by $-8.7 \pm 2.1\%$, reducing to $1.2 \pm 3.9\%$ when CT-based attenuation map of the couch was included in the reconstruction [163].

The PET reconstruction algorithm used may also have an impact on PET image quality. This study evaluated both the clinical standard OSEM reconstruction and a more novel Bayesian penalized-likelihood iterative image reconstruction (Q.Clear). This was to evaluate the potentially higher performing reconstruction in the radiotherapy setup but also enable comparisons to the literature using the clinical standard OSEM reconstruction. The Q.Clear reconstruction outperformed the OSEM reconstruction on the PET SNR, contrast recovery and background variability metrics.

This study acquired repeat measurements over multiple days. This was primarily to mitigate the impact of phantom preparation and positioning. However, the small standard errors do suggest that the PET-MR scanner performance was repeatable in terms of the image quality metrics evaluated. However, the repeatability of images would need to be assessed in patients though, since it is likely to be dominated by physiological and anatomic differences between imaging sessions (differences in bladder and bowel filling, patient setup and posture and so on) rather than variability in the scanner performance.

This study has evaluated the impact on PET-MR image quality of using the radiotherapy setup for pelvic cancer patients. Radiotherapy treatment positions for other anatomical sites (eg breast cancer, head and neck cancer) require different patient positions, coil supports and patient immobilisation devices. Further work would be required to investigate the impact on PET-MR image quality for those other anatomical sites.

Previous investigations have not included the anterior array coil in attenuation correction since this is what is routinely done in diagnostic imaging. Primarily this is due to the variable position and flexible shape of the anterior array coil being difficult to model accurately with attenuation maps since it will vary significantly in position and shape from patient to patient and cannot be directly imaged. Therefore ignoring the anterior coil avoids errors from incorrectly positioned attenuation maps and facilitates a simple workflow. However this study has shown that the anterior coil has a substantial and variable attenuation of between $6-12\%$, despite being characterised as a 'low attenuating' coil. The attenuation measured here is consistent with the mean reduction in activity of $-7.3\%$ reported for the same coil by [232]. They concluded that for qualitative diagnostic imaging this was not significant. But for radiotherapy dose painting this is problematic, since it requires accurate quantitative imaging [27]. On the other hand, an advantage of the radiotherapy setup compared to a diagnostic one is that the position and shape of the flexible anterior coil is the same for each patient since the coil bridge fixes the coil shape and height relative to the radiotherapy couch. This potentially means the anterior coil can be included in attenuation correction maps as long as the position of the centre of the coil is known relative to the centre of the image. This can easily be achieved by setting the scanner reference point to the centre of the coil and using the known height of the coil bridge relative to the posterior edge of the patient (the couch top). An issue with generating attenuation correction maps of MR coils containing high atomic number elements is that large streak and starvation artefacts are produced in kVCT images. This study has confirmed, as reported by Patrick et al [228], generating attenuation correction maps for MR coils can be done simply and robustly through MVCT imaging. In this study, the radiotherapy setup, once corrected, outperformed the standard diagnostic setup by 6.2±0.5% in activity measurements and its performance was within $\sim 2\%$ of the diagnostic setup without an anterior coil (i.e. an ideal, non-clinical, PET setup). This suggests that quantitative PET-MR is possible within the radiotherapy setup as long as the anterior coil is included within the attenuation map.

## 4.5 Conclusion

Acquiring PET-MR images in the radiotherapy planning position reduced MR image quality substantially, with a loss of MR SNR of 45%. The radiotherapy position also impacted PET image quality with reductions in measured activity to the diagnostic setup without anterior coil of $-17.7 \pm 0.1\%$, which reduced to $-2.7 \pm 0.1\%$ when attenuation correction map of the radiotherapy hardware was included. Contrast recovery curves were largely unchanged, suggesting qualitative PET image quality was not substantially affected. Noticeably the presence of the flexible anterior coil also had a significant and non-uniform effect on the PET attenuation. Including this coil in the attenuation correction map, which the radiotherapy setup enables, outperformed the standard diagnostic setup with anterior coil by 5.6%. The same impact was seen in PET SNR curves, where the radiotherapy setup with anterior coil corrected outperformed the diagnostic setup with anterior coil. This implies that accurate quantitative PET imaging is possible in the radiotherapy setup as long as appropriate attenuation correction is applied. The impact on PET-MR image quality will need to be considered when designing radiotherapy PET-MR imaging protocols.

# Chapter 5

# The Impact of using Attenuation Correction of Radiotherapy Hardware for PET-MR in Ano-Rectal Radiotherapy Patients

## 5.1 Introduction

The development of PET attenuation correction methods for PET-MR in the pelvic radiotherapy position potentially enables accurate quantitative PET to be utilised for radiotherapy planning. This has great possible benefits, including more accurate delineation of the GTV [21, 22], delineation of tumour sub-volumes for radiotherapy dose painting [23] and/or as a prognostic tool to identify poorer prognosis patients for dose escalation [141].

For anal cancers, $^{18}$F-FDG-PET has demonstrated significantly smaller GTVs compared to CT [21] and good correspondence with MR [22]. A study in rectum cancer patients showed reduced inter-observer variability for tumour delineations on $^{18}$F-FDG-PET-CT compared to CT alone [136]. PET imaging also has good potential for automatic delineation methods utilising the semiquantitative metric SUV [137], with automatic methods showing good agreement with manual contours [138] and better agreement with pathological analysis than CT or MR [235]. PET derived metabolic parameters such as the maximum SUV within a tumour (SUV$_{\text{max}}$) and TLG have also shown promise as prognostic factors for rectal cancers [141, 236].

Both accurate GTV delineation and radiotherapy dose painting and patient prognostics based on PET SUVs depend on high quality, quantitatively accurate PET imaging. This requires accurate attenuation correction of all objects traversed by the lines of response [237]. This needs to include the dedicated radiotherapy hardware, which is challenging for PET-MR since the radiotherapy hardware will non-uniformly attenuate the

PET signal [162] and will not be visible in the MR images. Chapter 4 demonstrated using phantoms that there is a reduction in PET-MR image quality from acquiring images in the radiotherapy position. In that chapter I also developed methods of incorporating the radiotherapy hardware into PET Attenuation Correction (AC) maps, with improvements in SUV accuracy from $-17.7 \pm 0.1$ % to $-2.7 \pm 0.1$ % in phantoms. The aim of this study was to test the feasibility of using these AC maps in ano-rectal radiotherapy patients and to determine the impact on GTV delineation and SUV measurements.

## 5.2 Materials and Methods

### 5.2.1 Patient Data Collection

17 patients enrolled in the Deep MR-only RT study (research ethics committee reference 20/LO/0583) who were planned for radical/neoadjuvant chemoradiotherapy for ano-rectal cancer and received a PET-MR scan were included in this sub-study. Exclusion criteria included contraindications for MR scanning, medical implants in the pelvic area (eg hip prostheses) and external contour greater than the scanner field of view. 10 female and 7 male patients were included with median age 64 years (range 49-76 years). Patients were diagnosed with rectal cancers (n=8) and anal cancers (n=9). 10 of these patients were taken from the evaluation cohort and 7 from the test cohort described in chapter 2 section 2.2.1 table 2.1.

All patients received a simultaneous PET-MR scan on a SIGNA PET/MR 3T scanner (version MP26 GE Healthcare, Waukesha, USA) after their radiotherapy planning CT scan and before their first treatment fraction. Patients were scanned in the radiotherapy treatment position on a flat couch-top with a coil bridge for the anterior MR coil as shown in figure 5.1. Patients were positioned to match their radiotherapy planning CT scan using a combined customisable foot and knee rest (Civco) and external lasers matched to patient tattoos. Immediately prior to entering the scan room patients emptied their bladder and drank 400 ml of water. The PET acquisition started 20 minutes (median, range 14-37 minutes) after patient drinking. The PET images were acquired 70 minutes (median, range 60-86 minutes) after injection with 3.5 $\mathrm{MBq\,kg^{-1}} \pm 10\%$ of $^{18}\mathrm{F-FDG}$ (one patient received 1.7 $\mathrm{MBq\,kg^{-1}}$). All patients had fasted for 6 hours prior to injection and had a measured blood glucose concentration of $< 10$ $\mathrm{mmol\,L^{-1}}$. The PET acquisition consisted of one 5 minute bed position with the patient tumour centred in the PET field of view. Images were reconstructed using a Bayesian penalized-likelihood iterative image reconstruction (Q.Clear) with a relative noise regularizing term factor of $\beta = 350$ [227] with point spread function correction and time of flight information.

MR images were acquired using the automatic dixon sequence used for the scanner-generated PET attenuation correction maps. This was a 3D sequence with a voxel size of $2.0 \times 2.0 \times 5.2$ mm$^3$ with 2.6 mm slice gaps, and a field of view $500 \times 500 \times 780$ mm$^3$.

The images were acquired with a repetition time 4.05 ms, echo times 2.232 ms (in-phase) and 1.116 ms (out-phase) and a receive bandwidth of 1302 Hz pixel$^{-1}$. An additional 3D T2-weighted turbo spin echo sequence was acquired as an anatomical reference for the PET image. This had a voxel size of $1.0 \times 1.0 \times 2.0$ mm$^3$, field of view $380 \times 304 \times 360$ mm$^3$, repetition time 2000 ms, echo time 148 ms and a receive bandwidth of 658 Hz pixel$^{-1}$.

All patients received contrast-enhanced CT scans (Sensation Open, Siemens, Erlangen, Germany) with a voxel size of $1.1 \times 1.1 \times 3$ mm$^3$ and a tube voltage of $V = 120$ kVp. Patients were imaged following routine bladder preparation consisting of an empty bladder 30 minutes prior to the scan, followed by drinking 400 ml of water, and bowel preparation consisting of the application of a micro-enema 60 minutes prior to the scan followed by bowel emptying. CT scans were acquired within 6 days (median, range 0-13 days) of the PET-MR acquisition.

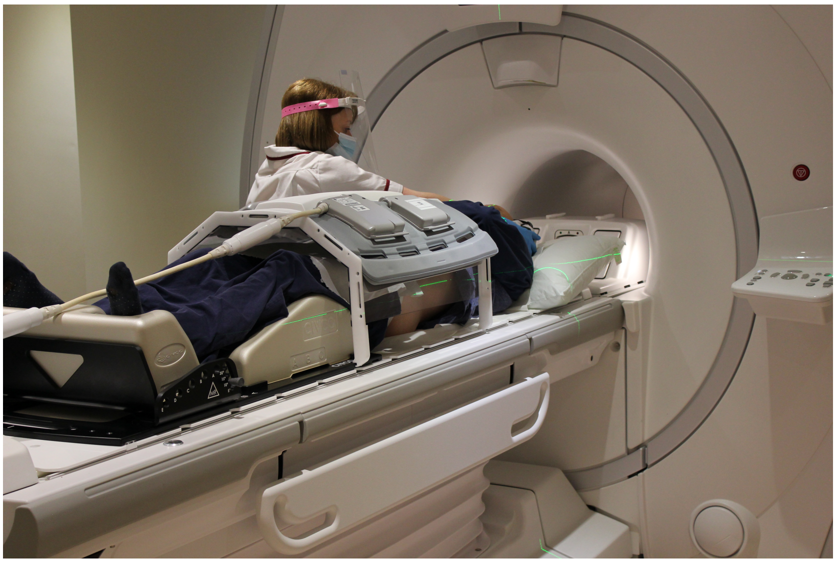
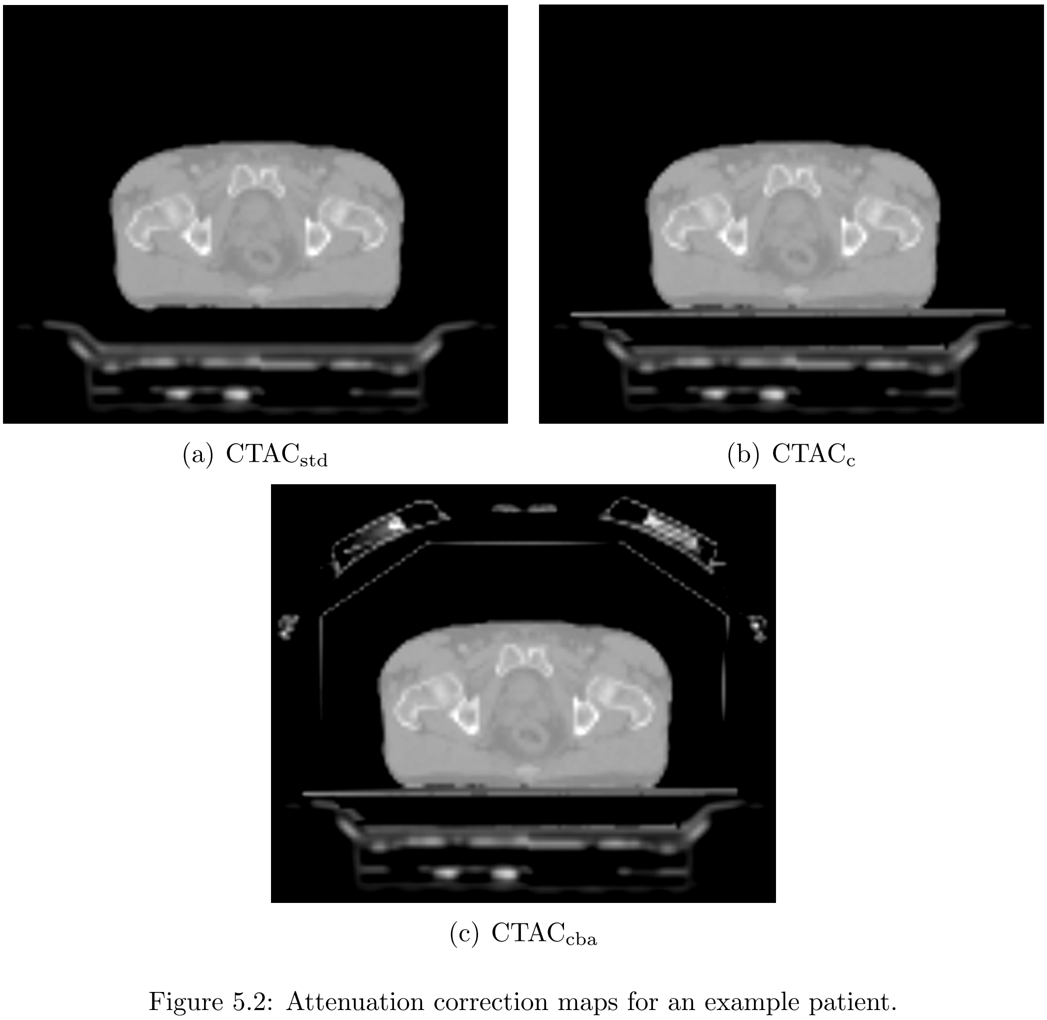


Figure 5.1: Example of patient setup showing the flat couch top, patient immobilisation device, coil bridge and anterior array coil.

### 5.2.2 Attenuation Correction Maps

AC maps can be divided into two components: a map of the patient and a map of all hardware components within the PET lines or response. For the purposes of this study the patient map needed to be kept consistent between all PET images. We decided to use the patient CT since CT is the gold standard source of patient AC. The CT was acquired in the same radiotherapy position as the PET-MR and so a rigid registration between the CT and the in-phase MR image in RayStation (v9B, RaySearch Laboratories, Stockholm, Sweden) was used. The external contour of the in-phase MR was automatically delineated using RayStation's function, and manually modified where necessary. The registered CT was cropped to the MR external contour, with any tissue outside the CT external contour but inside the MR external contour set to water density. Any air within the patient was automatically delineated and set to water density.

Three different hardware AC maps were used, each with the CT patient map: CTAC$_{\text{std}}$, CTAC$_{\text{c}}$ and CTAC$_{\text{cba}}$. CTAC$_{\text{std}}$ was automatically generated by the scanner and included the MR spine coil components within the scanner bed. CTAC$_{\text{c}}$ was the same as CTAC$_{\text{std}}$ but with the manual addition of a model of the radiotherapy couch placed abutting the patient posterior edge, as described in chapter 4. CTAC$_{\text{cba}}$ included CTAC$_{\text{c}}$ with the further manual addition of a model of the coil bridge and anterior coil. The coil bridge and anterior coil model was placed in the patient right-left and anterior-posterior directions using the measured distances to the radiotherapy couch. The inferior-superior position was calculated through landmarking the scanner table to the centre of the coil bridge and using the scanner table position during the PET acquisition, accessible through the private DICOM tag '`PET_table_z_position`'. Examples of the three attenuation maps are shown in figure 5.2. CTAC$_{\text{std}}$ would be the hardware AC map produced directly by the scanner without modification whereas CTAC$_{\text{cba}}$ would include all hardware within the PET lines of response. This study aimed to assess whether the improvement in PET accuracy from using CTAC$_{\text{cba}}$ would result in clinically significant differences in GTV delineation and SUV measurements or whether CTAC$_{\text{std}}$ was accurate enough for radiotherapy purposes.



(a) CTAC$_{\text{std}}$

(b) CTAC$_{\text{c}}$

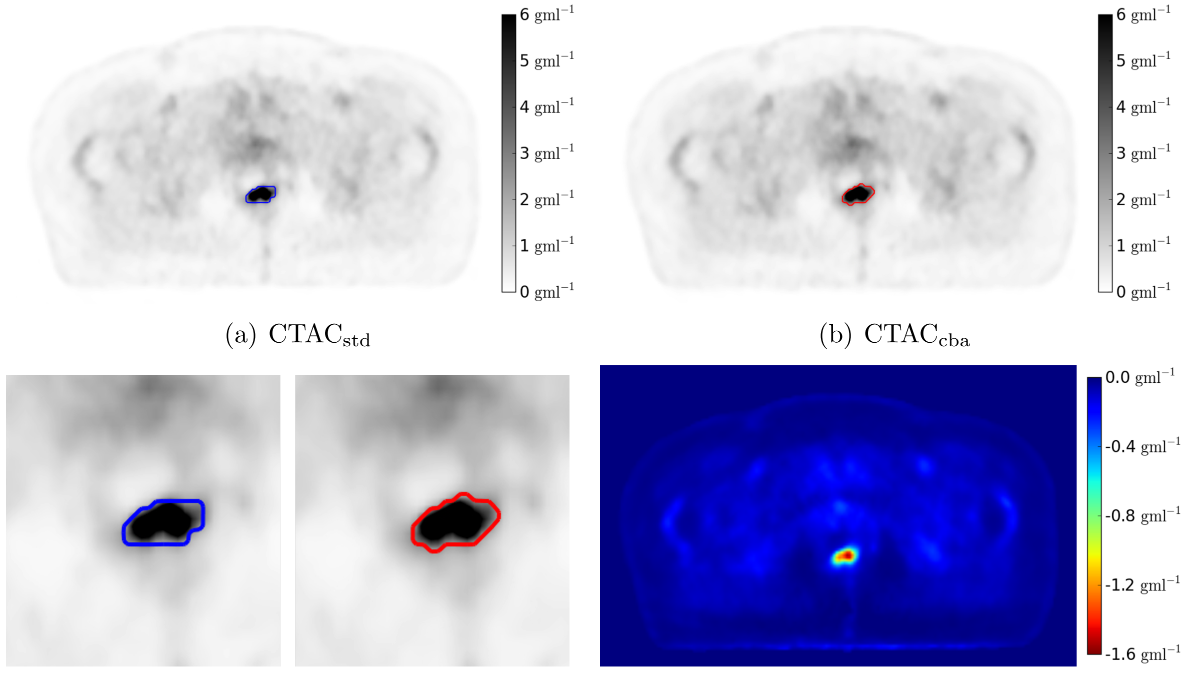
(c) CTAC$_{\text{cba}}$

Figure 5.2: Attenuation correction maps for an example patient.

### 5.2.3 Tumour Delineation

The $\mathrm{CTAC}_{\mathrm{std}}$ and $\mathrm{CTAC}_{\mathrm{cba}}$ PET images were independently contoured at least 7 weeks apart by an experienced consultant PET radiologist using RayStation. The image was automatically thresholded using a fixed SUV $= 2.5\,\mathrm{g\,ml^{-1}}$ [138] and the resultant volume manually adjusted by the radiologist as apprioriate to represent a gross tumour volume ($\mathrm{GTV}_{\mathrm{std}}^{\mathrm{man}}$ and $\mathrm{GTV}_{\mathrm{cba}}^{\mathrm{man}}$ for the $\mathrm{CTAC}_{\mathrm{std}}$ and $\mathrm{CTAC}_{\mathrm{cba}}$ images respectively). Primary and nodal volumes were delineated separately (GTVp and GTVn respectively). Examples of the PET images and manual GTV contours are shown in figure 5.3.

A threshold method was also used to automatically delineate the tumour on both $\mathrm{CTAC}_{\mathrm{std}}$ and $\mathrm{CTAC}_{\mathrm{cba}}$ images, referred to as $\mathrm{GTV}_{\mathrm{std}}^{\mathrm{thresh}}$ and $\mathrm{GTV}_{\mathrm{cba}}^{\mathrm{thresh}}$ respectively. A threshold value of 40% of the maximum SUV within the manual GTV contour of the relevant image was calculated and voxels with a SUV above that threshold were included in the contour using RayStation [21]. The thresholded contour was limited to be within a 0.5 cm expansion of the manual GTV contour of the relevant image to ensure physiological uptake was not included, except for patients (n=3) where the GTV abutted the bladder, where a 0.0 cm expansion was used in that direction.



(a) $\mathrm{CTAC}_{\mathrm{std}}$

(b) $\mathrm{CTAC}_{\mathrm{cba}}$

(c) Zoomed GTVs

(d) Difference Map

Figure 5.3: Example PET images reconstructed using the $\mathrm{CTAC}_{\mathrm{std}}$ (a) and $\mathrm{CTAC}_{\mathrm{cba}}$ (b) attenuation correction maps. The $\mathrm{GTVp}_{\mathrm{std}}^{\mathrm{man}}$ (a, blue contour) and $\mathrm{GTVp}_{\mathrm{cba}}^{\mathrm{man}}$ (b, red contour) are shown for image respectively and close-up versions in c). The per-pixel SUV difference map ($\mathrm{CTAC}_{\mathrm{std}}$-$\mathrm{CTAC}_{\mathrm{cba}}$) is also shown (d). The max difference shown corresponds to 10.7% of the $\mathrm{SUV}_{\mathrm{max}}$ of this patient's GTV.

### 5.2.4 Whole Image Analysis

The per pixel percentage difference in SUV for $\mathrm{CTAC}_{\mathrm{std}}$ and $\mathrm{CTAC}_{\mathrm{c}}$ compared to $\mathrm{CTAC}_{\mathrm{cba}}$ were calculated using MICE Toolkit (v1.0.8) [204]. An external contour was segmented on CTAC$_{\mathrm{cba}}$ using a threshold of 0.05 g ml$^{-1}$ and only differences within this external contour were included. A histogram of differences was calculated using 400 bins between $-100\%$ and $+100\%$ for each patient, and the mean difference within each bin over all patients determined.

The CTAC$_{\mathrm{cba}}$ PET image was used as the reference image for all analyses since the previous phantom study (chapter 4) had showed it had the smallest PET activity loss compared to a gold standard PET acquisition without radiotherapy hardware or anterior coil. The aim of this study was to assess whether this improvement in SUV accuracy translated into clinically relevant differences in tumour delineation and metabolic parameter measurements.

### 5.2.5 Tumour Delineation Analysis

The manual and thresholded GTV contours were compared between CTAC$_{\mathrm{std}}$ and CTAC$_{\mathrm{cba}}$ to determine the impact on radiotherapy target delineation of not including the radiotherapy hardware with the attenuation correction. The contours were compared using the following metrics: the volumetric Dice coefficient, the mean distance to agreement and the GTV volume, all calculated within RayStation. Due to the large variation between patients in GTV volume, the comparisons between CTAC$_{\mathrm{std}}$ and CTAC$_{\mathrm{cba}}$ PET images were performed as per-patient percentage differences (CTAC$_{\mathrm{std}}$ - CTAC$_{\mathrm{cba}}$) relative to the CTAC$_{\mathrm{cba}}$ result. The significance of these differences were evaluated using paired t-tests, with a Bonferroni corrected significance level of $p = 0.05/8 = 0.006$.

### 5.2.6 Metabolic Parameter Analysis

The manual and thresholded GTV contours were compared on metabolic parameters: SUV$_{\mathrm{max}}$, SUV$_{\mathrm{mean}}$ and TLG. TLG was defined as the multiplication of SUV$_{\mathrm{mean}}$ with GTV volume. These would not directly affect tumour delineation using PET, but have shown value as a prognostic factor for rectum patients [141] and so would have an impact on dose painting approaches or the personalisation of dose prescriptions based on the PET data. The large variation between patients in values meant the metabolic parameters were also evaluated as per-patient percentages differences. Statistical significance was assessed using paired t-tests with the same significance level ($p = 0.006$).

The impact on the prognostic value of PET imaging in the radiotherapy position of using CTAC$_{\mathrm{cba}}$ and CTAC$_{\mathrm{std}}$ was assessed using TLG according to the methods presented in [141,236]. Literature cut-off values were only available for rectal cancers so the anal cancer patients were not included in this analysis. The volume used in the TLG calculation was thresholded using either 30% (Ogawa et al.) or 50% (Choi et al.) of SUV$_{\mathrm{max}}$. Although neither study used the 40% of $SUV_{max}$ threshold used in this study, the thresholds were within 10% which was considered similar enough to apply the cut-off values. For Ogawa et al. TLG was determined for a combination of primary and nodal disease, whereas for Choi et al. only primary volumes were used. Therefore the TLG for the primary GTVs were compared to a cut-off value of 125.84 g (from Choi et al.) and the combined primary and nodal TLGs compared to a cut-off value of 341 g (from Ogawa et al.). TLG values were calculated using $\mathrm{CTAC_{std}}$ images and compared to the cut-offs, with patients recorded as good or poor prognosis. This was then repeated using the $\mathrm{CTAC_{cba}}$ images. Patients who were good prognosis when TLG was calculated using $\mathrm{CTAC_{std}}$ images but poor prognosis when TLG was calculated using $\mathrm{CTAC_{cba}}$ were recorded.

## 5.3 Results

### 5.3.1 Whole Image

The distribution of SUVs across the image in $\mathrm{CTAC_{std}}$ and $\mathrm{CTAC_{c}}$ were lower than $\mathrm{CTAC_{cba}}$. This is apparent in the histogram plots of differences between $\mathrm{CTAC_{std}}$ and $\mathrm{CTAC_{c}}$ to $\mathrm{CTAC_{cba}}$ (figure 5.4). The mean differences were $-13.8\%$ ($\mathrm{CTAC_{std}}$) and $-7.7\%$ ($\mathrm{CTAC_{c}}$).

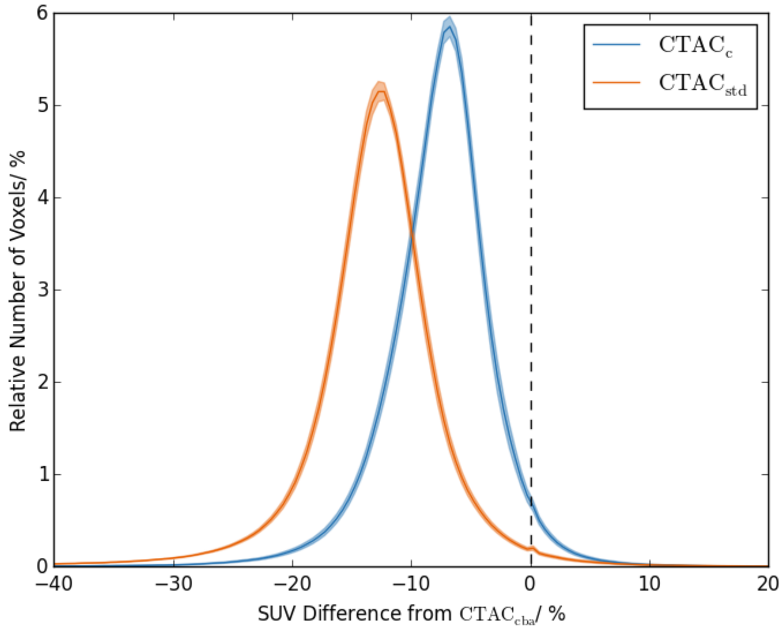


Figure 5.4: Histogram of relative number of voxels with percentage differences in SUV to $\mathrm{CTAC_{cba}}$ for $\mathrm{CTAC_{std}}$ (blue) and $\mathrm{CTAC_{c}}$ (orange). Relative number of voxels given as percentage of total voxels within patient external contour. Solid lines show mean counts over all patients for each bin, and shaded areas ± one standard error. The dashed line indicates zero difference.

### 5.3.2 Tumour Delineation Analysis

16 primary and 10 nodal GTVs were delineated. One patient was being treated post-surgery and had no primary GTV. The primary GTVs were larger and more FDG-avid than the nodal GTVs (figure 5.5). The manual primary GTV volumes were larger than the

thresholded volumes, $44.3 \pm 14.3$ cm$^3$ (mean ± standard error, range 2.4 cm$^3$,239.4 cm$^3$) and $18.9 \pm 5.8$ cm$^3$ (0.7 cm$^3$,95.7 cm$^3$) respectively. The nodal volumes were more similar, with the manual volumes being $15.6 \pm 6.9$ cm$^3$ (0.4 cm$^3$,66.7 cm$^3$ ) compared to $7.7 \pm 2.7$ cm$^3$ (0.8 cm$^3$,22.0 cm$^3$) for the thresholded volumes.

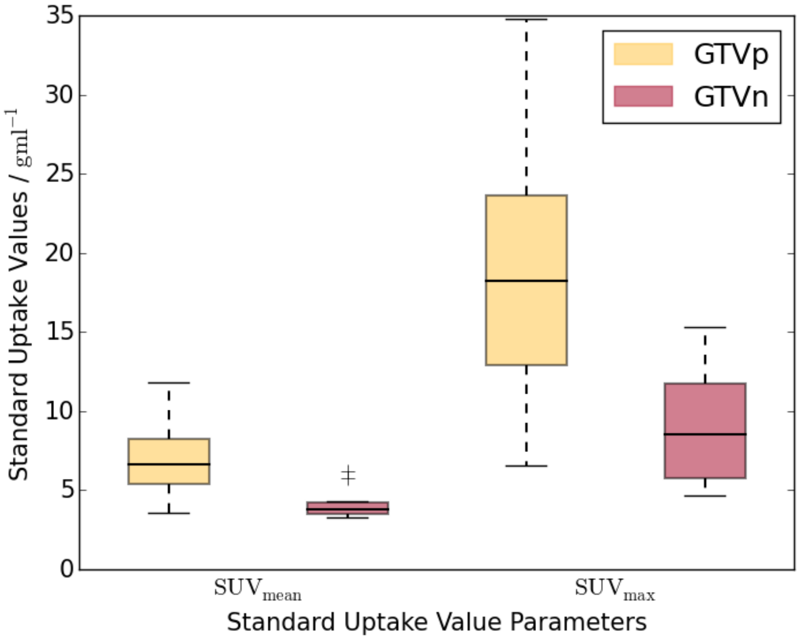


Figure 5.5: Boxplot of the values of SUV$_{\text{mean}}$ and SUV$_{\text{max}}$ for the primary (GTVp, yellow) and nodal (GTVn, red) GTVs. The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points. Data is shown for the manual delineations using the CTAC$_{\text{cba}}$ reconstructed PET images.

There was a difference in the manual GTV volumes between CTAC$_{\text{std}}$ and CTAC$_{\text{cba}}$, with the GTV$^{\text{man}}_{\text{std}}$ volumes being $-15.9 \pm 1.6\%$ (mean ± standard error, range −33.1%,−3.8%) different from the GTV$^{\text{man}}_{\text{cba}}$ volumes. This difference was not statistically significant ($p = 0.007$) over the whole cohort, but appeared to be larger for the less FDG-avid tumours (see figure 5.6). All volume differences greater than 13% occurred in GTVs with $SUV_{mean} \leq 8.5$ g ml$^{-1}$. However there remained a reasonable concordance between GTV$^{\text{man}}_{\text{std}}$ and GTV$^{\text{man}}_{\text{cba}}$ with a Dice coefficient of $0.89 \pm 0.01$ (0.77,0.97) and a mean distance to agreement of $0.65 \pm 0.06$ mm (0.14 mm,1.4 mm). The Dice coefficient also showed some dependence on SUV$_{\text{mean}}$, although less marked than the volume differences (see figure 5.7).

The threshold GTVs were much more similar, with GTV$^{\text{thresh}}_{\text{std}}$ volumes $-2.3 \pm 0.8\%$ (−13.7%,4.4%) different to GTV$^{\text{thresh}}_{\text{cba}}$ ($p = 0.07$, figure 5.6). Similarly, there was very good concordance between the threshold GTVs, the mean Dice coefficient was $0.98 \pm 0.00$ (0.93,1.00) and the mean distance to agreement was $0.12 \pm 0.02$ mm (0.00 mm,0.28 mm).

### 5.3.3 Metabolic Parameter Analysis

There was a substantial drop in the metabolic GTV parameters on the CTAC$_{\text{std}}$ images compared to CTAC$_{\text{cba}}$ images (see figure 5.8). The mean percentage difference for the

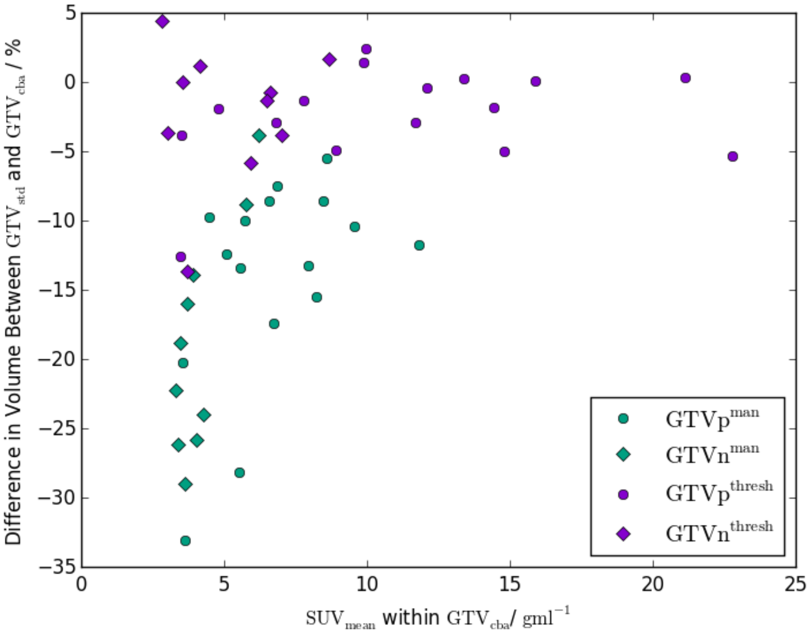
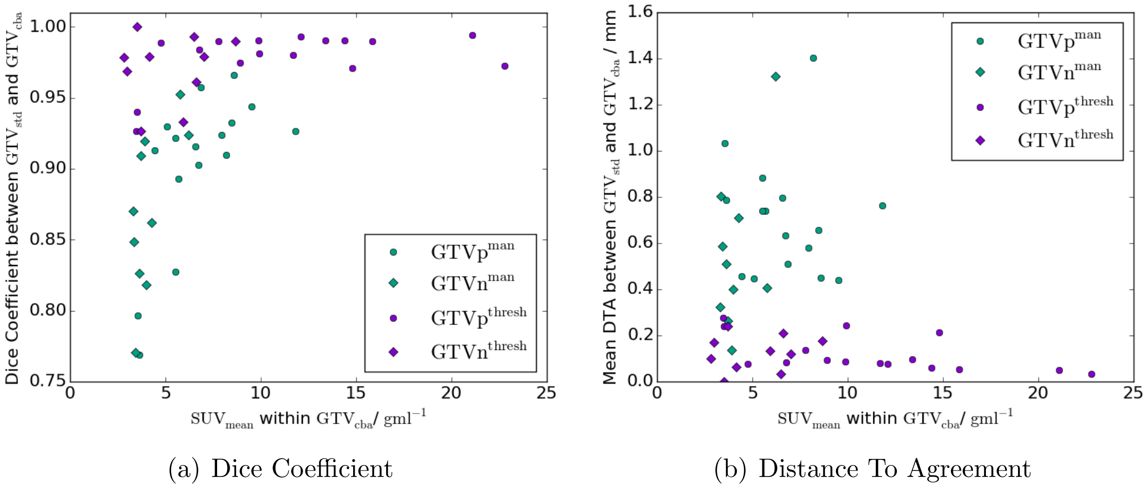


Figure 5.6: Plot of the difference in GTV volume between CTAC$_{std}$ to CTAC$_{cba}$ PET images as a function of the mean SUV within GTV$_{cba}$ . Both manual contours (green) and thresholded contours (purple) are shown. Primary GTVs are represented as circles and nodal GTVs as diamonds.



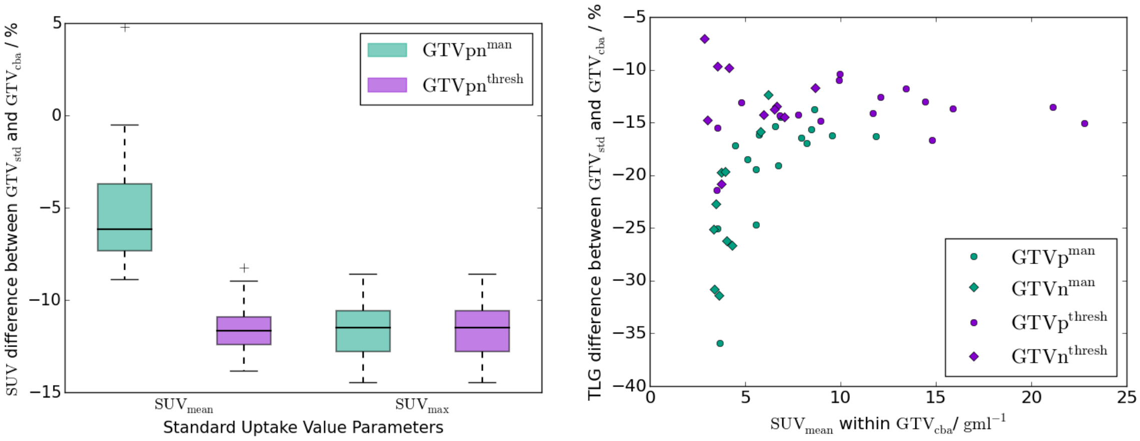
(a) Dice Coefficient

(b) Distance To Agreement

Figure 5.7: Plot of similarity metrics Dice coefficient (a) and mean distance to agreement (DTA, b) between GTV$_{std}$ and GTV$_{cba}$ as a function of SUV$_{mean}$. Manual contours (green) and thresholded contours (purple) are shown, with primary GTVs (circles) and nodal GTVs (diamonds) also distinguished.

manual contours of SUV$_{max}$ was $-11.5\pm0.3\%$ ($-14.5\%$,$-8.6\%$), SUV$_{mean}$ was $-5.2\pm0.6\%$ ($-8.9\%$,4.8%) both with $p < 0.001$, and TLG was $-20.5 \pm 1.2\%$ ($-35.9\%$,$-12.4\%$, $p = 0.005$). The equivalent values for the threshold contours were smaller, but still significant with SUV$_{max}$ being $-11.5 \pm 0.3\%$ ($-14.5\%$,$-8.6\%$, $p < 0.001$), SUV$_{mean}$ $-11.6 \pm 0.3\%$ ($-13.8\%$,$-8.2\%$, $p < 0.001$) and TLG $-13.7 \pm 0.6\%$ ($-21.4\%$,$-7.1\%$, $p = 0.003$).

Comparing the calculated TLG values to the TLG cut-off values gave 2/8 (using the Ogawa et al figure) or 5/8 (Choi et al. figure) rectum cancer patients in the poorer prognosis group. Importantly, one patient changed from the good prognosis to poor prognosis group when TLG was calculated using CTAC$_{cba}$ rather than CTAC$_{std}$, using



(a) Standard Uptake Value

(b) Total Lesion Glycolosis

Figure 5.8: (a) Box plot of the differences in SUV$_{\text{mean}}$ and SUV$_{\text{max}}$ between CTAC$_{\text{std}}$ and CTAC$_{\text{cba}}$ PET images for both primary and nodal GTVs. The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points. (b) shows the difference in TLG between CTAC$_{\text{std}}$ and CTAC$_{\text{cba}}$ images as a function of SUV$_{\text{mean}}$, with primary GTVs indicated as circles and nodal GTVs diamonds.For both plots manual GTVs are shown in green and thresholded GTVs in purple.

the Choi et al. cut-off value. This patient was not the patient who received the lower activity injection.

## 5.4 Discussion

PET-MR imaging has the potential to improve GTV delineation as well as enable dose painting and dose escalation treatment strategies for ano-rectal radiotherapy. This study has aimed to assess the impact of using a novel attenuation correction method on GTV delineation and GTV metabolic parameters for PET-MR imaging in the radiotherapy position in ano-rectal cancer patients.

The impact on manual GTV delineation was small. Although there were some larger volume differences ($> 15\%$) the similarity metrics still showed good ageement for most patients. The levels of agreement were similar to or better than inter-observer variability in GTV delineation in rectal cancer patients reported in the literature. Patel et al. reported PET-CT delineated primary GTVs had Dice coefficients of $0.81 \pm 0.03$ (mean ± standard error) and nodal GTVs $0.70 \pm 0.12$ [238]. Buijsen et al reported higher Dice coefficients, 0.90 for manual delineations and 0.96 for an automatic delineation using a source-to-background ratio method, also for rectal GTVs [239]. This implies that using CTAC$_{\text{std}}$ compared to CTAC$_{\text{cba}}$ does not introduce differences in manual GTV delineation larger than those introduced by inter-observer variability. No study has investigated the impact on ano-rectal target delineation using PET images acquired in the pelvic radiotherapy position so no comparisons with previous literature results could be made.

There was a marked dependence on mean SUV for the volume differences, with much larger differences for the less FDG-avid lesions (figure 5.6). This was probably due to the shallower gradients in SUV around the lower $\text{SUV}_{\text{mean}}$ GTVs meaning an 13.8% shift in SUV from using $\text{CTAC}_{\text{cba}}$ resulted in a larger volume expansion than in the more FDG-avid lesions. The Dice coefficient showed a similar if less pronounced trend with $\text{SUV}_{\text{mean}}$, with all values $< 0.90$ occuring for GTVs with $\text{SUV}_{\text{mean}} < 6$ gml$^{-1}$. This suggests that using $\text{CTAC}_{\text{cba}}$ may be more important in GTV delineation for less FDG-avid lesions. All except one of the manual nodal GTVs had $\text{SUV}_{\text{mean}} < 6$ gml$^{-1}$ (figure 5.5), so nodal delineations may require accurate attenuation correction to avoid under-segmentation.

The impact on the thresholded GTV delineations of using $\text{CTAC}_{\text{cba}}$ was much less than on the manual delineations, with small volume differences and high similarity metric scores. This was likely due to the fact that the threshold SUV was a relative value (40% of $\text{SUV}_{\text{max}}$), and so the $\sim 13\%$ shift in SUV changed both $\text{SUV}_{\text{max}}$ and the boundary voxels by approximately the same amount, resulting in a very similar volume. In contrast, the manual delineation used a fixed $SUV = 2.5$ gml$^{-1}$ threshold as the starting point for delineation, which means the increase in SUVs from using $\text{CTAC}_{\text{cba}}$ resulted in a larger volume delineated.

There was a much bigger impact from using $\text{CTAC}_{\text{cba}}$ rather than $\text{CTAC}_{\text{std}}$ on the metabolic parameters. There were statistically significant differences in $\text{SUV}_{\text{mean}}$, $\text{SUV}_{\text{max}}$ and TLG for both manual and thresholded GTVs. The differences in $\text{SUV}_{\text{mean}}$ for the thresholded volumes and $\text{SUV}_{\text{max}}$ for both manual and thresholded volumes were very similar to each other, with median differences similar to the $-13.8\%$ mean per-pixel SUV difference. The differences in $\text{SUV}_{\text{mean}}$ for the manual volumes were smaller and more variable. This was likely due to the changes in manual GTV volume with the two attenuation corrections, with the larger volumes on the $\text{CTAC}_{\text{cba}}$ images lowering the $\text{SUV}_{\text{mean}}$ and so partially offsetting the 13.8% increase in per-pixel SUVs. One GTV actually had a larger $\text{SUV}_{\text{mean}}$ in the $\text{CTAC}_{\text{std}}$ than the $\text{CTAC}_{\text{cba}}$. Examination of this volume indicated that the $\text{GTV}_{\text{cba}}^{\text{man}}$ extended over two more axial slices than $\text{GTV}_{\text{std}}^{\text{man}}$. This meant $\text{GTV}_{\text{cba}}^{\text{man}}$ included more lower SUV pixels, which reduced the $\text{SUV}_{\text{mean}}$ to 5% less than $\text{GTV}_{\text{std}}^{\text{man}}$, even though the $\text{SUV}_{\text{max}}$ in $\text{GTV}_{\text{cba}}^{\text{man}}$ was 10% higher than in $\text{GTV}_{\text{std}}^{\text{man}}$. This may also be in part due to the difficulty in identifying the primary tumour due to physiological uptake at the adjacent bowel. TLG was also significantly lower on the $\text{CTAC}_{\text{std}}$ images, with a similar dependence on $\text{SUV}_{\text{mean}}$ as the volume differences.

The clinical significance of these these statistically significant differences in metabolic parameters was difficult to determine. PET-CT SUV measurements have indicated a test-retest repeatability of $10 - 12\%$ in tumour SUVs when performed under carefully controlled conditions in a research setting [240]. In clinical diagnostic settings variability in SUVs is likely to be $15 - 20\%$ [179]. This is a similar order of variability as the error in SUVs reported in this study. However, the repeatability was determined using gold standard CT AC, and so failing to include the radiotherapy hardware in the AC map would generate an additional, systematic, bias to the SUV measurements. In addition, the SUV accuracy requirements for using SUV measurements for radiotherapy dose painting or treatment response assessment are higher than for routine clinical diagnostic purposes, suggesting the differences in metabolic parameters may be even more clinically significant in this context [241].

One way of investigating this is considering the use of PET metabolic parameters for treatment prognosis. This is a pre-cursor to using SUVs for dose painting or response assessment. Several studies have provided evidence that TLG measured in a pre-treatment $^{18}$F-FDG-PET scan are independent prognostic factors for disease-free and overall survival in rectal cancer patients [141, 236]. Using the the TLG cut-off value of Choi et al., 1/8 rectal cancer patients changed prognosis group when SUVs were calculated using $\text{CTAC}_{\text{cba}}$ instead of $\text{CTAC}_{\text{std}}$. If these prognostic factors are used to guide radiotherapy dose prescriptions, this indicates that acquiring accurate PET images which account for the attenuation of the radiotherapy hardware could be critical.

Only one study has assessed the impact on metabolic parameters when acquiring PET-MR images in the body radiotherapy position. Paulus et al. evaluated differences in three lung cancer patients scanned on a Siemens PET-MR scanner [156]. Images were acquired with and without the anterior array coil on a coil bridge but with the flat couch top in both cases. The differences in $\text{SUV}_{\text{mean}}$ and $\text{SUV}_{\text{max}}$ between no coil and bridge, and coil and bridge images, without attenuation correction, was $-10.0 \pm 2.4\%$ and $-11.1 \pm 2.0\%$ respectively. Including attenuation correction of the anterior coil and coil bridge reduced this to $-2.4 \pm 3.3\%$ ($\text{SUV}_{\text{mean}}$) and $-3.9 \pm 2.6\%$ ($\text{SUV}_{\text{max}}$). These results are not directly comparable to the results reported in this study because they included the flat couch top in the attenuation correction for both images (ie more similar to a comparison between $\text{CTAC}_{\text{cba}}$ and $\text{CTAC}_{\text{c}}$ rather than $\text{CTAC}_{\text{std}}$). The per-pixel difference in SUV in this study between $\text{CTAC}_{\text{cba}}$ and $\text{CTAC}_{\text{c}}$ of $-7.5\%$ agrees very well with the differences reported by Paulus et al. ( $-7.6\%$ for $\text{SUV}_{\text{mean}}$ and $-7.2\%$ for $\text{SUV}_{\text{max}}$).

A weakness of the methodology presented here is there has been no comparison to a patient acquisition without the radiotherapy hardware and anterior coil present. This would have provided a gold standard PET image to compare the performance of the different AC maps. However, this would also have introduced several confounding variables between the images in the radiotherapy and gold standard setups. These would include differences due to difficulties in registering patient images acquired in different setups and differences in SUV distribution due to imaging at a different time-point. In addition, it would have added significant imaging time for patients. Therefore it was decided to compare AC maps with and without the radiotherapy hardware included to assess the impact of changing the AC map. The prior phantom work in chapter 4 suggested that SUV measurements with the hardware included in the AC map were significantly closer to the gold standard than without, so it is reasonable to assume the same applies in patients.

The other major component of attenuation correction for PET-MR image reconstruction is accounting for the attenuation of the patient. The current vendor supplied method uses a Dixon MR sequence to segment the patient into fat, water and air tissue classes which are then assigned linear attenuation coefficients [242]. This has been shown to introduce SUV errors which are investigated in more detail in chapter 6. In this study this problem was avoided by using the registered radiotherapy planning CT image for patient attenuation correction. Improved methods for accounting for patient attenuation in PET-MR images are currently being investigated. From a radiotherapy perspective, algorithms used to generate synthetic CTs from MR images for MR-only radiotherapy are in clinical use [28, 81]. One of these algorithms has demonstrated improvements in patient PET attenuation correction compared to the previous Dixon-based method [31], although the magnitude of the difference in SUVs was less than half of the discrepancy reported here. This suggests that incorporating the radiotherapy hardware is more important for accurate PET quantification than accurate patient attenuation correction, although combining accurate attenuation correction maps of patient and radiotherapy hardware would result in the highest accuracy PET images. An investigation into patient attenuation correction and the combination with the results of this chapter are the focus of chapter 6.

## 5.5 Conclusion

Acquiring PET-MR images for radiotherapy planning requires patients to be imaged in the radiotherapy position on a flat couch with a coil bridge. Applying attenuation correction maps that incorporate this hardware was feasible in radiotherapy patients and resulted in a 13.8% increase in SUVs. This did not have a statistically significant change in GTV delineation, although differences were more pronounced for less FDG-avid volumes. It did have large differences in metabolic measurements, which were statistically significant ($p \leq 0.005$). These differences could be clinically relevant where metabolic parameter measurements are used for dose painting or treatment prognosis, as indicated by 1/8 rectum patients changing prognosis group when the radiotherapy hardware was included in the attenuation correction map. This suggests that it is important that attenuation correction of the radiotherapy hardware is incorporated if PET-MR images in the radiotherapy position are to be used for dose painting and treatment prognostication.

# Chapter 6

# Evaluating a Deep Learning sCT Algorithm for PET-MR Attenuation Correction in the Pelvis

## 6.1 Introduction

Simultaneous PET-MR enables high quality anatomic, functional and metabolic information to be acquired with high degrees of spatial alignment in the same imaging session [243]. This has potential benefits for improved staging and treatment response assessment in rectal cancer [244, 245] as well as improved GTV delineation [130] and identification for active tumour sub-volumes for dose painting [246]. This utilises the superb soft-tissue contrast of MR anatomical imaging, as well as its functional imaging ability such as DW-MR [22]. However, this comes at the cost of accurate attenuation correction of the patient, compared to PET-CT. Conventional MR images provide little signal from both low PET attenuating materials such as air and high PET attenuation materials such as cortical bone [247]. Therefore there is no one-to-one map possible from MR intensity values to linear attenuation coefficients [32]. The current vendor-supplied solution for the pelvis, MRAC, utilises a Dixon MR sequence to segment air, lung, fat and soft-tissue compartments, which are then assigned population values [242]. This however introduces PET attenuation errors through the omission of any bone information, with reported $\text{SUV}_{\text{max}}$ errors in soft-tissue lesions of $-6\%$ and in bone lesions of $-11\%$ [77].

This situation is very similar to the problem faced within MR-only radiotherapy, where MR cannot be used directly for radiotherapy dose calculations [28]. Therefore there is potential to use sCT algorithms designed for radiotherapy dose calculation for PET attenuation correction. This could facilitate a streamlined workflow with a single radiotherapy planning PET-MR examination as a ‘one-stop shop’ [30].

Two previous studies have investigated applying radiotherapy sCT algorithms for PET attenuation correction in the pelvis. Wallstén et al. used an atlas-based sCT derived from T2-weighted MR images for PET attenuation correction in 12 prostate cancer patients [31]. They reported reduced mean SUV differences in bone regions from $-17.7 \pm 8.4\%$ (MRAC) to $-4.2 \pm 5.7\%$ (sCTAC) which translated into significantly reduced SUV differences in the PET-avid prostate sub-volume of $-2.3\%$ (sCTAC) compared to $-5.9\%$ (MRAC, $p < 0.001$). Ahangari et al evaluated a Deep Learning sCT algorithm based on the Dixon MR sequence used for vendor-supplied MRAC for cervix radiotherapy patients [30]. The model was trained with 26 patients and evaluated on seven, with small mean differences in tumour $\mathrm{SUV}_{\mathrm{max}}$ of $-0.8 \pm 1.2\%$ (± standard error, range $-4.9\%$,4.7%, estimated from bar graph) and in $\mathrm{SUV}_{\mathrm{mean}}$ $-0.3 \pm 1.8\%$ ($-5.9\%$,7.4%).

However, to the best of the author's knowledge, no study has investigated applying a ZTE-based radiotherapy sCT for PET attenuation correction in ano-rectal cancer patients. ZTE imaging provides MR signal from bone. Since this is the primary deficiency in the current MRAC technique, ZTE-based sCT algorithms potentially could improve PET attenuation correction significantly [77]. The aim of this study was to apply a ZTE-based Deep Learning sCT algorithm developed for pelvic MR-only radiotherapy dose calculations (chapter 2) to PET-MR attenuation correction for ano-rectal cancer patients. Since the aim was to evaluate the equivalence in PET-MR attenuation correction between sCT and CT, the statistical analysis carried out would not be conventional superiority testing (as in chapter 5) but equivalence testing [248]. This has been applied in the MR-only radiotherapy literature [249] but has not been used previously for PET-MR attenuation correction analysis.

Successful use of the MR-only sCT algorithm for PET-MR attenuation correction, combined with attenuation correction of the radiotherapy hardware (chapters 4 and 5) and sCT radiotherapy dose accuracy (chapter 2), would enable a single PET-MR session to provide all the information required for accurate PET and MR imaging for GTV delineation and characterisation, OAR delineation and radiotherapy dose calculation in a PET-MR-only radiotherapy workflow.

## 6.2 Materials and Methods

### 6.2.1 Patient Data Collection

The study population consisted of 10 patients (four male and six female) who were all enrolled in the Deep MR-only RT study (research ethics committee reference 20/LO/0583) and received a PET-MR scan. Patients were diagnosed with anal cancer (n=6) stages T1/2N0M0-T2N1M0 and rectal cancer (n=4) stages T2N0M0-T3b/T4N0M0, and had a median age of 65 years (range 49-76). All patients were planned for radical/neoadjuvant chemoradiotherapy. Patients were excluded if they were contraindicated for MR scanning, had medical implants in the pelvic area (eg hip prostheses), were unable to fit inside the coil bridge or were unable to fast for 6 hours.

All patients received a simultaneous PET-MR scan on a SIGNA PET/MR 3T scanner (version MP26 GE Healthcare, Waukesha, USA) after their radiotherapy planning CT scan and before their first treatment fraction. Patients were scanned in the radiotherapy treatment position on a flat couch-top with a coil bridge supporting the anterior MR coil. Patients were setup in a combined customisable foot and knee rest (Civco), with their position adjusted to match external lasers to the radiotherapy patient tattoos. Patients emptied their bladder and drank 400 ml of water immediately before entering the scan room, with the PET acquisition starting 23 minutes (median, range 14-37 minutes) after patient drinking. All patients had fasted for 6 hours prior to injection and had a measured blood glucose concentration of $< 10$ mmol$\mathrm{L}^{-1}$. Patients were injected with 3.5 $\mathrm{MBqkg}^{-1} \pm$ 10% of $^{18}\mathrm{F} - \mathrm{FDG}$ (one patient received 1.7 $\mathrm{MBqkg}^{-1}$), with PET images starting to be acquired 73 minutes (median, range 60-86 minutes) post-injection. The PET acquisition consisted of one 5 minute bed position with the patient tumour centred in the PET field of view. Images were reconstructed using an OSEM algorithm with 4 iterations and 16 subsets and a 5.0 mm Gaussian filter using manufacturer provided offline reconstruction tool Duetto (version 2.17, GE Healthcare) in MatLab (Version 2017a, MathWorks, Natick, Massachusetts, USA). Point spread function correction and time of flight information were utilised. Images were reconstructed with a $60 \times 60$ cm$^2$ axial field of view, a $256 \times 256$ axial matrix and 89 slices with a slice thickness of 2.78 mm.

Two MR sequences were acquired: a novel ZTE sequence and the standard Dixon sequence used for the scanner-generated PET attenuation correction maps. The ZTE sequence was as described in chapter 2 section 2.2.1 table 2.2. The voxel size was $2.0 \times 2.0 \times 2.0$ mm$^3$, the duration of the sequence was 1.1 minutes and the acquisition started 29 minutes after the PET acquisition started (median, range 27-37 minutes). The Dixon sequence was as described in chapter 5, section 5.2.1. It had a voxel size of $2.0 \times 2.0 \times 5.2$ mm$^3$, with 2.6 mm slice gaps and an acquisition duration of 14.8 s. The Dixon sequence occurred concurrently with the start of the PET acquisition.

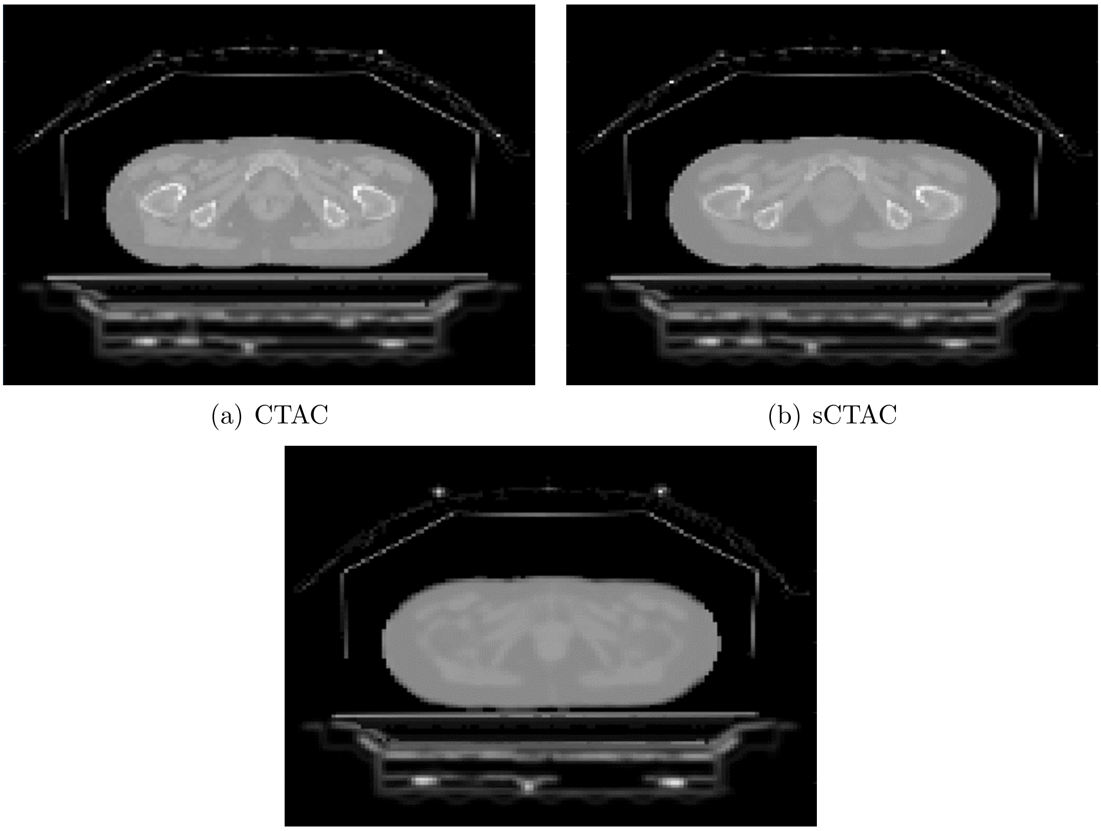
All patients received contrast-enhanced CT scans (Sensation Open, Siemens, Erlangen, Germany) in the radiotherapy planning positon with the same design of foot and knee rest and tattoo marks matched to external lasers. Images were acquired with a voxel size of $1.1 \times 1.1 \times 3$ mm$^3$ and a tube voltage of $V = 120$ kVp. Patients were imaged following routine bladder preparation consisting of an empty bladder 30 minutes prior to the scan, followed by drinking 400 ml of water, and bowel preparation consisting of the application of a micro-enema 60 minutes prior to the scan followed by bowel emptying. CT images were acquired within 6 days (median, range 5-13 days) of the PET-MR scan.

### 6.2.2 Attenuation Correction Maps

PET images were reconstructed with three different attenuation correction maps for each patient. All attenuation correction maps included the coil components within the scanner bed, the radiotherapy couch, coil bridge and anterior coil as described in chapter 5 and a model of the patient. The patient model varied between the different maps. The gold standard patient model (CTAC) consisted of the patient CT acquired in the same radiotherapy position and following the same bladder preparation protocol as the PET-MR and rigidly registered to the in-phase MR image in RayStation (v9B, RaySearch Laboratories, Stockholm, Sweden). The external contour of the in-phase MR was automatically delineated using RayStation's function, and manually modified where necessary. Modifications were required for all patients, typically in the $\sim 5$ most superior and inferior slices where reductions in MR signal caused the automatic contour to miss parts of the patient. The registered CT was cropped to the MR external contour, with any missing tissue set to water density. Any air within the patient was automatically delineated and set to water density. The CT was converted to 511 keV linear attenuation coefficient map using the PET-MR vendor-supplied calibration curve (GE Healthcare). The quality of the CT-MR registration was visually inspected for each patient.

A second map was generated using a sCT generated by a Deep Learning algorithm from the ZTE image, as described in chapter 2, section 2.2.1, without modification. The 10 patients in this study were selected from the 20 evaluation patients described in chapter 2 and so their images had not been used in the model training process. Although the sCT was derived from the ZTE image acquired in the same scanning session as the in-phase MR, it was acquired 29 minutes later (median, range 25-37 minutes). Therefore, in case of patient motion, the sCT was rigidly registered to the in-phase MR image (ie Dixon-based) in RayStation and cropped to the in-phase MR external contour, with any missing tissue set to water density. Similarly to the CT, any air within the sCT was automatically delineated and set to water density. This was converted to 511 keV linear attenuation coefficient map using the same calibration curve to produce the sCTAC.

The final map used the standard vendor-supplied (GE Healthcare) patient model derived from the automatic Dixon sequence (MRAC). This method segmented the MR into four tissue classes: air, lung, fat and soft-tissue, and assigned population-derived bulk density 511 keV linear attenuation coefficients to each class [242]. The MRAC was also cropped to the same in-phase MR external contour as the CT and sCT to ensure all attenuation correction maps had the same external contour. Examples of the three attenuation maps are shown in figure 6.1.



(a) CTAC

(b) sCTAC

(c) MRAC

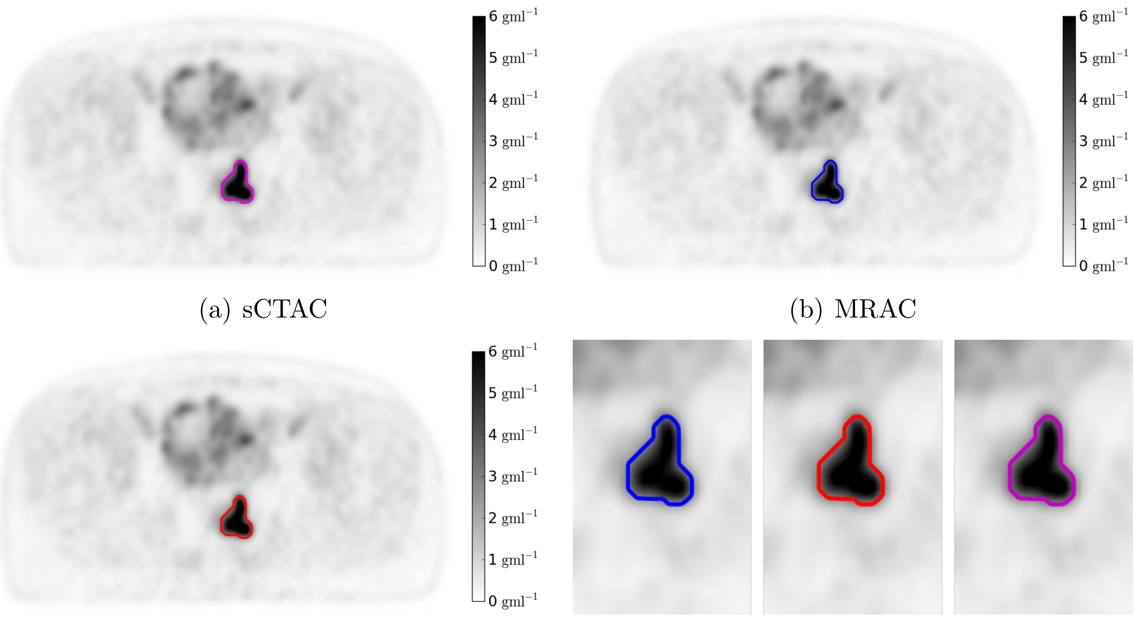
Figure 6.1: Attenuation correction maps for an example patient. All maps included the MR coils components, flat couch top and coil bridge. Patient models were based on CT (a), ZTE-derived sCT (b) and Dixon derived MR (c).

### 6.2.3 Tumour Delineation

The manual GTVs contoured on the $\text{CTAC}_{\text{cba}}$ PET images described in chapter 5 were used to create threshold contours on the CTAC, sCTAC and MRAC images. These manual GTVs were copied onto the sCTAC and MRAC PET images and an automatic GTV created as the volume encompassed within 40% of the maximum SUV within the manual GTV [21], within the manual GTV contour. Automatic contours were generated on CTAC, sCTAC and MRAC images and all did not extend beyond the boundary of the manual contour. The manual GTV contour was only used to generate the automatic GTV contours, which were then used in the subsequent analysis. Examples of the three PET images and automatic GTVs are shown in figure 6.2.

### 6.2.4 Data Analysis

The per pixel percentage difference in SUV for MRAC - CTAC and sCTAC - CTAC relative to CTAC were calculated using MICE Toolkit (v2021.2.1) [204]. Only differences within the CTAC external contour were included. This was automatically contoured using a threshold of 0.05 $\text{gml}^{-1}$ and the same contour applied to the sCTAC and MRAC images. Relative SUV differences were binned into 400 bins between $-100\%$ and $+100\%$



(a) sCTAC

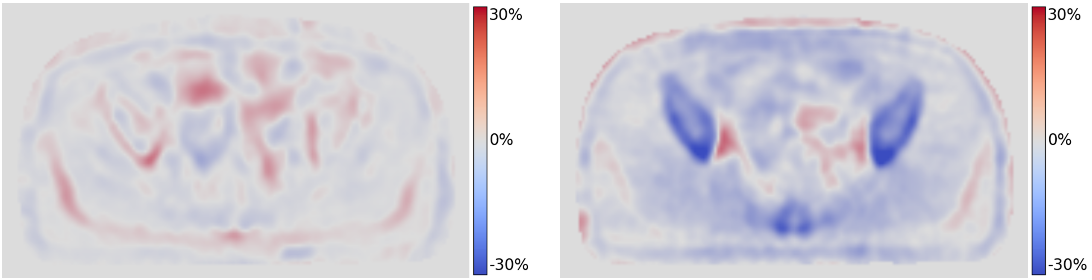
(b) MRAC

(c) CTAC

(d) Zoomed GTVs

Figure 6.2: Example PET images reconstructed using the sCTAC (a), MRAC (b) and CTAC (c) attenuation correction maps. The threshold GTV contour is shown in purple, blue and red respectively. Zoomed in pictures of the same GTVs are shown in (d). The patient was selected as having the sCTAC and MRAC SUV$_{\text{max}}$ differences closest to the mean differences.

for each patient, and the mean difference within each bin over all patients determined. An example whole image difference map is shown in figure 6.3.



(a) sCTAC

(b) MRAC

Figure 6.3: Example SUV difference maps to CTAC PET images for the sCTAC PET image (a) and MRAC PET image (b) for the same patient and slice as shown in figure 6.2.

A major discrepancy between the MRAC and gold standard CTAC is that the MRAC does not reproduce bone. Therefore SUV differences in the bone region were additionally investigated. A bone region of interest was automatically delineated on the CT using the ‘Bone ROI’ function in RayStation. This uses thresholding with all voxel $>$ 150 HU included and all voxels $<$ 50 HU excluded. A connected regions function is used to determine whether to include voxels with HU values in between the two thresholds. The resulting contour was expanded by 5 mm and then contracted by the same amount to remove small holes and smooth the overall contour. The per pixel percentage difference in SUV was calculated as described above but only within the region masked by the bone contour.

The similarity of the automatic GTV contours on sCTAC and MRAC PET images to CTAC was determined using the volumetric Dice similarity coefficient (DSC), the mean and maximum distances to agreement and the GTV volume, all calculated within RayStation. The accuracy of the calculation of a set of metabolic parameters on the sCTAC and MRAC GTVs was determined by comparing to CTAC measurements. The metabolic parameters assessed were: $\mathrm{SUV_{max}}$, $\mathrm{SUV_{mean}}$ and TLG. Large variation between patient SUVs meant the volume and metabolic comparisons between sCTAC and MRAC with CTAC were carried out as per-patient percentage differences relative to CTAC.

The SUV results were statistically tested for equivalence, with a null hypothesis that the sCTAC/MRAC PET images were different to the CTAC images. This is the opposite to conventional superiority testing (such as used in chapter 5) which aims to determine if differences are statistically significant and has a null hypothesis that the sCTAC/MRAC PET images are not different to CTAC images. Equivalence between MRAC/sCTAC and CTAC was assessed using two one sided t-tests for paired data [248]. Tests were done for differences in $\mathrm{SUV_{max}}$ and $\mathrm{SUV_{mean}}$, for both MRAC and sCTAC and both primary and nodal GTVs (ie 8 tests in all). A significance level of $p < 0.05$ was used, corrected for multiple testing by $p < 0.05/(8-1) = 0.007$ [250]. This multiple testing correction is specifically for equivalence testing, unlike the Bonferonni correction used in chapter 5.

Equivalence testing considers sCTAC/MRAC PET images clinically equivalent to CTAC images if the SUV differences are smaller than a pre-defined equivalence margin, which is the maximum difference that would be considered clinically unimportant. There were no reported equivalence margins for PET attenuation correction in the literature. Therefore, an equivalence margin was defined as the maximum difference that would not increase the overall literature PET-CT SUV uncertainty by $\geq 0.5\%$ (ie would be the same to two significant figures). The method assumed that the only additional SUV uncertainty from PET-MR compared to PET-CT was due to attenuation correction, which was independent of all other PET uncertainties. Therefore the attenuation correction uncertainty can be added in quadrature to the overall PET-CT SUV repeatability to calculate an overall PET-MR SUV uncertainty:

$$\Delta_{PETMR} = \sqrt{\Delta^2_{PETCT} + \Delta^2_{AC}}, \tag{6.1}$$

where $\Delta_{PETMR}$ is the overall PET-MR SUV uncertainty, $\Delta_{PETCT}$ the overall PET-CT uncertainty and $\Delta_{AC}$ is the attenuation correction uncertainty. The equivalence margin, equal to $\Delta_{AC}$, was defined such that $\Delta_{PETMR} - \Delta_{PETCT} < 0.5\%$. Literature values for PET-CT SUV repeatability were 30% ($\mathrm{SUV_{max}}$) and 20% ($\mathrm{SUV_{mean}}$), taken from a meta-analysis of 86 and 102 patients for $\mathrm{SUV_{max}}$ and $\mathrm{SUV_{mean}}$ respectively [251]. Using $\Delta_{PETCT} = 30\%$ for $\mathrm{SUV_{max}}$, this gives an equivalence margin of $\Delta_{AC} = 5\%$. Similarly for $\mathrm{SUV_{mean}}$, a PET-CT uncertainty of $\Delta_{PETCT} = 20\%$ requires an equivalence margin of $\Delta_{AC} = 4\%$. The SUV differences in $\mathrm{SUV_{max}}$ and $\mathrm{SUV_{mean}}$ between MRAC/sCTAC were tested for clinical equivalence using these margins.

## 6.3 Results

sCTs were successfully generated for each patient. The sCT to MR registrations were small but not negligible, with a range of $[-1.1, 4.0]$ mm, $[-6.9, -3.3]$ mm and $[0.8, 2.6]$ mm for the right-left, inferior-superior and posterior-anterior directions respectively. The pitch, roll and yaw angle ranges were $[-0.8, 0.3]^o$,$[-0.2, 1.0]^o$,$[-0.3, 0.4]^o$. There were 9 primary and 5 nodal GTVs contoured, (one patient had no primary following surgery before chemoradiotherapy).

### 6.3.1 Whole Image Analysis

The whole image SUVs in the sCTAC and MRAC reconstructed PET images were lower than those in the CTAC PET images with the mean difference being $-3.0\%$ for the MRAC and $-0.02\%$ for the sCTAC. The distributions of SUV differences were quite different, with the sCTAC SUV differences a much narrower distribution as well as closer to zero (see figure 6.4).

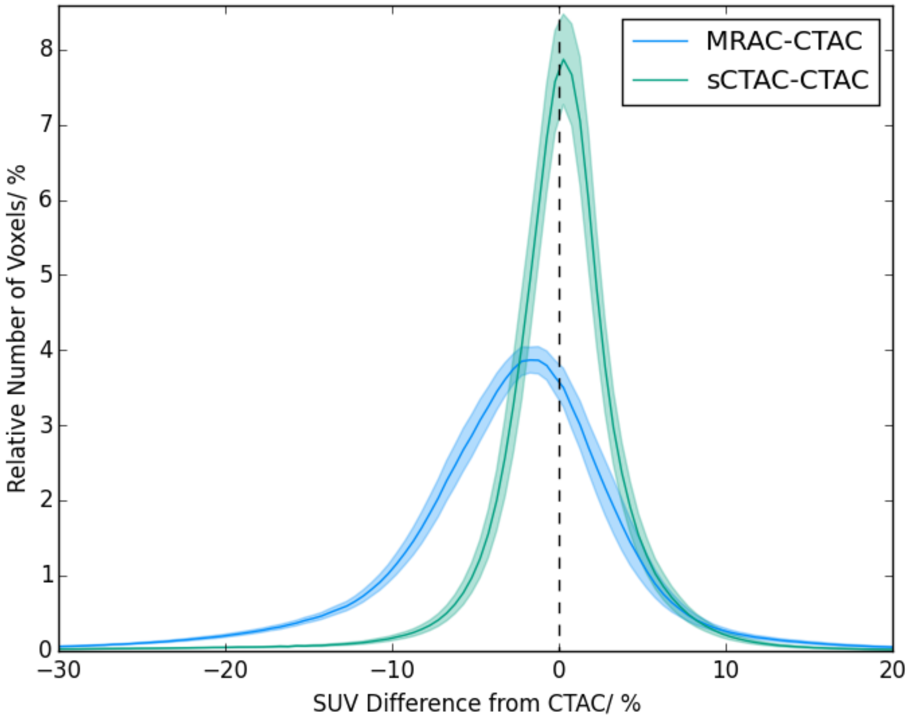


Figure 6.4: Histogram of relative number of voxels with percentage differences in SUV to CTACT for sCTAC (green) and MRAC (blue). Relative number of voxels given as percentage of total voxels within patient external contour. Solid lines show mean counts over all patients for each bin, and shaded areas ± one standard error. The dashed line indicates zero difference.

The differences within the bone mask were much larger than the whole image, with the mean MRAC difference being $-16.3\%$ and the sCTAC difference $-0.5\%$ (see figure 6.5).

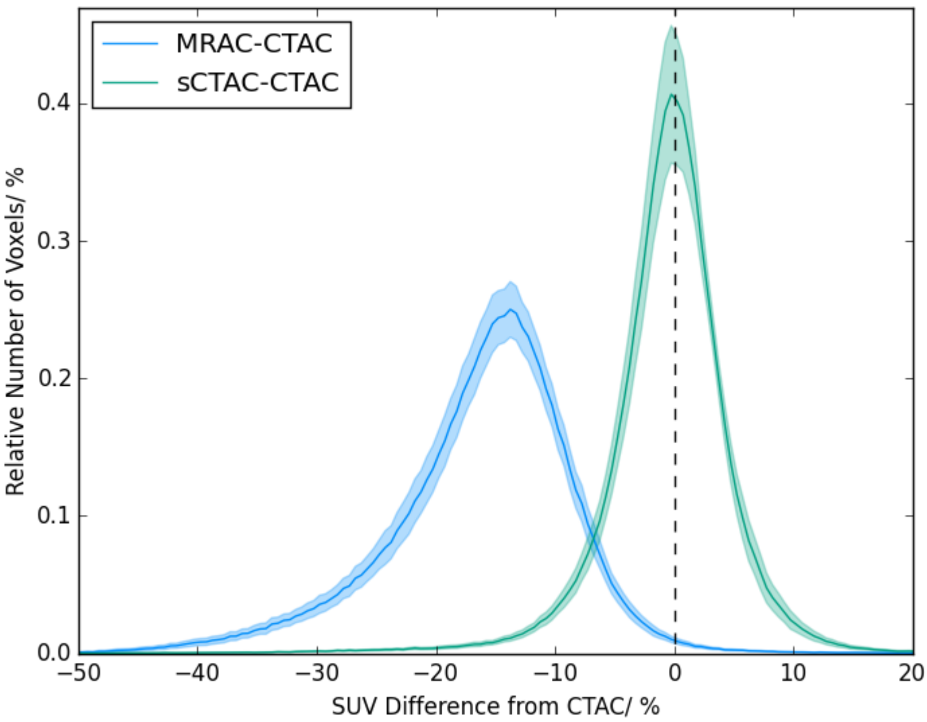


Figure 6.5: Histogram of relative number of voxels with percentage differences in SUV to CTAC for sCTAC (green) and MRAC (blue) within the bone region. Relative number of voxels given as percentage of total voxels within patient external contour. Therefore y-axis is in the same units as figure 6.4 but the scale is different. Solid lines show mean counts over all patients for each bin, and shaded areas ± one standard error. The dashed line indicates zero difference.

### 6.3.2 Tumour Delineation Analysis

The thresholded GTVs for the primary and nodal tumours were very similar to CTAC on both MRAC and sCTAC (see table 6.1). There was no difference between MRAC and sCTAC with the results agreeing within one standard error. Both MRAC and sCTAC volume differences were close to zero (within two standard errors).

Table 6.1: Delineation metrics for GTVs on MRAC and sCTAC. Volume indicates the volume difference between MRAC/sCTAC and CTAC, relative to the CTAC volume. All results given as mean ± standard error (minimum, maximum).

| Metric | GTV | MRAC | sCTAC |
|---|---|---|---|
| DSC | Primary | $0.990 \pm 0.002$ (0.978,0.994) | $0.992 \pm 0.002$ (0.983,0.998) |
| $\text{DTA}_{\text{mean}}$ [mm] | Primary | $0.06 \pm 0.01$ (0.02,0.13) | $0.06 \pm 0.01$ (0.01,0.11) |
| $\text{DTA}_{\text{max}}$ [mm] | Primary | $1.87 \pm 0.24$ (0.97,3.16) | $1.76 \pm 0.13$ (1.23,2.53) |
| Volume [%] | Primary | $0.72 \pm 0.5$ ($-0.4$,4.2) | $0.7 \pm 0.5$ ($-1.3$,3.2) |
| DSC | Nodal | $0.988 \pm 0.006$ (0.968,1.000) | $0.987 \pm 0.008$ (0.955,1.000) |
| $\text{DTA}_{\text{mean}}$ [mm] | Nodal | $0.04 \pm 0.01$ (0.00,0.07) | $0.04 \pm 0.02$ (0.0,0.11) |
| $\text{DTA}_{\text{max}}$ [mm] | Nodal | $1.4 \pm 0.4$ (0.0,2.34) | $1.4 \pm 0.4$ (0.0,2.0) |
| Volume [%] | Nodal | $0.8 \pm 0.3$ (0.0,1.6) | $1.1 \pm 0.6$ ($-0.1$,2.8) |

### 6.3.3 Metabolic Parameter Analysis

There were larger differences between MRAC and sCTAC in the metabolic parameters. In the primary tumours, the mean MRAC differences in SUV from CTAC were $-4.6 \pm 0.9\%$ ($-8.4\%$,$-1.3\%$) for SUV$_{\text{max}}$ and $-4.3 \pm 0.8\%$ ($-9.0\%$,$-1.6\%$) for SUV$_{\text{mean}}$. The sCTAC SUV differences were closer to zero and less dispersed, with differences of $1.0 \pm 0.8\%$ ($-1.7\%$,$6.3\%$) and $1.0 \pm 0.7\%$ ($-1.5\%$,$4.5\%$) for SUV$_{\text{max}}$ and SUV$_{\text{mean}}$ respectively (see figure 6.6). Equivalence testing for SUV$_{\text{max}}$ with a $\pm5\%$ equivalence margin gave $p = 0.33$ and $p < 0.001$ for the MRAC and sCTAC respectively. Using the corrected significance level of $p < 0.007$, this indicated the sCTAC was clinically equivalent to CTAC and the MRAC was not. For the SUV$_{\text{mean}}$ differences with a $\pm4\%$ equivalence margin, again the sCTAC was clinically equivalent to CTAC ($p < 0.001$) and the MRAC was not ($p = 0.21$).

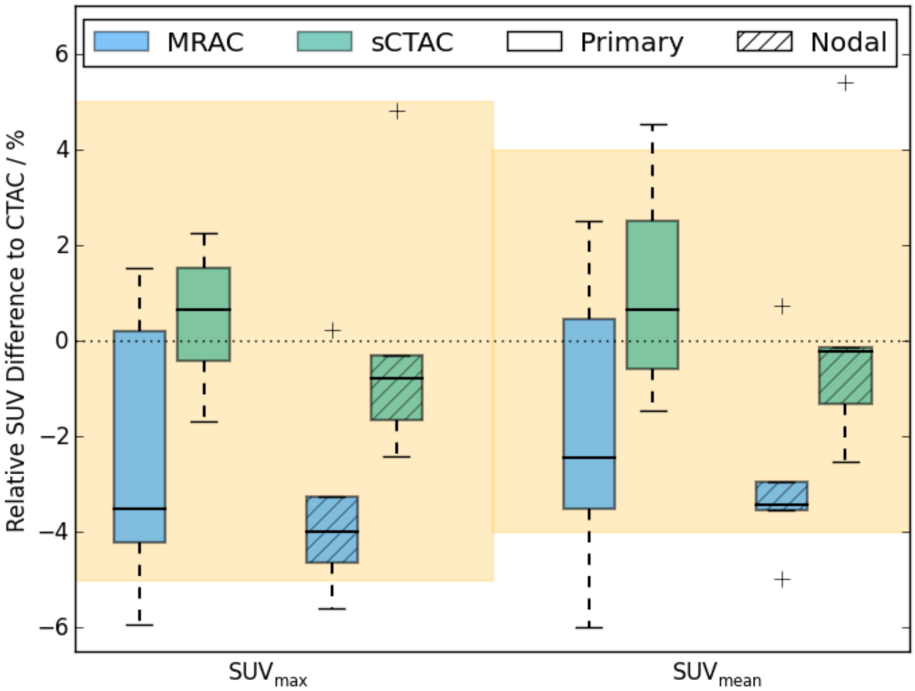


Figure 6.6: Boxplot of SUV differences to CTAC PET images for the MRAC (blue) and sCTAC (green) images. Solid bars indicate primary volumes (n=9) and hatched bars nodal volumes (n=5). The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points. The dotted line indicates zero difference and the yellow filled regions indicate the equivalence margins ($\pm5\%$ for SUV$_{\text{max}}$ and $\pm4\%$ for SUV$_{\text{mean}}$).

The SUV differences between MRAC and sCTAC were greater for the nodal volumes than the primary tumours. The MRAC had SUV$_{\text{max}}$ and SUV$_{\text{mean}}$ differences of $-6.2 \pm 1.3\%$ ($-9.4\%$,$-1.9\%$) and $-6.0 \pm 1.2\%$ ($-9.2\%$,$-2.0\%$). The sCTAC differences were substantially smaller, $-0.1 \pm 1.3\%$ ($-2.4\%$,$4.8\%$) for SUV$_{\text{max}}$ and $0.2 \pm 1.4\%$ ($-2.5\%$,$5.4\%$) for SUV$_{\text{mean}}$. The MRAC SUV parameters were all greater than two standard errors away from zero whereas the sCTAC parameter differences were within two standard errors for the primary tumours and one for the nodal tumours. Equivalence testing for SUV$_{\text{max}}$ with $\pm5\%$ equivalence margin gave not clinically equivalent for either MRAC ($p = 0.20$)

or sCTAC ($p = 0.009$), although sCTAC was approaching equivalence. SUV$_{\text{mean}}$ values (±4% margin) were also not clinically equivalent: $p = 0.23$ (MRAC) and $p = 0.01$ (sCTAC).

The mean TLG differences were also closer to zero for the sCTAC compared to MRAC, although the improvement was less than for the SUV differences. In the primary GTVs, TLG differences were $-3.6 \pm 0.9\%$ ($-9.2\%$,$-0.1\%$) and $1.7 \pm 0.9\%$ ($-1.2\%$,6.0%) for the MRAC and sCTAC respectively (see figure 6.7). For the nodal GTVs the differences were similar, $-5.2 \pm 1.5\%$ ($-9.2\%$,$-0.5\%$, MRAC) and $1.4 \pm 1.9\%$ ($-2.6\%$,8.4%).

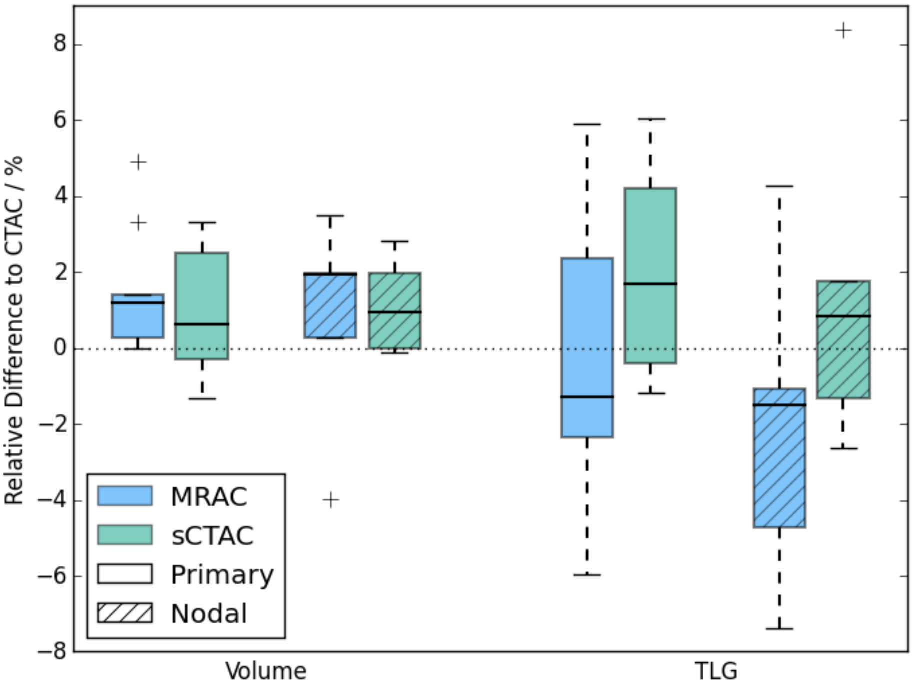


Figure 6.7: Boxplot of volume and TLG differences to CTAC PET images for the MRAC (blue) and sCTAC (green) images. Differences are relative to CTAC value. Solid bars indicate primary volumes (n=9) and hatched bars nodal volumes (n=5). The dotted line indicates zero difference. The rectangles indicate the interquartile range (IQR), with the horizontal black line the median value, the black whiskers the maximum (minimum) data point within $Q3 + 1.5IQR$ ($Q1 - 1.5IQR$) and the black crosses outlier data points.

## 6.4 Discussion

This study has assessed the use of a MR-only radiotherapy sCT algorithm for PET-MR attenuation correction in the pelvis compared to the current Dixon-based MRAC. Both methods were evaluated using CT-based attenuation correction. The sCTAC reduced the whole image SUV difference to CTAC to $-0.02\%$, compared to $-3.0\%$ for the MRAC. This did not translate into improvements in thresholded GTV delineation, with both MRAC and sCTAC having $DSC \geq 0.987$ for primary and nodal GTVs. However, differences in GTV metabolic parameters were larger, with differences to CTAC in SUV$_{\text{max}}$ being $1.0 \pm 0.8\%$ (sCTAC) rather than $-4.6 \pm 0.9\%$ (MRAC) for primary GTVs. SUV$_{\text{max}}$ calculated on the sCTAC was clinically equivalent to CTAC values within a ±5% equivalence margin for primary GTVs with $p < 0.001$ (using a corrected significance value $p < 0.007$), whereas the MRAC was not ($p = 0.33$).

The sCTAC reconstructed PET images had smaller SUV differences to CTAC over the whole image than the MRAC reconstructed images (figure 6.4). The mean sCTAC difference was very close to zero, and the distribution of differences tightly clustered around the mean. In contrast, the MRAC SUV difference was larger ($-3.0\%$) and had a substantially broader distribution of differences. The MRAC had substantially larger differences within the bone mask ($-16.3\%$ compared to $-0.5\%$). Although these only made up a small proportion of the overall voxels in the image, the absence of bone also caused the MRAC to over-estimate SUVs adjacent to the bone (figure 6.3). This suggests that the ZTE-based sCT model is accurately capturing the bone information.

The automatic GTV delineation using thresholds of 40% of $\text{SUV}_{\text{max}}$ were very similar between CTAC and both sCTAC and MRAC. The DSC results ($\geq 0.987$, with 1.0 indicating perfect agreement) were higher than the reported inter-observer variability of 0.96 in automated rectal cancer GTV delineation on the same PET image [239]. This suggests that both the sCTAC and MRAC provide sufficient SUV accuracy for accurate GTV delineation. This fits with results from chapter 5 where larger differences in SUV between PET images still resulted in small changes in GTV delineation.

The differences in patient attenuation produced larger differences in the metabolic parameter analysis. The sCTAC SUV differences to CTAC were smaller than MRAC for both $\text{SUV}_{\text{max}}$ and $\text{SUV}_{\text{mean}}$ for both primary and nodal GTVs, and also had smaller interquartile ranges except for the nodal $\text{SUV}_{\text{mean}}$ differences (figure 6.6). The equivalence analysis suggested that the sCTAC primary GTVs had clinically equivalent SUVs to CTAC (within $\pm5\%$ and $\pm4\%$ for $\text{SUV}_{\text{max}}$ and $\text{SUV}_{\text{mean}}$ respectively) whereas the MRAC SUVs were not. The nodal GTVs were not clinically equivalent for either sCTAC or MRAC, although for the sCTAC this is likely due to outlier results from one patient (see figure 6.6) and there only being five patients having nodal GTVs. Examining the sCT and CT for this patient showed there was a small discrepancy in the bone alignment on the few slices that the nodal GTV was on (it only had a volume of 1.4 $\text{cm}^3$). This resulted in more bone on the sCT for these slices than on the CT, resulting in an over-correction of the attenuation, leading to higher SUVs in the sCTAC image.

This highlights a limitation of this study which is that the CT image used as the gold standard was acquired on a different scanner and on a different day to the PET-MR image. Although both CT and PET-MR images were acquired in the same radiotherapy position with the same immobilisation following the same bladder preparation protocol, there could still be small discrepancies between patient setups, in both external and internal anatomy. These would be confounding differences due to misalignments rather than incorrect HU assignment in the sCT. This could potentially be improved through the use of a deformable registration between CT and MR, although this would not completely remove discrepancies. PET-CT also suffers from this problem to some extent, where although the PET and CT images are acquired in the same imaging session, they are separated in time. This can result in discrepancies due to gross patient motion or changes in internal anatomy (eg from bladder filling, see chapter 8).

The improvement in TLG differences using sCTAC compared to MRAC were smaller than for the SUV differences GTVs (figure 6.7). TLG is the product of the thresholded volume with $\text{SUV}_{\text{mean}}$ , and the MRAC over-estimated the volume and underestimated the $\text{SUV}_{\text{mean}}$. Therefore in the TLG calculation these errors partially cancelled out, improving the agreement with CTAC, although the sCTAC was still closer.

These results demonstrate that the sCTAC reconstructed PET images produce SUV differences that are clinically equivalent to CTAC for primary GTVs within equivalence margins of $\pm 4 - 5\%$ at the 90% confidence level (using a multiple testing corrected p-value of $p = 0.05/(8-1) = 0.007$ [250]). Equivalence testing is a well-established statistical methodology in clinical trials which aim to show that the test method has similar clinical effects to the control method, whilst having other benefits (such as lower cost, more efficient delivery etc.) [248]. However this method has not been used previously in evaluating PET-MR attenuation correction accuracy. Conventional superiority testing is sometimes used to claim equivalence if differences between the methods are not statistically significant (ie $p > 0.05$). However this does not demonstrate equivalence, but merely that the null hypothesis that the two methods are the same cannot be ruled out at the specified level of confidence. In contrast, equivalence testing demonstrates that the two methods are the same within a clinically defined equivalence margin at the specified confidence level. This is a much more robust way to determine equivalence [252].

Critical to the validity of equivalence testing is the equivalence margin used [248]. Unfortunately there is not a standard margin applied in the literature on PET-MR attenuation correction. The methodology used to derive the margins of $\pm 4 - 5\%$ rests on the assumptions that the attenuation correction method is independent of the other uncertainties in the PET imaging and that increases in the overall SUV uncertainty from 20%/30% to 20.4%/30.4% can be considered negligible. The overall PET-CT uncertainty was estimated as the test-retest repeatability values of SUV measurements reported in the literature. Different values have been given, with Velasquez et al. finding values as low as $10 - 12\%$ [240]. However this dataset underwent quality assurance and the repeatability analysis only included patients with sufficient scan quality and protocol compliance, which required rejecting one third of the study scans, suggesting SUV repeatability in clinical practice is likely to be higher [179]. A meta-analysis of five studies investigating the repeatability of PET-CT scans gave substantially higher values of 30% ($\text{SUV}_{\text{max}}$) and 20% ($\text{SUV}_{\text{mean}}$) [251]. This is likely to more accurately reflect clinical practice and so were used to determine the equivalence margin. A similar methodology could be applied to determine equivalence margins for TLG measurements, however evaluations of the overall repeatability of TLG measurement in ano-rectal cancer patients were not available in the literature.

The sCTAC results in this study compare well with previous published results on PET-MR attenuation correction in the pelvis. Shandiz et al. investigated using a short echo time sequence and automated image segmentation techniques to generate a bulk density sCT with five tissue classes (cortical bone, air cavity, fat, soft tissue and background) [253]. PET errors were estimated using simulated PET data for one healthy patient, with mean voxel-by-voxel SUV errors of $-14 \pm 15\%$, $4 \pm 6\%$, $8 \pm 13\%$ and $4 \pm 2\%$ in the bone, soft-tissue, fat and prostate regions. Bradshaw et al. used a Deep Learning model to generate a four tissue class sCT from T1 and T2 Dixon MR images from 12 patients [254]. This was evaluated on 16 FDG-avid lesions from five patients, with median $\text{SUV}_{\text{max}}$ differences of $-1\%$ (range $-4\%$,1% estimated from boxplot) and $\text{SUV}_{\text{mean}}$ differences $-1\%$ ($-4\%$,1%). The same ZTE sequence investigated here was used in combination with a Dixon MR images in a Deep Learning model with 10 training patients to generate sCTs [77]. Median $\text{SUV}_{\text{max}}$ differences for 30 bone lesions from 16 patients were $-1\%$ (range $-8\%$,3%, estimated from boxplot) and $-2\%$ ($-12\%$,5%) for 60 soft-tissue lesions.

The results in this study also demonstrated comparable or superior performance to other MR-only radiotherapy sCT algorithms which had been applied to PET attenuation correction. Wallstén et al. gave whole-image SUV differences of $-0.5\%$ and within-bone differences of $-4.2\%$ [31], which were larger than the $-0.02\%$ and $-0.5\%$ reported here. This translated into $\text{SUV}_{\text{mean}}$ differences in PET-avid lesions within the prostate of $-2.3\%$, again a larger difference than the $1.0 \pm 0.7\%$ reported in this study. Ahangari et al. found mean differences in $\text{SUV}_{\text{max}}$ of $-0.8 \pm 1.2\%$ (± standard error) and in $\text{SUV}_{\text{mean}}$ of $-0.3 \pm 1.8\%$ [30]. These were similar to the results reported here (absolute differences $\leq 1.0\%$).

Our MRAC results also show good agreement with the literature. Wallstén et al. found mean SUV differences within soft-tissue of $-3.6\%$ and within the bone region of $-17.7\%$ [31], very similar to the $-3.0\%$ and $-16.3\%$ found here. The $\text{SUV}_{\text{mean}}$ difference in the prostate PET-avid lesion was $-5.9\%$, similar to the $-4.3 \pm 0.8\%$ reported here. Leynes et al. found median $\text{SUV}_{\text{max}}$ differences in soft-tissue lesions of $-6\%$ ($-18\%$,4%) [77], which agrees within two standard errors with the $-4.6 \pm 0.9\%$ ($-8.4\%$,$-1.3\%$) reported in this study.

There are two aspects to consider when applying sCT algorithms to PET attenuation correction. On the one hand, PET is more sensitive to HU errors due to the lower energy of PET photons compared to those produced by megavoltage linear accelerators, so this makes PET attenuation correction more challenging for sCT algorithms. On the other hand, the overall uncertainty of SUV measurements is much higher than in radiotherapy dosimetry. The overall repeatability of SUV measurements is $20 - 30\%$ [251], whereas the overall uncertainty in radiotherapy dose delivered to the patient is $3 - 5\%$ [8]. Thus clinical equivalence in SUV accuracy could be achieved with differences of $4-5\%$ whereas the dose uncertainty of any individual component of the radiotherapy pathway has to be $\leq 1\%$ to not increase the overall dose uncertainty [8]. This suggests that sCT requirements for MR-only radiotherapy are more stringent than for PET attenuation correction, and so sCTs that are clinically acceptable for radiotherapy are likely to be able to be used for PET attenuation correction without modification. This agrees with the data found in this study and the two other studies applying radiotherapy developed sCTs to PET attenuation correction [30, 31].

An important consideration is how PET images acquired in the radiotherapy position on a PET-MR scanner utilising sCTAC and $AC_{\mathrm{cba}}$ of the radiotherapy hardware compared to PET images acquired on PET-CT. This chapter and chapters 4 & 5 suggest overall PET attenuation correction uncertainties of $\sim 4$ % compared to CT. However, these methods are likely to be less robust than PET-CT since they depend on models of the patient and the radiotherapy hardware whereas PET-CT provides a directly imaged attenuation correction map. For example, a patient with unusual anatomy or damage to the radiotherapy coil bridge could result in errors in the attenuation correction map. In contrast, a CT image provides attenuation correction information from the patient and radiotherapy hardware directly as they were at the time of CT imaging. This may have a discrepancy with the time of PET imaging, potentially resulting in incorrect attenuation due to organ motion from bladder filling (see chapter 8). But nonetheless, PET-CT is likely to provide more robust attenuation correction than PET-MR. This needs to set against the geometric improvements in image alignment between PET and MR images acquired simultaneously, which facilitates sub-volume analysis incorporating information from multiple images. PET-CT registered to MR on the other hand suffers from the MR-CT (and PET-CT) registration uncertainty, making PET-MR more geometrically robust. Future work could evaluate this further by comparing PET-MR and PET-CT images acquired in the same patient cohort in the radiotherapy position.

A limitation of this study was that only small numbers of patients were evaluated, especially for the nodal evaluation. This is likely to have prevented clinical equivalence in nodal SUV measurements being demonstrated due to the study being under-powered. In addition, measurements have only been made on one manufacturer's scanner in one centre, which was the same scanner on which the ZTE images used to the sCT model were acquired on. Evaluating the sCT algorithm on more patients acquired in different centres and on different scanners would enable the generalisability of this method to be tested.

## 6.5 Conclusions

A ZTE-based Deep Learning sCT algorithm for MR-only radiotherapy has been successfully applied for PET-MR attenuation correction. There were substantial reductions in SUV differences to gold standard CTAC, with mean whole-image differences being $-0.02\%$, compared to $-3.0\%$ for the current MRAC. The improvements in the bone regions were particularly large, $-0.5\%$ rather than $-16.3\%$. This had no impact on the accuracy of thresholded GTV delineation. However it did have a significant impact on metabolic parameters, with SUV differences in $\text{SUV}_{\text{max}}$ and $\text{SUV}_{\text{mean}}$ being $1.0 \pm 0.8\%$ and $1.0 \pm 0.7\%$ respectively rather than $-4.6 \pm 0.9\%$ and $-4.3 \pm 0.8\%$ for the MRAC. The SUV measurements in the primary GTVs were clinically equivalent to CTAC within $\pm 5\%$ and $\pm 4\%$ respectively ($p < 0.001$ for both), whereas MRAC measurements were not ($p = 0.33$ and $p = 0.21$). This suggests that PET images reconstructed using sCTAC substantially improve in SUV accuracy compared to current MRAC approaches and would enable highly accurate quantitative PET images to be acquired on a PET-MR scanner.

# Chapter 7

# Developing QA tests for simultaneous PET-MR imaging for radiotherapy planning

## 7.1 Introduction

Simultaneous PET-MR scanners enable acquiring both MR and PET functional information with high spatial alignment [255]. This has the potential to improve tumour delineation and to enable the metabolically active tumour sub-volumes to be identified with high accuracy through utilising the complementary information from both modalities [29]. Accurate delineation of these sub-volumes can enable dose painting strategies [15], which could potentially improve tumour control probability without changing treatment side-effects [12, 101]. A recent review concluded that using complementary information from Prostate Specific Membrane Antigen (PSMA)-PET, T2-weighted-MR and DW-MR for prostate boost volume delineation significantly improves the potential tumour control probability [152]. In addition, quantitative PET-MR metrics, such as PET SUV and DW-MR ADC, have shown potential as imaging biomakers for treatment prognosis and response monitoring [256]. Several studies have demonstrated the clinical feasibility of using PET-MR in the radiotherapy position for different treatment sites [30, 161, 257].

To use PET-MR imaging for radiotherapy planning, a comprehensive PET-MR QA programme needs to be developed. There are currently consensus guidelines on PET-MR QA for the diagnostic setting [164]. However, images used for radiotherapy planning must meet additional requirements compared to diagnostic imaging [154]. These include high geometric accuracy over the entire field of view [166], sufficient image quality for accurate delineation of tumour and organ at risk boundaries [119], a high degree of spatial alignment between images acquired in the same session [181] and high mechanical accuracy in couch and laser movements to ensure reproducible patient positioning [182].

In addition, if quantitative functional information is being used for automatic image segmentation or treatment response monitoring, then the accuracy and stability of these quantitative metrics needs to be assured [171, 178]. DW-MR is the most investigated functional MR technique for radiotherapy planning [105] and so a test of ADC accuracy is likely to be necessary. If other functional MR techniques are being used for radiotherapy planning (eg DCE-MR), then additional tests would also be required. This implies that a radiotherapy dedicated PET-MR QA programme needs to be developed. This would need to include the current recommendations for radiotherapy MR imaging: MR image quality, MR geometric accuracy and electromechanical accuracy tests, as well as PET-MR specific tests covering PET-MR alignment accuracy, DW-MR ADC accuracy and PET SUV accuracy.

Several studies have evaluated radiotherapy adapted PET-MR systems using QA phantoms. Paulus et al. investigated PET and MR image quality in a head and neck radiotherapy setup using uniform PET and MR phantoms [155]. Similarly, as described in chapter 4, we evaluated image quality in a pelvic radiotherapy setup using image quality phantoms. Methods to correct the PET attenuation from the radiotherapy setup [258] have also been evaluated. However, to the best of my knowledge, a comprehensive set of tests for a routine QA programme have not been evaluated in the literature previously. Therefore the aim of this study was to develop the tests needed for such a programme and to assess the repeatability of these tests and their stability over a 12 month period.

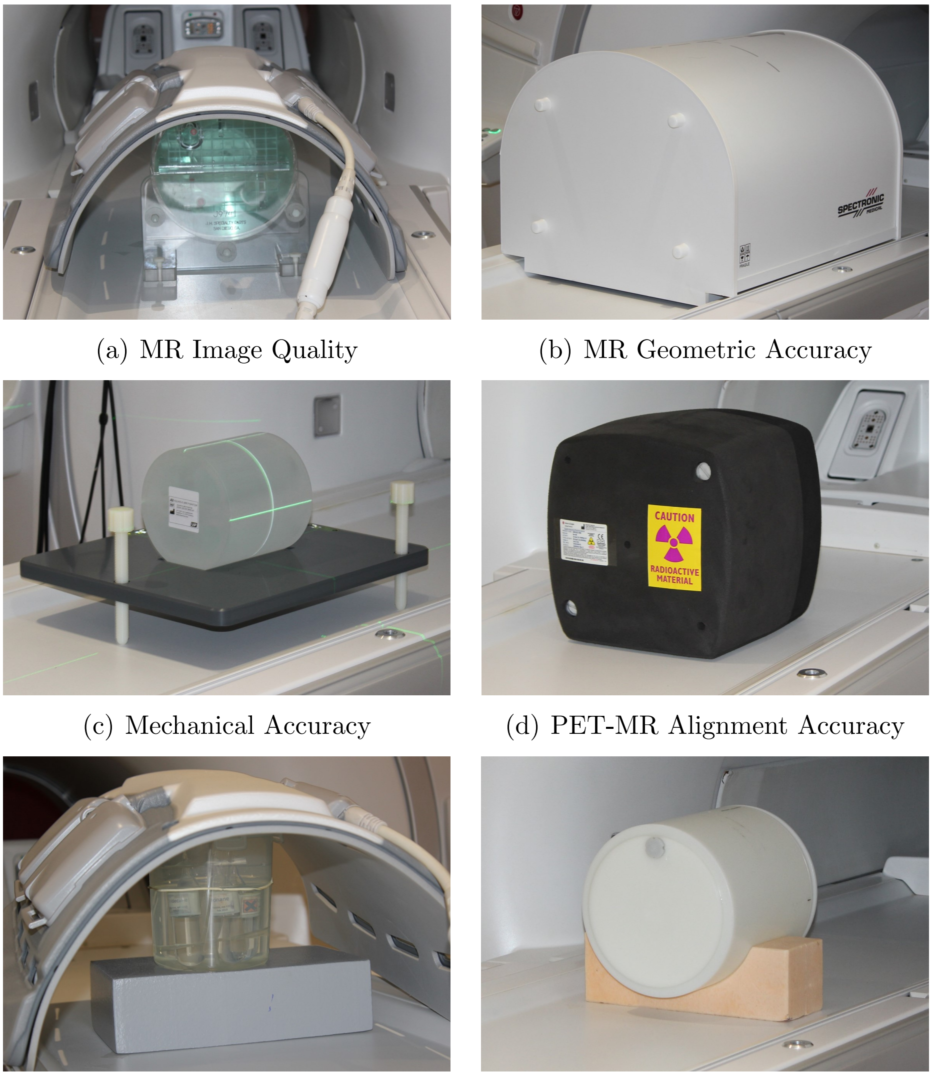
## 7.2 Materials and Methods

PET-MR radiotherapy QA tests were developed for MR image quality, MR geometric accuracy, electromechanical accuracy, PET-MR alignment accuracy, PET SUV accuracy and DW-MR ADC accuracy (figure 7.1). Repeatability was determined using three independent same-day measurements and stability through monthly measurements from October 2020 to September 2021, all acquired on a Signa 3T PET-MR scanner (version MP26, GE Healthcare, Milwaukee, USA) by the same observer. Measurements took approximately two hours with a further 15 minutes for image analysis. Repeatability and stability were assessed by the Standard Deviation (SD).

### 7.2.1 MR Image Quality

MR image quality was assessed with the ACR large image quality phantom [165]. Images were acquired using the in-built spine coil and anterior array coil used for pelvic imaging. Images were not acquired with the radiotherapy couch and coil bridge.

The recommended T1-weighted (ACR T1) and double-echo T2-weighted (ACR T2) axial spin echo sequences were acquired. Only the second echo images in the ACR T2 series were used. Images were analysed according to the ACR recommendations using in-house



(a) MR Image Quality

(b) MR Geometric Accuracy

(c) Mechanical Accuracy

(d) PET-MR Alignment Accuracy

(e) DW-MR ADC Accuracy

(f) PET SUV Accuracy

Figure 7.1: Photographs of the phantoms and setup used for each of the six QA tests evaluated. Each test was carried out three times on the same day using independent phantom setups (repeatability measurements) and once a month for 12 months (stability measurements). Images were acquired using the spine coil and anterior array coil for MR image quality and DW-MR ADC accuracy tests, and the in-built body coil/PET detector for the other tests.

developed Matlab software based upon open source software [229] and substantially modified to make it more accurate and robust. High-contrast resolution was assessed as the smallest diameter line of holes detectable in a horizontal or vertical array. Slice thickness accuracy was determined by imaged profile of two angled ramps. Slice position accuracy used crossed $45^o$ wedges at either end of the phantom. The image uniformity test was the ratio of near-minimum and near-maximum pixel values in the uniform phantom compartment. The ghosting ratio was the ratio of pixel values outside the phantom to those within the uniform compartment. The low-contrast object detection was the total number of visible 'spokes' of disks with decreasing contrast (5.1%-1.4%) and diameter (7.0-1.5 mm). This test was performed manually by a single observer using RadiAnt DICOM Viewer (version 4.6.9.18463, Medixant, Posnan, Poland). The geometric accuracy ACR test was omitted since it only covered a small FOV which was insufficient for radiotherapy purposes. Instead a dedicated large FOV geometric accuracy test was used (see next section).

### 7.2.2 MR Geometric Accuracy

MR Geometric accuracy was assessed using the large field of view GRADE phantom (Spectronic Medical, Helsingborg, Sweden). This consisted of $\sim 1,200$ spherical markers embedded in expanded foam in a grid pattern. Images were acquired with recommended 2D and 3D sequences and automatically analysed using the vendor provided software (version 1.0.46) which calculated the distortion shift of each marker as the absolute Euclidian distance in 3D from the marker position in the image to the known reference position [167].

The markers were grouped in concentric circles at increasing distances from the isocentre and the mean distortion in each group calculated. Repeatability was assessed by then calculating the mean and SD of those group means over the three repeats. In addition, each marker was uniquely identified and so the mean and SD of distortion for each marker over the three repeats was calculated ($\overline{D}_n \pm \sigma_n$ for marker $n$). A few markers on the periphery of the phantom were not identified in all three repeatability measurements and were excluded from the analysis (25/785 and 16/852 markers for the 2D and 3D sequences respectively). The mean SD of all markers over the three repeats was calculated using [167]

$$\overline{\sigma} = \frac{1}{N}\sum_{n=1}^{N}\sigma_n, \tag{7.1}$$

where $\sigma_n$ was the standard deviation of the $n^{th}$ marker and $N$ was the number of markers common to all images over the three repeats. The range in distortion for each marker over the three repeats was also calculated and the mean range over all markers determined. Monthly stability was assessed using the same methods. Similarly, markers not common to all 12 monthly measurements were excluded in the SD and range of distortion calculations (26 for the 2D and 23 for the 3D).

### 7.2.3 Mechanical Accuracy

The electromechnical performance of the scanner and external lasers was assessed using the Aquarius MRI phantom (LAP GmbH, Schwarzenbruck, Germany) and a plastic ruler. The Aquarius phantom was used to measure the alignment of the internal and external lasers with the scanner imaging plane. The phantom contained 10 cm long cross-planes in the transverse, sagittal and coronal axes surrounded by a copper sulphate solution (see figure 7.7). The phantom was aligned to the external lasers in all three planes, then aligned in the superior-inferior direction on the internal laser and shifted to isocentre. A 3D fast spin echo sequence centred on the scanner isocentre was acquired. The centre of the cross-planes and points ±5 cm in each axis were manually marked on the image in RayStation (v9b, RaySearch Laboratories, Stockholm, Sweden) and the transverse and rotational offsets in each plane from the image centre calculated using an in-house script.

The coincidence of the two lateral external lasers was assessed across the scanner couch using hand shielding. This used a piece of translucent paper held in the path of the lateral lasers such that both laser beams could be seen. The coincidence of the beams was assessed by shielding one of the beams with a hand to isolate one of the beams, and then removing that hand to visualise the two beams simultaneously. External laser movements were assessed by moving the relevant laser by a set amount and measuring the distance traversed with a ruler. The coincidence of the sagittal internal and external lasers was measured using a ruler. The electromechanical accuracy of the couch movements was assessed using a ruler fixed to the scanner couch aligned with the external sagittal laser. For both external laser and couch motions the final shift was back to the zero position to assess hysteresis.

### 7.2.4 PET-MR Alignment Accuracy

PET-MR alignment was assessed using the VQC phantom, provided by GE Healthcare for PET-MR alignment calibration. This consisted of five solid low activity ($\sim$ 0.7 MBq) $^{68}$Ge spheres in a geometric grid pattern, with each sphere having one MR-visible sphere superior and inferior of it. The phantom is designed so that the halfway point between each MR sphere pair should exactly coincide with the PET sphere. MR and PET images were acquired and analysed within RaySation using an in-house script which automatically identified all the PET and MR spheres. For each pair of MR spheres the point halfway between the two sphere centroids was calculated. A six-degrees-of-freedom rigid registration was performed to align these calculated MR points with the measured PET sphere centroids, giving a measure of PET-MR alignment.

### 7.2.5 DW-MR Apparent Diffusion Coefficient Accuracy

DW-MR ADC accuracy was assessed using an in-house phantom, consisting of three sealed vials containing 100 ml of n-nonane, n-undecane and n-tridecane respectively, surrounded by $\sim$ 800 ml of water. Images were acquired using the anterior array coil and spine coil. The phantom temperature was allowed to equilibrate with room temperature and measured before and after image acquisition. A single-shot Echo Planar Imaging DW-MR sequence was acquired with a single coronal slice through the vials using b-values 50 $\mathrm{s\,mm^{-2}}$ and 800 $\mathrm{s\,mm^{-2}}$, with two averages for $b = 800$ $\mathrm{s\,mm^{-2}}$ to improve SNR. Initial measurements with 8 b-values (50 $\mathrm{s\,mm^{-2}}$-1500 $\mathrm{s\,mm^{-2}}$) showed that the natural logarithm of signal intensity was highly linear with b-value, as expected for a free diffusion phantom [176]. Therefore only two b-values were acquired for speed.

The images were analysed in Medical Interactive Creative Environment (MICE) Toolkit (version 1.0.8, Umeå University, Sweden) [204]. Each vial was automatically segmented on the $b = 50$ s mm$^{-2}$ image and mean ADC calculated. Reference ADC values were determined from literature values as a function of temperature, and linearly interpolated from the measured phantom temperature. At 20°C these reference values were $1626 \times 10^{-6}$ mm$^2$ s$^{-1}$, $1009 \times 10^{-6}$ mm$^2$ s$^{-1}$ and $640 \times 10^{-6}$ mm$^2$ s$^{-1}$ for n-nonane, n-undecane and n-tridecane respectively [259]. The percentage difference in ADC between measured and reference was calculated.

### 7.2.6 PET Standard Uptake Value Accuracy

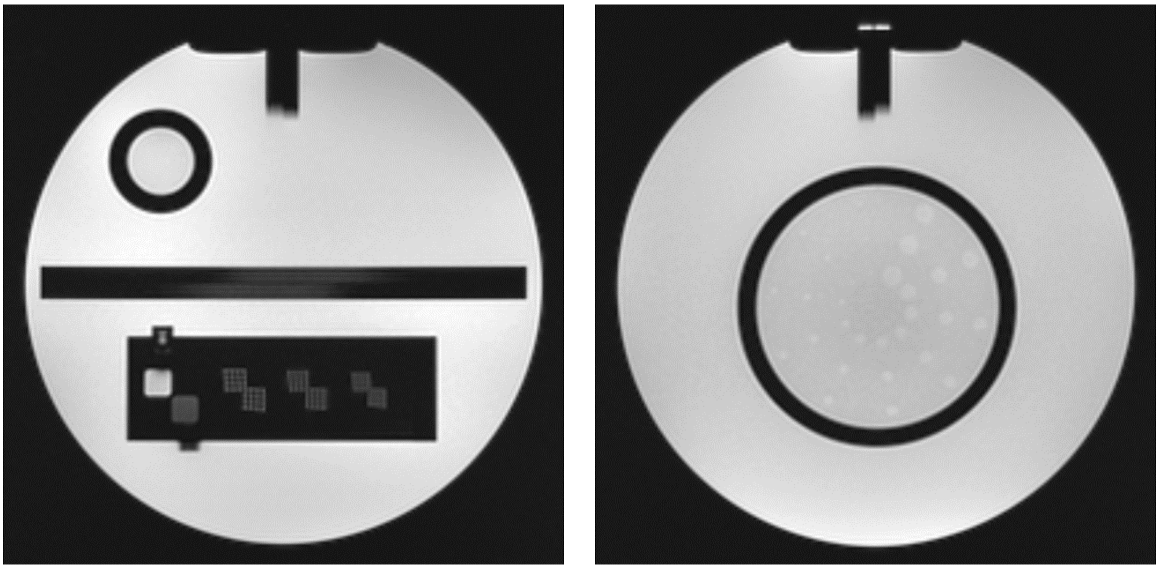
A uniform PET activity phantom was used to assess PET SUV accuracy containing $\sim$ 30 MBq of $^{18}$F$-$Fluorodeoxyglucose (FDG). A 10 minute single bed-position PET scan was acquired. The radiotherapy couch top and coil bridge and anterior array coil were not used. The PET image was reconstructed using an Ordered Subset Expectation Maximum reconstruction with 16 subsets and 4 iterations and a 5.0 mm Gaussian filter with resolution recovery and time of flight information.

Repeatability measurements were acquired by adding additional activity between scan 1 & 2 and scan 2 & 3 so that each acquisition had a unique activity fill. Three repeatability scans of the phantom were acquired on the same day, with the activity modified between scans 1 & 2 and 2 & 3. The initial activity in the phantom was 31.9 MBq at 13:45, measured with dose calibrator (CRC-15 PET, Capintec, New Jersey, USA). After scan 1, a small amount of FDG was then removed from the phantom into a syringe, an additional 8.9 MBq of FDG injected to the phantom and then as much as possible of the removed FDG returned. The residual activity of both syringes was measured and subtracted from the decay-corrected initial activity to determine the total activity in the phantom for the second repeat measurement, which was 34.5 MBq at 14:20. The process was repeated for scan 3 with 10.7 MBq of FDG added, giving a total activity at 14:57 of 38.0 MBq.

A CT image of the phantom was acquired for phantom attenuation correction. This is the only way to do attenuation correction in phantoms on the PET-MR. A non-attenuation corrected PET image of the phantom was acquired and rigidly registered to the phantom CT. The CT was resampled onto the PET image matrix in MICE Toolkit. The resampled CT and non-attenuation corrected PET were uploaded onto the scanner PET phantom library. On acquiring a PET scan of the phantom, the scanner automatically rigidly registered the acquired PET image to the phantom library PET image. This registration matrix was then used to align the phantom library CT to the newly acquired image, the CT was converted to 511 keV attenuation coefficients and then combined with an attenuation coefficient map of the spine coil and PET-MR couch. The combined attenuation correction map was used for the final PET image reconstruction (see figure 7.11). Images were automatically analysed using an in-house script within RayStation which placed a 18 cm diameter and 18 cm long cylindrical ROI at the phantom centre. The mean SUV within the ROI was calculated and the percentage difference to the reference SUV determined.

## 7.3 Results

Example images of the MR image quality test are shown in figure 7.2. All the MR image quality repeatability and stability measurements were within the recommended tolerance values except for the low-contrast detectability scores (table 7.1) and there were no monthly trends (figure 7.3, the spatial resolution results are not shown since every month had an identical result). The monthly stability means and SDs were very similar to the repeatability means and SDs for all tests except slice position, ghosting and image uniformity (T2).



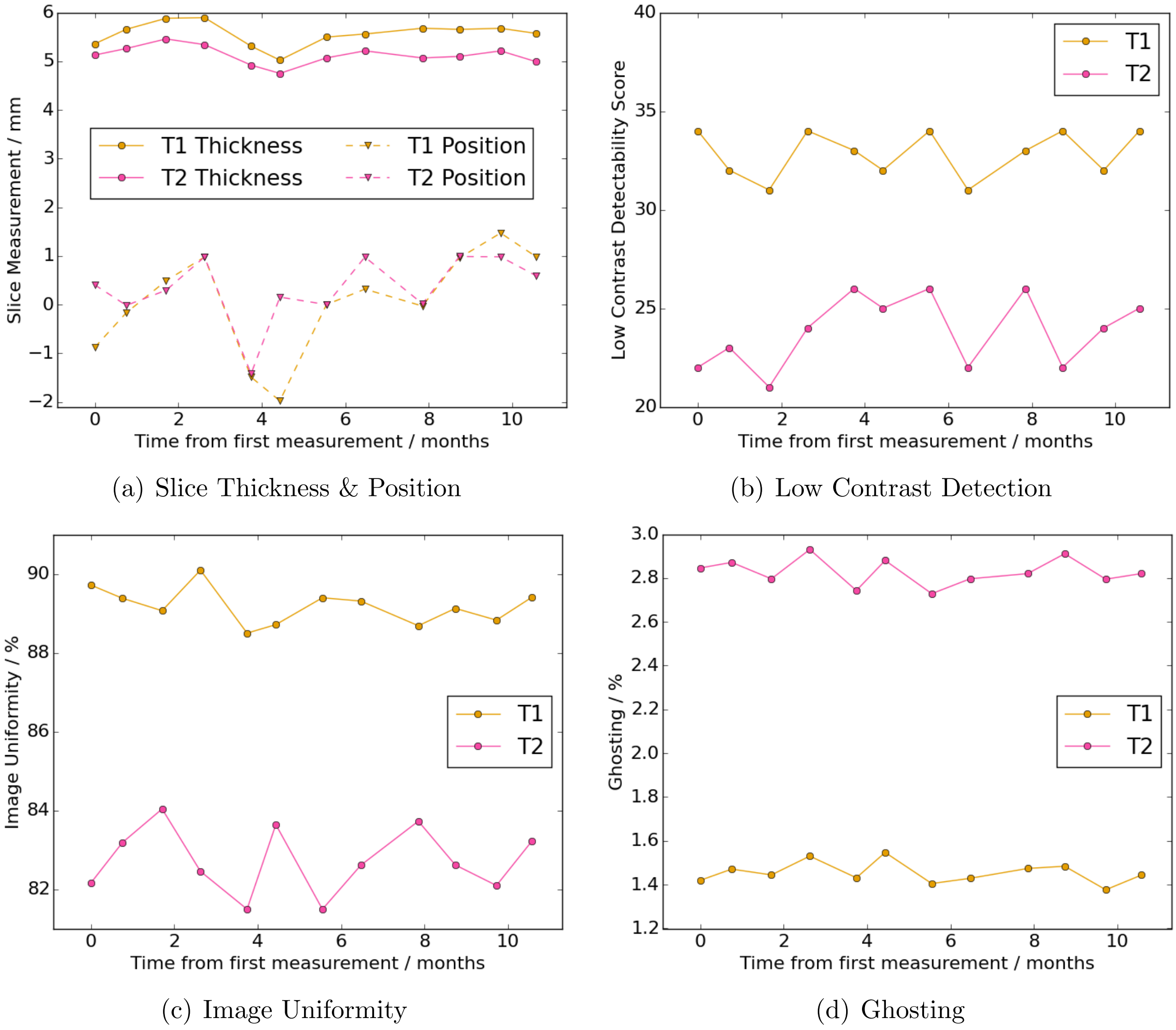
(a) Slice 1

(b) Slice 11

Figure 7.2: Example axial images of the American College of Radiologists phantom for the MR image quality test. Slices 1 (a) and 11 (b) are shown.

Example images for the MR geometric accuracy test are shown in figure 7.4. The distributions of distortions at different distances from the isocentre were similar between repeats and between months (figure 7.5). The mean monthly SDs of distortion were within 0.1 mm of the repeatability values (table 7.1).

The mechanical accuracy stability means agreed within one stability SD with the repeatability means (except for the external laser anterior-posterior offset, which agreed within 0.4 mm) and the stability SDs were within 0.2 mm or $0.1^o$ of the repeatability SDs (table 7.1). Example MR images of the Aquarius phantom for external laser offset measurements are shown in figure 7.7. There were no trends in the monthly measurements for the mechanical accuracy tests (figure 7.6).



(a) Slice Thickness & Position

(b) Low Contrast Detection

(c) Image Uniformity

(d) Ghosting

Figure 7.3: Monthly stability plots of the MR image quality test. The plots show slice thickness and position (a), low contrast detection (b), image uniformity (c) and image ghosting (d) measurements.

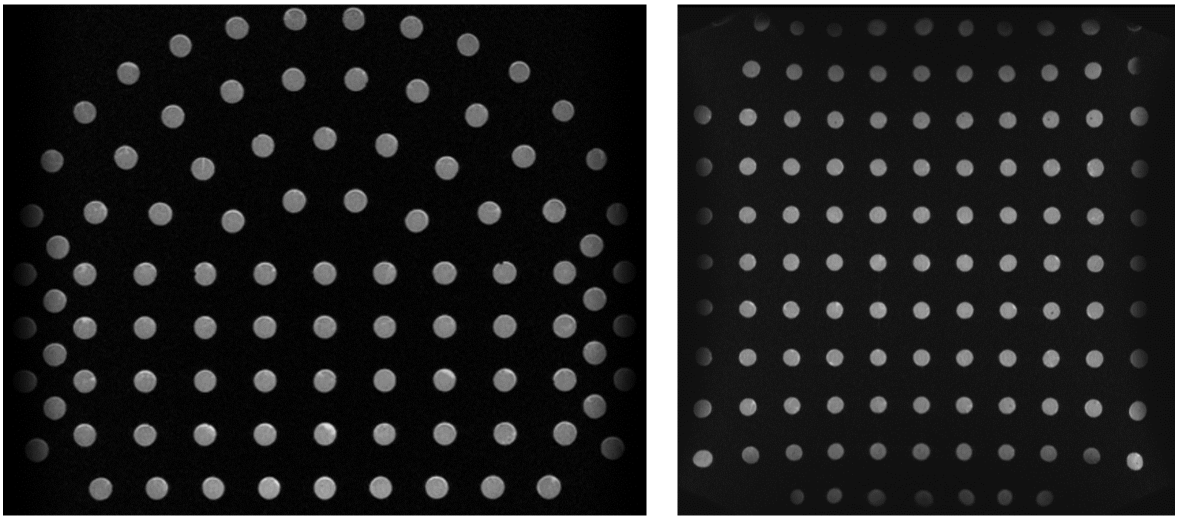
Similarly, the PET-MR alignment tests had very similar repeatability and monthly SDs, agreeing within 0.1 mm. Differences from zero were also small ($< 0.2$ mm and $< 0.2^o$). There were no monthly trends apparent (figure 7.6). Example MR and PET images for the PET-MR alignment test are shown in figure 7.8.

The DW-MR ADC stability measurements showed no monthly trends or trend with phantom temperature (figure 7.9). There were systematic discrepancies between the different alkanes, with the n-nonane vial always having a positive difference to the reference value and the n-undecane and n-tridecane vials being around zero difference. Example DW-MR images of the ADC accuracy phantom are shown in figure 7.10.

Example images of the PET phantom and corresponding AC map for the PET SUV accuracy tests are shown in figure 7.11. There was no monthly trend in the SUV accuracy measurements (figure 7.9), although all measured SUVs had a positive difference, suggesting a small systematic SUV over-estimation.

Table 7.1: The repeatability and stability results for all tests. T1/T2 refers to the ACR T1/T2 image series respectively. 2D/3D refers to the 2D and 3D geometric accuracy sequences respectively. $d$ indicates distance from the scanner isocentre. EL refers to the external lasers mounted on the laser bridge and IL to the scanner internal lasers.The reference column indicates the tolerance values from the ACR manual for the MR image quality test, the recommended tolerance for geometric distortion in MR for radiotherapy and the reference values for the other tests.

| **Test** | **Component** | **Reference** | **Repeatability** | | **Stability** | |
|---|---|---|---|---|---|---|
| | | | *Mean* | *SD* | *Mean* | *SD* |
| 1) MR Image Quality | Spatial Resolution (T1) [mm] | $\leq$ *1.0* | 1.0 | 0.0 | 1.0 | 0.0 |
| | Spatial Resolution (T2) [mm] | $\leq$ *1.0* | 1.0 | 0.0 | 1.0 | 0.0 |
| | Slice Thickness (T1) [mm] | *5 ± 0.7* | 5.3 | 0.2 | 5.6 | 0.2 |
| | Slice Thickness (T2) [mm] | *5 ± 0.7* | 4.9 | 0.2 | 5.1 | 0.2 |
| | Slice Position (T1) [mm] | $\leq$ *5* | 1.3 | 0.6 | 0.1 | 1.1 |
| | Slice Position (T2) [mm] | $\leq$ *5* | 1.4 | 0.5 | 0.3 | 0.7 |
| | Image Uniformity (T1) [%] | $\geq$ *82* | 88.8 | 0.3 | 89.1 | 0.5 |
| | Image Uniformity (T2) [%] | $\geq$ *82* | 85.0 | 0.1 | 82.7 | 0.8 |
| | Ghosting (T1) [%] | $\leq$ *3* | 0.7 | 0.05 | 1.5 | 0.05 |
| | Ghosting (T2) [%] | $\leq$ *3* | 0.9 | 0.04 | 2.8 | 0.06 |
| | Low-Contrast Detection (T1) | $\geq$ *37* | 35 | 1 | 33 | 1 |
| | Low-Contrast Detection (T2) | $\geq$ *37* | 26 | 1 | 24 | 2 |
| 2) MR Geometric Accuracy | SD of Distortion (2D) [mm] | - | 0.4 | - | 0.3 | - |
| | SD of Distortion (3D)[mm] | - | 0.2 | - | 0.3 | - |
| | Range of Distortion (2D) [mm] | - | 0.7 | 1.5 | 0.9 | 0.3 |
| | Range of Distortion (3D) [mm] | - | 0.4 | 0.2 | 0.9 | 0.3 |
| | Distortion $d < 10$ cm (2D) [mm] | $\leq$ *2.0* | 0.27 | 0.07 | 0.27 | 0.05 |
| | Distortion $10 \leq d < 15$ cm (2D)[mm] | $\leq$ *2.0* | 0.44 | 0.02 | 0.40 | 0.05 |
| | Distortion $15 \leq d < 20$ cm (2D) [mm] | $\leq$ *2.0* | 0.82 | 0.04 | 0.74 | 0.05 |
| | Distortion $20 \leq d < 25$ cm (2D) [mm] | $\leq$ *2.0* | 2.2 | 0.2 | 1.90 | 0.04 |
| | Distortion $d \geq 25$ cm (2D) [mm] | - | 7.5 | 0.7 | 4.7 | 0.1 |
| | Distortion $d < 10$ cm (3D) [mm] | $\leq$ *2.0* | 0.23 | 0.03 | 0.27 | 0.04 |
| | Distortion $10 \leq d < 15$ cm (3D) [mm] | $\leq$ *2.0* | 0.33 | 0.02 | 0.34 | 0.05 |
| | Distortion $15 \leq d < 20$ cm (3D) [mm] | $\leq$ *2.0* | 0.60 | 0.02 | 0.62 | 0.05 |
| | Distortion $20 \leq d < 25$ cm (3D) [mm] | $\leq$ *2.0* | 1.63 | 0.05 | 1.73 | 0.05 |
| | Distortion $d \geq 25$ cm (3D) [mm] | - | 5.99 | 0.03 | 4.85 | 0.04 |
| 3) Mechanical Accuracy | EL Right-Left Offset [mm] | *0.0* | 0.1 | 0.3 | 0.3 | 0.4 |
| | EL Ant-Post Offset [mm] | *0.0* | 0.0 | 0.3 | 0.4 | 0.1 |
| | EL Pitch Angle [$^o$] | *0.0* | −0.1 | 0.1 | 0.0 | 0.2 |
| | EL Roll Angle [$^o$] | *0.0* | 0.2 | 0.2 | 0.0 | 0.3 |
| | EL Yaw Angle [$^o$] | *0.0* | −0.1 | 0.1 | −0.1 | 0.2 |
| | EL Lateral Coincidence [mm] | *0.0* | 0.0 | 0.0 | 0.0 | 0.0 |
| | EL Movements [mm] | *0.0* | 0.0 | 0.1 | 0.0 | 0.1 |
| | EL-IL Right-Left Difference [mm] | *0.0* | 1.7 | 0.3 | 1.8 | 0.2 |
| | IL Sup-Inf Offset [mm] | *0.0* | 2.2 | 0.6 | 1.6 | 0.6 |
| | Couch Movements [mm] | *0.0* | 0.1 | 0.5 | 0.2 | 0.5 |
| 4) PET-MR Alignment | Right-Left Difference [mm] | *0.0* | 0.15 | 0.03 | −0.2 | 0.1 |
| | Ant-Post Difference [mm] | *0.0* | 0.12 | 0.02 | 0.0 | 0.1 |
| | Sup-Inf Difference [mm] | *0.0* | 0.02 | 0.07 | 0.2 | 0.1 |
| | Pitch Angle [$^o$] | *0.0* | 0.13 | 0.07 | −0.02 | 0.05 |
| | Roll Angle [$^o$] | *0.0* | −0.01 | 0.00 | 0.00 | 0.03 |
| | Yaw Angle [$^o$] | *0.0* | 0.01 | 0.05 | −0.07 | 0.07 |
| 5) DW-MR ADC Accuracy | Nonane Difference [%] | *0* | 2 | 1 | 3 | 1 |
| | Undecane Difference [%] | *0* | −2 | 2 | 0 | 1 |
| | Tridecane Difference [%] | *0* | −2 | 3 | 0 | 2 |
| 6) PET SUV Accuracy | SUV Difference [%] | *0.0* | 1.3 | 0.5 | 2.1 | 1.9 |



(a) Axial

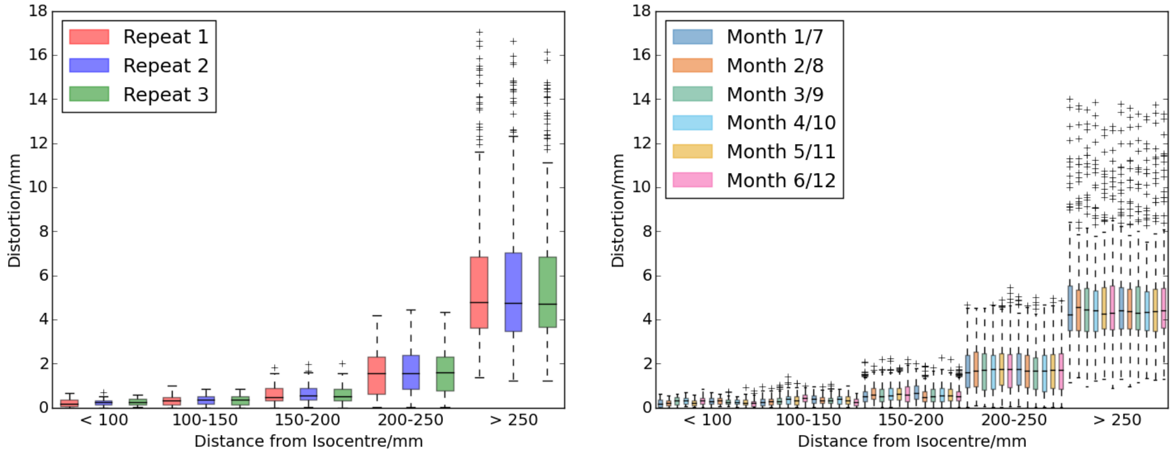
(b) Coronal

Figure 7.4: Example axial (a) and coronal (b) images of the GRADE phantom for the MR geometric accuracy test.

## 7.4 Discussion

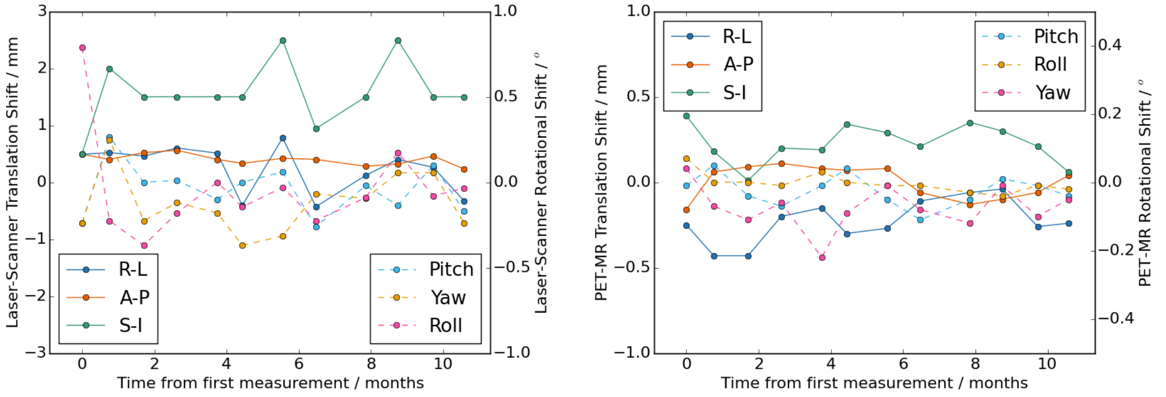
This study has developed phantoms and analysis software for PET-MR QA tests for radiotherapy, covering MR image quality, MR geometric distortion, mechanical accuracy, PET-MR alignment, DW-MR ADC accuracy and PET SUV accuracy. The tests appeared repeatable, with repeatability SDs $\leq$ 0.7 mm for all differences to reference in distance, $\leq 0.2^o$ for angle differences, $\leq$ 3% for percentage differences and 1 spoke for low-contrast detectability. The stability SDs were similar to repeatability, within 0.2 mm, $0.1^o$, 1 percentage point and 1 spoke, except for slice position and SUV accuracy. These were within 0.5 mm and 1.5 percentage points respectively. There were also no monthly trends, suggesting the tests were stable.

The MR image quality tests appeared repeatable with small repeatability SDs relative to the ACR tolerances. The monthly stability SDs were very similar, within 0.2 mm, 0.2 percentage points or 1 spoke, except for the T1 slice position and T2 image uniformity measurements. These were 0.5 mm and 0.7 percentage points larger respectively. This suggests there is more variation in scanner performance for these parameters, although the absence of any trends (see figure 7.3) implies that the performance is still stable. Adjeiwaah et al. investigated the repeatability of the ACR phantom using the same coil setup and scanner as used in this study [260]. The repeatability SDs reported agreed with those in this study within 0.1 mm (slice thickness), 0.9 mm (slice position), 0.5 percentage points (image uniformity) and 0.07 percentage points (ghosting). They also reported four year stability data, which showed the tests were stable over time in agreement with the results found here. All the MR image quality test results were within the ACR tolerance levels except for the low-contrast detectability score. This is likely due to the images not being acquired in a dedicated head and neck coil as recommended by ACR. Using



(a) Repeatability (b) Stability

Figure 7.5: Boxplots of the distributions of distortions at different distances from the isocentre for the 3D sequence for the repeatability (a) and monthly stability (b) measurements. The repeat and monthly measurements are displayed in the order acquired and shown as different colours. For the monthly measurements each colour represents two months, six months apart. The 2D sequence results showed very similar pattern.

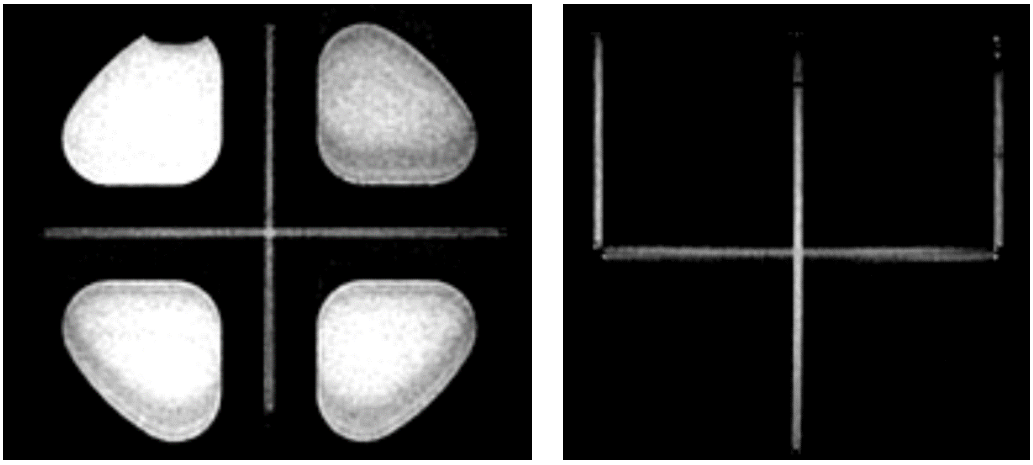


(a) Mechanical Accuracy (b) PET-MR Alignment

Figure 7.6: Monthly stability plot of mechanical accuracy (a) and PET-MR alignment (b) measurements. Both plots show translational (left axis, solid lines) and rotational (right axis, dashed lines) differences.

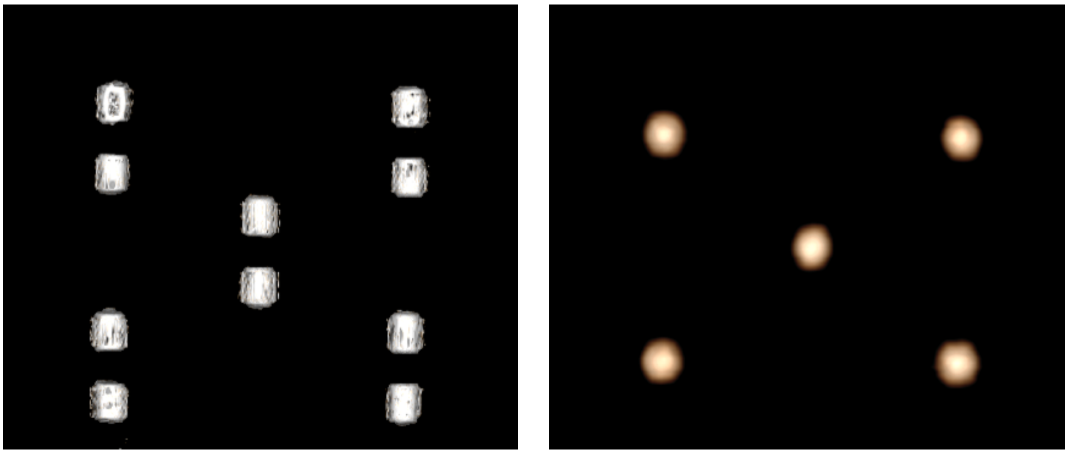
the in-built spine coil and anterior array coil ensured that the coils used for radiotherapy imaging were being tested.

The MR geometric accuracy tests appeared highly repeatable and stable, with both the repeatability and stability mean SDs of distortion being $\leq 0.4$ mm for 2D and 3D sequences. This is small compared to the 2 mm acceptable limit of distortion for MR-only radiotherapy [261]. The high repeatability of the test showed excellent agreement with a previous study using the same phantom on different MR scanners [167]. The stability results also showed good agreement with a study investigating MR geometric accuracy over 15 month period [168]. They reported SDs in mean distortion within concentric spheres at different distances from the scanner isocentre measured five times within that 15 month period that agreed within 0.04 mm of the SDs reported in this study. This suggests that the repeatability and stability of MR geometric accuracy test reported here



(a) Axial (b) Coronal

Figure 7.7: Example axial (a) and coronal (b) images of the Aquarius phantom for the mechanical accuracy test.



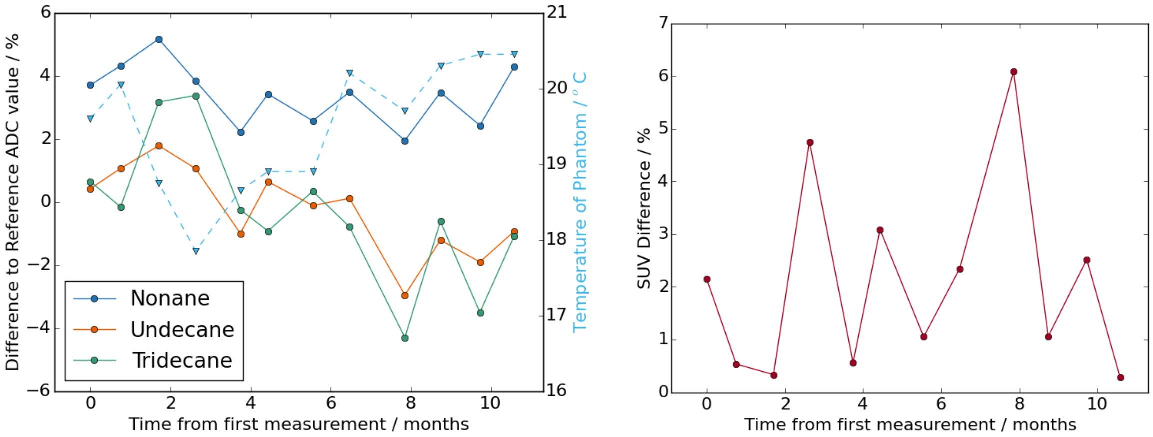
(a) MR (b) PET

Figure 7.8: Example MR (a) and PET (b) images of the VQC phantom for the PET-MR alignment test. Images are shown as 3D renderings as viewed from anterior to the phantom.

is equivalent to those reported in the literature as suitable for clinical use.

The mechanical accuracy tests were highly repeatable and stable over time, with all SDs being $\leq 0.6$ mm and all mean differences being within one SD of zero except two. These were small compared to the recommended tolerances of ±2 mm [182]. The larger differences were the external-internal laser difference and the internal laser offset to the scanner isocentre. Given the high coincidence between the external laser and scanner isocentre ($0.1 \pm 0.3$ mm) this indicates that the larger differences were due to the internal laser being misaligned by $\sim 2$ mm. This is within the scanner specification but confirms the requirement of external lasers for radiotherapy patient setup.

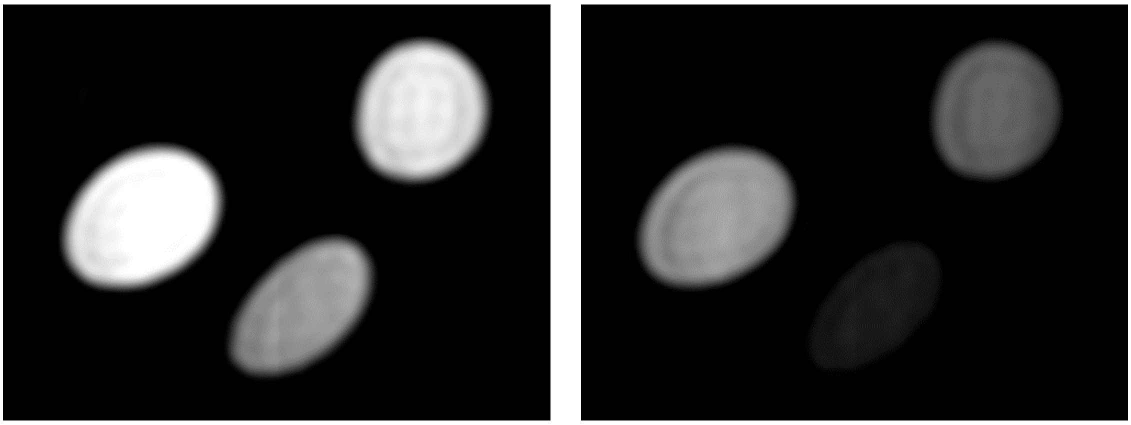
The PET-MR alignment accuracy test was highly repeatable, with repeatability SDs $< 0.1$ mm and $< 0.1^o$. The stability SDs were larger in most directions, but still very small (0.1 mm and $< 0.1^o$) indicating a very stable system. There is not a recognised clinical tolerance since PET-MR is an emerging modality, however both SDs were significantly



(a) ADC Accuracy

(b) SUV Accuracy

Figure 7.9: Monthly stability plot of percentage difference in mean ADC value to temperature-corrected literature reference value for each vial (a) and plot of percentage difference in mean SUV within phantom to reference activity (b). Plot (a) also shows measured temperature of phantom (blue dashed line).



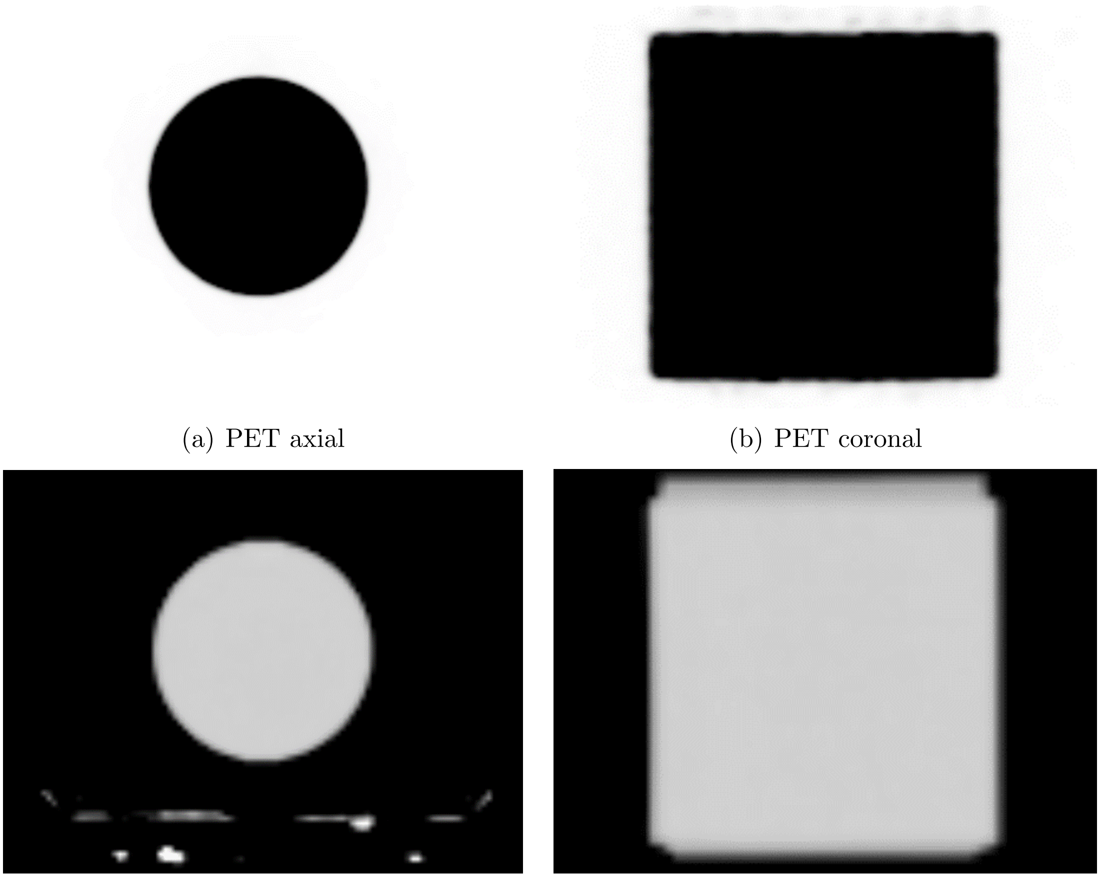
(a) $b = 50\ \mathrm{s\,mm^{-2}}$

(b) $b = 800\ \mathrm{s\,mm^{-2}}$

Figure 7.10: Example coronal images of the in-house DW-MR phanto for the ADC accuracy test. Images are shown for the $b = 50\ \mathrm{s\,mm^{-2}}$ (a) and $b = 800\ \mathrm{s\,mm^{-2}}$ (b).

smaller than the typical PET voxel dimensions of 2-3 mm. High PET-MR alignment was expected because the mechanical positions of the PET and MR imaging systems did not change during this study. However, given a major benefit of PET-MR scanners is the high intrinsic spatial alignment of the images, regular testing PET-MR alignment is important for providing assurance. To the best of the authors' knowledge this has not been published in the literature, although it is recommended to regularly assess PET-MR alignment for diagnostic PET-MR QA [164].

The DW-MR ADC accuracy test also demonstrated good repeatability, with SDs of differences $\leq 3\%$. This compares well with values of other ADC free-diffusion phantoms. Chenevert et al. reported an ADC phantom consisting of distilled water held at $T = 0$ °C in an ice-water bath was repeatable to within ±5% [173]. The stability measurements had similar SDs (agreeing within 1 percentage point with the repeatability SDs) and agreed with the repeatability measurements within one repeatability SD. This suggests



(a) PET axial

(b) PET coronal

(c) AC map axial

(d) AC map coronal

Figure 7.11: Example PET axial (a) and coronal (b) images of the uniform flood phantom for the PET SUV accuracy test. Also shown are the corresponding AC maps, (c) and (d), of the phantom. The axial AC map shows the PET couch and spine coil elements that were included in the AC map.

the scanner performance was stable over the period measured. The vial containing nonane appeared to have a small but systematic bias. This may be due to the location of the nonane vial being further from the image centre than the other two vials, which can influence ADC measurements [172]. The stability results also show good agreement with similar studies. Winfield et al. reported high stability of ADC measurements for a cylinder phantom containing five plastic tubes with sucrose solutions ranging from $0\% - 20\%$ sucrose in an ice-water bath at $T = 0$ °C [169]. Measurements were made on three different scanners at a frequency of once every two to three months for approximately 20 months, with coefficients of variation $\leq 4\%$ for each scanner. This suggests that the evaluated DW-MR ADC accuracy test was similarly repeatable and stable to other phantoms reported in the literature.

Finally, the PET SUV accuracy test showed substantially larger monthly variation (SD 1.9%) to the repeatability SD (0.5%). This suggests variation in calibrator and scanner performance over time had a larger impact on SUV accuracy then the inherent uncertainty of the test measurement itself. There did also appear to be a bias in the results, with all differences being positive, suggesting the scanner systematically over-estimated the SUVs in the phantom. This may be due to small errors in the scanner applied attenuation correction maps of the PET-MR couch and MR receive coil, which are not present when the PET system undergoes quarterly calibration. However, this effect was small with all differences to the reference value being within 6% and 10/12 measurements being with 3%. This is consistent with studies indicating drifts in calibrator performance over time of approximately 4% [262]. One study investigating longitudinal changes in PET SUV measurements for six different scanners reported a large variability in practice, with changes in SUV differences ranging from 0.3% to 58.6% [263]. Half of the scanners reported changes in SUV differences greater than the 6% range measured here. This highlights the importance of regular PET SUV accuracy measurements to ensure high quality, quantitative PET images.

A potential limitation of this study was that the MR image quality and PET SUV accuracy tests were not acquired with the radiotherapy flat couch and coil bridge. This was done because the radiotherapy hardware causes a significant drop in MR SNR of 45% and PET activity of 17.7% (see chapter 4). This would have reduced the sensitivity of the tests to detect changes in MR/PET performance, which was their primary aim. However, not doing so did mean that the full clinical setup was not evaluated on a monthly basis. This is unlikely to be relevant since the impact on MR SNR and PET activity loss would not change unless the radiotherapy hardware was damaged, which would be readily apparent from visual inspection. On the other hand changes in MR and PET performance, eg due to coil elements or detector arrays failing, can cause subtle image degredation and be difficult to detect from routine use of images [260]. Therefore it was considered that the higher sensitivity of the setup without using radiotherapy hardware was more relevant for routine QA tests.

A limitation of this study was that only one scanner from one manufacturer was evaluated. Future work will evaluate the generalisability of these QA tests to scanners from other manufacturers in other centres. This data, combined with the repeatability and stability data reported here and considerations of the clinical impact of variations from ideal performance, will then be used to generate QA tolerances and test frequencies to enable a comprehensive QA programme to be proposed.

## 7.5 Conclusions

Tests for a comprehensive PET-MR radiotherapy QA programme have been developed and assessed for repeatability and stability. All the tests appeared repeatable and stable over a 12-month period, although monthly variation was larger than test repeatability for PET SUV accuracy and two of the MR image quality tests. Future work will use this data to derive appropriate tolerance levels and test frequencies, which combined with these tests will form a QA programme. This will enable high-quality, robust PET-MR imaging to be used for radiotherapy planning.

# Chapter 8

# Tracking Organ Motion in the Pelvis for Radiotherapy Planning

## 8.1 Introduction

The use of anatomical Magnetic Resonance (MR) imaging for pelvic radiotherapy planning is rapidly increasing due to the increased contouring precision and accuracy from its superb soft-tissue contrast. MR is a very flexible imaging modality able to acquire multiple different anatomical image contrasts and functional information such as DW-MR and DCE-MR imaging. Utilising multiple image contrasts enables sequences to be optimised for different tasks, potentially improving contouring accuracy still further. Functional imaging also has great potential for pelvic radiotherapy planning for delineating tumour sub-volumes for radiotherapy dose painting [23] and for monitoring treatment response.

However this multi-parametric imaging approach takes time to acquire, typically $\geq$ 20 minutes of acquisition time. This is a significant issue for MR imaging for pelvic radiotherapy due to organ motion, primarily from bladder filling. The rate at which the bladder fills is quite variable with studies reporting mean filling rates between $5 \pm 3$ cm$^3$ min$^{-1}$ ($\pm$ standard deviation) [183] and $9 \pm 3$ cm$^3$ min$^{-1}$ [185]. The mean bladder volume at planning of prostate radiotherapy patients from a meta-analysis of bladder filling studies was 271 cm$^3$ [190]. Therefore these bladder filling rates correspond to changes of $2-3\%$ min$^{-1}$.

For radiotherapy treatment planning it is critical that the patient anatomy in the planning image(s) matches with the patient anatomy on treatment. Often pelvic radiotherapy pathways control bladder filling through the use of a standardised bladder preparation protocol, typically consisting of the patient emptying their bladder, drinking a set volume of water (typically $300 - 500$ cm$^3$ [190]) and then being imaged/treated a set time later (typically 30-60 minutes [190]). For CT-based radiotherapy both the acquisition of a planning image and the delivery of radiotherapy treatment only take a few minutes, during which time bladder volume changes are small. However the longer acquisition times of MR mean there will be larger changes during the imaging session, $\geq 40\%$ volume changes over MR acquisition times $\geq 20$ minutes, which means images acquired at the start/end of the protocol do not reproduce the patient anatomy that will be found at treatment.

For example, a MR-only prostate MR protocol may include a small FOV high-resolution T2-weighted sequence for prostate delineation, a large FOV T2-weighted sequence for OAR delineation and on-treatment image matching with CBCT [222], a DW-MR sequence and a DCE-MR sequence for delineating tumour sub-volumes for dose painting and/or monitoring treatment response [105], and a dedicated sequence for sCT generation if required. Assuming approximately 5 minutes per sequence, the total protocol duration is 25 minutes, during which time the bladder volume will have increased by $125 - 225$ cm$^3$. This is a very large change, $\pm 23 - 40\%$ if the mean bladder volume occurred halfway through imaging. Bladder motion substantially impacts other organs in the pelvis such as cervix, prostate, rectum and small bowel [185, 186] and so even for non-bladder cancer patients this large bladder filling motion renders images acquired at the start/end of the imaging protocol less representative of treatment and potentially unusable. To prevent this it is recommended to keep imaging acquisition times as short as possible and limit the number of sequences acquired [264]. However this means bladder filling is a significant factor limiting the use of MR to its full potential for radiotherapy planning in the pelvis.

Advanced MR image reconstruction techniques such as compressed sensing [265] and Deep Learning noise reduction [201] have the potential to significantly decrease MR acquisition times, reducing this problem substantially. However MR images in the radiotherapy setup have low SNR (see chapter 4) and so using these advanced image reconstruction techniques to improve SNR rather than reduce acquisition time would also have significant clinical benefits. Another possible method to remove this problem would be to compensate for bladder filling through using deformable registration of images acquired early or later in the session to the central image. However, deformable registration is challenging when the appearance of the two images are very different [266], such as T2-weighted MR and DWI images. Since the whole point of acquiring additional images is to generate different information, deformable registration by itself is unlikely to be successful.

A potential solution to this problem is to acquire a brief motion tracking MR sequence interleaved throughout the other MR sequences in the MR imaging protocol. If the protocol is timed so that the centre of the imaging session matches the bladder preparation time used for treatment (as recommended in ref. [264]), then this sequence can be used to correct MR images acquired earlier or later in the protocol to the centre time through deformable registrations. These registrations would be between images with very similar appearance and so would be much more likely to be accurate. This method could provide all the images within the protocol as if they were acquired simultaneously at the time matching the bladder preparation time, and so accurately representing the patient anatomy as it will be on treatment. The aim of this chapter was to develop such a method to track and correct for organ motion due to bladder filling and to evaluate it in healthy volunteers.

## 8.2 Materials and Methods

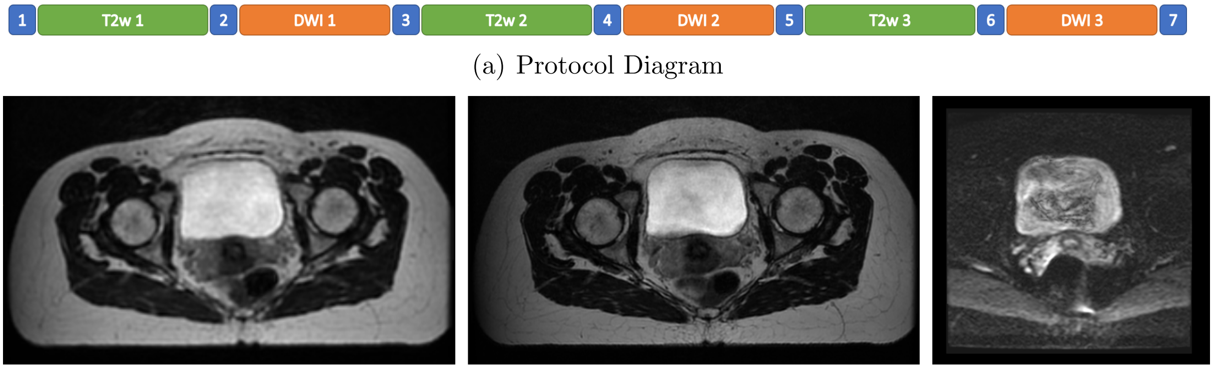
### 8.2.1 Participant Data Collection

Nine healthy volunteers (5 male, 4 female) with a median age of 35 years (range 26,50) were scanned on a SIGNA PET/MR 3T scanner (version MP26 GE Healthcare, Waukesha, USA). Ethical approval for the study was granted by Newcastle University's Research Ethics Committee (study number 1907/2223) and all participants gave informed consent. Participants were scanned in the radiotherapy treatment position on a flat couch-top with a coil bridge for the anterior MR coil as described in chapter 2. Immediately prior to entering the scan room participants emptied their bladder and drank 400 cm$^3$ of water. Each participant attended for three repeat visits separated by at least one day. The median gap between first and last visit was 15 days (range 7,15). The bladder preparation and imaging protocols were identical for each visit.

All participants underwent an imaging protocol consisting of three MR sequences referred to as tracker, T2w and DWI, interleaved as as shown in figure 8.1a. The tracker sequence was a T2-weighted 3D turbo spin echo with a field of view $380 \times 304 \times 360$ mm$^3$, voxel size $2.0 \times 2.0 \times 2.5$ mm$^3$ , repetition time 1500 ms, echo time 135 ms and a receive bandwidth of 1302 Hz pixel$^{-1}$. The sequence was acquired with compressed sense factor of 2. The T2w sequence was also a 3D T2-weighted turbo spin echo sequence with a matched field of view, voxel size $1.0 \times 1.0 \times 2.0$ mm$^3$, repetition time 2000 ms, echo time 148 ms and a receive bandwidth of 658 Hz pixel$^{-1}$. The DWI sequence was a single-shot echo planar imaging sequence with b-values 100 smm$^{-2}$, 500 s mm$^{-2}$ and 1000 s mm$^{-2}$. The repetition time was 3000 ms and echo time 69.8 ms. The axial field of view varied between $200 \times 200$ mm$^2$ and $270 \times 270$ mm$^2$ with corresponding voxel sizes of $2.1 \times 1.6 \times 4.0$ mm$^3$ to $2.8 \times 2.1 \times 4.0$ mm$^3$. This gave the smallest voxel size whilst ensuring the participant anatomy in the anterior-posterior direction was covered. The number of slices was also varied from 22-26 to ensure anatomical coverage. The tracker and T2w sequences were centred on the femoral heads and DWI sequence centred on the cervix (female participants) or prostate (male). The acquisition times were 1.0, 6.2 and $5.5 - 6.1$ minutes for the tracker, T2w and DWI sequences respectively.

### 8.2.2 Automatic Contouring

The Deep Learning automatic contouring algorithm described in chapter 2 was used to generate automatic contours for the bladder, rectum and femoral heads (all participants)



(a) Protocol Diagram

(b) Example Images

Figure 8.1: Diagram of the MR imaging protocol (a). The tracker, T2w and DWI sequences are shown as blue, green and orange boxes respectively. The length of the boxes are proportional to the acquisition time. b) shows example images of the tracker (left), T2w (middle) and DWI (right).

and prostate (male participants only) organs on the T2w images. Each contour was manually reviewed by a clinical scientist and rated on the same 5-point scale used for the clinical scoring of the automatic contours:

1. Delete: The contour is not acceptable for clinical use, complete recontouring needed
2. Major: The contour is not acceptable for clinical use, significant correction needed
3. Intermediate: The contour is not acceptable for clinical use, some correction needed
4. Minor: The contour is acceptable for clinical use, but minor corrections could be done
5. Accept: The contour is acceptable for clinical use at it is

Contours rated $< 4$ (ie not clinically acceptable) were manually modified.

### 8.2.3 Analysing Organ Motion

The volume and centre point of the bladder on each T2w image was calculated. The change in volume and centre point as a function of time after drinking in each imaging session was determined. The bladder volume as a function of time after drinking for each participant and visit was fitted with a least squares linear regression to calculate a bladder fill rate (gradient) and the bladder volume at 30 minutes post-drinking.

Other organs will also move as the bladder fills due to the change in bladder volume. This change was evaluated through comparing the contours for each organ on the T2w 1 image to the T2w 2 image contours using the DSC and $\text{DTA}_{\text{mean}}$ delineation metrics. Similarly, the contours on the T2w 3 image were compared to the T2w 2 image contours using the same metrics. The degree of similarity between T2w 1 or T2w 3 and T2w 2 would capture the change in organ position and shape with the change in bladder filling.

The intra-participant repeatability of the bladder filling was also assessed. A modified

Bland-Altman analysis was carried out to assess the mean difference in bladder fill rate and bladder volume at 30 minutes post-drinking [267]. The 95% limits of agreement for both of these parameters were calculated. These indicate the interval around the per participant mean bladder fill rate within which 95% of bladder fill rates for different visits for the same participant would occur.

### 8.2.4 Developing a Model to Compensate for Organ Motion

The aim of the tracker images was to correct images acquired early or late in the scanning session to the middle point, which would be timed to match the bladder filling protocol used for treatment. Therefore deformable registrations between tracker 1 and tracker 3 (deformation D1→3), tracker 1 and 4 (D1→4) and tracker 2 and 4 (D2→4) were carried out to assess which performed best in correcting T2w 1 image to the T2w 2 image. Similarly, deformable registrations from tracker 5 to 3 (D5→3), tracker 6 to 3 (D6→3) and tracker 6 to 4 (D6→4) were carried out for correcting the T2w 3 image. All deformable image registrations were carried out in RayStation (v9a) using the hybrid intensity and structure based algorithm with the correlation coefficient as the similarity measure and an isotropic resolution 2.5 mm voxel$^{-1}$ [268].

### 8.2.5 Evaluating a Model to Compensate for Organ Motion

If the organ motion compensation model worked perfectly it would deform contours on images acquired early or late in the scanning session so that they would be exactly as if the image had been acquired at the middle point of the scan. Therefore, the model was evaluated by acquiring three T2w images at different time points, contouring each of the images, and deforming the T2w 1 and T2w 3 images to the T2w 2 image using the tracker-based deformations. If these deformations were accurate, the deformed T2w 1 and T2w 3 contours should exactly match the T2w 2 contours (assuming the automatic contours were accurate, see below). The similarity of the deformed T2w 1 and T2w 3 contours with the T2w 2 contours was assessed with the delineation metrics DSC and $\mathrm{DTA}_{\mathrm{mean}}$.

Firstly, the best performing deformation was determined using the bladder contour. The T2w 1 contour was deformed using each of the first three deformable registrations and the deformed contour compared to the T2w 2 image contours using the delineation metrics. The registration with the highest DSC and lowest $\mathrm{DTA}_{\mathrm{mean}}$ was selected for the following evaluation. A similar process was used with the bladder contour on the T2w 3 image and the second three deformable registrations.

Secondly, the selected deformations were used to deform all the automatic contours on the T2w 1 and T2w 3 images and the deformed contours compared to the T2w 2 contours using the same delineation metrics. These delineation results were assessed next to the delineation results of the T2w 1 and T2w 3 contours being compared to the T2w 2 contours without deformation (described above).

This evaluation relied on using the automatic MR contours that had not been reviewed by an oncologist. Therefore the distribution of DSC and $\text{DTA}_{\text{mean}}$ results from chapter 2 comparing automatic contour to manual contours on a 20-patient cohort were used as a measure of the automatic contour uncertainty. This provided a reference to assess the performance of the motion compensation model.

The aim was to carry out a similar analysis of the performance of the motion compensation model using contours on the DWI images and in addition investigate the impact on ADC values. Unfortunately, due to clinical time constraints, it was not possible for DWI contours to be produced and so that analysis was not performed. Future work will aim to investigate this.

## 8.3 Results

All participants followed the drinking protocol and images were acquired at the planned times after drinking (figure 8.2). Automatic contours were successfully created for all organs on all images (see figure 8.3). Upon review by the clinical scientist, manual adjustment was required for 8/81, 7/81, 0/162 and 2/45 contours for the bladder, rectum, femoral heads and prostate organs respectively. Clinically unacceptable bladder contours only occurred for small bladder volumes ($< 50$ cm$^3$) and unacceptable rectum errors were normally due to an incorrect superior border.

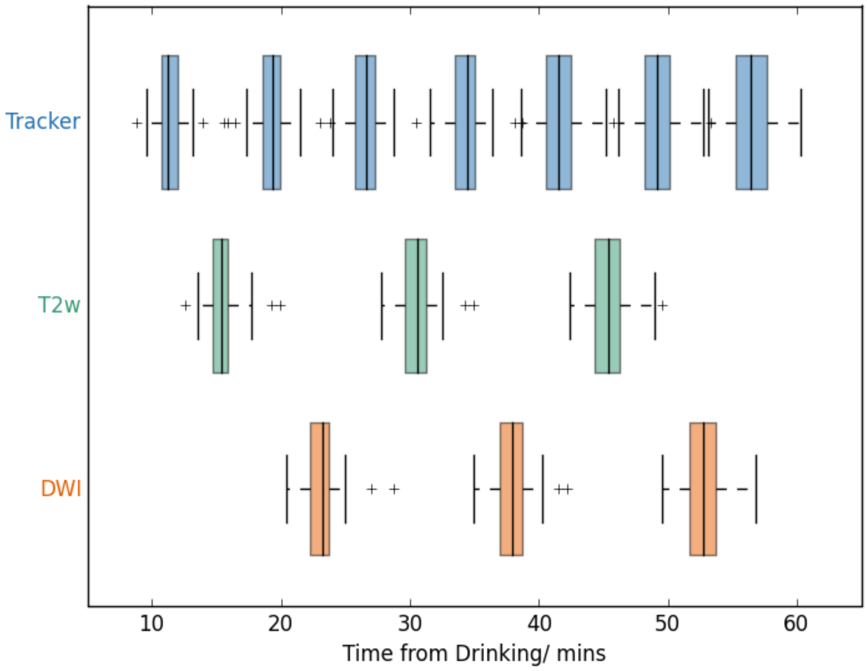
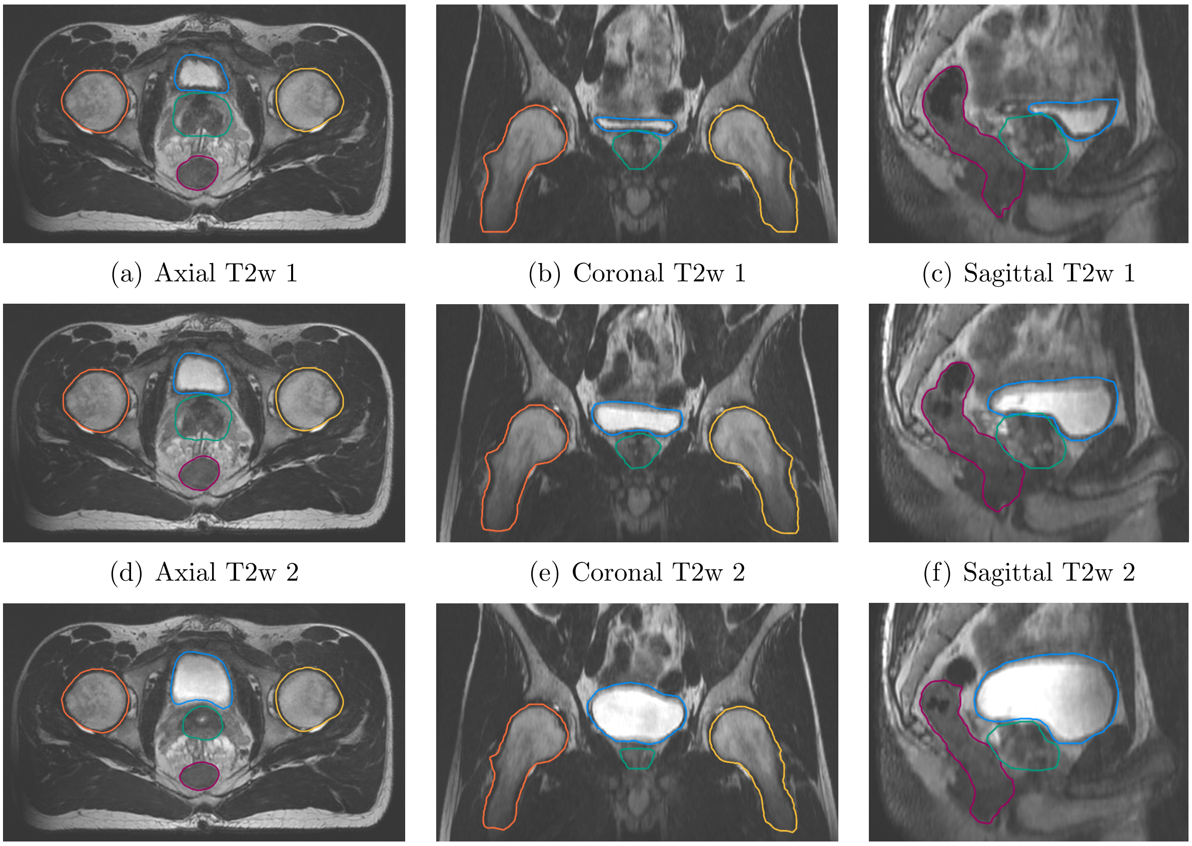


Figure 8.2: Boxplots of time differences from drinking time to middle of acquisition for the different sequences.



(a) Axial T2w 1

(b) Coronal T2w 1

(c) Sagittal T2w 1

(d) Axial T2w 2

(e) Coronal T2w 2

(f) Sagittal T2w 2

(g) Axial T2w 3

(h) Coronal T2w 3

(i) Sagittal T2w 3

Figure 8.3: Example T2-weighted images and contours for a representative participant and visit. Contours shown are bladder (blue), rectum (purple), left femoral head (yellow), right femoral head (orange) and prostate (green). The time after drinking for these images were 16.4, 31.0 and 45.7 minutes for the T2w 1, T2w 2 and T2w 3 images respectively. The participant and visit were selected through having the closest bladder volume at 30 minutes to the mean of all participants.

### 8.3.1 Organ Motion Analysis

The mean bladder fill rate was $6 \pm 2$ cm$^3$ min$^{-1}$ (± standard deviation, range 2 cm$^3$ min$^{-1}$, 9 cm$^3$ min$^{-1}$). The mean bladder volume at 30 minutes was 148±84 cm$^3$ (36 cm$^3$,341 cm$^3$). The increase in bladder volume was well fitted with a least squares regression (see figure 8.4). As the bladder filled the centre of the mass of the bladder also moved. For most participants the shift in the superior-inferior direction was the largest, although for several participants shifts in the anterior-posterior and right-left directions were of a similar magnitude (figure 8.5).

The change in bladder filling impacted other organs as well. This can be seen by the comparison of contours on images T2w 1 and T2w 3 to T2w 2 (figure 8.6). Although the bladder showed the largest DTA$_{\text{mean}}$ values (T2w 1: $4.3 \pm 0.3$ mm,T2w 3: $6.5 \pm 0.4$ mm, mean ± standard deviation), the prostate had changes of $1.6 \pm 0.2$ mm (T2w 1) and $2.0 \pm 0.3$ mm (T2w 3) and the rectum $1.7 \pm 0.2$ mm (T2w 1) and $1.9 \pm 0.2$ mm (T2w 3). As expected, the femoral heads showed the smallest differences (means $< 0.9$ mm). An example participant shown in figure 8.7 illustrates the impact of bladder filling on the movement of other organs in the pelvis.

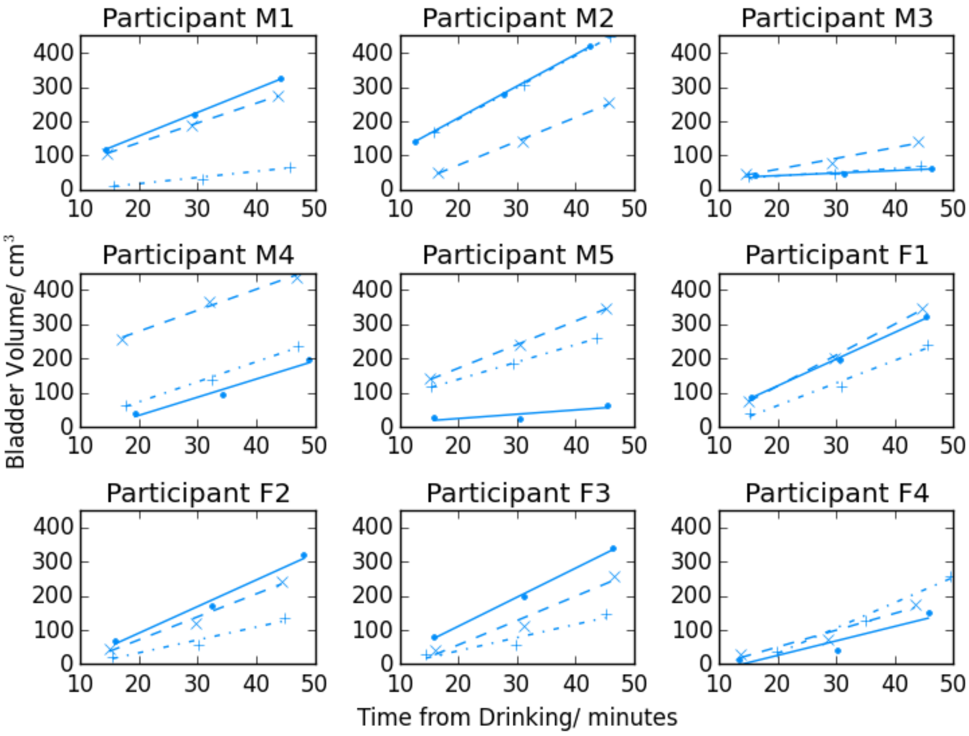


Figure 8.4: Plots of bladder volumes as a function of time from drinking for each participant visit. Bladder volumes for visits one, two and three are indicated by circle, cross and plus markers respectively. Least squares regression lines for visits one, two and three are shown by solid, dashed and dash-dotted lines respectively.

There was significant variability in bladder fill rates and 30-minute volumes both between participants and for the same participant between visits (figure 8.4). The 95% limits of agreement were $\pm 3.8$ $\mathrm{cm^3\,min^{-1}}$ for bladder fill rates and $\pm 156$ $\mathrm{cm^3}$ for the 30-minute volumes. Similarly the mean standard deviation of bladder fill rates over the three visits was 1.8 $\mathrm{cm^3\,min^{-1}}$ and of 30-minute volumes 70 $\mathrm{cm^3}$.

## 8.3.2 Evaluation of Model for Organ Motion Compensation

Figure 8.8 shows the distributions of DSC and $\mathrm{DTA_{mean}}$ results for each of the deformable registrations. Both for correcting for early images (smaller bladder volumes) and late images (larger bladder volumes), deformations which used tracker images that were on different sides of the T2w images (ie D1→4 and D6→3) performed worst. For the T2w 3 image the D6→4 deformation was better than the D5→3 deformation. The situation was less clear for the T2w 1, where the D2→4 deformation was only marginally better than the D1→3. Therefore, the deformations chosen for subsequent analysis were D2→4 for the T2w 1 image and D6→4 for the T2w 3 image.

An example of the impact of the deformable registration on compensating for bladder filling can be seen in figure 8.7.

The DSC results for the D2→4 and D6→4 deformations are shown in figure 8.9, with the

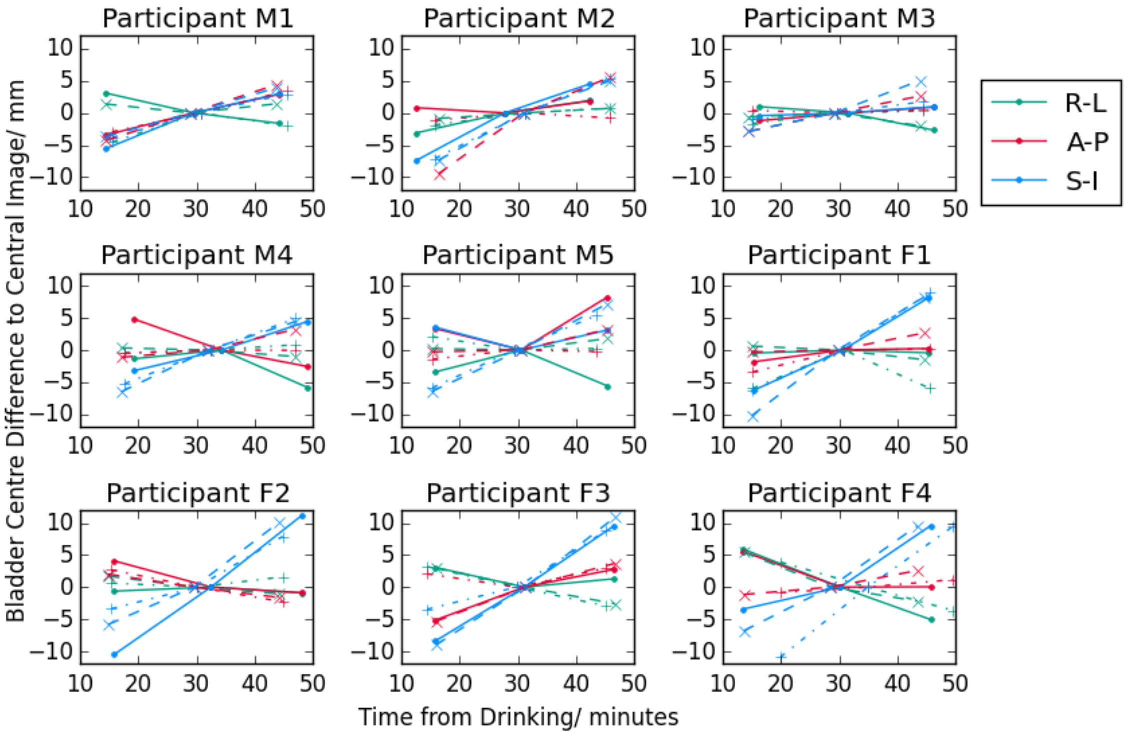


Figure 8.5: Plots of bladder centres relative to the bladder centre for the central image (T2w 2) as a function of time from drinking. Movements in the right-left (R-L), anterior-posterior (A-P) and superior-inferior (S-I) direction are shown by green, red and blue markers respectively. Results for visits one, two and three are indicated by circle markers and solid lines, cross markers and dashed lines and plus markers and dash-dotted lines respectively.

equivalent $\text{DTA}_{\text{mean}}$ results in figure 8.10. There was a clear improvement from the deformations for the bladder contour with a mean difference in DSC of $\Delta_{D2\to4} = 0.20 \pm 0.02$ $(-0.05,0.42)$ and $\Delta_{D6\to4} = 0.18 \pm 0.02$ $(0.03,0.32)$ for the D2→4 and D6→4 deformations respectively. The $\text{DTA}_{\text{mean}}$ showed a similar improvement, $\Delta_{D2\to4} = 1.6 \pm 0.5$ mm $(-3.2$ mm,$5.0$ mm$)$ and $\Delta_{D6\to4} = 5.1 \pm 0.4$ mm $(0.3$ mm,$9.5$ mm$)$. There were smaller improvements for the femoral heads ($\Delta\text{DTA}_{\text{mean}} \leq 0.4$ mm) and little or no improvement for the rectum and prostate.

Finally, although no quantitative analysis of the DWI images was carried out, figure 8.11 shows an example of the application of the model to DW-MR images. There were significant discrepancies from the ADC map from DWI 3 image compared to the DWI 2 image. But the application of the tracker deformation visually substantially improved the agreement between DWI 2 and the deformed DWI 3 image.

## 8.4 Discussion

This study has analysed organ motion in the pelvis due to bladder filling and developed and evaluated a method for compensating for this motion using fast tracker images and deformable registrations. If successful, this would enable multiple MR sequences to be acquired for radiotherapy planning even though the total acquisition time was $\geq 20$ mins.

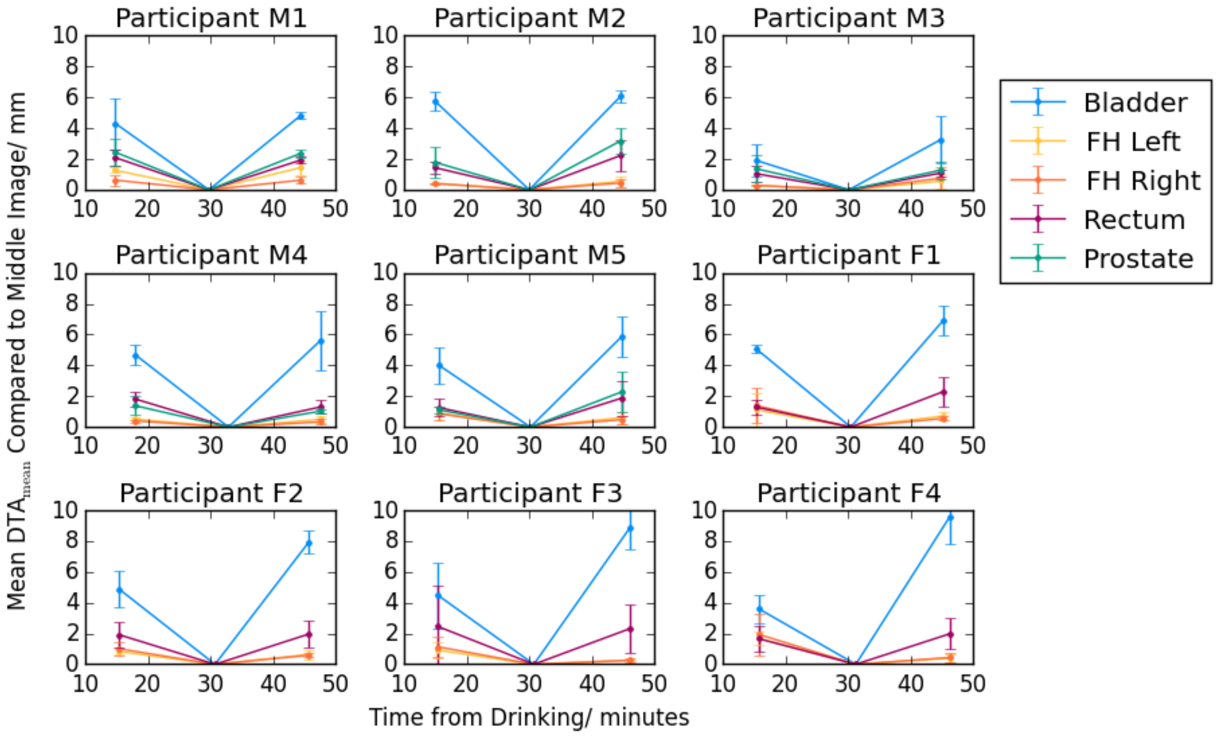
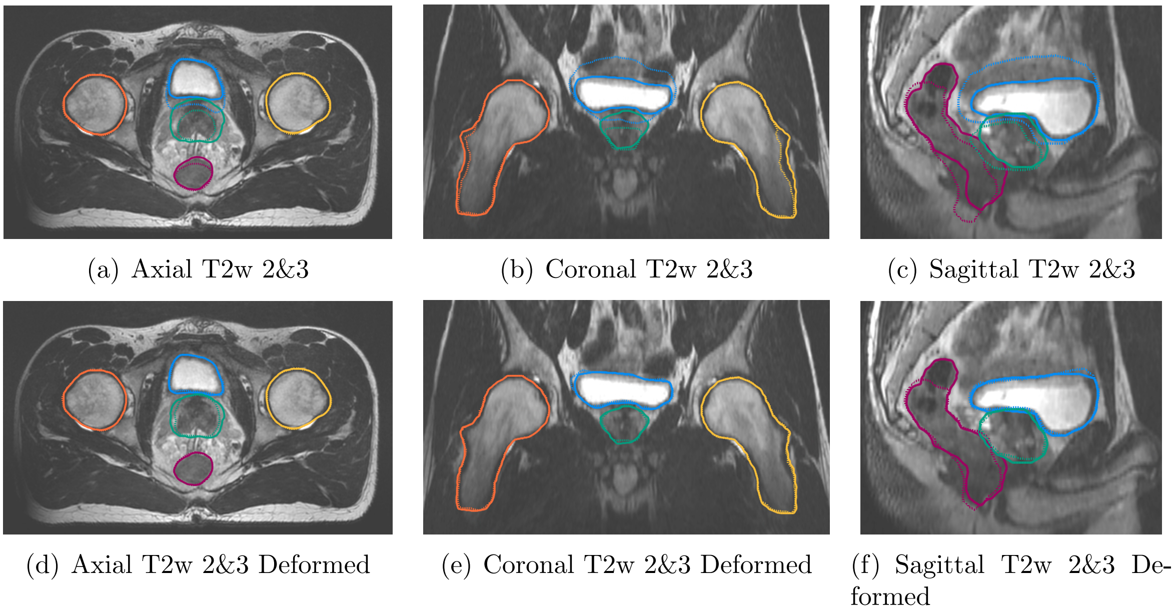


Figure 8.6: Plots of DTA$_{\text{mean}}$ of contours on image T2w 1 and T2w 3 compared to T2w 2 as function of time from drinking. The following organs are shown: bladder (blue), rectum (purple), femoral heads (yellow and orange for left and right respectively) and prostate (green). Values are given as the mean over three visits, with the error bars showing ± one standard deviation.



(a) Axial T2w 2&3

(b) Coronal T2w 2&3

(c) Sagittal T2w 2&3

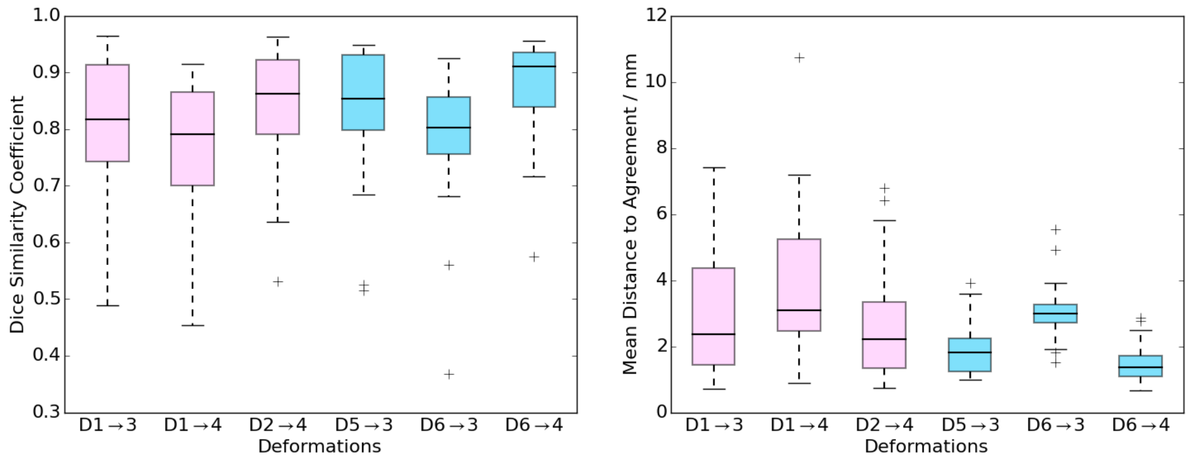
(d) Axial T2w 2&3 Deformed

(e) Coronal T2w 2&3 Deformed

(f) Sagittal T2w 2&3 Deformed

Figure 8.7: Top row: example contours from the T2w 2 image (solid lines) and T2w 3 image (dotted lines) on the T2w 2 image for the same participant in figure 8.3. Contours shown are bladder (blue), rectum (purple), left femoral head (yellow), right femoral head (orange) and prostate (green). Bottom row: same contours but with the T2w 3 image deformed by the D6→4 registration (dotted lines). The differences between solid and dotted contours have been substantially reduced.

The bladder filling rate was $6 \pm 2$ cm$^3$ min$^{-1}$, which lies within the ranges reported in the literature. Hynds et al reported bladder filling rates following bladder prep of $5 \pm 3$ cm$^3$ min$^{-1}$ in 30 prostate radiotherapy patients prior to treatment, which agrees well



(a) DSC

(b) $\text{DTA}_{\text{mean}}$

Figure 8.8: a) Boxplot of the DSC results comparing T2w 1 (pink) and T2w 3 (blue) bladder contours to T2w 2 contours.

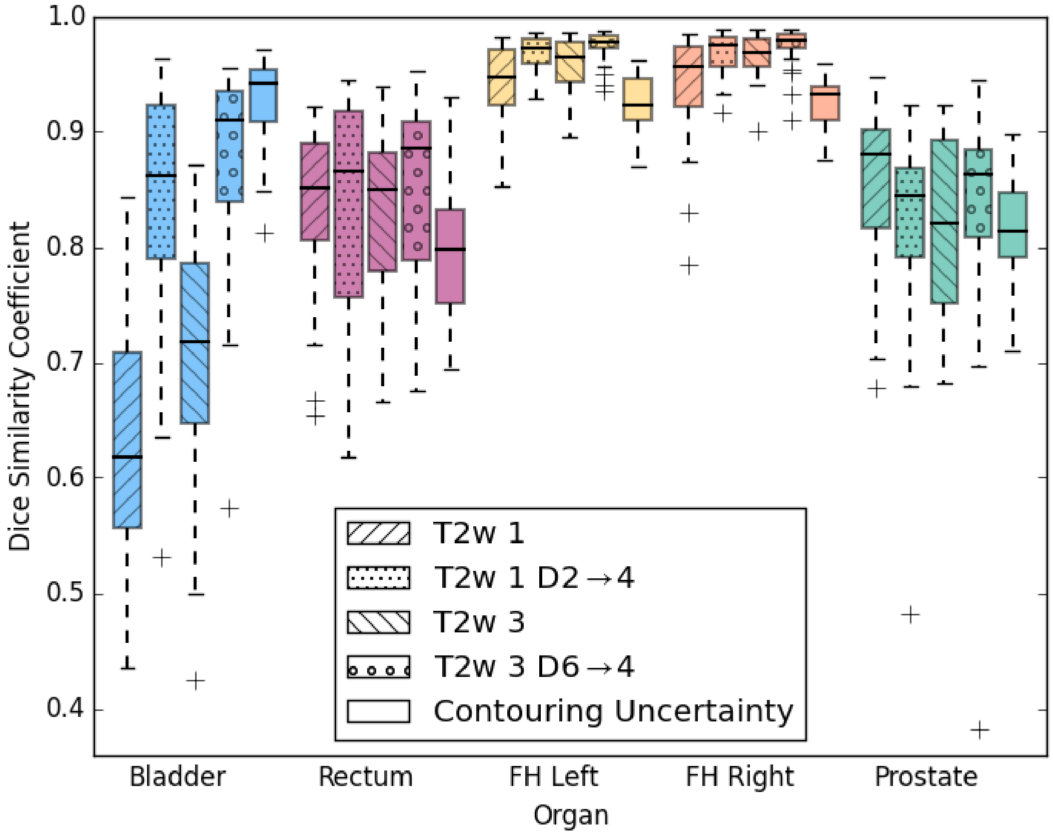


Figure 8.9: Boxplot of the DSC results comparing T2w 1 and T2w 3 contours to T2w 2 contours for each organ (left diagonal and right diagonal hatches respectively). Also shown is the same comparison after deformation with the 2→4 (dotted hatch) and 6→4 (circled hatch) registrations. Finally the contouring uncertainty is shown as the boxplot of the automatic contours compared to manual contours from a different patient cohort as described in chapter 2.

with this study [183]. Lotz et al. found higher bladder filling rates, $9 \pm 3$ $\text{cm}^3\,\text{min}^{-1}$, although the reported range ($2 - 15$ $\text{cm}^3\,\text{min}^{-1}$) encompassed all the bladder filling rates here ($2 - 9$ $\text{cm}^3\,\text{min}^{-1}$) [185]. McBain et al. found much lower bladder rates, $0.9 \pm 0.7$ $\text{cm}^3\,\text{min}^{-1}$ (range $0.2, 2.0$ $\text{cm}^3\,\text{min}^{-1}$), although this was following one hour fasting

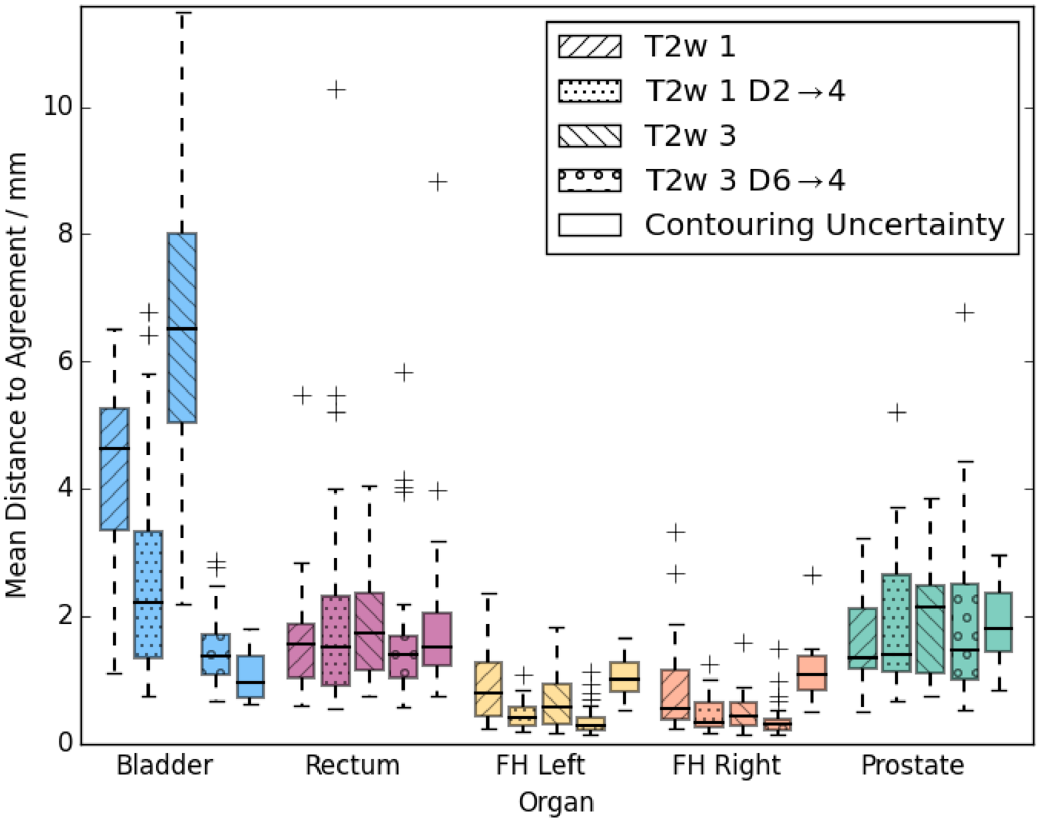


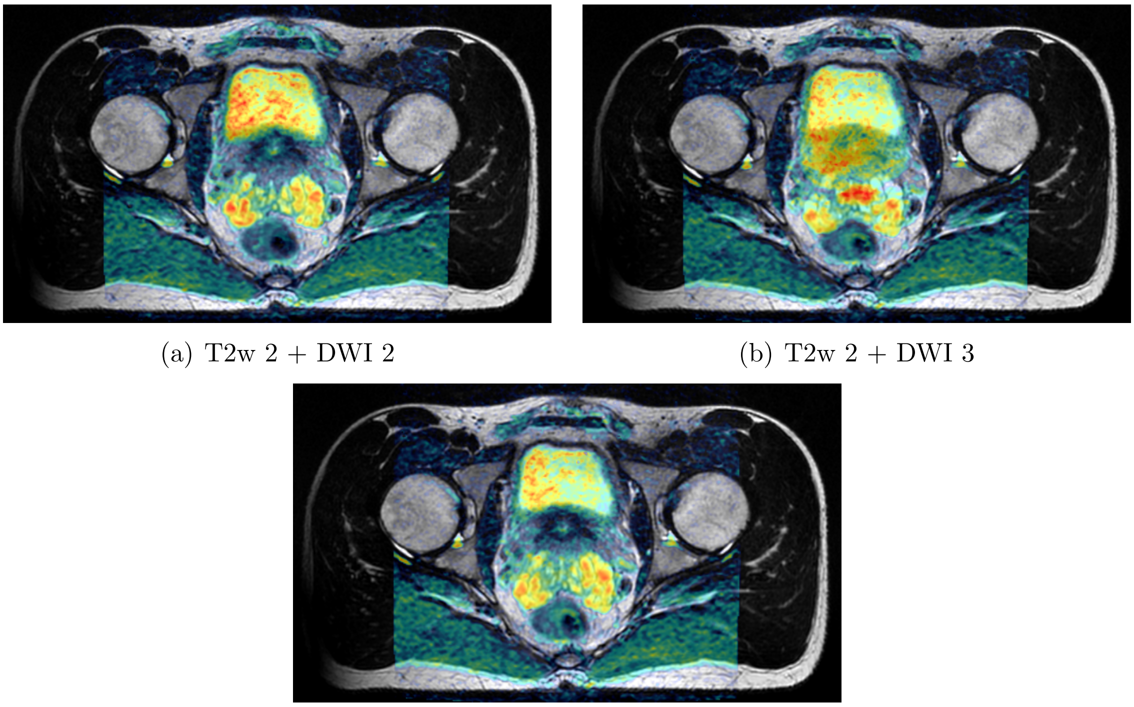
Figure 8.10: Boxplot of the DTA$_{\text{mean}}$ results comparing T2w 1 and T2w 3 contours to T2w 2 contours for each organ (left diagonal and right diagonal hatches respectively). Also shown is the same comparison after deformation with the 2→4 (dotted hatch) and 6→4 (circled hatch) registrations. The contouring uncertainty is shown as the boxplot of the automatic contours compared to manual contours from a different patient cohort (described in chapter 2).

from fluids and directly post-voiding bladder [188]. Similar to Lotz et al., we found the increase in bladder volume with time was well-fitted with a least-squares regression line.

The mean bladder volume at 30 minutes was 148 ± 84 cm$^3$, which is lower than the 271 cm$^3$ mean volume from all studies in a meta-analysis of bladder volumes for prostate radiotherapy, although within the mean ± standard deviation of ±149 cm$^3$ [190]. Incomplete bladder emptying is also a symptom of prostate cancer which may have contributed to the larger volumes in the prostate cancer patients [269].That meta-analysis reported a wide range of mean bladder volumes, 159 cm$^3$ to 367 cm$^3$, indicating wide variation between patients in agreement with this study.

There was also significant variability between visits for the same participant, both in bladder fill rate and bladder volume at 30 minutes. This too is in agreement with the literature [183, 270], suggesting that bladder filling rate and volume does not depend simply on bladder preparation protocol and particular patient anatomy but also on hydration status and possibly other factors [185]. This remains an ongoing challenge for accurate radiotherapy in the pelvis [186].

Visually the motion compensation method appeared to substantially reduce the impact of bladder motion (figures 8.7 and 8.11). This is clearly seen for the bladder contour



(a) T2w 2 + DWI 2

(b) T2w 2 + DWI 3

(c) T2w 2 + DWI 3 Deformed

Figure 8.11: Example ADC maps (shown in colour wash) overlain on the T2w 2 image for the participant in figure 8.3. a) shows the ADC map from the DWI 2 image (7.3 minutes between centre of DWI 2 acquisition and centre of T2w 2 acquisition). b) shows the ADC map from the DWI 3 image (21.9 minutes between centre of acquisitions). c) shows the ADC map from the DWI 3 image having been deformed by D6→4 registration. Each ADC map is shown on the same colour wash scale.

with improvements in $\text{DTA}_{\text{mean}}$ of 1.6 mm and 5.1 mm. However the impact on prostate and rectum contours appeared minimal. This could potentially be due to the contouring uncertainty of the automatic contouring algorithm. This was estimated using the results from chapter 2 of the comparison between automatic and manual contours in a different patient group. For the bladder contour both the motion-induced contour differences and the improvements from the application of the compensation method were substantially higher than the contouring uncertainty. This implies that the evaluation is accurately capturing the improvement. However, for the rectum and prostate contours the motion-induced differences were similar to the contouring uncertainty, which means it is impossible to determine whether differences were due to organ motion or inaccuracies in the automatic contouring. Although all contours were reviewed and amended where considered necessary, this was not done by an appropriately trained clinician and so some contouring inaccuracies were likely to remain. Surprisingly the model appeared to reduce motion-induced differences in the femoral head contours, even though it would be expected that the femoral heads would be unaffected by bladder filling. Potentially there were small changes in the femoral heads due to participants slightly shifting position during the scan, and the deformable registration was correcting those. However, the differences were all less than the contouring uncertainty of the automatic contouring algorithm, so it is impossible to conclude this was a genuine effect.

The D6→4 deformations resulted in better agreement between the T2w 3 and T2w 2 bladder contours than the D2→4 deformations did for the T2w 1 and T2w 2 contours ($\text{DTA}_{\text{mean}}$ of 1.5 ± 0.1 mm compared to 2.6 ± 0.3 mm). This may suggest that the deformation algorithm worked better going from larger volumes to smaller ones, although it may also reflect the increased variabilities in the automatic contouring algorithm for small bladder volumes.

The evaluation of the motion compensation method had several limitations. Firstly, the use of automatic contours that had not been reviewed by a trained clinician, as discussed above. Secondly, the motion-induced differences in contours and the subsequent improvement from the application of the compensation method were evaluated using delineation metrics DSC and $\text{DTA}_{\text{mean}}$. However these have limited correlation with clinical rating (as discussed in chapter 2). Future work could use clinical rating by an appropriate clinician to evaluate the effectiveness of the model, which would be a more clinically relevant evaluation method.

A potential advantage of the approach presented here is that the tracker images are available for clinician review to assess the accuracy of the deformations. This is important because deformable registration is a non-anatomical model, with constraints that are not valid in all situations and per-patient review of the deformations is important within a clinical pathway [266]. The model used here enables the a clinician to review the deformations on the tracker model to visually assess them before applying them to other image sets, therefore facilitating clinical implementation.

Future work will quantitatively evaluate the impact of the motion compensation model on DW-MR images as well. This is important to demonstrate that the motion compensation method can accurately handle images which look very different to anatomical MR images. The impact of using the tracker-derived deformable registration versus simply deformably registering the DWI to the T2w image will also be assessed. Finally, the motion compensation model also provides a patient-specific assessment of how the patient anatomy changes with bladder filling. The contours on the central T2w 2 image could be deformed to earlier and later time points by applying the inverse of the deformations used to correct images at those earlier and later time points. This patient-specific organ motion model could then be used with a robust treatment planning approach to generate treatment plans robust to the expected organ motion, without having to use large population based PTV margins [271]. This could potentially result in improved target coverage and/or reduced dose to normal tissues. Future work will develop a method of generating these patient-specific motion models and evaluate the effectiveness of robust treatment planning incorporating them. However the effectiveness of such an approach may be limited by the large intra-participant variation in bladder filling found in this study and reported in the literature. So critical to that evaluation will be investigating the impact of intra-participant variation in bladder filling.

Another important application that will be investigated in the future is using the motion compensation model on a simultaneous PET-MR scanner. PET imaging would also benefit from long acquisition times through improved image quality and/or reducing the injected patient activity. A simultaneous PET-MR scanner would be able to acquire PET images throughout the $\geq$ 20 minutes of MR acquisition time, providing a substantial increase in PET signal compared to typical PET acquisitions of 2-5 minutes. However, the PET images would also be affected by changes in bladder filling, and so the images could be sub-divided into smaller $\sim$ 5 minute images that could be corrected using the MR tracker-based deformable registrations. This would provide improved quality PET images without the impact of bladder filling. An alternative approach would be to use the PET images themselves to track bladder motion, since there is always a large amount of activity within the bladder contents. Future work will evaluate both of these alternatives.

## 8.5 Conclusion

A method for tracking and correcting organ motion in the pelvis has been developed. Organ motion in the pelvis due to bladder motion is substantial. Initial results of the motion compensation method are promising but require further evaluation using clinician rating. Use of this method could enable using multiple MR sequences for pelvic radiotherapy planning, facilitating the use of the full potential of MR for radiotherapy. It would also provide a patient-specific motion model that could be used for robust planning approaches.

# Chapter 9

# Conclusions and Future Work

This thesis aimed to overcome the scientific and technical barriers to MR-only and PET-MR radiotherapy planning for pelvic cancers. The methods developed and/or evaluated in this thesis and how they have enabled a PET-MR-only radiotherapy pathway for pelvic cancers are summarised in figure 9.1. These were designed to fit with the general radiotherapy workflow, summarised in figure 1.1.

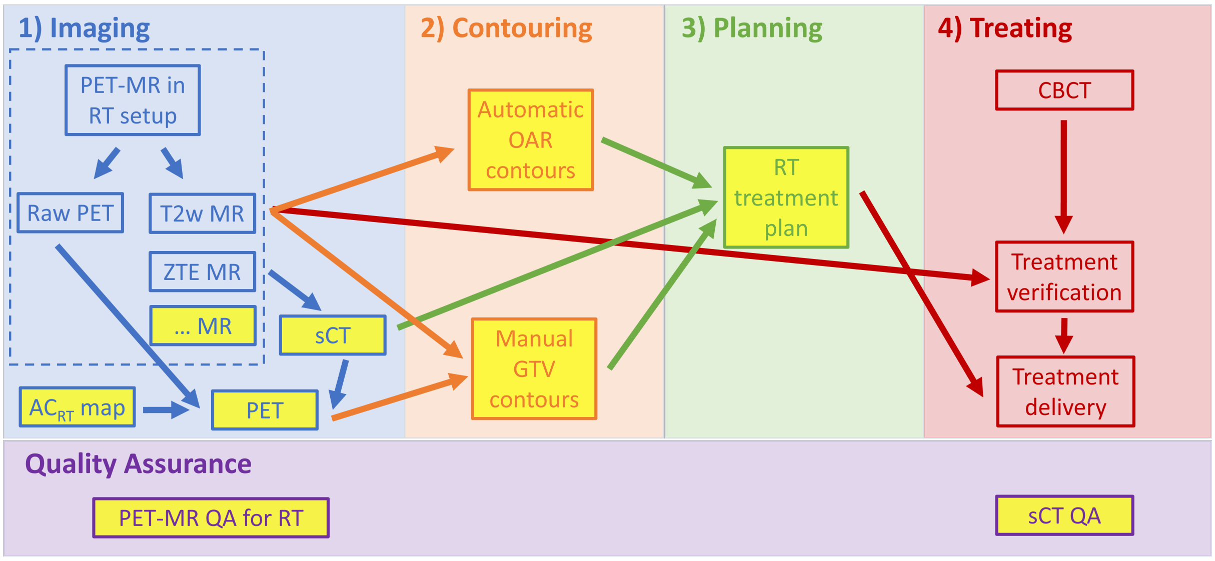


Figure 9.1: Diagram of the PET-MR-only radiotherapy pathway, with arrows indicating dataflows, showing the methods developed and/or evaluated in this thesis which have enabled the PET-MR-only radiotherapy pathway (highlighted in yellow). 1) Imaging, a PET AC method for the RT hardware has been developed and evaluated (chapters 4 & 5) and combined with a ZTE-based sCT for patient AC (chapter 6), and a method of tracking and compensating organ motion developed which enables multiple MR sequences to be acquired (chapter 8). 2) Contouring, an automatic MR-based OAR contouring algorithm has been evaluated (chapter 2) and the impact on PET GTV delineation from the RT hardware investigated (chapter 5). 3) Planning, the sCT has been evaluated for radiotherapy dose calculations (chapter 2). Finally, QA methods for ensuring PET-MR image quality for RT planning (chapter 7) and the dose calculation accuracy of sCT (chapter 3) have been developed.

Chapter 2 described the evaluation of the dose calculation accuracy of a ZTE-based Deep Learning sCT algorithm. Dose calculation accuracy was high, with mean dose differences to the PTV D98% being $\leq 0.6\%$ for both the cylindrical PTV plans and the clinical treatment plans. This suggests that the sCT dose not significantly increase the overall dosimetric uncertainty of the radiotherapy pathway, indicating it would be suitable for clinical use. An automatic MR-based contouring algorithm was also evaluated for pelvic OARs. Doses to automatic and manual contours from clinically representative treatment plans were compared, with DVH differences $< 1.0$ Gy (or 1 percentage point in relative volume) for the bladder femoral heads, penile bulb, rectum and urethra organs. Median clinical ratings for these organs, and the pelvis body, were considered acceptable, although only the femoral heads and pelvis body were acceptable for all patients. This suggests the algorithm is sufficiently accurate for clinical use for prostate and ano-rectal radiotherapy treatment planning as long as the contours are manually reviewed and modified where necessary. Together, the sCT and automatic contouring algorithms enable a streamlined MR-only radiotherapy pathway to be implemented, efficiently reducing geometric uncertainties without increasing dosimetric uncertainties. In the future these algorithms could be evaluated on larger multi-centre datasets to ensure these results are generalisable. This could also specifically include investigating the sensitivity of the algorithms to variations in the input images such as resolution, fields of view and SNR.

A patient-specific QA test of the sCT dose accuracy is important to provide clinical confidence in MR-only radiotherapy pathways. Chapter 3 presented the evaluation of such a test using the dose calculated on the first-fraction CBCT with a highly automated method. There was a small systematic difference between sCT and CBCT, indicating that asymmetric dose tolerances of $[-2.0\%, 1.0\%]$ would be appropriate clinically. Future work could validate these tolerance levels to detect deliberate errors such as not applying 3D distortion correction post-processing to the image.

In chapter 4 the impact on PET-MR image quality of acquiring images in the radiotherapy setup was evaluated using standard image quality phantoms. Both PET and MR images were adversely affected, with PET activity losses of $-17.7 \pm 0.1\%$ and MR SNR reductions of 45%. A method of accounting for the PET attenuation from the radiotherapy hardware was devised, which substantially reduced the PET activity losses to $-2.7 \pm 0.1\%$ and out-performed the diagnostic setup with MR anterior coil in place by 5.6%. A further development would be to investigate methods of improving the MR SNR in the radiotherapy position, such as utilising Deep Learning MR reconstruction methods or light-weight blanket MR receive coils that would remove the need for coil bridges.

The PET attenuation correction method developed in chapter 4 was then applied to ano-rectal radiotherapy patients in chapter 5. Similar improvements in PET activity between images reconstructed with and without attenuation correction of the radiotherapy hardware were observed (differences of 13.8%). These improvements did not translate into statistically significant differences in tumour delineation, but did significantly change PET metabolic parameters such as $\text{SUV}_{\text{max}}$ and TLG. These differences could be clinically significant for radiotherapy dose painting and treatment prognostication where accurate quantitative PET is important. This suggests that attenuation correction of the radiotherapy hardware is feasible and likely to be important for the use of PET-MR for pelvic radiotherapy.

Another barrier to the use of PET-MR for radiotherapy planning is accurate attenuation correction of the patient, which was the focus of chapter 6. This evaluated the same sCT images described in chapter 2 when used for PET attenuation correction, compared to gold standard CT-based attenuation correction. Tumour delineation on the sCTAC PET images was very similar to the current clinical standard MRAC method, indicating again no significant change for delineation accuracy. However, metabolic parameters $\text{SUV}_{\text{max}}$ and $\text{SUV}_{\text{mean}}$ within the primary GTVs on sCTAC were statistically equivalent to those calculated using CTAC within ±5% and ±4% respectively ($p < 0.001$). In contrast the MRAC-derived metabolic parameters were not equivalent ($p = 0.33$ and $p = 0.21$). This suggests that sCT based attenuation correction enables accurate quantitative PET images, substantially improving on the current MRAC method. This study was limited by small patient numbers, so further evaluation in larger multi-centre patient cohorts would provide a more complete evaluation of the sCTAC.

Essential to the use of PET-MR in the radiotherapy planning is a QA programme focused on radiotherapy imaging requirements. Chapter 7 described the development and evaluation of the QA tests needed for such a programme. All the tests were repeatable and stable over a 12-month period suggesting they would be appropriate for a QA programme. Future work will derive appropriate tolerance levels and test frequencies, enabling a comprehensive QA programme to be produced. This will underpin the use of high-quality, robust PET-MR imaging for radiotherapy planning.

Finally, the true potential of PET-MR-only radiotherapy rests on using the range of PET-MR images available. However, imaging sessions $\geq 20$ minutes become challenging for pelvic radiotherapy because of the change in internal anatomy over that duration of image acquisition due to bladder filling. Chapter 8 attempted to overcome this barrier by developing a MR method for tracking and correcting organ motion using deformable registrations between repeated brief tracking sequences interleaved throughout the MR acquisition. Initial results appear promising but require further evaluation from clinicians to fully assess the utility of the method. Future work will evaluate this method fully for anatomical MR, DW-MR and PET acquisitions in the pelvis. Successful validation would enable the full potential of PET-MR images to be utilised for radiotherapy planning in the pelvis.

Looking to the future, there remain a number of hurdles that still need to be overcome before MR-only and PET-MR can be widely translated into clinical pratice. MR-only radiotherapy is the closest, with several centres worldwide clinically treating patients using MR-only radiotherapy methods [71,213]. Commerical sCT algorithms are available for an expanding range of clinical sites, including male and female pelvis [67], head and neck [272] and brain [198]. So far however, MR-only radiotherapy has only been used to treat patients with prostate cancer [81]. This may be due to concerns around using the sCT for image matching to on-treatment CBCT images, with the linac manufacturers so far not enabling MR images to be used for this purpose [222]. Further research on MR-only radiotherapy outside the prostate, especially methods of on-treatmet image verification, are needed. For all clinical sites there also still remains a lack of clinical evidence demonstrating equivalence or superiority in outcomes for MR-only radiotherapy compared to the current standard of care [81]. Further research in cohort studies and randomised trials is required to generate this clinical evidence.

PET-MR imaging for radiotherapy planning remains significantly further from routine clinical use. The methods developed and evaluated in this thesis have enabled accurate, quantitative PET-MR imaging in the pelvic radiotherapy position. The key next steps are to develop and validate ways of utilising this quantitative PET-MR for radiotherapy planning, such as treatment stratification and dose painting. O'Connor et al have published a quantitative imaging biomarker roadmap with three stages of development: 1) imaging biomarker evaluated in pre-clinical and observational clinical settings, 2) imaging biomarker established as a reliable metric which can be uesd in large-scale studies and 3) imaging biomarker clinically implemented in patient care [273]. Currently PET-MR imaging for radiotherapy remains in the first stage, with method developments and small-scale exploratory studies. Therefore the next steps for PET-MR in radiotherapy research is to focus on moving to stage 2), with the establishing of combined PET-MR metrics with demonstrated reliability and relevance. The research areas to focus on for PET-MR in radiotherapy planning differ depending on its role in treatment stratification and response monitoring, or in dose painting.

For treatment stratification and response monitoring, the first step is to develop a combined PET-MR imaging metric. A principle advantage of simultaneous PET-MR scanners is the high spatial alignment between PET and MR images, which is particularly important in the pelvis due to the impact of organ motion (as demonstrated in chapter 8). This means a sub-volume analysis of the distribution of PET SUV and DW-MR ADC values within the tumour can be utilised, rather than whole-tumour metrics as is required when registering PET-CT and MR images from different imaging sessions. It is biologically plausible that the response of a tumour to treatment is governed by its most active sub-volumes rather than the tumour as a whole, which potentially makes a combined PET-MR utilising a sub-volume analysis more accurate. Therefore developing such combined PET-MR metrics should be a focus of future research efforts in PET-MR for radiotherapy. The second step would be to quantify the test-retest repeatability of these developed PET-MR metric(s) [274]. This is essential to interpreting PET-MR measurements as it enables determining whether differences between patients reflect true differences in tumour activity or just the uncertainties involved in the PET-MR measurement. Repeatability measurements should be done both in phantoms and in patient cohorts [274]. The third step involves evaluating the value of PET-MR metrics for predicting patient's response to treatment, both prior to and during treatment. And finally this predictive value needs to be demonstrated to be generalisable across different treatment centres and with sufficient patient numbers to be statistically sound [275]. Successful completion of this research programme would enable PET-MR imaging for radiotherapy treatment stratification and response monitoring to move past the first translational gap into stage 2), ready to be used in large-scale studies that test treating patients on the basis of PET-MR derived treatment stratification strategies.

For dose painting, there are a similar set of steps that future research programmes would need to cover. Firstly, a method of localising the tumour sub-volume to receive a radiotherapy boost using PET and MR imaging data needs to be developed. Similar to treatment stratificaiton, the high spatial alignment of the PET-MR scanners gives this an intrinsic advantage which should be exploited. Secondly, this method needs to be evaluated by comparisons to histopathology analysis [275]. Good alignment between PET-MR identified tumour sub-volumes and histopathology would provide strong evidence of the accuracy of PET-MR based sub-volume contouring and motivate using the images for dose painting. There are a number of challenges regarding histopathology validation of imaging techniques, including the difficulty of accurately registering *ex vivo* samples with *in vivo* imaging [276] and quantifying the inter-observer variability in histopathology delineations [277]. However, progress has been made in developing robust methods of validating imaging identified sub-volumes with histopathology, especially in the prostate [110, 276]. Histopathology analysis also has challenges for sites where surgery is not routinely used, such as anus and gynaecological cancers. Thirdly, the test-retest repeatability of PET-MR sub-volume delineation needs to be established [274]. Fourthly, the prescribed dose to the boost volume determined using radiobiological modelling to predict tumour control probability and planning studies to determine safety in dose to OARs. Finally, the safety of this approach would need to be assessed in phase I clinical trials. Together this research programme would cross the translational gap to stage two, enabling PET-MR based dose painting to be assessed in phase III randomised controlled trials.

There has been one phase III trial evaluting DW-MR and DCE-MR based sub-volume dose painting in prostate cancer which has recently reported results [278]. This has demonstrated significantly improved biochemical disease free survival in the boost dose arm (92% compared to 85%) with no significant differences in treatment toxicity. This is a really promising result for the dose painting strategy in general, and motivates continuing research in this area following the steps laid out above.

In summary, this thesis has sought to overcome the technical barriers to MR-only and PET-MR radiotherapy planning. These two approaches can be combined to produce a PET-MR-only pathway, a 'one-stop shop' PET-MR imaging session providing high quality anatomical and functional MR images, quantitative PET images, sCT images and automatic OAR contours. This wealth of imaging information would be available within a streamlined patient pathway and without the challenges of registering images from multiple patient setups and with intra-session differences from bladder filling corrected. Future research would focus on developing and validating methods of robustly and repeatably incorporating this image information for tumour delineation, radiotherapy dose painting and treatment response monitoring. All this would be based on the accurate quantitative PET-MR images enabled by the methods developed in this thesis.

# References

[1] B. Benton, C. Norton, J. O. Lindsay, S. Dolan, and H. J. N. Andreyev. Can Nurses Manage Gastrointestinal Symptoms Arising from Pelvic Radiation Disease? *Clin Oncol*, 23(8):538–551, October 2011.

[2] Kirsten AL Morris and Najib Y. Haboubi. Pelvic radiation therapy: Between delight and disaster. *World J Gastrointest Surg*, 7(11):279–288, November 2015.

[3] Rashmi Jadon, Emma Higgins, Louise Hanna, Mererid Evans, Bernadette Coles, and John Staffurth. A systematic review of dose-volume predictors and constraints for late bowel toxicity following pelvic radiotherapy. *Radiat Oncol*, 14(1):57, April 2019.

[4] Emile NJT van Lin, Jurgen J Fütterer, Stijn WTPJ Heijmink, Lisette P van der Vight, Aswin L Hoffmann, Peter van Kollenburg, HenkJan J Huisman, Tom WJ Scheenen, J Alfred Witjes, Jan Willem Leer, et al. IMRT boost dose planning on dominant intraprostatic lesions: Gold marker-based three-dimensional fusion of CT with dynamic contrast-enhanced and 1 H-spectroscopic MRI. *Int J Radiat Oncol Biol Phys*, 65(1):291–303, 2006.

[5] Ting Yu, Qiongwen Zhang, Tianying Zheng, Huashan Shi, Yang Liu, Shijian Feng, Meiqin Hao, Lei Ye, Xueqian Wu, and Cheng Yang. The Effectiveness of Intensity Modulated Radiation Therapy versus Three-Dimensional Radiation Therapy in Prostate Cancer: A Meta-Analysis of the Literatures. *PLoS One*, 11(5):e0154499, May 2016.

[6] Lara Hathout, Terence M. Williams, and Salma K. Jabbour. The Impact of Novel Radiation Treatment Techniques on Toxicity and Clinical Outcomes In Rectal Cancer. *Curr Colorectal Cancer Rep*, 13(1):61–72, February 2017.

[7] Barbara Segedin and Primoz Petric. Uncertainties in target volume delineation in radiotherapy–are they relevant and what can we do about them? *Radiology and Oncology*, 50(3):254–262, 2016.

[8] David Thwaites. Accuracy required and achievable in radiotherapy dosimetry: Have modern technology and techniques changed our views? *J Phys Conf Ser*, 444:012006, 2013.

[9] CF Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *J Med Phys*, 33(4):136–140, 2008.

[10] V Grégoire, TR Mackie, W De Neve, M Gospodarowicz, JA Purdy, M van Herk,

and A Niemierko. State of the art on dose prescription, reporting and recording in Intensity-Modulated Radiation Therapy (ICRU report No. 83). *Cancer Radiother*, 15(6-7):555–559, 2011.

[11] David Bernstein, Alexandra Taylor, Simeon Nill, and Uwe Oelfke. New target volume delineation and PTV strategies to further personalise radiotherapy. *Phys Med Biol*, 66(5):055024, February 2021.

[12] James M. Galvin and Wilfried De Neve. Intensity Modulating and Other Radiation Therapy Devices for Dose Painting. *J Clin Oncol*, 25(8):924–930, March 2007.

[13] Gustavo Arruda Viani, Eduardo Jose Stefano, and Sergio Luis Afonso. Higher-Than-Conventional Radiation Doses in Localized Prostate Cancer Treatment: A Meta-analysis of Randomized, Controlled Trials. *Int J Radiat Oncol Biol Phys*, 74(5):1405–1418, August 2009.

[14] Piet Dirix, Vincent Vandecaveye, Frederik De Keyzer, Sigrid Stroobants, Robert Hermans, and Sandra Nuyts. Dose painting in radiotherapy for head and neck squamous cell carcinoma: Value of repeated functional imaging with 18F-FDG PET, 18F-fluoromisonidazole PET, diffusion-weighted MRI, and dynamic contrast-enhanced MRI. *J Nucl Med*, 50(7):1020–1027, 2009.

[15] Uulke A van der Heide, Antonetta C Houweling, Greetje Groenendaal, Regina GH Beets-Tan, and Philippe Lambin. Functional MRI for radiotherapy dose painting. *Magn Reson Imaging*, 30(9):1216–1223, 2012.

[16] D Thorwarth. Functional imaging for radiotherapy treatment planning: Current status and future directions—a review. *Br J Radiol*, 88(1051):20150056, 2015.

[17] VS Khoo and DL Joon. New developments in MRI for target volume delineation in radiotherapy. *Br J Radiol*, 79(SP1):S2–15, 2006.

[18] Coen Rasch, Isabelle Barillot, Peter Remeijer, Adriaan Touw, Marcel van Herk, and Joos V Lebesque. Definition of the prostate in CT and MRI: A multi-observer study. *International Journal of Radiation Oncology* Biology* Physics*, 43(1):57–66, 1999.

[19] Uulke A. van der Heide, Johannes G. Korporaal, Greetje Groenendaal, Stefan Franken, and Marco van Vulpen. Functional MRI for tumor delineation in prostate radiation therapy. *Imaging Med*, 3(2):219, April 2011.

[20] The Royal College of Radiologists, Royal College of Physicians of London, Royal College of Physicians and Surgeons of Glasgow, Royal College of Physicians of Edinburgh, British Nuclear Medicine Society, and Administration of Radioactive Substances Advisory Committee. Evidence-based indications for the use of PET-CT in the United Kingdom 2016. *Clin Radiol*, 71(7):e171–e188, July 2016.

[21] Marco Krengli, Maria E. Milia, Lucia Turri, Eleonora Mones, Maria C. Bassi, Barbara Cannillo, Letizia Deantonio, Gianmauro Sacchetti, Marco Brambilla, and Eugenio Inglese. FDG-PET/CT imaging for staging and target volume delineation in conformal radiotherapy of anal carcinoma. *Radiat Oncol*, 5(1):10, February 2010.

[22] Espen Rusten, Bernt Louni Rekstad, Christine Undseth, Ghazwan Al-Haidari, Bettina Hanekamp, Eivor Hernes, Taran Paulsen Hellebust, Eirik Malinen, and Marianne Grønlie Guren. Target volume delineation of anal cancer based on magnetic resonance imaging or positron emission tomography. *Radiat Oncol*, 12(1):147, September 2017.

[23] Constantinos Zamboglou, Benedikt Thomann, Khodor Koubar, Peter Bronsert, Tobias Krauss, Hans C. Rischke, Ilias Sachpazidis, Vanessa Drendel, Nasr Salman, Kathrin Reichel, Cordula A. Jilg, Martin Werner, Philipp T. Meyer, Michael Bock, Dimos Baltas, and Anca L. Grosu. Focal dose escalation for prostate cancer using 68Ga-HBED-CC PSMA PET/CT and MRI: A planning study based on histology reference. *Radiat Oncol*, 13(1):81, May 2018.

[24] Ines Joye, Annelies Debucquoy, Christophe M. Deroose, Vincent Vandecaveye, Eric Van Cutsem, Albert Wolthuis, André D'Hoore, Xavier Sagaert, Mu Zhou, Olivier Gevaert, and Karin Haustermans. Quantitative imaging outperforms molecular markers when predicting response to chemoradiotherapy for rectal cancer. *Radiother Oncol*, 124(1):104–109, July 2017.

[25] Tufve Nyholm, Morgan Nyberg, Magnus G Karlsson, and Mikael Karlsson. Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments. *Radiat Oncol*, 4(54), 2009.

[26] Adalsteinn Gunnlaugsson, Emilia Persson, Christian Gustafsson, Elisabeth Kjellén, Petra Ambolt, Silke Engelholm, Per Nilsson, and Lars E. Olsson. Target definition in radiotherapy of prostate cancer using magnetic resonance imaging only workflow. *Phys Imaging Radiat Oncol*, 9:89–91, January 2019.

[27] Markus Alber and Daniela Thorwarth. Multi-modality functional image guided dose escalation in the presence of uncertainties. *Radiother Oncol*, 111(3):354–359, 2014.

[28] Emily Johnstone, Jonathan J Wyatt, Ann M Henry, Susan C Short, David Sebag-Montefiore, Louise Murray, Charles G Kelly, Hazel M McCallum, and Richard Speight. A systematic review of synthetic Computed Tomography generation methodologies for use in Magnetic Resonance Imaging – only radiation therapy. *Int J Radiat Oncol Biol Phys*, 100(1):199–217, 2018.

[29] Daniela Thorwarth, Sara Leibfarth, and David Mönnich. Potential role of PET/MRI in radiotherapy treatment planning. *Clin Transl Imaging*, 1(1):45–51, 2013.

[30] Sahar Ahangari, Naja Liv Hansen, Anders Beck Olin, Trine Jakobi Nøttrup, Heidi Ryssel, Anne Kiil Berthelsen, Johan Löfgren, Annika Loft, Ivan Richter Vogelius, Tine Schnack, Bjoern Jakoby, Andreas Kjaer, Flemming Littrup Andersen, Barbara Malene Fischer, and Adam Espe Hansen. Toward PET/MRI as one-stop shop for radiotherapy planning in cervical cancer patients. *Acta Oncol*, 60(8):1045–1053, August 2021.

[31] Elin Wallstén, Jan Axelsson, Joakim Jonsson, Camilla Thellenberg Karlsson, Tufve Nyholm, and Anne Larsson. Improved PET/MRI attenuation correction in the

pelvic region using a statistical decomposition method on T2-weighted images. *Eur J Nucl Med Mol Imaging Phys*, 7(1):68, November 2020.

[32] Jens M Edmund and Tufve Nyholm. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol*, 12(1):28, 2017.

[33] Florian Wiesinger, Mikael Bylund, Jaewon Yang, Sandeep Kaushik, Dattesh Shanbhag, Sangtae Ahn, Joakim H. Jonsson, Josef A. Lundman, Thomas Hope, Tufve Nyholm, Peder Larson, and Cristina Cozzini. Zero TE-based pseudo-CT image conversion in the head and its application in PET/MR attenuation correction and MR-guided radiation therapy planning. *Magn Reson Med*, 80(4):1440–1451, 2018.

[34] Jaewon Yang, Florian Wiesinger, Sandeep Kaushik, Dattesh Shanbhag, Thomas A. Hope, Peder E. Z. Larson, and Youngho Seo. Evaluation of Sinus/Edge-Corrected Zero-Echo-Time–Based Attenuation Correction in Brain PET/MRI. *J Nucl Med*, 58(11):1873–1879, January 2017.

[35] M. Weiger and K. P. Pruessmann. MRI with Zero Echo Time. In *eMagRes*. American Cancer Society, 2012.

[36] Romain Froidevaux, Markus Weiger, David O. Brunner, Benjamin E. Dietrich, Bertram J. Wilm, and Klaas P. Pruessmann. Filling the dead-time gap in zero echo time MRI: Principles compared. *Magn Reson Med*, 79(4):2036–2045, 2018.

[37] Florian Wiesinger, Laura I. Sacolick, Anne Menini, Sandeep S. Kaushik, Sangtae Ahn, Patrick Veit-Haibach, Gaspar Delso, and Dattesh D. Shanbhag. Zero TE MR bone imaging in the head. *Magn Reson Med*, 75(1):107–114, 2016.

[38] Simone Mastrogiacomo, Weiqiang Dou, John A. Jansen, and X. Frank Walboomers. Magnetic Resonance Imaging of Hard Tissues and Hard Tissue Engineered Bio-substitutes. *Mol Imaging Biol*, April 2019.

[39] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3–4:100004, September 2019.

[40] J. Ker, L. Wang, J. Rao, and T. Lim. Deep Learning Applications in Medical Image Analysis. *IEEE Access*, 6:9375–9389, 2018.

[41] Philippe Meyer, Vincent Noblet, Christophe Mazzara, and Alex Lallement. Survey on deep learning for radiotherapy. *Comput Biol Med*, 98:126–146, 2018.

[42] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J Digit Imaging*, 32(4):582–596, August 2019.

[43] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[44] Chenyang Shen, Dan Nguyen, Zhiguo Zhou, Steve B. Jiang, Bin Dong, and Xun Jia. An introduction to deep learning in medical physics: Advantages, potential,

and challenges. *Phys Med Biol*, 65(5):05TR01, March 2020.

[45] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Med Image Anal*, 42:60–88, December 2017.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241. Springer International Publishing, 2015.

[47] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Trans Med Imaging*, 35(5):1160–1169, May 2016.

[48] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, October 2016.

[49] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, pages 240–248. Springer International Publishing, 2017.

[50] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*, 11(2):178–189, 2004.

[51] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging*, 25(11):1451–1461, 2006.

[52] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, March 2019.

[53] Emilia Persson, Christian Gustafsson, Fredrik Nordström, Maja Sohlin, Adalsteinn Gunnlaugsson, Karin Petruson, Niina Rintelä, Kristoffer Hed, Lennart Blomqvist, Björn Zackrisson, et al. MR-OPERA–A Multi-center/multi-vendor validation of MRI-only prostate treatment planning using synthetic CT images. *Int J Radiat Oncol Biol Phys*, 99(3):692–700, 2017.

[54] Jonathan Wyatt and Hazel McCallum. Applying a commercial atlas-based synthetic Computed Tomography algorithm to patients with hip prostheses for prostate Magnetic Resonance-only radiotherapy. *Radiother Oncol*, 133:100–105, April 2019.

[55] Carl Siversson, Fredrik Nordström, Terese Nilsson, Tufve Nyholm, Joakim Jonsson, Adalsteinn Gunnlaugsson, and Lars E Olsson. Technical Note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm. *Med Phys*, 42(10):6090–6097, 2015.

[56] M Maspero, PR Seevinck, G Schubert, MA Hoesl, B van Asselen, MA Viergever, JJW Lagendijk, GJ Meijer, and CAT van den Berg. Quantification of confounding factors in MRI-based dose calculations as applied to prostate IMRT. *Phys Med Biol*, 62(3), 2017.

[57] N Papanikolaou, J Battista, A Boyer, C Kappas, E Klein, T Mackie, M Sharpe, and J Van Dyk. Tissue inhomogeneity corrections for megavoltage photon beams. AAPM Report No. 85. *AAPM Reports*, 2004.

[58] Daniel A. Low, William B. Harms, Sasa Mutic, and James A. Purdy. A technique for the quantitative evaluation of dose distributions. *Med Phys*, 25(5):656–661, 1998.

[59] Jonathan J Wyatt, Jason A Dowling, Charles G Kelly, Jill McKenna, Emily Johnstone, Richard Speight, Ann Henry, Peter B Greer, and Hazel M McCallum. Investigating the generalisation of an atlas-based synthetic-CT algorithm to another centre and MR scanner for prostate MR-only radiotherapy. *Phys Med Biol*, 62(24):N548, 2017.

[60] Jonathan Lambert, Peter B Greer, Fred Menk, Jackie Patterson, Joel Parker, Kara Dahl, Sanjiv Gupta, Anne Capp, Chris Wratten, Colin Tang, et al. MRI-guided prostate radiation therapy planning: Investigation of dosimetric accuracy of MRI-based dose planning. *Radiother Oncol*, 98(3):330–334, 2011.

[61] Stan J. Hoogcarspel, Joanne M. Van der Velden, Jan J. W. Lagendijk, Marco van Vulpen, and Bas W. Raaymakers. The feasibility of utilizing pseudo CT-data for online MRI based treatment plan adaptation for a stereotactic radiotherapy treatment of spinal bone metastases. *Phys Med Biol*, 59(23):7383–7391, November 2014.

[62] M Köhler, T Vaara, MV Grootel, R Hoogeveen, R Kemppainen, and S Renisch. MR-only simulation for radiotherapy planning. *Philips White Paper*, 2015.

[63] Siemens. MR-only RT planning for the brain and pelvis with Synthetic CT. *Siemens White Paper*, 2019.

[64] W T Dixon. Simple proton spectroscopic imaging. *Radiology*, 153(1):189–194, October 1984.

[65] Holger Eggers, Bernhard Brendel, Adri Duijndam, and Gwenael Herigault. Dual-echo Dixon imaging with flexible choice of echo times. *Magn Reson Med*, 65(1):96–107, 2011.

[66] Daniela Thorwarth, Corinna Warschburger, David Monnich, Ulrich Grosse, Matthias Kundel, Konstantin Nikolaou, Daniel Zips, Daniel Wegener, and Arndt-

Christan Müller. Synthetic CT generation for the pelvic region based on DIXON-MR sequences: Workflow, dosimetric quality and daily patient positioning. *MReadings: MR in RT*, 2019.

[67] Neelam Tyagi, Sandra Fontenla, Jing Zhang, Michelle Cloutier, Mo Kadbi, Jim Mechalakos, Michael Zelefsky, Joe Deasy, and Margie Hunt. Dosimetric and workflow evaluation of first commercial synthetic CT software for clinical use in pelvis. *Phys Med Biol*, 62(8):2961, 2017.

[68] Shupeng Chen, Hong Quan, An Qin, Seonghwan Yee, and Di Yan. MR image-based synthetic CT for IMRT prostate treatment planning and CBCT image-guided localization. *J Appl Clin Med Phys*, 17(3):236–245, 2016.

[69] Jason A Dowling, Jidi Sun, Peter Pichler, David Rivest-Hénault, Soumya Ghose, Haylea Richardson, Chris Wratten, Jarad Martin, Jameen Arm, Leah Best, et al. Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences. *Int J Radiat Oncol Biol Phys*, 93(5):1144–1153, 2015.

[70] Joshua Kim, Kim Garbarino, Lonni Schultz, Kenneth Levin, Benjamin Movsas, M Salim Siddiqui, Indrin J Chetty, and Carri Glide-Hurst. Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy. *Radiat Oncol*, 10(1):1, 2015.

[71] Mikko Tenhunen, Juha Korhonen, Mika Kapanen, Tiina Seppälä, Lauri Koivula, Juhani Collan, Kauko Saarilahti, and Harri Visapää. MRI-only based radiation therapy of prostate cancer: Workflow and early clinical experience. *Acta Oncol*, pages 1–6, 2018.

[72] Juha Korhonen, Mika Kapanen, Jani Keyriläinen, Tiina Seppälä, and Mikko Tenhunen. A dual model HU conversion from MRI intensity values within and outside of bone segment for MRI-based radiotherapy treatment planning of prostate cancer. *Med Phys*, 41(1):011704, 2014.

[73] Mika Kapanen and Mikko Tenhunen. T1/T2*-weighted MRI provides clinically relevant pseudo-CT density data for the pelvic bones in MRI-only based radiotherapy treatment planning. *Acta Oncol*, 52(3):612–618, 2013.

[74] Adam Johansson, Mikael Karlsson, and Tufve Nyholm. CT substitute derived from MRI sequences with ultrashort echo time. *Med Phys*, 38(5):2708–2714, May 2011.

[75] Lei Xiang, Qian Wang, Dong Nie, Lichi Zhang, Xiyao Jin, Yu Qiao, and Dinggang Shen. Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. *Med Image Anal*, 47:31–44, July 2018.

[76] Jie Fu, Yingli Yang, Kamal Singhrao, Dan Ruan, Fang-I. Chu, Daniel A. Low, and John H. Lewis. Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. *Med Phys*, 46(9):3788–3798, 2019.

[77] Andrew P. Leynes, Jaewon Yang, Florian Wiesinger, Sandeep S. Kaushik, Dat-

tesh D. Shanbhag, Youngho Seo, Thomas A. Hope, and Peder E. Z. Larson. Zero-Echo-Time and Dixon Deep Pseudo-CT (ZeDD CT): Direct Generation of Pseudo-CT Images for Pelvic PET/MRI Attenuation Correction Using Deep Convolutional Neural Networks with Multiparametric MRI. *J Nucl Med*, 59(5):852–858, January 2018.

[78] Matteo Maspero, Mark H. F. Savenije, Anna M. Dinkla, Peter R. Seevinck, Martijn P. W. Intven, Ina M. Jurgenliemk-Schulz, Linda G. W. Kerkmeijer, and Cornelis A. T. van den Berg. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol*, 63(18):185001, September 2018.

[79] David Bird, Michael G. Nix, Hazel McCallum, Mark Teo, Alexandra Gilbert, Nathalie Casanova, Rachel Cooper, David L. Buckley, David Sebag-Montefiore, Richard Speight, Bashar Al-Qaisieh, and Ann M. Henry. Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning. *Radiother Oncol*, 156:23–28, March 2021.

[80] Gyu Sang Yoo, Huan Minh Luu, Heejung Kim, Won Park, Hongryull Pyo, Youngyih Han, Ju Young Park, and Sung-Hong Park. Feasibility of Synthetic Computed Tomography Images Generated from Magnetic Resonance Imaging Scans Using Various Deep Learning Methods in the Planning of Radiation Therapy for Prostate Cancer. *Cancers*, 14(1):40, January 2022.

[81] David Bird, Ann M. Henry, David Sebag-Montefiore, David L. Buckley, Bashar Al-Qaisieh, and Richard Speight. A Systematic Review of the Clinical Implementation of Pelvic Magnetic Resonance Imaging–Only Planning for External Beam Radiation Therapy. *Int J Radiat Oncol Biol Phys*, 105(3):479–492, November 2019.

[82] S Webb. The physical basis of IMRT and inverse planning. *Br J Radiol*, 76(910):678–689, October 2003.

[83] Carlos E. Cardenas, Jinzhong Yang, Brian M. Anderson, Laurence E. Court, and Kristy B. Brock. Advances in Auto-Segmentation. *Semin Radiat Oncol*, 29(3):185–197, July 2019.

[84] Gregory Sharp, Karl D. Fritscher, Vladimir Pekar, Marta Peroni, Nadya Shusharina, Harini Veeraraghavan, and Jinzhong Yang. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys*, 41(5):050902, May 2014.

[85] Sharif Elguindi, Michael J. Zelefsky, Jue Jiang, Harini Veeraraghavan, Joseph O. Deasy, Margie A. Hunt, and Neelam Tyagi. Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy. *Phys Imaging Radiat Oncol*, 12:80–86, October 2019.

[86] Anna Kuisma, Iiro Ranta, Jani Keyriläinen, Sami Suilamo, Pauliina Wright, Marko Pesola, Lizette Warner, Eliisa Löyttyniemi, and Heikki Minn. Validation of automated magnetic resonance image segmentation for radiation therapy planning in

prostate cancer. *Physics and Imaging in Radiation Oncology*, 13:14–20, January 2020.

[87] Charlens Alvarez, Fabio Martínez, and Eduardo Romero. A multiresolution prostate representation for automatic segmentation in magnetic resonance images. *Med Phys*, 44(4):1312–1323, 2017.

[88] David Pasquier, Thomas Lacornerie, Maximilien Vermandel, Jean Rousseau, Eric Lartigau, and Nacim Betrouni. Automatic Segmentation of Pelvic Structures From Magnetic Resonance Images for Prostate Cancer Radiotherapy. *Int J Radiat Oncol Biol Phys*, 68(2):592–600, June 2007.

[89] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation. *Med Phys*, 46(4):1707–1718, 2019.

[90] Jordan Wong, Allan Fong, Nevin McVicar, Sally Smith, Joshua Giambattista, Derek Wells, Carter Kolbeck, Jonathan Giambattista, Lovedeep Gondara, and Abraham Alexander. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol*, 144:152–158, March 2020.

[91] D. Roach, M. G. Jameson, J. A. Dowling, M. A. Ebert, P. B. Greer, A. M. Kennedy, S. Watt, and L. C. Holloway. Correlations between contouring similarity metrics and simulated treatment outcome for prostate radiotherapy. *Phys Med Biol*, 63(3):035001, 2018.

[92] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, Christopher Kelly, Alan Karthikesalingam, Carlton Chu, Dawn Carnell, Cheng Boon, Derek D'Souza, Syed Ali Moinuddin, DeepMind Radiographer Consortium, Hugh Montgomery, Geraint Rees, Mustafa Suleyman, Trevor Back, Cían Hughes, Joseph R. Ledsam, and Olaf Ronneberger. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv:1809.04430 [physics, stat]*, October 2020.

[93] Ying Song, Junjie Hu, Qiang Wu, Feng Xu, Shihong Nie, Yaqin Zhao, Sen Bai, and Zhang Yi. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy. *Radiother Oncol*, 145:186–192, April 2020.

[94] M. B. Altman, J. A. Kavanaugh, H. O. Wooten, O. L. Green, T. A. DeWees, H. Gay, W. L. Thorstad, H. Li, and S. Mutic. A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Phys Med Biol*, 60(13):5199–5209, June 2015.

[95] The Royal College of Radiologists. Radiotherapy target volume definition and peer review. *Royal College of Radiologists*, BFCO(17)(2), 2017.

[96] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol*, 13:1–6, January 2020.

[97] Minsong Cao, Bradley Stiehl, Victoria Y. Yu, Ke Sheng, Amar U. Kishan, Robert K. Chin, Yingli Yang, and Dan Ruan. Analysis of Geometric Performance and Dosimetric Impact of Using Automatic Contour Segmentation for Radiotherapy Planning. *Front Oncol*, 10:1762, 2020.

[98] Ward van Rooij, Max Dahele, Hugo Ribeiro Brandao, Alexander R. Delaney, Berend J. Slotman, and Wilko F. Verbakel. Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation. *Int J Radiat Oncol Biol Phys*, 104(3):677–684, July 2019.

[99] W. Jeffrey Zabel, Jessica L. Conway, Adam Gladwish, Julia Skliarenko, Giulio Didiodato, Leah Goorts-Matthews, Adam Michalak, Sarah Reistetter, Jenna King, Keith Nakonechny, Kyle Malkoske, Muoi N. Tran, and Nevin McVicar. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. *Pract Radiat Oncol*, 11(1):e80–e89, January 2021.

[100] Femke Vaassen, Colien Hazelaar, Richard Canters, Stephanie Peeters, Steven Petit, and Wouter van Elmpt. The impact of organ-at-risk contour variations on automatically generated treatment plans for NSCLC. *Radiother Oncol*, 163:136–142, October 2021.

[101] Finn Edler von Eyben, Timo Kiljunen, Aki Kangasmaki, Kalevi Kairemo, Rie von Eyben, and Timo Joensuu. Radiotherapy Boost for the Dominant Intraprostatic Cancer Lesion—A Systematic Review and Meta-Analysis. *Clin Genitourin Cancer*, 14(3):189–197, June 2016.

[102] R Luypaert, S Boujraf, S Sourbron, and M Osteaux. Diffusion and perfusion MRI: Basic physics. *Eur J Radiol*, 38(1):19–27, 2001.

[103] Donald W. McRobbie, Elizabeth A. Moore, Martin J. Graves, and Martin R. Prince. *MRI from Picture to Proton*. Cambridge University Press, second edition, 2006.

[104] Roland Bammer. Basic principles of diffusion-weighted imaging. *Eur J Radiol*, 45(3):169–184, 2003.

[105] Lars E. Olsson, Mikael Johansson, Björn Zackrisson, and Lennart K. Blomqvist. Basic concepts and applications of functional magnetic resonance imaging for radiotherapy of prostate cancer. *Phys Imaging Radiat Oncol*, 9:50–57, January 2019.

[106] Hugh Harvey and Nandita M deSouza. The role of imaging in the diagnosis of primary prostate cancer. *J Clin Urol*, 9(2_suppl):11–17, 2016.

[107] Cher Heng Tan, Wei Wei, Valen Johnson, and Vikas Kundra. Diffusion-Weighted MRI in the Detection of Prostate Cancer: Meta-Analysis. *Am J Roentgenol*, 199(4):822–829, October 2012.

[108] Andrea Delli Pizzi, Raffaella Basilico, Roberta Cianci, Barbara Seccia, Mauro Timpani, Alessandra Tavoletta, Daniele Caposiena, Barbara Faricelli, Daniela Gabrielli, and Massimo Caulo. Rectal cancer MRI: Protocols, signs and future perspectives radiologists should consider in everyday clinical practice. *Insights Imaging*, 9(4):405–412, August 2018.

[109] Christina Tsien, Yue Cao, and Thomas Chenevert. Clinical applications for diffusion magnetic resonance imaging in radiotherapy. *Semin Radiat Oncol*, 24(3):218–226, 2014.

[110] E. J. Alexander, J. R. Murray, V. A. Morgan, S. L. Giles, S. F. Riches, S. Hazell, K. Thomas, S. A. Sohaib, A. Thompson, A. Gao, D. P. Dearnaley, and N. M. DeSouza. Validation of T2- and diffusion-weighted magnetic resonance imaging for mapping intra-prostatic tumour prior to focal boost dose-escalation using intensity-modulated radiotherapy (IMRT). *Radiother Oncol*, 141:181–187, December 2019.

[111] Evelyn M. Monninkhof, Juliette W. L. van Loon, Marco van Vulpen, Linda G. W. Kerkmeijer, Floris J. Pos, Karin Haustermans, Laura van den Bergh, Sofie Isebaert, Gill M. McColl, Robert Jan Smeenk, Juus Noteboom, Iris Walraven, Petra H. M. Peeters, and Uulke A. van der Heide. Standard whole prostate gland radiotherapy with and without lesion boost in prostate cancer: Toxicity in the FLAME randomized controlled trial. *Radiother Oncol*, 127(1):74–80, April 2018.

[112] T. Metens, D. Miranda, J. Absil, and C. Matos. What is the optimal b value in diffusion-weighted MR imaging to depict prostate cancer at 3T? *Eur Radiol*, 22(3):703–709, March 2012.

[113] Andrew B. Rosenkrantz, Nainesh Parikh, Andrea S. Kierans, Max Xiangtian Kong, James S. Babb, Samir S. Taneja, and Justin M. Ream. Prostate Cancer Detection Using Computed Very High b-value Diffusion-weighted Imaging: How High Should We Go? *Acad Radiol*, 23(6):704–711, June 2016.

[114] Yingqiu Song, Beth Erickson, Xiaojian Chen, Guiling Li, Gang Wu, Eric Paulson, Paul Knechtges, and Li X. Allen. Appropriate magnetic resonance imaging techniques for gross tumor volume delineation in external beam radiation therapy of locally advanced cervical cancer. *Oncotarget*, 9(11):10100–10109, January 2018.

[115] Antoine Schernberg, Corinne Balleyguier, Isabelle Dumas, Sébastien Gouy, Alexandre Escande, Enrica Bentivegna, Philippe Morice, Eric Deutsch, Christine Haie-Meder, and Cyrus Chargari. Diffusion-weighted MRI in image-guided adaptive brachytherapy: Tumor delineation feasibility study and comparison with GEC-ESTRO guidelines. *Brachytherapy*, 16(5):956–963, September 2017.

[116] Francisco Donato, Daniel N. Costa, Qing Yuan, Neil M. Rofsky, Robert E. Lenkinski, and Ivan Pedrosa. Geometric Distortion in Diffusion-weighted MR Imaging of the Prostate—Contributing Factors and Strategies for Improvement. *Acad Radiol*, 21(6):817–823, June 2014.

[117] Harriet C Thoeny, Frederik De Keyzer, and Ann D King. Diffusion-weighted MR

imaging in the head and neck. *Radiology*, 263(1):19–32, 2012.

[118] Gabriel Nketiah, Kirsten M. Selnæs, Elise Sandsmark, Jose R. Teruel, Brage Krüger-Stokke, Helena Bertilsson, Tone F. Bathen, and Mattijs Elschot. Geometric distortion correction in prostate diffusion-weighted MRI and its effect on quantitative apparent diffusion coefficient analysis. *Magn Reson Med*, 79(5):2524–2532, 2018.

[119] Gary P Liney and Marinus A Moerland. Magnetic resonance imaging acquisition techniques for radiotherapy planning. *Semin Radiat Oncol*, 24:160–168, 2014.

[120] Christopher Nguyen, Ali-Reza Sharif-Afshar, Zhaoyang Fan, Yibin Xie, Sidney Wilson, Xiaoming Bi, Lucas Payor, Rola Saouaf, Hyung Kim, and Debiao Li. 3D high-resolution diffusion-weighted MRI at 3T: Preliminary application in prostate cancer patients undergoing active surveillance protocol for low-risk prostate cancer. *Magn Reson Med*, 75(2):616–626, 2016.

[121] Emine Ulku Saritas, Charles H. Cunningham, Jin Hyung Lee, Eric T. Han, and Dwight G. Nishimura. DWI of the spinal cord with reduced FOV single-shot EPI. *Magn Reson Med*, 60(2):468–473, 2008.

[122] Brent A. Warndahl, Eric A. Borisch, Akira Kawashima, Stephen J. Riederer, and Adam T. Froemming. Conventional vs. reduced field of view diffusion weighted imaging of the prostate: Comparison of image quality, correlation with histology, and inter-reader agreement. *Magn Reson Imaging*, 47:67–76, April 2018.

[123] Cornelia Brendle, Petros Martirosian, Nina F. Schwenzer, Sascha Kaufmann, Stephan Kruck, Ulrich Kramer, Mike Notohamiprodjo, Konstantin Nikolaou, and Christina Schraml. Diffusion-weighted imaging in the assessment of prostate cancer: Comparison of zoomed imaging and conventional technique. *Eur J Radiol*, 85(5):893–900, May 2016.

[124] Kolja M. Thierfelder, Michael K. Scherr, Mike Notohamiprodjo, Jakob Weiß, Olaf Dietrich, Ullrich G. Mueller-Lisse, Josef Pfeuffer, Konstantin Nikolaou, and Daniel Theisen. Diffusion-weighted MRI of the Prostate: Advantages of Zoomed EPI with Parallel-transmit-accelerated 2D-selective Excitation Imaging. *Eur Radiol*, 24(12):3233–3241, December 2014.

[125] Wenchuan Wu and Karla L Miller. Image formation in diffusion MRI: A review of recent technical developments. *J Magn Reson Imaging*, 46(3):646–662, 2017.

[126] Nan-kuei Chen, Arnaud Guidon, Hing-Chiu Chang, and Allen W. Song. A robust multi-shot scan strategy for high-resolution diffusion weighted MRI enabled by multiplexed sensitivity-encoding (MUSE). *NeuroImage*, 72:41–47, May 2013.

[127] Angela Tong, Gregory Lemberskiy, Chenchan Huang, Krishna Shanbhogue, Thorsten Feiweier, and Andrew B. Rosenkrantz. Exploratory study of geometric distortion correction of prostate diffusion-weighted imaging using B0 map acquisition. *J Magn Reson Imaging*, 50(5):1614–1619, 2019.

[128] Hossein Jadvar. FDG PET in Prostate Cancer. *PET Clinics*, 4(2):155–161, April 2009.

[129] Sahar Mirpour, Joyce C. Mhlanga, Prashanti Logeswaran, Gregory Russo, Gustavo Mercier, and Rathan M. Subramaniam. The Role of PET/CT in the Management of Cervical Cancer. *Am J Roentgenol*, 201(2):W192–W205, July 2013.

[130] Alba Fiorentino, Riccardo Laudicella, Elisa Ciurlia, Salvatore Annunziata, Valentina Lancellotta, Paola Mapelli, Carmelo Tuscano, Federico Caobelli, Laura Evangelista, Lorenza Marino, Natale Quartuccio, Michele Fiore, Paolo Borghetti, Agostino Chiaravalloti, Maria Ricci, Isacco Desideri, and Pierpaolo Alongi. Positron emission tomography with computed tomography imaging (PET/CT) for the radiotherapy planning definition of the biological target volume: PART 2. *Crit Rev Oncol Hematol*, 139:117–124, July 2019.

[131] Aamer Mahmud, Raymond Poon, and Derek Jonker. PET imaging in anal canal cancer: A systematic review and meta-analysis. *Br J Radiol*, 90(1080):20170370, October 2017.

[132] Miriam K. Rutegård, Malin Båtsman, Jan Axelsson, Patrik Brynolfsson, Fredrik Brännström, Jörgen Rutegård, Ingrid Ljuslinder, Lennart Blomqvist, Richard Palmqvist, Martin Rutegård, and Katrine Riklund. PET/MRI and PET/CT hybrid imaging of rectal cancer - description and initial observations from the RECTOPET (REctal Cancer trial on PET/MRI/CT) study. *Cancer Imaging*, 19(1):52, July 2019.

[133] Jacqueline Esthappan, Sasa Mutic, Robert S Malyapa, Perry W Grigsby, Imran Zoberi, Farrokh Dehdashti, Tom R Miller, Walter R Bosch, and Daniel A Low. Treatment planning guidelines regarding the use of CT/PET-guided IMRT for cervical carcinoma with positive paraaortic lymph nodes. *Int J Radiat Oncol Biol Phys*, 58(4):1289–1297, March 2004.

[134] Lilie L. Lin, Sasa Mutic, Daniel A. Low, Richard LaForest, Milos Vicic, Imran Zoberi, Tom R. Miller, and Perry W. Grigsby. Adaptive brachytherapy treatment planning for cervical cancer using FDG-PET. *Int J Radiat Oncol Biol Phys*, 67(1):91–96, January 2007.

[135] Ying Zhang, Jing Hu, Jianping Li, Ning Wang, Weiwei Li, Yongchun Zhou, Junyue Liu, Lichun Wei, Mei Shi, Shengjun Wang, Jing Wang, Xia Li, and Wanling Ma. Comparison of imaging-based gross tumor volume and pathological volume determined by whole-mount serial sections in primary cervical cancer. *OncoTargets Ther*, July 2013.

[136] Marco Krengli, Barbara Cannillo, Lucia Turri, Paolo Bagnasacco, Laura Berretta, Teresa Ferrara, Mario Galliano, Sergio Gribaudo, Antonella Melano, Fernando Munoz, Piera Sciacero, Vassiliki Tseroni, Maria Chiara Bassi, Marco Brambilla, and Eugenio Inglese. Target Volume Delineation for Preoperative Radiotherapy of Rectal Cancer: Inter-Observer Variability and Potential Impact of FDG-PET/CT Imaging. *Technol Cancer Res Treat*, 9(4):393–398, August 2010.

[137] Giovanni Lucignani. SUV and segmentation: Pressing challenges in tumour assessment and treatment. *Eur J Nucl Med Mol Imaging*, 36(4):715–720, April 2009.

[138] Ellen Day, James Betler, David Parda, Bodo Reitz, Alexander Kirichenko, Seyed Mohammadi, and Moyed Miften. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. *Med Phys*, 36(10):4349–4358, October 2009.

[139] Letizia Deantonio, Maria Elisa Milia, Tiziana Cena, Gianmauro Sacchetti, Carola Perotti, Marco Brambilla, Lucia Turri, and Marco Krengli. Anal cancer FDG-PET standard uptake value: Correlation with tumor characteristics, treatment response and survival. *Radiol Med*, 121(1):54–59, January 2016.

[140] Jose G. Bazan, Albert C. Koong, Daniel S. Kapp, Andrew Quon, Edward E. Graves, Billy W. Loo, and Daniel T. Chang. Metabolic Tumor Volume Predicts Disease Progression and Survival in Patients with Squamous Cell Carcinoma of the Anal Canal. *J Nucl Med*, 54(1):27–32, January 2013.

[141] Shimpei Ogawa, Michio Itabashi, Chisato Kondo, Mitsuru Momose, Shuji Sakai, and Shingo Kameoka. Prognostic Value of Total Lesion Glycolysis Measured by 18F-FDG-PET/CT in Patients with Colorectal Cancer. *Anticancer Res*, 35(6):3495–3500, June 2015.

[142] Ephraim E. Parent and David M. Schuster. Update on 18F-Fluciclovine PET for Prostate Cancer Imaging. *J Nucl Med*, 59(5):733–739, January 2018.

[143] Sandi A. Kwee, Gregory P. Thibault, Richard S. Stack, Marc N. Coel, Bungo Furusato, and Isabell A. Sesterhenn. Use of Step-Section Histopathology to Evaluate 18F-Fluorocholine PET Sextant Localization of Prostate Cancer. *Mol Imaging*, 7(1):7290.2008.00002, January 2008.

[144] Jeremie Calais, Minsong Cao, and Nicholas G. Nickols. The Utility of PET/CT in the Planning of External Radiation Therapy for Prostate Cancer. *J Nucl Med*, 59(4):557–567, January 2018.

[145] Michael Pinkawa, Richard Holy, Marc D. Piroth, Jens Klotz, Sandra Nussen, Thomas Krohn, Felix M. Mottaghy, Martin Weibrecht, and Michael J. Eble. Intensity-Modulated Radiotherapy for Prostate Cancer Implementing Molecular Imaging with 18F-Choline PET-CT to Define a Simultaneous Integrated Boost. *Strahlenther Onkol*, 186(11):600–606, November 2010.

[146] Michael Pinkawa, Marc D. Piroth, Richard Holy, Jens Klotz, Victoria Djukic, Nuria Escobar Corral, Mariana Caffaro, Oliver H. Winz, Thomas Krohn, Felix M. Mottaghy, and Michael J. Eble. Dose-escalation using intensity-modulated radiotherapy for prostate cancer - evaluation of quality of life with and without 18F-choline PET-CT detected simultaneous integrated boost. *Radiat Oncol*, 7(1):14, January 2012.

[147] Yu Kuang, Lili Wu, Emily Hirata, Kyle Miyazaki, Miles Sato, and Sandi A. Kwee. Volumetric Modulated Arc Therapy Planning for Primary Prostate Cancer With Selective Intraprostatic Boost Determined by 18F-Choline PET/CT. *Int J Radiat Oncol Biol Phys*, 91(5):1017–1025, April 2015.

[148] Yong-il Kim, Gi Jeong Cheon, Jin Chul Paeng, Jeong Yeon Cho, Cheol Kwak,

Keon Wook Kang, June-Key Chung, Euishin Edmund Kim, and Dong Soo Lee. Usefulness of MRI-assisted metabolic volumetric parameters provided by simultaneous 18F-fluorocholine PET/MRI for primary prostate cancer characterization. *Eur J Nucl Med Mol Imaging*, 42(8):1247–1256, July 2015.

[149] Isabel Syndikus, Joachim Kwok-Chiu Chan, Thelma Rowntree, Laura Howard, and John Staffurth. Hypofractionated dose painting IMRT using 20 fractions for intermediate to high-risk localized prostate cancer: Two-year outcome data (BIOPROP20, NCT02125175). *J Clin Oncol*, 37(7_suppl):59–59, February 2019.

[150] Constantinos Zamboglou, Gesche Wieser, Steffen Hennies, Irene Rempel, Simon Kirste, Martin Soschynski, Hans Christian Rischke, Tobias Fechter, Cordula A. Jilg, Mathias Langer, Philipp T. Meyer, Michael Bock, and Anca-Ligia Grosu. MRI versus 68Ga-PSMA PET/CT for gross tumour volume delineation in radiation treatment planning of primary prostate cancer. *Eur J Nucl Med Mol Imaging*, 43(5):889–897, May 2016.

[151] Constantinos Zamboglou, Vanessa Drendel, Cordula A. Jilg, Hans C. Rischke, Teresa I. Beck, Wolfgang Schultze-Seemann, Tobias Krauss, Michael Mix, Florian Schiller, Ulrich Wetterauer, Martin Werner, Mathias Langer, Michael Bock, Philipp T. Meyer, and Anca L. Grosu. Comparison of $^{68}$ Ga-HBED-CC PSMA-PET/CT and multiparametric MRI for gross tumour volume detection in patients with primary prostate cancer based on slice by slice comparison with histopathology. *Theranostics*, 7(1):228–237, 2017.

[152] Constantinos Zamboglou, Matthias Eiber, Thomas R. Fassbender, Matthias Eder, Simon Kirste, Michael Bock, Oliver Schilling, Kathrin Reichel, Uulke A. van der Heide, and Anca L. Grosu. Multimodal imaging for radiation therapy planning in patients with primary prostate cancer. *Phys Imaging in Radiat Oncol*, 8:8–16, October 2018.

[153] P Brandmaier, S Purz, K Bremicker, M Höckel, H Barthel, R Kluge, T Kahn, O Sabri, and P Stumpp. Simultaneous [18F] FDG-PET/MRI: Correlation of apparent diffusion coefficient (ADC) and standardized uptake value (SUV) in primary and recurrent cervical cancer. *PLoS One*, 10(11):e0141684, 2015.

[154] Eric S Paulson, Beth Erickson, Chris Schultz, and X Allen Li. Comprehensive MRI simulation methodology using a dedicated MRI scanner in radiation oncology for external beam radiation treatment planning. *Med Phys*, 42(1):28–39, 2015.

[155] Daniel H Paulus, Daniela Thorwath, Holger Schmidt, and Harald H Quick. Towards integration of PET/MR hybrid imaging into radiation therapy treatment planning. *Med Phys*, 41(7):072505, 2014.

[156] Daniel H Paulus, Mark Oehmigen, Johannes Grueneisen, Lale Umutlu, and Harald H Quick. Whole-body hybrid imaging concept for the integration of PET/MR into radiation therapy treatment planning. *Phys Med Biol*, 61(9):3504, 2016.

[157] Maria A Schmidt and Geoffrey S Payne. Radiotherapy planning using MRI. *Phys*

*Med Biol*, 60(22):R323, 2015.

[158] Bernhard Gruber, Martijn Froeling, Tim Leiner, and Dennis WJ Klomp. RF coils: A practical guide for nonphysicists. *J Magn Reson Imaging*, 48(3):590–604, 2018.

[159] M McJury, A O'Neill, M Lawson, C McGrath, A Grey, W Page, and JM O'Sullivan. Assessing the image quality of pelvic MR images acquired with a flat couch for radiotherapy treatment planning. *Br J Radiol*, 84(1004):750–755, 2011.

[160] Margaret E. Daube-Witherspoon, Joel S. Karp, Michael E. Casey, Frank P. DiFilippo, Horace Hines, Gerd Muehllehner, Vilim Simcic, Charles W. Stearns, Lars-Eric Adam, Steve Kohlmyer, and Vesna Sossi. PET Performance Measurements Using the NEMA NU 2-2001 Standard. *J Nucl Med*, 43(10):1398–1409, January 2002.

[161] René M Winter, Sara Leibfarth, Holger Schmidt, Kerstin Zwirner, David Mönnich, Stefan Welz, Nina F Schwenzer, Christian la Fougère, Konstantin Nikolaou, Sergios Gatidis, et al. Assessment of image quality of a radiotherapy-specific hardware solution for PET/MRI in head and neck cancer patients. *Radiother Oncol*, 128(3):485–491, 2018.

[162] Patrik Brynolfsson, Jan Axelsson, August Holmberg, Joakim H. Jonsson, David Goldhaber, Yiqiang Jian, Fredrik Illerstam, Mathias Engström, Björn Zackrisson, and Tufve Nyholm. Technical Note: Adapting a GE SIGNA PET/MR scanner for radiotherapy. *Med Phys*, 45(8):3546–3550, 2018.

[163] Stephan Witoszynskyj, Piotr Andrzejewski, Dietmar Georg, Marcus Hacker, Tufve Nyholm, Ivo Rausch, and Barbara Knäusl. Attenuation correction of a flat table top for radiation therapy in hybrid PET/MR using CT- and 68Ge/68Ga transmission scan-based $\mu$-maps. *Phys Med*, 65:76–83, September 2019.

[164] Alejandra Valladares, Sahar Ahangari, Thomas Beyer, Ronald Boellaard, Zacharias Chalampalakis, Claude Comtat, Laura DalToso, Adam E. Hansen, Michel Koole, Jane Mackewn, Paul Marsden, Johan Nuyts, Francesco Padormo, Ronald Peeters, Sebastian Poth, Esteban Solari, and Ivo Rausch. Clinically Valuable Quality Control for PET/MRI Systems: Consensus Recommendation From the HYBRID Consortium. *Front Phys*, 7, 2019.

[165] American College of Radiology. Magnetic resonance imaging quality control manual. Technical report, 2015.

[166] Mika Kapanen, Juhani Collan, Annette Beule, Tiina Seppälä, Kauko Saarilahti, and Mikko Tenhunen. Commissioning of MRI-only based treatment planning procedure for external beam radiotherapy of prostate. *Magn Reson Med*, 70(1):127–135, 2013.

[167] Jonathan Wyatt, Stephen Hedley, Emily Johnstone, Richard Speight, Charles Kelly, Ann Henry, Susan Short, Louise Murray, David Sebag-Montefiore, and Hazel McCallum. Evaluating the repeatability and set-up sensitivity of a large field of view distortion phantom and software for magnetic resonance-only radiotherapy. *Phys Imaging Radiat Oncol*, 6:31–38, 2018.

[168] Iiro Ranta, Reko Kemppainen, Jani Keyriläinen, Sami Suilamo, Samuli Heikkinen,

Mika Kapanen, and Jani Saunavaara. Quality assurance measurements of geometric accuracy for magnetic resonance imaging-based radiotherapy treatment planning. *Phys Med*, 62:47–52, June 2019.

[169] Jessica M. Winfield, David J. Collins, Andrew N. Priest, Rebecca A. Quest, Alan Glover, Sally Hunter, Veronica A. Morgan, Susan Freeman, Andrea Rockall, and Nandita M. deSouza. A framework for optimization of diffusion-weighted MRI protocols for large field-of-view abdominal-pelvic imaging in multicenter studies. *Med Phys*, 43(1):95–110, 2016.

[170] Ioannis Lavdas, Kevin Behan, Annie Papadaki, Donald W. McRobbie, and Eric Aboagye. A phantom for diffusion-weighted MRI (DW-MRI). *J Magn Reson Imaging*, 38(1):173–179, October 2013.

[171] Kathryn E Keenan, Maureen Ainslie, Alex J Barker, Michael A Boss, Kim M Cecil, Cecil Charles, Thomas L Chenevert, Larry Clarke, Jeffrey L Evelhoch, Paul Finn, et al. Quantitative magnetic resonance imaging phantoms: A review and the need for a system phantom. *Magn Reson Med*, 79(1):48–61, 2018.

[172] Dariya Malyarenko, Craig J Galbán, Frank J Londy, Charles R Meyer, Timothy D Johnson, Alnawaz Rehemtulla, Brian D Ross, and Thomas L Chenevert. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *J Magn Reson Imaging*, 37(5):1238–1246, 2013.

[173] Thomas L Chenevert, Craig J Galbán, Marko K. Ivancevic, Susan E. Rohrer, Frank J Londy, Thomas C. Kwee, Charles R Meyer, Timothy D Johnson, and Alnawaz Rehemtulla. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. *J Magn Reson Imaging*, 34(4):938–987, 2011.

[174] Sergios Gatidis, Holger Schmidt, Petros Martirosian, and Nina F. Schwenzer. Development of an MRI phantom for diffusion-weighted imaging with independent adjustment of apparent diffusion coefficient values and T2 relaxation times. *Magn Reson Med*, 72(2):459–463, 2014.

[175] Wen Lu, Hou Jing, Zhou Ju-Mei, Nie Shao-Lin, Cao Fang, Yu Xiao-Ping, Lu Qiang, Zeng Biao, Zhu Su-Yu, and Hu Ying. Intravoxel incoherent motion diffusion-weighted imaging for discriminating the pathological response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. *Sci Rep*, 7(1):1–9, August 2017.

[176] Damien J. McHugh, Feng-Lei Zhou, Ian Wimpenny, Gowsihan Poologasundarampillai, Josephine H. Naish, Penny L. Hubbard Cristinacce, and Geoffrey J. M. Parker. A biomimetic tumor tissue phantom for validating diffusion-weighted MRI measurements. *Magn Reson Med*, 80(1):147–158, 2018.

[177] Brent Foster, Ulas Bagci, Awais Mansoor, Ziyue Xu, and Daniel J Mollura. A review on segmentation of positron emission tomography images. *Comput Biol Med*, 50:76–96, 2014.

[178] Lei Xing. Quality Assurance of Positron Emission Tomography/Computed Tomography for Radiation Therapy. *Int J Radiat Oncol Biol Phys*, 71(1, Supplement):S38–S42, May 2008.

[179] Paul E. Kinahan and James W. Fletcher. Positron Emission Tomography-Computed Tomography Standardized Uptake Values in Clinical Practice and Assessing Response to Therapy. *Semin Ultrasound CT MR*, 31(6):496–505, December 2010.

[180] Francesca Gallivanone, Irene Carne, Matteo Interlenghi, Daniela D'Ambrosio, Maurizia Baldi, Daniele Fantinato, and Isabella Castiglioni. A Method for Manufacturing Oncological Phantoms for the Quantification of 18F-FDG PET and DW-MRI Studies. *Contrast Media Mol Imaging*, 2017, 2017.

[181] D. Thorwarth, T. Beyer, R. Boellaard, D. De Ruysscher, A. Grgic, J. A. Lee, U. Pietrzyk, B. Sattler, A. Schaefer, W. van Elmpt, W. Vogel, W. J. G. Oyen, and U. Nestle. Integration of FDG- PET/CT into external beam radiation therapy planning. *Nuklearmedizin*, 51(4):140–153, 2012.

[182] Sasa Mutic, Jatinder R. Palta, Elizabeth K. Butker, Indra J. Das, M. Saiful Huq, Leh-Nien Dick Loo, Bill J. Salter, Cynthia H. McCollough, and Jacob Van Dyk. Quality assurance for computed-tomography simulators and the computed-tomography-simulation process: Report of the AAPM Radiation Therapy Committee Task Group No. 66. *Med Phys*, 30(10):2762–2792, 2003.

[183] S Hynds, C K McGarry, D M Mitchell, S Early, L Shum, D P Stewart, J A Harney, C R Cardwell, and J M O'Sullivan. Assessing the daily consistency of bladder filling using an ultrasonic Bladderscan device in men receiving radical conformal radiotherapy for prostate cancer. *Br J Radiol*, 84(1005):813–818, September 2011.

[184] Joos V. Lebesque, Allison M. Bruce, A. P. Guus Kroes, Adriaan Touw, Tarek Shouman, and Marcel van Herk. Variation in volumes, dose-volume histograms, and estimated normal tissue complication probabilities of rectum and bladder during conformal radiotherapy of T3 prostate cancer. *Int J Radiat Oncol Biol Phys*, 33(5):1109–1119, December 1995.

[185] Heidi T. Lotz, Marcel van Herk, Anja Betgen, Floris Pos, Joos V. Lebesque, and Peter Remeijer. Reproducibility of the bladder shape and bladder shape changes during filling. *Medical Physics*, 32(8):2590–2597, 2005.

[186] Emma C. Fields and Elisabeth Weiss. A practical review of magnetic resonance imaging for the evaluation and management of cervical cancer. *Radiat Oncol*, 11(1):15, February 2016.

[187] Arne Grün, Michael Kawgan-Kagan, David Kaul, Harun Badakhshi, Carmen Stromberger, Volker Budach, and Dirk Böhmer. Impact of bladder volume on acute genitourinary toxicity in intensity modulated radiotherapy for localized and locally advanced prostate cancer. *Strahlenther Onkol*, 195(6):517–525, June 2019.

[188] Catherine A. McBain, Vincent S. Khoo, David L. Buckley, Jonathan S. Sykes, Melanie M. Green, Richard A. Cowan, Charles E. Hutchinson, Christopher J. Moore,

and Patricia M. Price. Assessment of Bladder Motion for Clinical Radiotherapy Practice Using Cine–Magnetic Resonance Imaging. *Int J Radiat Oncol Biol Phys*, 75(3):664–671, November 2009.

[189] Samuel Fransson, David Tilly, Anders Ahnesjö, Tufve Nyholm, and Robin Strand. Intrafractional motion models based on principal components in Magnetic Resonance guided prostate radiotherapy. *Phys Imaging Radiat Oncol*, 20:17–22, October 2021.

[190] Hsiao-Hsuan Chen, Kun-Sheng Lin, Pei-Tzu Lin, Liang-Tseng Kuo, Chiung-Chen Fang, and Ching-Chi Chi. Bladder volume reproducibility after water consumption in patients with prostate cancer undergoing radiotherapy: A systematic review and meta-analysis. *Biomed J*, November 2020.

[191] Michael J. Zelefsky, Diane Crean, Gig S. Mageras, Olga Lyass, Laura Happersett, C. Clifton Ling, Steven A. Leibel, Zvi Fuks, Sarah Bull, Hanne M. Kooy, Marcel van Herk, and Gerald J. Kutcher. Quantification and predictors of prostate position variability in 50 patients evaluated with multiple CT scans during conformal radiotherapy. *Radiother Oncol*, 50(2):225–234, February 1999.

[192] Helen A. McNair, Linda Wedlake, Irene M. Lips, Jervoise Andreyev, Marco Van Vulpen, and David Dearnaley. A systematic review: Effectiveness of rectal emptying preparation in prostate cancer patients. *Pract Radiat Oncol*, 4(6):437–447, November 2014.

[193] Helen A. McNair, Linda Wedlake, Gerard P. McVey, Karen Thomas, Jervoise Andreyev, and David P. Dearnaley. Can diet combined with treatment scheduling achieve consistency of rectal filling in patients receiving radiotherapy to the prostate? *Radiother Oncol*, 101(3):471–478, December 2011.

[194] S Yahya, A Zarkar, E Southgate, P Nightingale, and G Webster. Which bowel preparation is best? Comparison of a high-fibre diet leaflet, daily microenema and no preparation in prostate cancer patients treated with radical radiotherapy to assess the effect on planned target volume shifts due to rectal distension. *Br J Radiol*, 86(1031):20130457, November 2013.

[195] Jonathan J. Wyatt, Rachel A. Pearson, Christopher P. Walker, Rachel L. Brooks, Karen Pilling, and Hazel M. McCallum. Cone beam computed tomography for dose calculation quality assurance for magnetic resonance-only radiotherapy. *Phys Imaging Radiat Oncol*, 17:71–76, January 2021.

[196] Jonathan J. Wyatt, Elizabeth Howell, Maelene Lohezic, Hazel M. McCallum, and Ross J. Maxwell. Evaluating the image quality of combined positron emission tomography-magnetic resonance images acquired in the pelvic radiotherapy position. *Phys Med Biol*, 66(3):035018, January 2021.

[197] Jonathan J. Wyatt, Hazel M. McCallum, and Ross J. Maxwell. Developing quality assurance tests for simultaneous Positron Emission Tomography – Magnetic Resonance imaging for radiotherapy planning. *Phys Imaging Radiat Oncol*, 22:28–35,

April 2022.

[198] A. Gonzalez-Moya, S. Dufreneix, N. Ouyessad, C. Guillerminet, and D. Autret. Evaluation of a commercial synthetic computed tomography generation solution for magnetic resonance imaging-only radiotherapy. *J Appl Clin Med Phys*, 22(6):191–197, 2021.

[199] Liesbeth Vandewinckele, Michaël Claessens, Anna Dinkla, Charlotte Brouwer, Wouter Crijns, Dirk Verellen, and Wouter van Elmpt. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. *Radiother Oncol*, 153:55–66, December 2020.

[200] M Amin, S Edge, and F Greene. *AJCC Cancer Staging Manual.* American Joint Committee on Cancer, 8th edition edition, 2017.

[201] R. Marc Lebel. Performance characterization of a novel deep learning-based MR image reconstruction pipeline. *arXiv:2008.06559 [cs, eess]*, August 2020.

[202] Sandeep Kaushik, Mikael Bylund, Cristina Cozzini, Dattesh Shanbhag, Steven F. Petit, Jonathan J. Wyatt, Marion I. Menzel, Carolin Pirkl, Bhairav Mehta, Vikas Chauhan, Kesavadas Chandrasekharan, Joakim Jonsson, Tufve Nyholm, Florian Wiesinger, and Bjoern Menze. Region of Interest focused MRI to Synthetic CT Translation using Regression and Classification Multi-task Network. *arXiv:2203.16288 [physics]*, March 2022.

[203] Rich Caruana. Multitask Learning. *Mach Learn*, 28(1):41–75, July 1997.

[204] T. Nyholm, M. Berglund, P. Brynolfsson, and J. Jonsson. EP-1533: ICE-Studio - An Interactive visual research tool for image analysis. *Radiother Oncol*, 115:S837, April 2015.

[205] Marie E Korsholm, Line W Waring, and Jens M Edmund. A criterion for the reliable use of MRI-only radiotherapy. *Radiat Oncol*, 9(1):16, 2014.

[206] Douglas H. Brand, Alison C. Tree, Peter Ostler, Hans van der Voet, Andrew Loblaw, William Chu, Daniel Ford, Shaun Tolan, Suneil Jain, Alexander Martin, John Staffurth, Philip Camilleri, Kiran Kancherla, John Frew, Andrew Chan, Ian S. Dayes, Daniel Henderson, Stephanie Brown, Clare Cruickshank, Stephanie Burnett, Aileen Duffton, Clare Griffin, Victoria Hinder, Kirsty Morrison, Olivia Naismith, Emma Hall, Nicholas van As, D. Dodds, E. Lartigau, S. Patton, A. Thompson, M. Winkler, P. Wells, T. Lymberiou, D. Saunders, M. Vilarino-Varela, P. Vavassis, T. Tsakiridis, R. Carlson, G. Rodrigues, J. Tanguay, S. Iqbal, M. Winkler, S. Morgan, A. Mihai, A. Li, O. Din, M. Panades, R. Wade, Y. Rimmer, J. Armstrong, M. Panades, and N. Oommen. Intensity-modulated fractionated radiotherapy versus stereotactic body radiotherapy for prostate cancer (PACE-B): Acute toxicity findings from an international, randomised, open-label, phase 3, non-inferiority trial. *Lancet Oncol*, 20(11):1531–1543, November 2019.

[207] David I Pryor, Sandra L Turner, Keen Hun Tai, Colin Tang, Giuseppe Sasso, Marcus Dreosti, Henry H Woo, Lee Wilton, and Jarad M Martin. Moderate hypofractiona-

tion for prostate cancer: A user's guide. *J Med Imaging Radiat Oncol*, 62(2):232–239, 2018.

[208] The Royal College of Radiologists. National rectal cancer intensity-modulated radiotherapy (IMRT) guidance. *Royal College of Radiologists*, BFCO(21)(1):34, 2021.

[209] Thomas Zilli, Sandra Jorcano, Samuel Bral, Carmen Rubio, Anna M.E. Bruynzeel, Angelo Oliveira, Ufuk Abacioglu, Heikki Minn, Zvi Symon, and Raymond Miralbell. Once-a-week or every-other-day urethra-sparing prostate cancer stereotactic body radiotherapy, a randomized phase II trial: 18 months follow-up results. *Cancer Med*, 9(9):3097–3106, 2020.

[210] R Muirhead, RA Adams, DC Gilbert, M Harrison, R Glynne-Jones, D Sebag-Montefiore, MA Hawkins, and Vernon Hospital. National guidance for IMRT in anal cancer. *Royal College of Radiologists*, 1(3):26, 2015.

[211] Jason Fiege, Boyd McCurdy, Peter Potrebko, Heather Champion, and Andrew Cull. PARETO: A novel evolutionary optimization approach to multiobjective IMRT planning. *Med Phys*, 38(9):5217–5229, September 2011.

[212] Geert Wortel, Dave Eekhout, Emmy Lamers, René van der Bel, Karen Kiers, Terry Wiersma, Tomas Janssen, and Eugène Damen. Characterization of automatic treatment planning approaches in radiotherapy. *Phys Imaging Radiat Oncol*, 19:60–65, July 2021.

[213] Neelam Tyagi, Michael J. Zelefsky, Andreas Wibmer, Kristen Zakian, Sarah Burleson, Laura Happersett, Aleksi Halkola, Mo Kadbi, and Margie Hunt. Clinical experience and workflow challenges with magnetic resonance-only radiation therapy simulation and planning for prostate cancer. *Phys Imaging Radiat Oncol*, 16:43–49, October 2020.

[214] Amy Walker, Gary Liney, Peter Metcalfe, and Lois Holloway. MRI distortion: Considerations for MRI based radiotherapy treatment planning. *Australas Phys Eng Sci Med*, 37:103–113, 2014.

[215] Tarraf Torfeh, Rabih Hammoud, Gregory Perkins, Maeve McGarry, Souha Aouadi, Azim Celik, Ken-Pin Hwang, Joseph Stancanello, Primoz Petric, and Noora Al-Hammadi. Characterization of 3D geometric distortion of magnetic resonance imaging scanners commissioned for radiation therapy planning. *Magn Reson Imaging*, 34(5):645–653, 2016.

[216] Robin L. Stern, Robert Heaton, Martin W. Fraser, S. Murty Goddu, Thomas H. Kirby, Kwok Leung Lam, Andrea Molineu, and Timothy C. Zhu. Verification of monitor unit calculations for non-IMRT clinical radiotherapy: Report of AAPM Task Group 114. *Med Phys*, 38(1):504–530, 2011.

[217] Jens M Edmund, Daniel Andreasen, Faisal Mahmood, and Koen Van Leemput. Cone beam computed tomography guided treatment delivery and planning verification for magnetic resonance imaging only radiotherapy of the brain. *Acta Oncol*, 54(9):1496–1500, 2015.

[218] Valentina Giacometti, Raymond B King, Christina E Agnew, Denise M Irvine, Suneil Jain, Alan R Hounsell, and Conor K McGarry. An evaluation of techniques for dose calculation on cone beam computed tomography. *Br J Radiol*, 92(1096):20180383, February 2019.

[219] Emilia Palmér, Emilia Persson, Petra Ambolt, Christian Gustafsson, Adalsteinn Gunnlaugsson, and Lars E. Olsson. Cone beam CT for QA of synthetic CT in MRI only for prostate patients. *J Appl Clin Med Phys*, 19(6):44–52, September 2018.

[220] Shifeng Chen, Quynh Le, Yildirim Mutaf, Wei Lu, Elizabeth M. Nichols, Byong Yong Yi, Tish Leven, Karl L. Prado, and Warren D. D'Souza. Feasibility of CBCT-based dose with a patient-specific stepwise HU-to-density curve to determine time of replanning. *J Appl Clin Med Phys*, 18(5):64–69, 2017.

[221] David Dearnaley, Isabel Syndikus, Helen Mossop, Vincent Khoo, Alison Birtle, David Bloomfield, John Graham, Peter Kirkbride, John Logue, Zafar Malik, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. *The Lancet Oncology*, 17(8):1047–1060, 2016.

[222] Jonathan J. Wyatt, Rachel L. Brooks, Dean Ainslie, Emily Wilkins, Elizabeth Raven, Karen Pilling, Rachel A. Pearson, and Hazel M. McCallum. The accuracy of Magnetic Resonance – Cone Beam Computed Tomography soft-tissue matching for prostate radiotherapy. *Phys Imaging Radiat Oncol*, 12:49–55, October 2019.

[223] J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[224] Grace Kim, Todd Pawlicki, and Gary Luxton. SU-FF-I-32: Performance and Stability of the Varian Cone-Beam CT System. *Med Phys*, 33(6), 2006.

[225] Irina Fotina, Johannes Hopfgartner, Markus Stock, Thomas Steininger, Carola Lütgendorf-Caucig, and Dietmar Georg. Feasibility of CBCT-based dose calculation: Comparative analysis of HU adjustment techniques. *Radiother Oncol*, 104(2):249–256, August 2012.

[226] National Electrical Manufacturers Association. Standards publication NU 2-2007: Performance measurements of positron emission tomographs. *Technical Report*, 2007.

[227] Steve Ross. Q.Clear. *GE Healthcare White Paper*, 2014.

[228] John C. Patrick, R. Terry Thompson, Aaron So, John Butler, David Faul, Robert Z. Stodilka, Slav Yartsev, Frank S. Prato, and Stewart Gaede. Technical Note: Comparison of megavoltage, dual-energy, and single-energy CT-based $\mu$-maps for a four-channel breast coil in PET/MRI. *Med Phys*, 44(9):4758–4765, 2017.

[229] Jidi Sun, Michael Barnes, Jason Dowling, Fred Menk, Peter Stanwell, and Peter B Greer. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. *Australas Phys Eng Sci Med*, 38(1):39–46, 2015.

[230] A. J. McCann, A. Workman, and C. McGrath. A quick and robust method for

measurement of signal-to-noise ratio in MRI. *Phys Med Biol*, 58(11):3775–3790, May 2013.

[231] Susanne Ziegler, Bjoern W. Jakoby, Harald Braun, Daniel H. Paulus, and Harald H. Quick. NEMA image quality phantom measurements and attenuation correction in integrated PET/MR hybrid imaging. *Eur J Nucl Med Mol Imaging Phys*, 2(1):18, August 2015.

[232] Scott D. Wollenweber, Gaspar Delso, Timothy Deller, David Goldhaber, Martin Hüllner, and Patrick Veit-Haibach. Characterization of the impact to PET quantification and image quality of an anterior array surface coil for PET/MR imaging. *Magn Reson Mater Phy*, 27(2):149–159, April 2014.

[233] Alexander M. Grant, Timothy W. Deller, Mohammad Mehdi Khalighi, Sri Harsha Maramraju, Gaspar Delso, and Craig S. Levin. NEMA NU 2-2012 performance studies for the SiPM-based ToF-PET component of the GE SIGNA PET/MR system. *Med Phys*, 43(5):2334–2343, 2016.

[234] Silje Kjærnes Øen, Lars Birger Aasheim, Live Eikenes, and Anna Maria Karlberg. Image quality and detectability in Siemens Biograph PET/MRI and PET/CT systems—a phantom study. *Eur J Nucl Med Mol Imaging Phys*, 6(1):16, August 2019.

[235] Jeroen Buijsen, Jørgen van den Bogaard, Marco H. M. Janssen, Frans C. H. Bakers, Stephanie Engelsman, Michel Öllers, Regina G. H. Beets-Tan, Marius Nap, Geerard L. Beets, Philippe Lambin, and Guido Lammering. FDG-PET provides the best correlation with the tumor specimen compared to MRI and CT in rectal cancer. *Radiother Oncol*, 98(2):270–276, February 2011.

[236] Byung Wook Choi, Sungmin Kang, Sung Uk Bae, Woon Kyung Jeong, Seong Kyu Baek, Bong-Il Song, Kyoung Sook Won, and Hae Won Kim. Prognostic value of metabolic parameters on 18F-fluorodeoxyglucose positron tomography/computed tomography in classical rectal adenocarcinoma. *Sci Rep*, 11(1):12947, June 2021.

[237] Timothy G. Turkington. Attenuation correction in hybrid positron emission tomography. *Semin Nucl Med*, 30(4):255–267, October 2000.

[238] Deep A. Patel, Stephanie T. Chang, Karyn A. Goodman, Andrew Quon, Brian Thorndyke, Sanjiv S. Gambhir, Alex McMillan, Billy W. Loo, and Albert C. Koong. Impact of Integrated PET/CT on Variability of Target Volume Delineation in Rectal Cancer. *Technol Cancer Res Treat*, 6(1):31–36, February 2007.

[239] Jeroen Buijsen, Jørgen van den Bogaard, Hiska van der Weide, Stephanie Engelsman, Ruud van Stiphout, Marco Janssen, Geerard Beets, Regina Beets-Tan, Philippe Lambin, and Guido Lammering. FDG–PET–CT reduces the interobserver variability in rectal tumor delineation. *Radiother Oncol*, 102(3):371–376, March 2012.

[240] Linda M. Velasquez, Ronald Boellaard, Georgia Kollia, Wendy Hayes, Otto S. Hoekstra, Adriaan A. Lammertsma, and Susan M. Galbraith. Repeatability of 18F-FDG PET in a Multicenter Phase I Study of Patients with Advanced Gastrointestinal

Malignancies. *J Nucl Med*, 50(10):1646–1654, October 2009.

[241] Daniela Thorwarth. Radiotherapy treatment planning based on functional PET/CT imaging data. *Nucl Med Rev*, 15(C):43–47, 2012.

[242] Axel Martinez-Möller, Michael Souvatzoglou, Gaspar Delso, Ralph A. Bundschuh, Christophe Chefd'hotel, Sibylle I. Ziegler, Nassir Navab, Markus Schwaiger, and Stephan G. Nekolla. Tissue Classification as a Potential Approach for Attenuation Correction in Whole-Body PET/MRI: Evaluation with PET/CT Data. *J Nucl Med*, 50(4):520–526, April 2009.

[243] Sara Leibfarth, Urban Simoncic, David Mönnich, Stefan Welz, Holger Schmidt, Nina Schwenzer, Daniel Zips, and Daniela Thorwarth. Analysis of pairwise correlations in multi-parametric PET/MR data for biological tumor characterization and treatment individualization strategies. *Eur J Nucl Med Mol Imaging*, 43(7):1199–1208, 2016.

[244] Onofrio A. Catalano, Susanna I. Lee, Chiara Parente, Christy Cauley, Felipe S. Furtado, Robin Striar, Andrea Soricelli, Marco Salvatore, Yan Li, Lale Umutlu, Lina Garcia Cañamaque, David Groshar, Umar Mahmood, Lawrence S. Blaszkowsky, David P. Ryan, Jeffrey W. Clark, Jennifer Wo, Theodore S. Hong, Hiroko Kunitake, Liliana Bordeianou, David Berger, Rocco Ricciardi, and Bruce Rosen. Improving staging of rectal cancer in the pelvis: The role of PET/MRI. *Eur J Nucl Med Mol Imaging*, 48(4):1235–1245, April 2021.

[245] Barbara J. Amorim, Theodore S. Hong, Lawrence S. Blaszkowsky, Cristina R. Ferrone, David L. Berger, Liliana G. Bordeianou, Rocco Ricciardi, Jeffrey W. Clark, David P. Ryan, Jennifer Y. Wo, Motaz Qadan, Mark Vangel, Lale Umutlu, David Groshar, Lina G. Cañamaques, Debra A. Gervais, Umar Mahmood, Bruce R. Rosen, and Onofrio A. Catalano. Clinical impact of PET/MR in treated colorectal cancer patients. *Eur J Nucl Med Mol Imaging*, 46(11):2260–2269, October 2019.

[246] Alice M. Couwenberg, Johannes P. M. Burbach, Maaike Berbee, Miangela M. Lacle, René Arensman, Mihaela G. Raicu, Frank J. Wessels, Joanne Verdult, Jeanine Roodhart, Onne Reerink, Sieske Hoendervangers, Jeroen Buijsen, Heike I. Grabsch, Apollo Pronk, Esther C. J. Consten, Anke B. Smits, Joost T. Heikens, Ane L. Appelt, Wilhelmina M. U. van Grevenstein, Helena M. Verkooijen, and Martijn P. W. Intven. Efficacy of Dose-Escalated Chemoradiation on Complete Tumor Response in Patients with Locally Advanced Rectal Cancer (RECTAL-BOOST): A Phase 2 Randomized Controlled Trial. *Int J Radiat Oncol Biol Phys*, 108(4):1008–1018, November 2020.

[247] Mikael Karlsson, Magnus G Karlsson, Tufve Nyholm, Christopher Amies, and Björn Zackrisson. Dedicated magnetic resonance imaging in the radiotherapy clinic. *Int J Radiat Oncol Biol Phys*, 74(2):644–651, 2009.

[248] Esteban Walker and Amy S. Nowacki. Understanding Equivalence and Noninferiority Testing. *J Gen Intern Med*, 26(2):192–196, February 2011.

[249] Matteo Maspero, Marcus D. Tyyger, Rob H. N. Tijssen, Peter R. Seevinck, Martijn

P. W. Intven, and Cornelis A. T. van den Berg. Feasibility of magnetic resonance imaging-only rectum radiotherapy with a commercial synthetic computed tomography generation solution. *Phys Imaging Radiat Oncol*, 7:58–64, July 2018.

[250] Carolyn Lauzon and Brian Caffo. Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests. *Am Stat*, 63(2):147–154, May 2009.

[251] Adrianus J. de Langen, Andrew Vincent, Linda M. Velasquez, Harm van Tinteren, Ronald Boellaard, Lalitha K. Shankar, Maarten Boers, Egbert F. Smit, Sigrid Stroobants, Wolfgang A. Weber, and Otto S. Hoekstra. Repeatability of 18F-FDG Uptake Measurements in Tumors: A Metaanalysis. *J Nucl Med*, 53(5):701–708, May 2012.

[252] B Jones, P Jarvis, JA Lewis, and AF Ebbutt. Trials to assess equivalence: The importance of rigorous methods. *Br Med J*, 313(7048):36, 1996.

[253] Mehdi Shirin Shandiz, Hamid Saligheh Rad, Pardis Ghafarian, Khadijeh Yaghoubi, and Mohammad Reza Ay. Capturing Bone Signal in MRI of Pelvis, as a Large FOV Region, Using TWIST Sequence and Generating a 5-Class Attenuation Map for Prostate PET/MRI Imaging. *Mol Imaging*, 17:1536012118789314, January 2018.

[254] Tyler J. Bradshaw, Gengyan Zhao, Hyungseok Jang, Fang Liu, and Alan B. McMillan. Feasibility of Deep Learning–Based PET/MR Attenuation Correction in the Pelvis Using Only Diagnostic MR Images. *Tomography*, 4(3):138–147, September 2018.

[255] Serena Monti, Carlo Cavaliere, Mario Covello, Emanuele Nicolai, Marco Salvatore, and Marco Aiello. An Evaluation of the Benefits of Simultaneous Acquisition on PET/MR Coregistration in Head/Neck Imaging. *J Healthc Eng*, 2017, 2017.

[256] Michaela Daniel, Piotr Andrzejewski, Alina Sturdza, Katarina Majercakova, Pascal Baltzer, Katja Pinker, Wolfgang Wadsak, Markus Mitterhauser, Richard Pötter, Petra Georg, et al. Impact of hybrid PET/MR technology on multiparametric imaging and treatment response assessment of cervix cancer. *Radiother Oncol*, 125(3):420–425, 2017.

[257] Anders B. Olin, Adam E. Hansen, Jacob H. Rasmussen, Claes N. Ladefoged, Anne K. Berthelsen, Katrin Håkansson, Ivan R. Vogelius, Lena Specht, Anita B. Gothelf, Andreas Kjaer, Barbara M. Fischer, and Flemming L. Andersen. Feasibility of Multiparametric Positron Emission Tomography/Magnetic Resonance Imaging as a One-Stop Shop for Radiation Therapy Planning for Patients with Head and Neck Cancer. *Int J Radiat Oncol Biol Phys*, 108(5):1329–1338, December 2020.

[258] Leticia Taeubert, Yannick Berker, Bettina Beuthien-Baumann, Aswin L. Hoffmann, Esther G. C. Troost, Marc Kachelrieß, and Clarissa Gillmann. CT-based attenuation correction of whole-body radiotherapy treatment positioning devices in PET/MRI hybrid imaging. *Phys Med Biol*, 65(23):23NT02, November 2020.

[259] P.s. Tofts, D. Lloyd, C.a. Clark, G.j. Barker, G.j.m. Parker, P. McConville, C. Baldock, and J.m. Pope. Test liquids for quantitative MRI measurements of self-

diffusion coefficient in vivo. *Magn Reson Med*, 43(3):368–374, March 2000.

[260] Mary Adjeiwaah, Anders Garpebring, and Tufve Nyholm. Sensitivity analysis of different quality assurance methods for magnetic resonance imaging in radiotherapy. *Phys Imaging Radiat Oncol*, 13:21–27, January 2020.

[261] Joseph Weygand, Clifton David Fuller, Geoffrey S Ibbott, Abdallah SR Mohamed, Yao Ding, Jinzhong Yang, Ken-Pin Hwang, and Jihong Wang. Spatial Precision in Magnetic Resonance Imaging–Guided Radiation Therapy: The Role of Geometric Distortion. *Int J Radiat Oncol Biol Phys*, 95(4):1304–1316, 2016.

[262] Catherine M. Lockhart, Lawrence R. MacDonald, Adam M. Alessio, Wendy A. McDougald, Robert K. Doot, and Paul E. Kinahan. Quantifying and Reducing the Effect of Calibration Error on Variability of PET/CT Standardized Uptake Value Measurements. *J Nucl Med*, 52(2):218–224, February 2011.

[263] Robert K. Doot, Larry A. Pierce, Darrin Byrd, Brian Elston, Keith C. Allberg, and Paul E. Kinahan. Biases in Multicenter Longitudinal PET Standardized Uptake Value Measurements. *Transl Oncol*, 7(1):48–54, February 2014.

[264] Richard Speight, Michael Dubec, Cynthia L. Eccles, Ben George, Ann Henry, Trina Herbert, Robert I. Johnstone, Gary P. Liney, Hazel McCallum, and Maria A. Schmidt. IPEM topical report: Guidance on the use of MRI for external beam radiotherapy treatment planning. *Phys Med Biol*, 66(5):055025, February 2021.

[265] Kieren G Hollingsworth. Reducing acquisition time in clinical MRI by data undersampling and compressed sensing reconstruction. *Phys Med Biol*, 60(21):R297–R322, October 2015.

[266] Kristy K Brock, Sasa Mutic, Todd R McNutt, Hua Li, and Marc L Kessler. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys*, 44(7):e43–e76, 2017.

[267] Mark Jones, Annette Dobson, and Sue O'Brian. A graphical method for assessing agreement with the mean between multiple observers using continuous measures. *Int J Epidemiol*, 40(5):1308–1313, October 2011.

[268] Ola Weistrand and Stina Svensson. The ANACONDA algorithm for deformable image registration in radiotherapy. *Med Phys*, 42(1):40–53, 2015.

[269] J. Sausville and M. Naslund. Benign prostatic hyperplasia and prostate cancer: An overview for primary care physicians. *Int J Clin Pract*, 64(13):1740–1745, 2010.

[270] Karin Braide, Jon Kindblom, Ulrika Lindencrona, Marianne Månsson, and Jonas Hugosson. The value of a bladder-filling protocol for patients with prostate cancer who receive post-operative radiation: Results from a prospective clinical trial. *Acta Oncol*, 58(4):463–468, April 2019.

[271] Jan Unkelbach, Markus Alber, Mark Bangert, Rasmus Bokrantz, Timothy C. Y. Chan, Joseph O. Deasy, Albin Fredriksson, Bram L. Gorissen, Marcel van Herk, Wei Liu, Houra Mahmoudzadeh, Omid Nohadani, Jeffrey V. Siebers, Marnix Witte,

and Huijun Xu. Robust radiotherapy planning. *Phys Med Biol*, 63(22):22TR02, November 2018.

[272] Emilia Palmér, Anna Karlsson, Fredrik Nordström, Karin Petruson, Carl Siversson, Maria Ljungberg, and Maja Sohlin. Synthetic computed tomography data allows for accurate absorbed dose calculations in a magnetic resonance imaging only workflow for head and neck radiotherapy. *Phys Imaging Radiat Oncol*, 17:36–42, January 2021.

[273] James P. B. O'Connor, Eric O. Aboagye, Judith E. Adams, Hugo J. W. L. Aerts, Sally F. Barrington, Ambros J. Beer, Ronald Boellaard, Sarah E. Bohndiek, Michael Brady, Gina Brown, David L. Buckley, Thomas L. Chenevert, Laurence P. Clarke, Sandra Collette, Gary J. Cook, Nandita M. deSouza, John C. Dickson, Caroline Dive, Jeffrey L. Evelhoch, Corinne Faivre-Finn, Ferdia A. Gallagher, Fiona J. Gilbert, Robert J. Gillies, Vicky Goh, John R. Griffiths, Ashley M. Groves, Steve Halligan, Adrian L. Harris, David J. Hawkes, Otto S. Hoekstra, Erich P. Huang, Brian F. Hutton, Edward F. Jackson, Gordon C. Jayson, Andrew Jones, Dow-Mu Koh, Denis Lacombe, Philippe Lambin, Nathalie Lassau, Martin O. Leach, Ting-Yim Lee, Edward L. Leen, Jason S. Lewis, Yan Liu, Mark F. Lythgoe, Prakash Manoharan, Ross J. Maxwell, Kenneth A. Miles, Bruno Morgan, Steve Morris, Tony Ng, Anwar R. Padhani, Geoff J. M. Parker, Mike Partridge, Arvind P. Pathak, Andrew C. Peet, Shonit Punwani, Andrew R. Reynolds, Simon P. Robinson, Lalitha K. Shankar, Ricky A. Sharma, Dmitry Soloviev, Sigrid Stroobants, Daniel C. Sullivan, Stuart A. Taylor, Paul S. Tofts, Gillian M. Tozer, Marcel van Herk, Simon Walker-Samuel, James Wason, Kaye J. Williams, Paul Workman, Thomas E. Yankeelov, Kevin M. Brindle, Lisa M. McShane, Alan Jackson, and John C. Waterton. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*, 14(3):169–186, March 2017.

[274] Amita Shukla-Dave, Nancy A. Obuchowski, Thomas L. Chenevert, Sachin Jambawalikar, Lawrence H. Schwartz, Dariya Malyarenko, Wei Huang, Susan M. Noworolski, Robert J. Young, Mark S. Shiroishi, Harrison Kim, Catherine Coolens, Hendrik Laue, Caroline Chung, Mark Rosen, Michael Boss, and Edward F. Jackson. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *J Magn Reson Imaging*, 49(7):e101–e121, 2019.

[275] Oliver J. Gurney-Champion, Faisal Mahmood, Marcel van Schie, Robert Julian, Ben George, Marielle E. P. Philippens, Uulke A. van der Heide, Daniela Thorwarth, and Kathrine R. Redalen. Quantitative imaging for radiotherapy purposes. *Radiother Oncol*, 146:66–75, May 2020.

[276] Kristina Sandgren, Erik Nilsson, Angsana Keeratijarut Lindberg, Sara Strandberg, Lennart Blomqvist, Anders Bergh, Bengt Friedrich, Jan Axelsson, Margareta Ögren, Mattias Ögren, Anders Widmark, Camilla Thellenberg Karlsson, Karin Söderkvist,

Katrine Riklund, Joakim Jonsson, and Tufve Nyholm. Registration of histopathology to magnetic resonance imaging of prostate cancer. *Phys Imaging Radiat Oncol*, 18:19–25, April 2021.

[277] Joana Caldas-Magalhaes, Nicolien Kasperts, Nina Kooij, Cornelis AT van den Berg, Chris HJ Terhaard, Cornelis PJ Raaijmakers, and Marielle EP Philippens. Validation of imaging with pathology in laryngeal cancer: Accuracy of the registration methodology. *International Journal of Radiation Oncology* Biology* Physics*, 82(2):e289–e298, 2012.

[278] Linda G. W. Kerkmeijer, Veerle H. Groen, Floris J. Pos, Karin Haustermans, Evelyn M. Monninkhof, Robert Jan Smeenk, Martina Kunze-Busch, Johannes C. J. de Boer, Jochem van der Voort van Zijp, Marco van Vulpen, Cédric Draulans, Laura van den Bergh, Sofie Isebaert, and Uulke A. van der Heide. Focal Boost to the Intraprostatic Tumor in External Beam Radiotherapy for Patients With Localized Prostate Cancer: Results From the FLAME Randomized Phase III Trial. *J Clin Oncol*, 39(7):787–796, March 2021.