Synthetic-CT quality assessment for MRI-only based treatment planning in radiation therapy

Hilda Chourak

COLLEGE DOCTORAL BRETAGNE

MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

THESE DE DOCTORAT DE .

L'UNIVERSITE DE RENNES

ECOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Electronique
Spécialité : AST - Signal, Image, Vision

Par
Hilda CHOURAK

Synthetic-CT quality assessment for MRI-only based treatment planning in radiation therapy

En co-direction internationale avec *The Australian e-Health Research Centre*, Australia

Thèse présentée et soutenue à Rennes, le 2 Novembre 2023
Unité de recherche : LTSI – UMR 1099

Rapporteurs avant soutenance :

Ninon Burgos	Chercheure CNRS / HDR, Institut du cerveau, Paris, France
Martin Ebert	Professor and medical physicist, University of Western Australia, Perth, Australia

Composition du Jury :

Présidente :	Juliette Thariat	PU/PH, LPCC, Université de Caen et Centre François Baclesse, France
Examinateurs :	Nick Reynaert	Professeur et Physicien médical, Institut Jules Bordet, Bruxelles, Belgique
	Juliette Thariat	PU/PH, LPCC, Université de Caen et Centre François Baclesse, France
Rapporteurs :	Ninon Burgos	Chercheure CNRS/HDR, Institut du cerveau, Paris, France
	Martin Ebert	Professor, Medical physicist, University of Western Australia, Perth, Australia
Dir. de thèse :	Renaud De Crevoisier	PU/PH, LTSI et Centre Eugène Marquis, Rennes, France
Co-dir. de thèse :	Jason Dowling	A/Prof and Research team leader, AEHRC, CSIRO, Herston, Australia

Invité(s)

Oscar Acosta	Encadrant de thèse, MCU/HDR, LTSI, Université de Rennes, France
Jean-Claude Nunes	Encadrant de thèse, MCU/HDR, LTSI, Université de Rennes, France
Anaïs Barateau	PhD et physicienne médicale, LTSI et Centre Eugène Marquis, Rennes, France

Acknowledgments

This thesis would not have been possible without the support and help of several people, to whom I would like to express my gratitude.

I would like to thank first the reviewers **Ninon Burgos** and **Martin Ebert**, for their careful reading of my thesis and their insightful comments. I also extend my gratitude to all the members of the jury, including the examiners **Nick Reyneart** and **Juliette Thariat,** for their presence, whether in person or remotely, and for the remarks they provided during the defense. These feedbacks will contribute to the improvement of my work. Thanks to **Juliette Thariat**, who presided brilliantly over the jury on the day of my defense.

My sincere acknowledgment goes to my research directors, **Renaud de Crevoisier** and **Jason Dowling,** as well as **Oscar Acosta** and **Jean-Claude Nunes** for accompanying and supporting me throughout this doctorate.
Thank you, **Renaud** for your good advices. **Jason**, it has been a pleasure to collaborate with you, thank you for your patience and our interesting discussions. **Jean-Claude**, I would like to thank you for your involvement and your thoroughness. **Oscar**, I enjoyed working with you. Thank you for your trust, your cheerfulness and your optimism.

My research would not have been possible without **Peter Greer** and **Caroline Lafond**. Thank you for your invaluable help with radiotherapy and medical physics, making my work more clinically relevant.
Thank you, **Pierre Fontaine**, for taking the time to introduce me to the use of radiomics for medical image analysis. Thanks to **Anaïs Barateau**, whom I invited to join my thesis jury, for her expertise in medical physics and her consistently relevant and instructive feedback on my work. I would like also to express my gratitude to **Jérémy Beaumont**, for his help and for giving me enough confidence to enrol for this doctorate, and thus make this possible. Thank to my colleagues at CSIRO and LTSI for the good time we shared and our stimulating discussions.

I would like to thank the administrative teams at LTSI and CSIRO for their invaluable help with the formalities associated with my PhD, particularly with regard to the difficulties and uncertainties involved in moving between France and Australia during the COVID period.

These past 2 years would not have been the same without my friends **Lorraine** and **Filip**, who made my time in Australia unforgettable, and made me discover great places and new hobbies. I'm glad I met you folks.

Finally, I would like to thank my parents and my siblings, **Leïla**, **Siegfried**, and **Dimitri**, for their unconditional support. Thank for being by my side, no matter where I am.

Table of contents

Résumé en français

La radiothérapie externe est un traitement local du cancer qui utilise des rayonnements ionisants pour détruire les cellules cancéreuses. Les faisceaux de haute énergie produits par un accélérateur linéaire sont dirigés vers le volume de la tumeur tout en minimisant l'exposition aux rayonnements des organes voisins à risque afin de prévenir les dommages. L'efficacité de la radiothérapie externe repose sur l'effet différentiel de la réparation des dommages à l'ADN entre les tissus sains et la tumeur. La quantité de rayonnement délivrée, appelée dose, est exprimée en Gray (Gy), qui est définie comme l'énergie déposée par unité de masse (joule par kilogramme). Le radiothérapeute prescrit une dose totale pour la tumeur, généralement administrée à raison de 2 Gy par séance. La stratégie de traitement dépend de l'emplacement et du stade du cancer. Par exemple, dans les tumeurs à un stade précoce, l'objectif principal est de minimiser la toxicité en limitant le volume cible et la dose globale. En revanche, pour les tumeurs plus avancées, l'accent est mis sur l'amélioration du contrôle local de la maladie en augmentant la dose et/ou en combinant la radiothérapie avec des médicaments radio-sensibilisants. Les traitements hypofractionnés, caractérisés par une dose plus élevée par séance et une réduction du nombre total de séances, gagnent en importance en raison d'avantages tels que l'amélioration du confort du patient, la rentabilité et le ciblage précis du cancer avec des volumes cibles plus petits.

La tomodensitométrie (TDM ou scanner CT) permet d'accéder aux densités électroniques des tissus, essentiel pour des calculs de dose précis, et est donc la modalité d'imagerie utilisée en routine pour la planification dosimétrique du traitement. S'ensuit plusieurs séances d'irradiation durant lesquelles le traitement est délivré. Concernant les techniques de délivrance du traitement, les plus répandues sont maintenant la radiothérapie conformationnelle par modulation d'intensité (RCMI, en anglais IMRT pour intensity modulated radiation thearpy) et l'arcthérapie volumique (VMAT). Avec l'IMRT, plusieurs angles de faisceau sont utilisés et l'intensité de chaque faisceau peut être modulée à l'aide de collimateurs multi-lames (MLC), ce qui permet de créer des profils de dose complexes. Contrairement à l'IMRT, qui comprend généralement moins de 10 angles de faisceau à champ fixe, la VMAT inclut un grand nombre de directions de faisceau à partir d'une trajectoire en arc et délivre des doses de manière dynamique pendant la rotation du dispositif. Dans le cas d'un cancer de la prostate, les séances s'étalent sur 5 à 8 semaines en fractionnement standard, à raison de 5 séances par semaine. La tomographie volumique à faisceau conique (CBCT), système d'imagerie 3D-kV en salle, permet le positionnement du patient sous l'accélérateur linéaire (par recalage rigide CBCT-CT).

L'imagerie par résonance magnétique (IRM) est quant à elle largement utilisée pour le diagnostic du cancer, car elle offre un contraste tissulaire supérieur au CT et CBCT sans induire de rayonnements ionisants. Cette modalité d'imagerie permet une délinéation du volume

cible et des organes à risque plus précise que sur CT dans le cadre d'un cancer de la prostate. L'IRM a suscité un vif intérêt au sein de la communauté de la radiothérapie externe, en particulier avec le développement récent d'IRM-linac qui intègrent un scanner IRM avec un accélérateur linéaire. Cependant, l'une des principales limites de l'IRM est son incapacité à fournir des informations sur la densité électronique des tissus, essentielle pour un calcul précis de la dose. Pour pallier cette limitation, plusieurs méthodes ont été proposées dans la littérature pour générer des CT synthétiques (sCT). Les sCT (en unités Hounsfield) reproduisent un CT en étant créés à partir de l'IRM. Les méthodes de génération de sCT peuvent être regroupées en trois catégories : les méthodes par assignement de densité, celles basées sur la génération d'atlas, et enfin les méthodes utilisant des modèles de machine-learning. Parmi ces dernières, les méthodes d'apprentissage profond se sont révélées extrêmement prometteuses en termes de performances et de rapidité d'exécution. L'IRM pourrait ainsi remplacer le CT pour toute planification initiale de radiothérapie externe, ainsi que pour du monitoring de dose dans le cadre d'une radiothérapie adaptative avec un IRM-linac.

Malgré les avantages que présente l'utilisation de l'IRM pour la planification de traitement en radiothérapie, son intégration en clinique fait face à un défi majeur : le manque de mesures d'évaluation standardisées. Actuellement, l'évaluation de ces méthodes repose sur des métriques s'appuyant sur la comparaison d'intensités à une image de référence, et nécessitent donc un CT comme vérité terrain. Les métriques les plus fréquemment employées sont l'erreur absolue moyenne (MAE en anglais pour mean absolute error), l'erreur moyenne (mean error, ou ME en anglais) et le « peak signal-to-noise ratio »(PSNR). Des mesures basées sur la perception telles que la mesure de l'indice de similarité structurelle (SSIM) et la SSIM multi-échelle sont aussi couramment utilisées. La précision du calcul de dose résultant des sCT est aussi souvent évaluée en comparant la distribution de dose sur sCT à celle attendue (obtenue à partir du CT de planification). Les méthodes d'évaluation dosimétriques sont principalement la comparaison des histogrammes dose-volume, les différences de dose par voxel (absolues ou relatives) ou encore l'analyse gamma.

Ces mesures fournissent une évaluation globale et offrent un aperçu limité de la précision de l'image générée dans le contour externe du patient ou de volumes délinéés. Elles dépendent aussi de la précision du recalage entre IRM et CT de planification, et de l'existence du CT de planification. Or, à terme, ce CT ne sera plus acquis. Par conséquent, il est impératif de développer des méthodes de contrôle de qualité des sCT robustes qui non seulement permettent l'administration précise du traitement, mais offrent également une analyse approfondie des limites de chaque méthode de génération de sCT.

Ce travail de thèse a permis l'évaluation spatiale des méthodes de génération de sCT et propose des stratégies pour évaluer la qualité de sCT généré au jour le jour. Ces stratégies sont destinées à être incluses dans le workflow clinique afin d'assurer la fiabilité des techniques de planification basée sur IRM.

Dans le premier chapitre de ce manuscrit, le processus standard du traitement par radiothérapie externe est présenté ainsi que les avantages d'une planification de traitement sur IRM. La deuxième partie de ce chapitre traite des méthodes de pointe pour générer un sCT à partir de l'IRM, notamment à l'aide de modèles d'apprentissage profond (comme décrit dans notre article état de l'art publié dans *Physica Medica*). Aussi, un aperçu des différentes méthodes de contrôle de la qualité impliquées dans la radiothérapie basée sur IRM est présenté.

Le deuxième chapitre décrit les objectifs de la thèse, et présente le cadre général dans lequel s'inscrivent les différentes composantes des travaux de cette thèse.

Le troisième chapitre présente la mise en œuvre d'une méthodologie permettant d'évaluer la précision des méthodes de génération de sCT à différents niveaux. Cela comprend une évaluation des erreurs à différentes échelles : dans l'ensemble du pelvis, par organe, et enfin, par voxel. Afin de permettre une analyse par voxel précise, les images ont été recalées de façon non rigide via une approche robuste impliquant le calcul des descriptions structurelles des organes. La méthodologie présentée est fondée sur l'analyse d'une base de données patients et met en exergues les sous-régions anatomiques ou les méthodes de génération tendent à systématiquement sous ou surévaluer les valeurs d'intensités. Les résultats sont présentés dans deux articles. Le premier article, publié à la suite de la présentation de l'étude a *l'International symposium on biomedical imaging (ISBI),* décrit la méthode utilisée pour évaluer quatre approches de génération de sCT en termes de métriques d'image. Le deuxième article, publié dans *Frontiers in Oncology*, va plus loin dans l'analyse statistique et comprend une évaluation dosimétrique.

Le chapitre précédent se concentre sur la détection de sous régions ou les erreurs de prédictions d'unité Hounsfield (UH) sont significatives pour différentes méthodes de génération de sCT en examinant à la fois les résultats image et dosimétrique. Cependant, ces évaluations sont indépendantes et l'impact des erreurs d'UH sur la dose n'est pas trivial. En effet, si une méthode a tendance à échouer dans une région spécifique, quelles seraient les conséquences pour le traitement, s'il y en a ? Ainsi, le chapitre 4 présente une analyse de sensibilité comme outil pour répondre à cette question. L'analyse suit la méthode de screening de Morris et explore la corrélation entre les changements d'intensité dans différentes structures anatomiques et la dose dans le volume cible afin d'identifier les régions où les erreurs d'UH ont un impact plus important sur la dose. Aussi, cette étude évalue l'influence d'un artefact (représentant un volume d'erreur) sur la dose délivrée au centre du volume cible, en tenant compte de trois critères : la taille de l'artefact, son emplacement par rapport au volume cible et la variation d'UH dans le volume d'erreur. Il en ressort qu'une erreur même faible (25 UH) répartie dans l'ensemble du pelvis aura un impact significatif sur la dose à l'isocentre, alors que des erreurs dans la vessie auront impact négligeable. La taille du volume d'erreur ainsi que la variation d'intensité et sa position par rapport aux rayons

incidents jouent un rôle majeur sur le calcul de dose. Il est important de noter que ces résultats dépendent du traitement prescrit (IMRT ou VMAT par exemple), et qu'il serait pertinent de prendre en compte les sous régions d'erreurs détectées par l'analyse par voxel précédemment présentée pour une méthode de génération donnée. L'approche présentée dans ce chapitre est donc à adapter au cas par cas.

Cette étude a été publiée par le journal *Physical and Engineering Sciences in Medicine* en octobre 2023.

Le chapitre 5 propose deux méthodes pour évaluer un sCT patient lorsqu'aucun CT de planification n'est disponible. Ces méthodes s'appuient sur de statistiques extraites d'une cohorte de CT. Cette évaluation est effectuée au niveau du voxel pour la première étude (comme décrit dans un article publié suite à une conférence internationale *IEEE 2022, Digital image computing : techniques et applications (DICTA)*), ainsi qu'au niveau global et par organe en utilisant la sélection des caractéristiques radiomics pour la seconde. La première met en avant les voxels dont les valeurs d'UH sont significativement différentes de celles de la cohorte de référence pour chaque voxel. La précision de cette méthode dépend de la qualité des processus de recalage mis en place : celui permettant la création d'un atlas de référence à partir de la cohorte de CT, mais aussi celui permettant la comparaison du nouveau sCT à cet atlas. Les résultats pour les patients ayant une anatomie trop éloignée de celles des patients constituant la cohorte de référence peuvent donc être biaisés. Aussi, le recalage peut être couteux en temps. La seconde méthode permet d'obtenir un score illustrant la cohérence de l'image générée dans son ensemble et par organe, selon des caractéristiques images pertinentes et sélectionnées via un algorithme de machine learning (forêts aléatoires conditionnelles). Elle permet également de se libérer des biais induit par le recalage et n'est pas couteuse en temps de calcul.

Dans le chapitre 6, un résumé des méthodes d'évaluation de sCT est présenté, ainsi que des recommandations pour mettre en place un protocole visant à intégrer le contrôle de la qualité de chaque sCT dans un workflow clinique. Ce protocole souligne l'importance de la mise en place de contrôle aux différentes étapes de la planification et implique l'évaluation : de l'IRM dans un premier temps, puis du sCT et enfin du calcul de dose. Ceci induit l'utilisation de métriques subjectives d'abord, mais aussi de métriques objectives de complexité variables présentés dans les chapitres précédents.

Cette thèse est par définition multidisciplinaire, car elle s'inscrit dans une problématique de traitement d'image dans une perspective clinique avec un aspect dosimétrique (relatif à la physique médicale). Aussi, les méthodes présentées sont applicables aux sCT générés à partir de CBCT dans le cadre d'une radiothérapie adaptative pour les centres équipés d'accélérateur linéaire standards, puisqu'elles sont indépendantes de la modalité d'imagerie utilisée pour la création de sCT.

Ces études ont été réalisées sur un ensemble de données d'imagerie rétrospective de patients atteints d'un cancer de la prostate localisé traités à l'hôpital Calvary Mater à Newcastle, Australie.

Ces travaux ont été co-financés par la Région Bretagne, France (bourse ARED) et le CSIRO, Australie (collaborative project agreement). Afin de permettre la lecture de ce manuscrit de thèse aux co-financeurs australiens de ce projet, la suite de ce manuscrit est rédigée en anglais.

Acronyms

AE	Absolute error
AI	Artificial intelligence
APE	Absolute percent error
ART	Adaptive radiation therapy
CBCT	Cone beam computerised tomography
CNN	Convolutional neural network
CT	Computed tomography
DLM	Deep-learning model
DSC	Dice score coefficient
DVH	Dose-volume histogram
E	Error
EBRT	External beam radiation therapy
GAN	Generative adversarial network
Gy	Gray (dose unit)
HD	Hausdorff distance
HU	Hounsfield Units
IGRT	Image guided radiation therapy
IMRT	Intensity-modulated radiotherapy
IQA	Image quality assessment
LINAC	Linear particle accelerator
MAE	Mean absolute error
MAPE	Mean absolute percent error
MASD	Mean absolute surface distance
ME	Mean error
MRI	Magnetic resonance imaging
MSE	Mean square error
MS-SSIIM	Multi-scale structural similarity index
NCC	Normalized cross-correlation
NMI	Normalized mutual information
OAR	Organs at risk
PSNR	Peak signal-to-noise ratio
PTV	Planning target volume

QA	Quality assessment
QC	Quality control
RMSE	Root mean square error
ROI	Region of interest
RSD(AE)	Relative Standard Deviation of Absolute Error
RT	Radiation therapy
SBRT	Stereotactic body radiation therapy
sCT	synthetic-CT
SSIM	Structural similarity metric
TPS	Treatment planning system
VIF	Visual information fidelity
vMAE	Voxel-wise mean absolute error
vMAPE	Voxel-wise mean absolute percent error
VMAT	Volumetric arc therapy
vME	Voxel-wise mean error
VOI	Volume of interest

Introduction

External beam radiotherapy is a local cancer treatment that utilizes ionizing radiation to destroy cancer cells. The high-energy beams produced by a linear accelerator are directed towards the tumour volume while minimizing radiation exposure to nearby organs at risk in order to prevent harm. In the context of external beam radiotherapy, computed tomography (CT) scans provide access to the electron densities of tissues, which are essential for accurate dose calculations. Cone Beam Computerized Tomography (CBCT), on the other hand, enables patient positioning under the linear accelerator (planning CBCT-CT registration) as well as real-time monitoring of the tumour during treatment. However, these imaging modalities suffer from limited soft tissue contrast and expose patients to additional radiation.

In contrast, Magnetic Resonance Imaging (MRI), widely used for cancer diagnosis, offers superior tissue contrast without the need for ionizing radiation. This imaging modality holds significant potential for precise delineation of the target volume and organs at risk [1] and dose targeting[2]. MRI has gained particular interest in external beam radiotherapy, especially with the recent development of MRI-linac machines that integrate an MRI scanner with a linear accelerator. However, a key limitation of MRI is its inability to provide electron density information essential for accurate dose calculation. To overcome this limitation, several methods have been proposed in the literature for generating synthetic CT scans (sCT) from MRI[3], [4]. Among these, deep learning methods have shown tremendous promise in terms of both performance and computational efficiency[5]. Despite their favourable characteristics, the integration of MRI into the radiotherapy workflow faces a major challenge—the lack of standardized assessment metrics. Currently, evaluation of these methods relies on full-reference intensity-based metrics [6], which require a CT as ground truth, such as mean absolute error (MAE), mean error (ME), and peak signal-to-noise ratio (PSNR). Additionally, perception-based metrics like the structural similarity index measure (SSIM) and the multiscale SSIM are commonly employed. However, these metrics provide a global assessment and offer limited insight into the agreement within the patient's body contour or individual organs. Therefore, it is imperative to develop robust quality control methods that not only facilitate accurate treatment delivery but also enable thorough analysis of method limitations and shortcomings.

This thesis primarily aims to investigate areas where sCT generation tends to be less accurate and proposes strategies to assess the quality of daily generated sCT. These strategies are intended to be included in the clinical workflow to ensure the safe application of MRI-only techniques.

In the first chapter of this manuscript, the standard process of external beam radiotherapy treatment is presented, along with a comparison to an MRI-only workflow. The second part of this chapter discusses the state-of-the-art methods for generating sCT from MRI using deep-learning models (as described in an article published in *Physica Medica*). Additionally,

an overview of the different quality control steps involved in MRI-only radiotherapy is presented.

The second chapter outlines the main objectives of the thesis, in addition to an overall framework of thesis components.

The third chapter presents research on a methodology to assess the accuracy of sCT generation methods at various levels. This includes a standard error evaluation in the whole pelvis, followed by an organ-wise error assessment, and finally, the implementation of voxel-wise analysis. These findings are presented in two papers. The first paper, published in the *IEEE International Symposium on Biomedical Imaging (ISBI),* presents the method used to assess four sCT generation approaches in terms of image quality. The second paper, published in *Frontiers in Oncology*, delves into the statistical analysis and includes the assessment of dose calculations.

Chapter 4 explores the correlation between localized HU errors and the dose at the center of the target volume using the Morris screening method. This study has been submitted to *Physical and Engineering Sciences in Medicine*.

Chapter 5 proposes two methods to assess patient-specific sCT without ground truth, based on statistics extracted from a cohort of CT scans. This assessment is done at a voxel level (as described in a paper published in the *IEEE 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*), as well as at a global and organ-wise level using radiomics feature selection.

In Chapter 6, a summary of methods and metrics for assessing sCT will be provided, along with recommendations for a protocol aimed at integrating daily sCT quality control into the clinical workflow.

Finally, a conclusion summarizes the work conducted, highlights its limitations, and outlines future research directions.

These studies were performed on a retrospective imaging dataset from localised prostate cancer patients who were treated at the Calvary Mater Hospital in Newcastle, Australia.

References

[1] E. S. Paulson, B. Erickson, C. 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*, vol. 42, no. 1, pp. 28–39, Jan. 2015, doi: 10.1118/1.4896096.

[2] E. Johnstone *et al.*, “Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 100, no. 1, pp. 199–217, Jan. 2018, doi: 10.1016/j.ijrobp.2017.08.043.

[3] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” *Med Phys*, 2021, doi: 10.1002/mp.15150.

[4] J. M. Edmund and T. Nyholm, “A review of substitute CT generation for MRI-only radiation therapy,” *Radiation Oncology*, vol. 12, no. 1, p. 28, Dec. 2017, doi: 10.1186/s13014-016-0747-y.

[5] M. Boulanger *et al.*, “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[6] G. Wang, Y. Zhang, X. Ye, and X. Mou, “Image quality assessment,” in *Machine Learning for Tomographic Imaging*, in 2053-2563. IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

Chapter 1: Context of MRI for radiation therapy

This chapter aims to provide the clinical context by introducing the principles of standard external beam radiation therapy (EBRT). It discusses the benefits of magnetic resonance imaging (MRI) and deep learning in EBRT. Furthermore, it introduces the state-of-the-art methods for dose calculation using synthetic-CT generated from MRI. We published a review of deep learning-based approaches in the journal *Physica Medica* [1]. Additionally, the chapter provides an overview of the various quality control steps involved in MRI-only radiotherapy.

External beam radiation therapy

Radiotherapy is a treatment prescribed for more than two-thirds of patients with all types of cancer, i.e around 300,000 patients per year in France. EBRT involves delivering ionising radiation to the tumour to damage the DNA of cancer cells, mainly through double-strand breaks, thereby inhibiting their ability to multiply. These ionising rays primarily consist of high-energy photon beams (MV) or, less commonly, electron beams (MeV), and proton beams. These beams are administered using a linear particle accelerator (LINAC) while the patient remains immobilised on the treatment table with restraints. To ensure tolerance of healthy tissues, the radiation is delivered in divided sessions, typically one session per day for 5 days per week, usually over a period of 5 to 8 weeks.

The effectiveness of radiotherapy is based on the differential effect of DNA damage repair between healthy tissue and the tumour. The amount of radiation delivered, known as the dose, is expressed in Grays (Gy), which is defined as the energy deposited per unit of mass (Joule per kilogram). The radiation therapist prescribes a total dose for the tumour, typically delivered at a rate of 2 Gy per session. The treatment strategy depends on the cancer's location and stage. For instance, in early-stage tumours, the primary objective is to minimise toxicity by limiting the target volume and overall dose. In contrast, for more advanced tumours, the focus shifts to enhancing local disease control by escalating the dose and/or combining radiation with radiosensitising drugs. Hypofractionated treatments, characterised by higher dose per session and reduced total sessions, are gaining prominence due to advantages such as improved patient comfort, cost-effectiveness, and precise cancer targeting with smaller target volumes.

Standard workflow

There are four stages involved in a standard EBRT procedure (Figure 1.1): 1) acquisition of a planning CT scan, 2) delineation of target volume(s) and organs at risk (OARs), 3) dosimetric

planning, and 4) treatment. The following paragraphs provide a detailed description of each step.

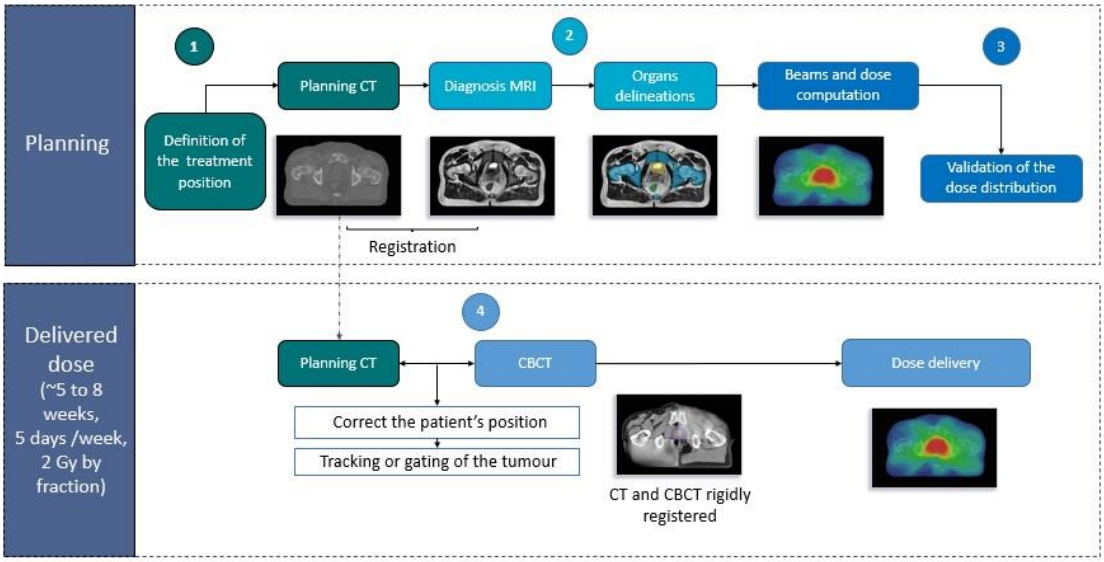

Figure 1.1: Standard external beam radiation therapy workflow.
The planning process is divided into three steps. The first step (1) involves acquiring the planning CT scan in the treatment position. In the second step (2), an MRI is obtained to provide improved visualisation of soft tissues, enabling more accurate delineation of the target volume and organs at risk. Both modalities are then registered to align the contours with the planning CT. In the third step (3), the beam parameters are determined by an expert, ensuring compliance with the dose constraints prescribed by the radiation therapist, who subsequently validates the dose distribution. A CBCT scan is acquired at each session to accurately align the patient with the planning CT and reproduce the treatment position as precisely as possible prior to dose delivery.

1- Planning CT

The acquisition of a computed tomography (CT) image is the initial stage in EBRT. Prior to acquiring the image, the patient is positioned using restraints to establish a reference position for the treatment. This position should be both comfortable for the patient and reproducible. External markers, such as tattoos on the patient's skin or markings on the restraints, are utilised to replicate this position accurately. The length of the acquisition (head-foot axis) varies based on the anatomical location. It should encompass the entire target volume and the OARs, in accordance with the considered dosimetric criteria.

Limitations of the CT scanning is that it is an imaging technique that involves exposure to ionising radiation and has low contrast in the soft tissue, impeding clear visualisation of the tumour volume.

2- Organs at risk and tumour delineation

Manual delineation of the OARs and the target volume, which comprises the tumour and its extensions, has traditionally been performed based on the planning CT scan. However, due to the limited ability of CT to accurately distinguish soft tissues, diagnostic MRI is utilised in

addition to enhance the accuracy of the delineations. To achieve this, the diagnostic MRI is first registered with the planning CT, and then the delineations of the volumes of interest from the MRI are transferred to the planning CT. The two imaging modalities offer complementary information, with CT providing better visualisation of bone tissue, while MRI enables improved visualisation of soft tissue.

CT-MRI registration introduces uncertainties into the radiotherapy workflow. MRI scans are not acquired in the exact treatment position, posing challenges for accurate CT-MRI registration. Additionally, anatomical variations between the two acquisitions further complicate this stage. For instance, in prostate cancer cases, these variations can be attributed to varying bladder volumes or the presence of gas in the rectum or intestines, resulting in reported discrepancies of up to 2 mm in calculations for prostate cancer patients[2].

MRI-based planning could help to reduce these uncertainties without requiring additional radiation exposure, as MRI is a non-ionising imaging modality. Moreover, it has the potential to lower the overall treatment cost by eliminating the need for multiple scans.

3- Dosimetric planning

Dose planning is performed by a dosimetrist, radiation therapist, or medical physicist on a treatment planning system (TPS). Figure 1.2 shows an example of dose prescription for a prostate cancer on a TPS. This process utilises the planning CT and delineations of the target volume and OARs. The objective is to determine the optimal radiation pattern, including the number of beams and their incidences, in order to deliver the prescribed dose to the planning target volume (PTV) while minimising radiation exposure to OARs. Various irradiation techniques are available, with intensity-modulated radiotherapy (IMRT) and volumetric arc therapy (VMAT) being the most common nowadays. IMRT involves the use of fixed beams, a constant flow rate, and a single multi-leaf collimator (MLC) that conforms each beam to the shape of the PTV. Unlike IMRT, which normally employs fewer than ten fixed-field beam angles, VMAT uses numerous beam directions from an arc trajectory and delivers doses dynamically while the gantry rotates. The operator selects the beam parameters during the planning process, known as direct planning. More advanced techniques of IMRT or VMAT, require inverse planning[3]. In this case, dosimetric objectives are provided to the TPS, which then determines the optimal solution. Intensity modulated treatments can be delivered using fixed beams (IMRT) or arcs (VMAT).

Dosimetric planning must comply with dose constraints for the target volume, the dose prescription specified by the radiation therapist, and the OARs. National and international recommendations regarding dosimetric constraints are considered for OARs. These constraints may include maximum dose, maximum average dose, or maximum percentage of volume that can receive a certain dose, depending on the type of OARs[4]. Intensity-modulated techniques allow for precise dose sculpting around the PTV while minimising radiation exposure to OARs. This technique results in high dose gradients. Another technique,

known as stereotactic body radiation therapy (SBRT), involves delivering higher radiation doses to the tumour in a reduced number of treatment sessions (hypofractionation). It is often used as an alternative to surgery.

Dose calculation in radiotherapy relies on having information about the electron density of tissues, which can be obtained from CT scans but not from MRI. Electron density of tissues is crucial information for treatment planning as it affects how they interact with radiation. It can be derived from the Hounsfield Units (HU) of CT scans, where HU values are assigned to specific tissues based on their radiodensity compared to that of water (e.g., water has a HU value of 0, while dense bone may have HU values above 1000, and air -1000). The electron density is obtained through a calibration curve.

To enable treatment planning to be based on MRI images, it is thus essential to convert MRI into synthetic-CT, to ensure the accuracy of the generated data.

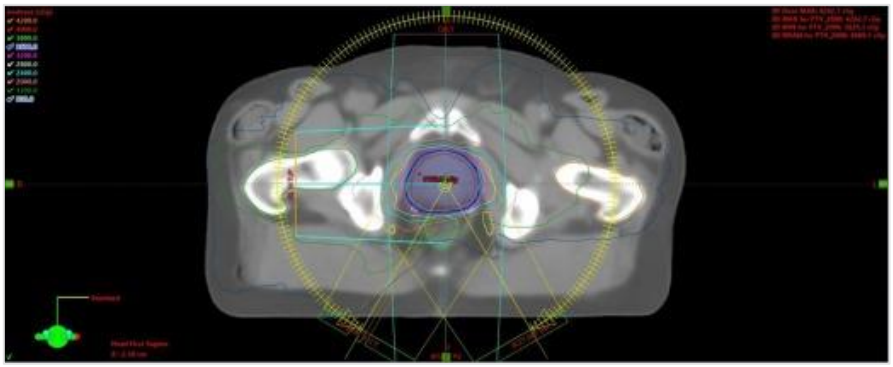

Figure 1.2: Axial view of the 3D treatment plan for a prostate cancer.
(Figure from the book "Biomedical Image synthesis and simulation", chapter 20 [5]).

4- Treatment delivery

At each treatment session, the patient is positioned under the linear accelerator in the same position as during the planning scan. Image-guided radiation therapy (IGRT) is commonly employed to precisely deliver radiation to the tumour target while minimising exposure to healthy surrounding tissues. It involves the utilisation of imaging technologies before or during the treatment session to accurately locate the tumour and make necessary adjustments to the radiation beams.

In IGRT, two techniques can be used: on-board 2D imaging called kV-kV (two 2D images at different angles acquired with low energy imaging systems (kV)) or 3D imaging called CBCT (Cone Beam Computerised Tomography). The CBCT image or the two kV-kV images are then registered with the planning CT scan using the target volume and/or bone tissue. The treatment table is adjusted based on the geometric transformation parameters obtained during the registration process (only translations or translations and rotations).

One of the main limitations of this stage is the low contrast of soft tissues in CBCT images. This limitation makes it challenging to visualise the tumour and results in imprecise patient positioning when relying on soft tissues. CBCT imaging has poor image quality, including the presence of numerous artifacts and low resolution, which adds complexity to the CBCT-CT registration process. Moreover, CBCT is an imaging technique that involves exposure to ionising radiation. To improve patient positioning without additional radiation exposure, an ideal solution would be to use MRI. This can be achieved with a machine combining an MRI and a linear accelerator, known as an MRI-LINAC[6].
Table 1.1 summarises the advantages and drawbacks of standard image-guided radiotherapy treatments compared to MRI-based radiotherapy treatments.

Table 1.1: Advantages and drawbacks of standard radiation treatment planning and MRI-based treatment planning.

		Advantages	***Drawbacks***
Standard image-guided radiation therapy (planning CT and CBCT)		- Easy access to electron density - Spatial resolution of the CT - Fast acquisition CT /CBCT - Registration CT/CBCT - Systems accessibility	- Poor soft tissue contrast, leading to an overestimation of volume of interest - Anatomical imaging only - Ionising modality
MRI-based radiation therapy	***Planning MRI and treatment with a standard LINAC***	- Better soft tissue contrast (more accurate delineations) - Variety of sequences - Functional imaging - Non-ionising	- No access to electron density - Distortions in MRI images - Registration MRI/CBCT - Patients with contraindications - Limited availability of MRI for radiotherapy departments
	MRI-LINAC	- Better soft tissue contrast (more accurate delineations) - Variety of sequences - Non-ionising - Functional imaging - Daily adaptation	- No access to electron density - Distortions in MRI images - Patients with contraindications - Cost

Adaptive radiotherapy

Adaptive radiotherapy (ART) involves a feedback loop during standard radiotherapy treatment, where the initially defined treatment plan is modified to account for inter- or intra-fraction anatomical changes[7]. The goal of ART is to ensure optimal dosimetric coverage of the target volume during treatment and/or to limit the dose to OARs in the presence of anatomical variations from the planning CT scan.
Various ART strategies have been implemented in radiotherapy departments. Offline ART focuses on adapting the treatment plan between sessions, considering inter-fraction anatomical modifications. This strategy does not consider changes that occur within a treatment session. On the other hand, online ART involves adapting the treatment plan during the session while the patient is on the treatment table. This strategy allows for real-time adjustments to account for random changes. The last ART strategy is inline or real-time ART, which monitors movements of the target volume during irradiation using a "real-time" imaging system and compensates for them by adjusting either the multi-leaf collimator (MLC) or the source with the CyberKnife (Accuray)[8], [9]. The implementation of an ART strategy depends on factors such as the location of the tumour and the types of movement/anatomical variations involved. However, not all patients can benefit from ART due to limitations in human and technical resources, as well as the lack of formal clinical evidence demonstrating its benefits. The emergence of MRI-LINACs, which combine MRI and particle accelerators, would make ART more accessible. However, calculating a dose from MRI images is not a straightforward task. The challenges and methods for dose calculation from MRI images will be discussed later in this chapter.

Treatment planning from MRI: state-of-the-art and challenges

In external radiotherapy, X-ray imaging (CT-scan and CBCT) serves as the reference modality for treatment planning and target volume positioning before irradiation. CT provides access to the electronic density of tissues, which is necessary for dose calculation. CBCT allows the patient to be positioned under the linear accelerator by registration with the planning CT, and it also enables tumour gating or tracking during treatment. Although X-ray imaging has advantages such as accurate representation of bone tissue, it provides poor contrast between soft tissues, leading to imprecise definition of the volumes of interest, including tumours and OARs. Additionally, it is an irradiating modality that can increase the risk of radiation-induced cancers due to repeated CBCT image acquisitions, which raises concerns[10].

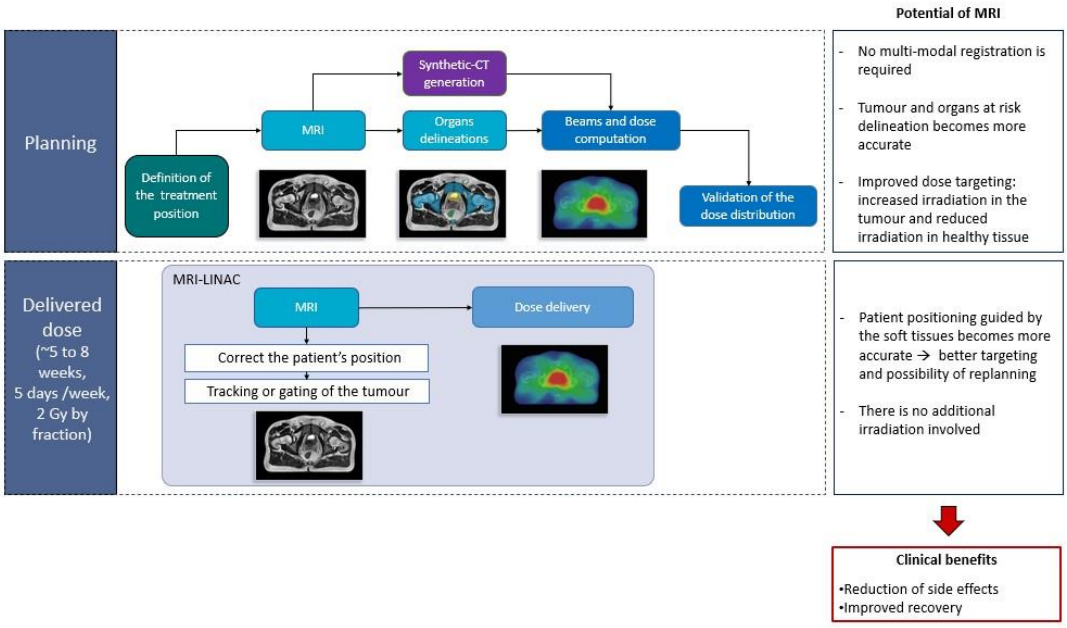

Figure 1.3: MRI-only workflow and its potential benefits

MRI, commonly used for diagnostic purposes, offers the potential to overcome these drawbacks (Figure 1.3). Compared to X-ray imaging, MRI provides better contrast between soft tissues, leading to improved tumour targeting and organ at risk definition. MRI is a non-irradiating modality that provides both morphological and functional information. To leverage these advantages during treatment, new treatment technologies combining a linear accelerator and MRI (MRI-LINAC) have been developed. These technologies allow for precise dose delivery with real-time tracking and gating of the tumour based on MRI images [6]. However, calculating the dose from MRI alone remains a crucial challenge in radiotherapy[11], [12]. The main challenge lies in the fact that MRI does not easily provide access to tissue electron densities, which are necessary for dose calculation.

Standard MRI sequences used in clinical practice (e.g., T1 and T2 weighted) do not capture the signal from tissues with short transverse relaxation times, such as collagen and bone cortex, resulting in poor visibility of these structures in the images and their inability to be distinguished from air. Furthermore, MRI image intensities are not always uniform in homogeneous anatomical structures.

To generate a dose distribution from an MRI scan, the strategy involves creating a substitute CT scan, known as a synthetic-CT (sCT) or pseudo-CT (pCT), using three categories of methods: bulk-density [13]–[15], atlas-based [16]–[22], and machine learning [23]–[29], including deep-learning based methods[30]–[39]. In recent years, deep learning-based approaches have gained popularity due to their ability to automatically generate accurate sCTs with minimal computation time. This section briefly presents the three main strategies for generating sCT and provides an overview of the state-of-the-art methods using deep learning approaches, which have been published in *Physica Medica*[1].

The accuracy of methods from these different categories has previously been compared in Largent et al.[26], [40].

Synthetic-CT generation using bulk-density based methods

Bulk-density methods involve delineating volumes of interest on the patient's MRI, either manually or automatically, and then assigning a homogeneous electron density to each of these volumes. These methods have shown encouraging results and have been the first to be integrated into commercial devices, either alone like in the first release of MRCAT (MR for Calculating Attenuation, Ingenia 3T MR Scanner, Philips Healthcare, Cleveland, OH, USA)[14], [41] or combined with a multi-atlas-based registration method to create bony contours separated into cortical bone and trabecular bone (SyngoVia software platform , Siemens Healthineers, Erlangen, Germany)[42].
The density assignment approach is a simple method, but it can be tedious, time-consuming, and dependent on the operator (leading to inter-operator variability in manual segmentations). Automatic segmentation could be considered for certain tumour locations, but it may be computationally expensive in terms of computing time. Additionally, tissues with short transverse relaxation times are not visible on the images obtained from standard MRI sequences, which may limit the accuracy of delineations. The lack of consideration for tissue heterogeneity is not recommended, especially for certain tumour locations such as bone[43].

Synthetic-CT generation using atlas-based methods

The initial methods for generating atlas-based sCTs involved mapping the patient's MRI to a CT scan (atlas) and transferring the HU values from the atlas onto the patient's MRI to obtain the electron density of the tissues[22]. These methods had limitations, as the atlas used was not representative of the population, and the use of a single multimodal registration was insufficient to account for complex anatomies. Moreover, due to the challenges in distinguishing soft tissues in CT, the registration of this atlas with MRI introduced uncertainties for this type of tissue. To address these issues, a more accurate and robust methodology was developed. This methodology involved iteratively recalibrating and averaging the MRI and CT images from a cohort to construct a representative "MRI-CT" atlas, which was then matched to the patient's MRI[17], [19] . However, this methodology was not very robust when dealing with patients who exhibited significant anatomical differences from the atlas.

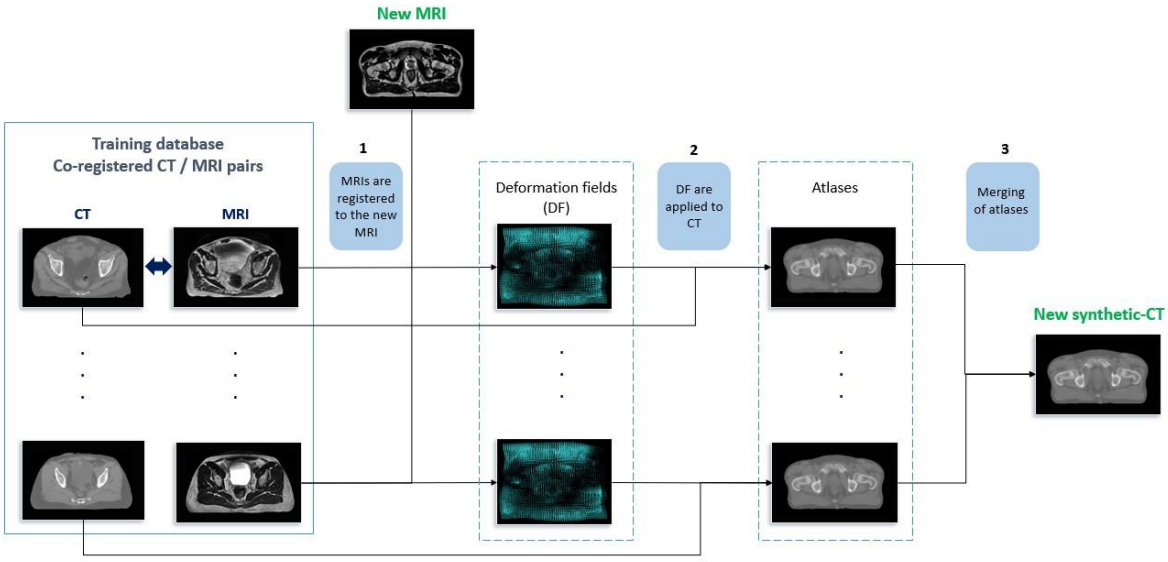

Figure 1.4: Multi-atlas MRI dose calculation method. (1) After intra-patient registration of each pair of MRI-CT images in the database, the atlas MRIs in the training dataset are registered with the new patient's MRI. (2) The deformation fields obtained from the registration step are then applied to the CT scans in the database, aligning these images with the patient's MRI. (3) For each voxel, the intensities of the registered CT atlases are averaged and weighted based on the local similarity between the patient's MRI and the base MRIs. The resulting image is the synthetic-CT for the patient, on which the dose calculation will be performed.

Further modifications were eventually made by incorporating multiple atlases, resulting in the method being divided into two stages (Figure 1.4). In the first stage, after performing deformable intra-patient registration for each pair of MRI-CT images in a cohort, the cohort's MRI was registered with the patient's MRI. The geometric transformations obtained from this registration were then applied to the CT images within the cohort. In the second stage, the re-registered CTs were merged to generate a sCT [16], [18], [20], [21]. The results were deemed clinically acceptable.

In the literature, various CT fusion strategies have been proposed for multi-atlas methods. The most encountered strategy is the weighted average of CT intensities [20], [21]. However, it has the drawback of smoothing the sCT intensities. Consequently, the weighted average has been substituted with the median value [16] or a shape recognition method [18]. The latter, when compared to an atlas method using a weighted mean to merge the CTs, did not show significant improvement in terms of dose but had a longer calculation time (98 min versus 120 min per patient, based on an atlas base consisting of 6 patients).

In summary, atlas methods offer several advantages, such as being fully automated and applicable to any type of anatomical structure, from pelvis[20], to head and neck[16], [21]. They yield satisfactory dosimetric results by considering tissue heterogeneity. However, atlas methods have limitations, including long calculation times due to the successive stages of deformable registration. Additionally, these methods are highly reliant on the quality of intra-patient CT-MRI registration, which can be complex due to anatomical variations between acquisition sessions (MRI and CT), especially in the pelvic area. Furthermore, they depend on

the inter-patient registration of the MRI base with the patient's MRI. As a result, atlas methods are not robust when significant anatomical differences exist between the patient and the images used for the atlas database.

Synthetic-CT generation using model-fitting approaches

Synthetic-CT generation methods based on model-fitting, or statistical learning techniques, have the following objectives: 1) to model the relationships between CT HU values and MRI intensities of a training cohort using algorithms, and 2) to apply these algorithms to a new MRI image, the target, in order to predict the corresponding HU values for the sCT.

Kapanen et al.[44] were the first to employ polynomial regression in generating sCTs, as documented in the literature. To minimise variations in patient positioning between the two modalities, bone registration was utilised to align the MRIs and CTs. Subsequently, the pelvic bones were delineated on the CT images, and regions of interest (ROIs) were manually placed on these delineations (small spheres with a few millimetres in diameter) and propagated onto the MRI images. The image database was then divided into a learning cohort and a validation cohort. A second-order polynomial regression was employed to establish the relationship between CT and MRI intensities within the pelvic bone. This regression was trained using ROIs derived from the CT and MRI scans in the training cohort. To generate the corresponding sCTs for the MRIs in the validation cohort, the regression model was applied to the pelvic bones in the MRIs, assigning an electron density of 1 (0 HU) to other tissue types.
To enhance the outcomes of the aforementioned study, Korhonen et al.[45] proposed the inclusion of manual soft tissue classification into three classes (muscle, urine, fat), in addition to the regression model.
According to the authors, the dosimetric results appeared satisfactory. However, one drawback of this method is the placement of ROIs and the manual segmentation of bony tissues, which are time-consuming and not highly reproducible due to inter-observer variability, particularly for anatomies more complex than the prostate.
Patch-based methods, originally developed for image segmentation[46], have more recently been proposed for sCT generation[23], [24], [26], [28], [29]. These methods can be categorised into two types:

1- Approaches that involve non-overlapping partitioning of MRI-CT images from a training cohort, followed by a step to model the relationships between MRI and CT intensities.
2- "Non-local mean" approaches.

Both of these methods require a multimodal intra-patient CT-MRI registration and a monomodal inter-patient MRI registration of the learning cohort.
The first approach involves calculating image descriptors from the MRI scans of the cohort. These image descriptors, along with their corresponding CT scans, are divided into non-

overlapping patches. Patches located at the same position for each patient are grouped together to form a stack. Algorithms are then trained on these stacks to establish the relationships between MRI and CT voxels at specific locations. These trained models are subsequently applied to the patient's MRI to generate the corresponding sCT. Huynh et al. were the first to propose this methodology for sCT generation[28]. Structured random forests were used as the algorithms to model the relationships between MRI and CT voxels using image and CT descriptor stacks. These models were then applied to the patient's MRI to generate their sCT. The evaluation results of the generated images were found to be consistent with those reported in the literature, although no detailed dosimetric studies were conducted as part of this work.

The second approach employs sliding windows to extract patches from MRI and CT images with overlap. The MRI and CT images are registered using affine transformations, both from the training cohort and the patient's MRI. Patch extraction is limited to the surrounding area of the current voxel. For a specific patch from the patient's MRI, the k nearest patches from the MRIs in the training cohort are selected based on Euclidean distance. The CT patches corresponding to these selected patches are combined to generate the sCT. Andreasen et al. introduced this methodology for sCT generation in the pelvic region and brain[23], [29]. Largent et al. [40] (Figure 1.5) went further by including a multipoint-wise aggregation scheme to generate the sCT patches.

The advantage of patch-based methods is that they can generate accurate sCTs without requiring complex inter-patient deformable registration. By using affine registration instead of deformable registration, patch-based methods are computationally less intensive compared to atlas-based methods. However, the effectiveness of these approaches relies on the quality of the affine registration and the efficiency of the k-nearest neighbour search method employed. Additionally, the quality of the deformable intra-patient CT-MRI registration also impacts the performance of patch-based methods. It's important to note that the computation time for patch-based methods is still too high to allow real-time dose calculation in a standard radiotherapy workflow or adaptive radiotherapy.

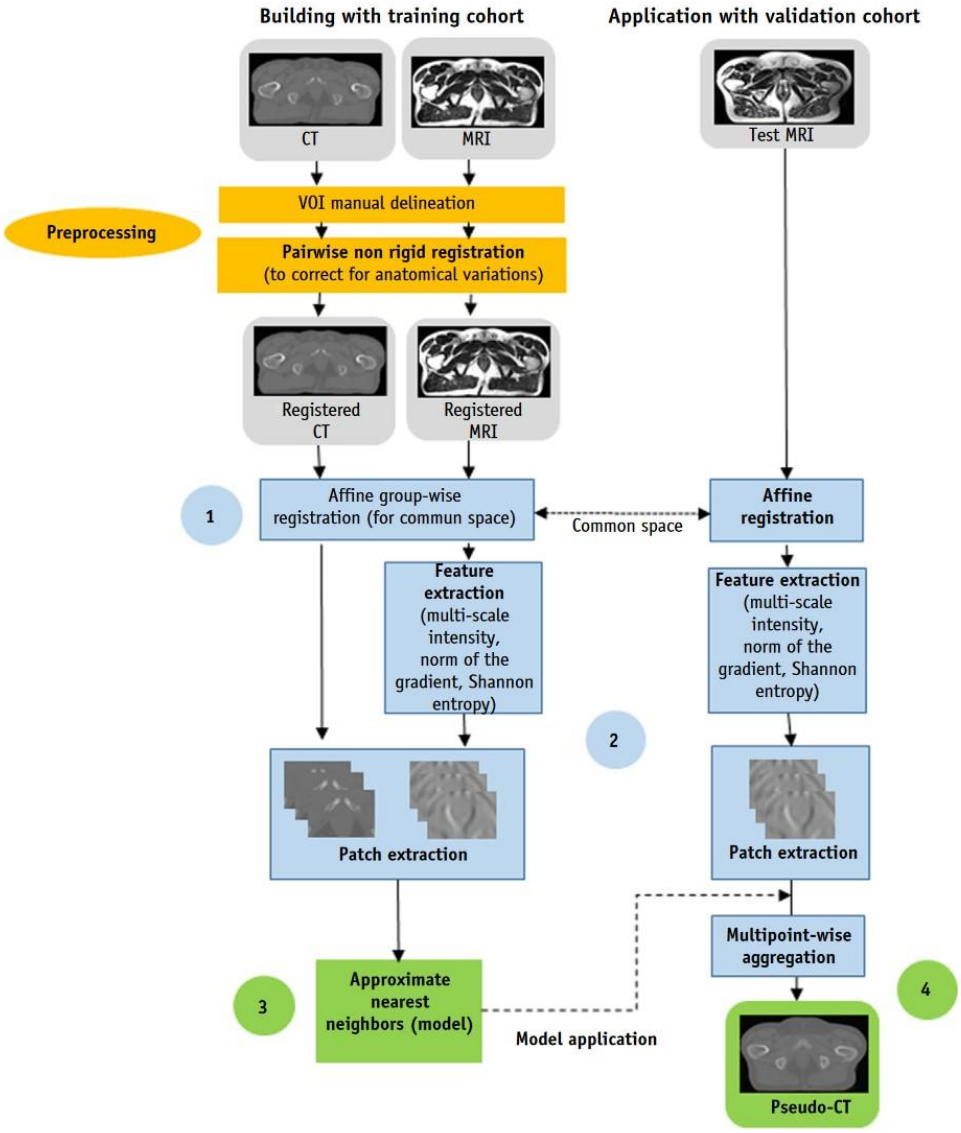

Figure 1.5:Patch-based method workflow for synthetic-CT generation from MRI. (Figure from Largent et al., published in "International Journal of Radiation Oncology, Biology, Physics" [40]) The building part involves training an approximate near-neighbour model using patches from the training CT and MRI (steps 1-3). In the application part, the trained model is used to generate a pseudo-CT from a test MRI (step 4).

Synthetic-CT generation using deep-learning models (DLMs)

Deep learning is a subcategory of machine learning methods. However, due to the growing interest in these approaches, the generation of synthetic-CT using deep learning-based models will be treated separately.

Artificial intelligence (AI) encompasses techniques that seek to automate cognitive tasks performed by humans using machines. Statistical learning or machine learning is a subset of AI. The process consists of a learning phase and an application phase.

During the learning phase, the model's optimal parameters are iteratively determined using relevant data and optimisation algorithm such as the Adam optimization algorithm. This data

is extracted from a dataset that is relevant to the problem being solved. The goal of this phase is to optimise the model for subsequent application to new data in order to perform the desired task.

Neural networks are a specific type of machine learning method that aims to mimic the functioning of the human brain. These networks are composed of interconnected artificial neurons, which are mathematical functions. When these neurons are connected, they enable the network to carry out tasks in a similar way to biological neurons and synapses. The architecture of a neural network refers to how these artificial neurons are connected to each other.

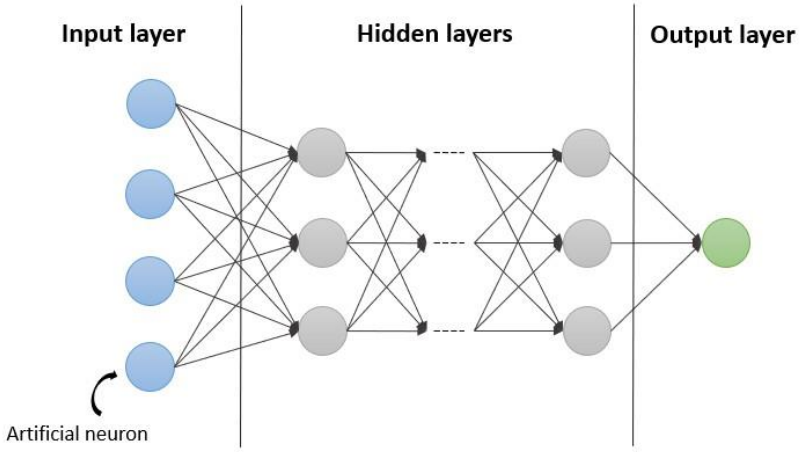

Figure 1.6: Multilayers perceptron architecture

Figure 1.6 illustrates the architecture of a neural network known as a multilayer perceptron, which is the most well-known architecture. The first layer of this network reads the input data, which are the explanatory variables of the model. The hidden layers capture the relationships between the input data and the data to be predicted. The output layer represents the result of the network, providing its prediction or estimate of the data to be predicted.

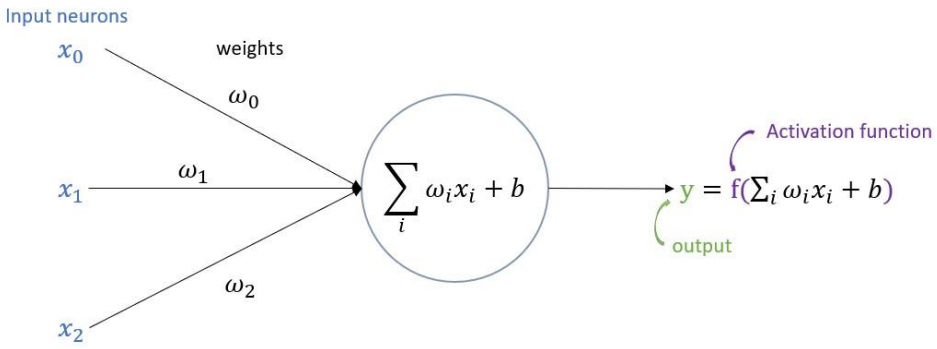

Figure 1.7: Illustration of an artificial neuron

Figure 1.7 presents a mathematical representation of an artificial neuron. The signals x_0, x_1 and x_2 represent the neuron's input data, which come from the previous layers. These signals are weighted by ω_0, ω_1 and ω_2 respectively. During training, these weights are adjusted to enable the network to effectively predict the output data. The weighted sum of these signals $\sum_i(\omega_i x_i)$ is calculated by adding a bias term b. An activation function f is then applied to the result of this sum to obtain the neuron's output data. This activation function represents a threshold at which the neuron emits an output signal.
Deep learning is a category of neural networks that utilise a large number of hidden layers. These algorithms have recently been proposed for generating sCTs from MRIs used in radiotherapy[1], [12]. One advantage of deep learning methods is that they do not require deformable inter-patient registration. However, in most cases, they still rely on multimodal intra-patient CT-MRI registration[47].

The state of the art of dose calculation methods from MRI images using DLMs will be presented here in the form of an article. This article, published in the journal *Physica Medica in 2021,* was written in collaboration with Marion Boulanger and Safaa Tahri.
As part of this review, I was involved in collecting and sorting the publications presented, while examining the metrics used in these studies. I also revised the manuscript.

Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review

M. Boulanger[a], Jean-Claude Nunes[a,*], H. Chourak[a,c], A. Largent[b], S. Tahri[a], O. Acosta[a], R. De Crevoisier[a], C. Lafond[a], A. Barateau[a]

[a]Univ. Rennes 1, CLCC Eug`ene Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
[b]Developing Brain Institute, Department of Diagnostic Imaging and Radiology, Children's National Hospital, Washington, DC, USA
[c]CSIRO Australian e-Health Research Centre, Herston, Queensland, Australia

Physica Medica 89 (2021) 265–281

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review

M. Boulanger [a], Jean-Claude Nunes [a,*], H. Chourak [a,c], A. Largent [b], S. Tahri [a], O. Acosta [a], R. De Crevoisier [a], C. Lafond [a], A. Barateau [a]

[a] Univ. Rennes 1, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
[b] Developing Brain Institute, Department of Diagnostic Imaging and Radiology, Children's National Hospital, Washington, DC, USA
[c] CSIRO Australian e-Health Research Centre, Herston, Queensland, Australia

ARTICLE INFO

Keywords:
Deep learning
MRI
Synthetic-CT
Radiation therapy
Dose calculation

ABSTRACT

Purpose: In radiotherapy, MRI is used for target volume and organs-at-risk delineation for its superior soft-tissue contrast as compared to CT imaging. However, MRI does not provide the electron density of tissue necessary for dose calculation. Several methods of synthetic-CT (sCT) generation from MRI data have been developed for radiotherapy dose calculation. This work reviewed deep learning (DL) sCT generation methods and their associated image and dose evaluation, in the context of MRI-based dose calculation.

Methods: We searched the PubMed and ScienceDirect electronic databases from January 2010 to March 2021. For each paper, several items were screened and compiled in figures and tables.

Results: This review included 57 studies. The DL methods were either generator-only based (45% of the reviewed studies), or generative adversarial network (GAN) architecture and its variants (55% of the reviewed studies). The brain and pelvis were the most commonly investigated anatomical localizations (39% and 28% of the reviewed studies, respectively), and more rarely, the head-and-neck (H&N) (15%), abdomen (10%), liver (5%) or breast (3%). All the studies performed an image evaluation of sCTs with a diversity of metrics, with only 36 studies performing dosimetric evaluations of sCT.

Conclusions: The median mean absolute errors were around 76 HU for the brain and H&N sCTs and 40 HU for the pelvis sCTs. For the brain, the mean dose difference between the sCT and the reference CT was <2%. For the H&N and pelvis, the mean dose difference was below 1% in most of the studies. Recent GAN architectures have advantages compared to generator-only, but no superiority was found in term of image or dose sCT uncertainties. Key challenges of DL-based sCT generation methods from MRI in radiotherapy is the management of movement for abdominal and thoracic localizations, the standardization of sCT evaluation, and the investigation of multicenter impacts.

Introduction

In radiation therapy, computed tomography (CT) is the standard imaging modality for treatment planning. Magnetic resonance imaging (MRI) is a complementary modality to CT providing better soft-tissue contrast without irradiation. MRI improves the delineation accuracy of the target volume and/or organs at risk (OARs) in the brain, head-and-neck (H&N), and lung or prostate radiotherapy [1–3]. However, MRI does not provide information on the electron density of the tissue, requires for accurate dose calculation. Most of the literature has proposed the generation of synthetic-CT (sCT) images for MRI-based dose planning. sCT (or pseudo-CT) is a synthetic image in Hounsfield Units (HU) generated from MRI data.

The methods for generating sCTs can be divided into three categories: bulk density, atlas-based and machine learning (ML) methods (including classical ML methods and deep learning methods [DLMs]). The bulk density methods consist of segmenting MRI images into several classes (usually air, soft-tissue, and bone). Each of these delineated volumes is assigned a homogeneous electron density, and the dose can then be calculated. This method has several drawbacks: it is tedious, time-consuming, operator-dependent, and does not consider tissue heterogeneity [4–8]. The atlas-based methods involve complex, non-

* Corresponding author.
E-mail address: jean-claude.nunes@univ-rennes1.fr (J.-C. Nunes).

https://doi.org/10.1016/j.ejmp.2021.07.027
Received 11 February 2021; Received in revised form 15 July 2021; Accepted 19 July 2021
Available online 30 August 2021

rigid registrations of one or several co-registered MRI-CT atlases with a target MRI. This registration step is followed by a fusion step to generate the sCT. The drawbacks of this method are the lack of robustness in the case of large anatomical variations and the need for computationally intensive pairwise registrations [4,5,9,10]. Among the classical ML methods, the patch-based methods (such as [4]) can be decomposed into four steps. The first step is interpatient rigid or affine registration with MR images. These methods involve inter-patient registration, feature extraction, and patch partitioning during the training step. The training patches closest to the patches of the target MRI are then selected for aggregation to generate the sCT [4]. The main drawbacks of this method are the imprecise interpatient registration and calculation time.

DLMs are models comprising multiple processing layers that learn multiscale representations of data through multiple levels of abstraction [11]. These methods have recently been introduced in radiotherapy for applications, including image segmentation, image processing and reconstruction, image registration, treatment planning, and radiomics [12–19]. DLMs have been proposed for sCT generation from MRI. They were trained to model the relationships between HU CT values and MRI intensities. Once the optimal DL parameters are estimated, the model can be applied to a test MRI to generate its corresponding sCT. DLMs have the advantage of being fast for sCT generation, and some do not require deformable inter-patient registration (only intra-patient registration) such as in [20].

Two reviews, both published in 2018, have already summarized sCT generation methods from MRI [21,22], they focused only on the bulk density, atlas-based, and voxel methods and did not include recent DLMs. Other studies have listed sCT generation methods from MRI in the context of MR-only radiotherapy [2,23–25]. More recently, Wang et al. [26] proposed a review on medical imaging synthesis using DL and Spadea and Maspero et al. [27] a review on sCT generation with DLM from MR, CBCT and PET images.

This study aimed to review literature studies using DLMs for MRI-based dose calculation in radiation therapy. This paper reviews the DL networks (with the loss functions), the image and dose endpoints for evaluation and the results per anatomical localization.

Materials and methods

We searched the PubMed and ScienceDirect electronic databases from January 2010 to March 2021 (date of first online release) using the following keywords: "deep learning", "substitute CT" or "pseudo CT" or "computed tomography substitute" or "synthetic CT", "MRI" or "MR" or "magnetic resonance imaging", "radiation therapy" or "radiotherapy". Mesh terms used in PubMed were: "radiotherapy", "Magnetic Resonance Imaging", and "deep learning". The search string on PubMed was: "MRI" AND "radiotherapy" AND ("GAN" OR "CNN" OR "deep learning" OR "machine learning" OR "U-Net" OR "neural network") NOT "radiomics" NOT "chemotherapy" NOT "brachytherapy" NOT "Positron Emission Tomography Computed Tomography" NOT "chemoradiotherapy" NOT "segmentation" NOT "reconstruction". We only retained original research papers (no abstract, no review paper) that reported data obtained from humans, were written in English, and addressed DL sCT generation from MRI in radiotherapy.

For each paper, we screened: anatomical localization, MR device, MR sequence, pre or post-treatment, use of registration, number of patients included in the study, type of DL network, loss functions, number of patients for training step, number of patients for evaluation step, main image and dose evaluation results. Tables per anatomical localization (brain, H&N, breast-liver-abdomen, and pelvis) were created to compile these information.

Results

Fig. 1 summarizes the number of DL studies for sCT generation from MRI in radiation therapy per year and anatomical localization. The first study was published in 2016 [28] and, at the time of manuscript submission, a total of 57 articles meeting the selection criteria had been published. Some studies investigated sCT generation for several anatomical localizations [29–33].

In total, 24 studies were based on brain data, 9 on H&N data, 2 on breast data, 3 on liver data, 6 on abdomen data, and 18 on pelvic data.

A. Common deep learning networks for sCT generation from MRI

Deep learning, as a mainstream of ML method, uses trainable computational models containing multiple processing components with adjustable parameters to learn a representation of data. Many DL network architectures have been developed, depending on specific applications or learning data. Several reviews have detailed the DL network architectures for radiotherapy or medical imaging [12,26,27,34–37]. The DL architecture for sCT generation from MRI can be roughly divided into two classes: generator-only and generative adversarial network (GAN) and its variants (such conditional-GAN, Least square GAN and cycle-GAN). Fig. 2 shows the hierarchy of the DL architectures.

1. Generator-only models

i. Basic concepts of convolutional neural networks (CNN)

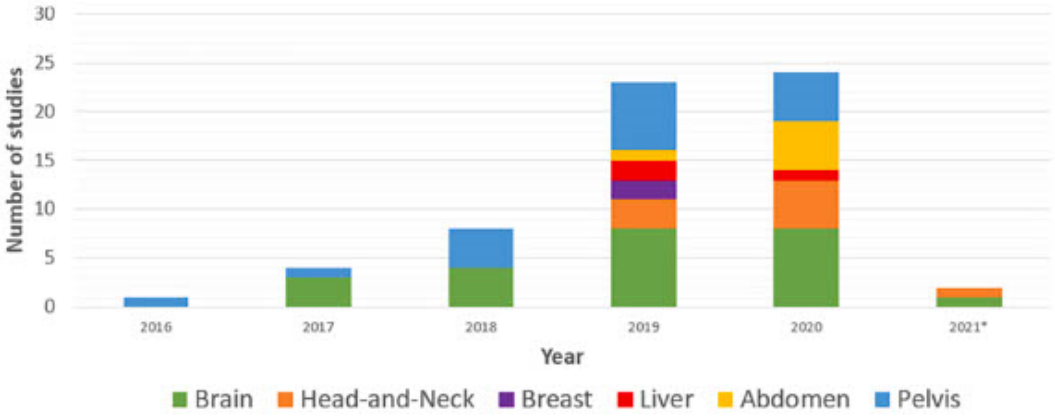

Fig. 1. Numbers of publications on deep learning methods for synthetic-CT generation from MRI in radiation therapy per year and anatomical localization. *: ongoing year (to March 2021), number of studies at the time of publication.

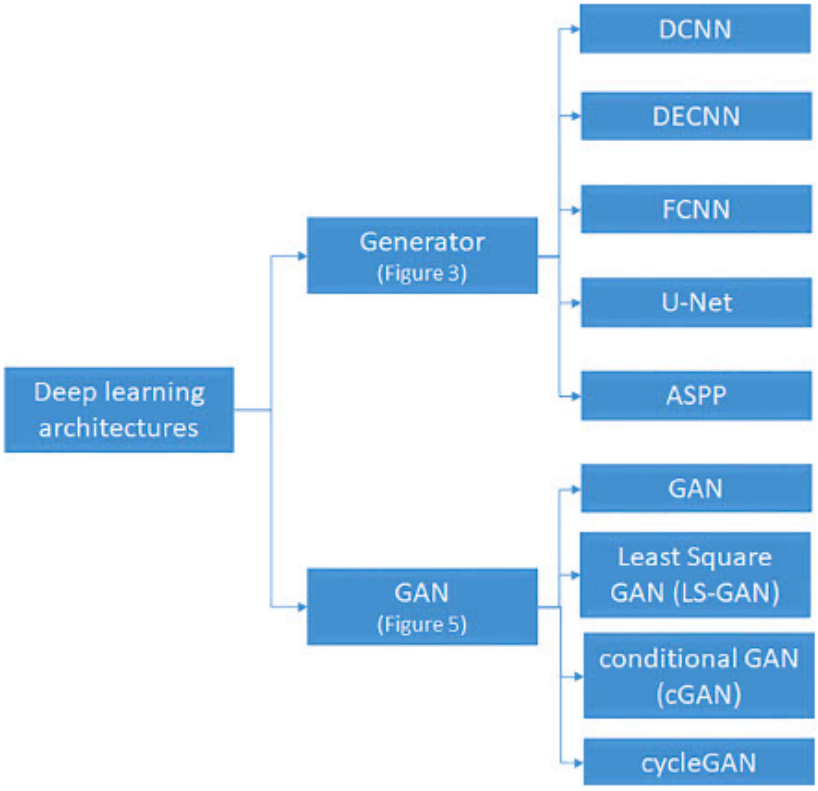

Fig. 2. Hierarchy of deep learning architectures. Deep learning (DL) architectures can roughly be divided into generator-only and generative adversarial network (GAN). In generator-only different DL architectures are included such as deep convolutional neural network (DCNN), deep embedding CNN (DECNN), fully CNN (FCNN), U-Net, or atrous spatial pyramid pooling (ASPP). GAN family includes GAN, and its most popular variants: Least Square GAN (LS-GAN) conditional GAN (cGAN), and cycle-GAN.

For image applications, a convolutional neural network (CNN, or ConvNet) is a popular class of deep neural networks using a set of convolution kernels/filters for detecting image features. A CNN consists of an input layer, multiple hidden layers and an output layer. The hidden layers include layers that perform convolutions with trainable kernels. Nonlinear activation functions (Rectified Linear Units (ReLU) [38], Leaky-RELU [39], Parametric-ReLU (PreLU) or exponential linear unit (ELU) [40]) play a crucial role in discriminative capabilities of the deep neural networks. The ReLU layer preserves the input otherwise is the most commonly used activation layer due to its computational simplicity, representational sparsity, and linearity. It is commonly to periodically insert a pooling layer between successive convolutional layers in a CNN architecture. Pooling layers allow to reduce the dimension (subsampling) of the feature maps. These maps are generated by following the convolutional operations. The pooling methods performs down-sampling by dividing the input into rectangular pooling regions and computing the average, the maximum, or the minimum of each region represented by the filter (mean pooling, max-pooling, min-pooling). Batch normalization [41] layers are inserted after a convolutional or fully connected layer to improve the convergence of the loss function during gradient descent (optimizer). It prevents the problem of vanishing gradient from arising and significantly reduces the time required for network convergence. After several convolution and pooling layers, the CNN generally ends with several fully connected layers. Dropout is one of the most promising techniques for regularization of CNN. Softmax layer is typically the final output layer in a neural network that performs multi-class classification (for example: object recognition).

ii. Generator-only models

The generator model can be considered as representing a complex end-to-end mapping function that transforms an input MR image to its corresponding CT image. During the training phase, the generator tries to minimize an objective function called a loss function (voxel-wise loss function L_G), which is an intensity-based similarity measurement between the generated image (sCT) and the corresponding ground truth image (real CT). Fig. 3 presents the global architecture of generator-only model.

In sCT generation from MRI, the generator architectures are generally based on convolution encoder-decoder networks (CED). In the literature, the variants of generator model include deep CED network [42], deep embedding CNN (DECNN) or Embedded Net [30], fully convolutional network (FCN) [28], U-Net [20,42–57 56,58,59], efficient CNN (eCNN) model [60], ResNet [61], SE-ResNet [61,62], and DenseNet [63]. Fig. 4 presents some architectures of CED-based generators (Fig. 4).

The CED network consists of a paired encoder and decoder networks. CED have been extensively used in DL literature thanks its excellent performance. In the encoding part, low-level feature maps are down-sampled to high-level feature maps. In the decoding part, the high-level feature maps are upsampled to low-level feature maps using the transposed convolutional layer to construct the prediction image (sCT).

The encoder network uses a set of combined 2D convolution filtering (no dilated convolutions) for detecting image features, followed by normalization (instance [66] or batch normalization [41]), a nonlinear activation function (ReLU [38], LeakyRELU [39], or PreLU), and max-pooling.

The decoder path combines the feature and spatial information through a sequence of symmetrical transpose convolutional layers (up-convolutions), up-sampling operators, concatenate layer (concatenations with high-resolution features), and convolutional layers with a ReLU activation function.

The most well-known and popular CED variants for biomedical

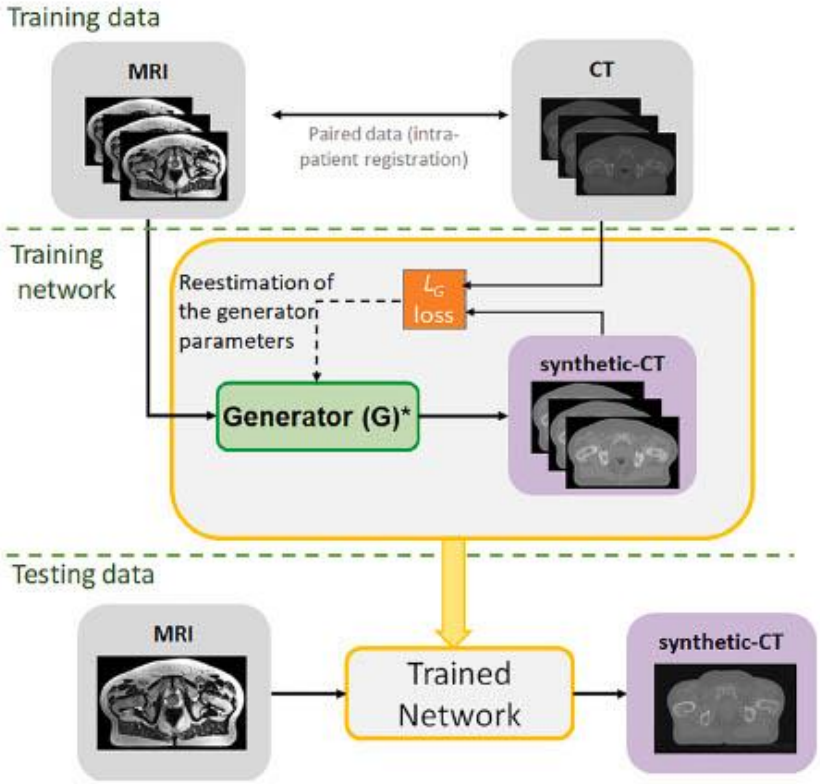

Fig. 3. Illustration of generator-only model. *: Generator model varies according to networks. The generator models often based on convolution encoder-decoder networks (CED) are trained to produce synthetic CTs (sCTs) from MRI. For this purpose, a single loss function L_G between MRI and registered CT images is computed. In the testing step, for a new given test patient, the MRI goes through the trained network to obtain the corresponding sCT.

image applications is the U-shaped CNN (U-Net) architecture proposed by Ronneberger et al. [67]. The U-Net [67] has a CED structure with direct skip connections between the encoder and decoder. Han et al. were the first to publish a sCT study with a U-Net architecture [44] that is similar to Ronneberger's model. This 2D U-net model directly learns a mapping function to convert a 2D MR grayscale image to its corresponding 2D sCT image. Han et al. study [44] differs from the original U-net since the three fully connected layers were removed. Thus, the number of parameters is reduced by 90%, and the final model is easier to train. In Wang et al. [46], the U-net model used batch normalization [41] and leaky ReLU, which was different from the classical U-net [67].

The DECNN model proposed by Xiang et al. [30] is derived by inserting multiple embedding blocks into the U-net architecture. This embedding strategy helps to backpropagate the gradients in the CNN and also provides easier and more effective training of the end-to-end mapping from MR to CT with faster convergence.

The efficient CNN (eCNN) model [60] was built based on the encoder-decoder networks in the U-Net model [67] where the convolutional layers were replaced with the building structures (aiming at extracting image features from the input MRI).

Some generative models use dilated convolutions called "atrous convolution" (rather than conventional convolutions) that expands the receptive field without loss of resolution or coverage [68]. Wolterink et al. [68] used a dilated CNN capturing larger anatomical context to differentiate between tissues with similar intensities on MR.

The ResNet architecture [61] has three convolutional layers (containing convolution operations, a batch normalization layer, a ReLU) activation function, followed by nine residual blocks (containing convolutional layers, batch normalization layers, and ReLU activation function) with fully connected layers. HighRes-net [69] consists of a CED architecture with residual connections, normalization layers, and rectified linear unit (ReLU) activations [38] using high-resolution ground truth (no pooling layers) as supervision with few trainable parameters [43]. The atrous spatial pyramid pooling (ASPP) generator [56] employs atrous or dilated convolution and is implemented in a similar U-Net architecture. The ASPP module permits a reduction in the total number of trainable parameters (almost divided by 4).

FCN better preserves the neighborhood information in the generated sCT images [28]. Compared to the conventional CNN, the pooling layers are not used in this task of image-to-image translation [28]. FCNs can simplify and speed network learning and inference and make the learning problem much easier. However, Fully connected layers are incredibly computationally expensive.

The deep CED network [42] consists of a combined encoder network

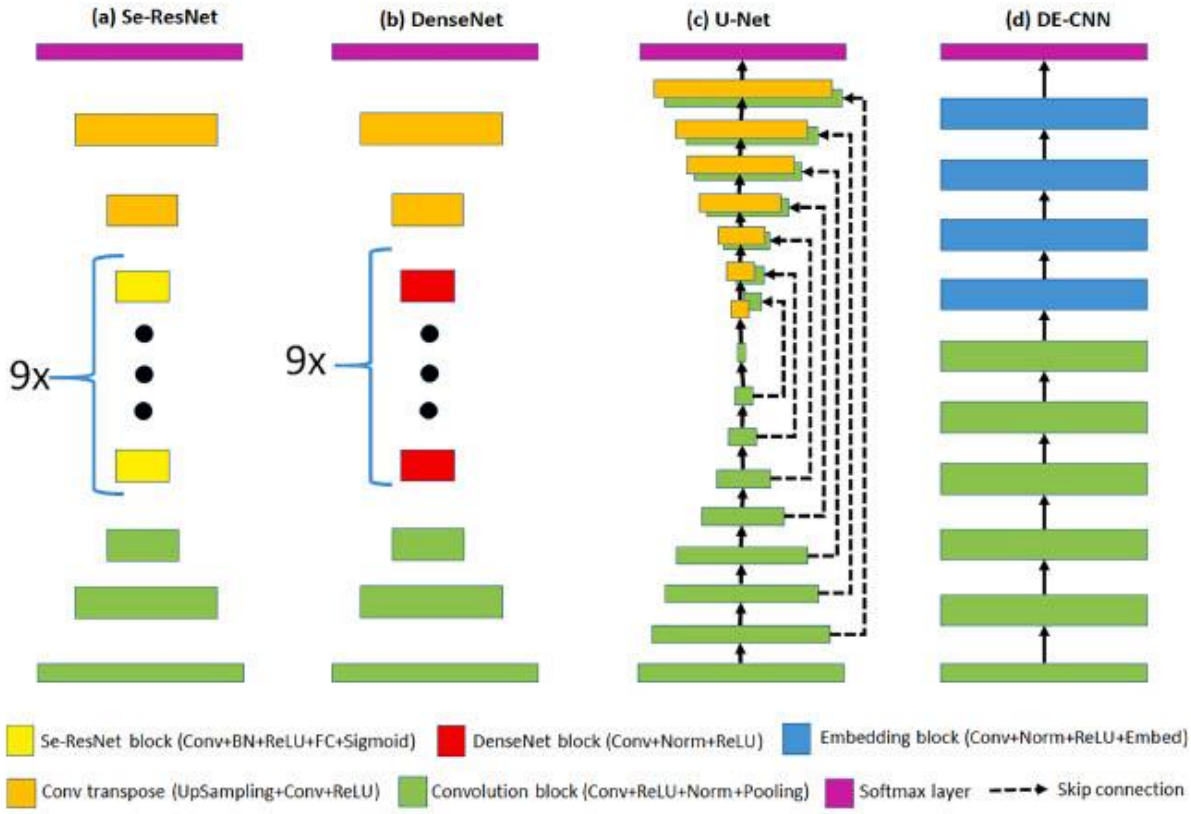

Fig. 4. Representation of generator architecture for U-Net [64] and adapted implementations of DenseNet [63], SE-ResNet [65], and Embedded Net [30]. The size of the boxes indicates the relative resolutions of the feature maps. The green boxes represent convolutional layers and orange boxes represent transposed convolutional layers. Yellow, red and blue boxes represent the SE-ResNet, DenseNet, and Embedded blocks. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(the popular Visual Geometry Group [VGG] 16-layer net model) and a decoder network (reversed VGG16) with multiple symmetrical shortcut connections between layers.

Twenty-nine state-of-the-art sCT image generation methods have adopted a generator-only network [20,28,30,42–57,70–79]. The loss functions L_G evaluating sCT and real CTs used in these generative models are:

- the mean square error (MSE), the L2-norm, or the Euclidean norm: only for sCT [20,42,46,47,55,57,78], for sCT and embedding blocks [30],
- the mean absolute error (MAE), mean absolute deviation (MAD), or L1-norm [43–45,49,52,53,70,71],
- a combined MAE and MSE loss [48],
- perceptual loss [20] based on VGG (the output of the 7th VGG16 convolutional layer).

The use of L2 distance as a loss function tends to produce blurry results. Perceptual loss is used to capture the discrepancy between the high frequency components within an image.

One limitation of generative models based on CNN is that they may lead to blurry results due to generally misalignment between MR and CT [80].

2. Generative adversarial network (GAN)

The following section summarizes GAN-based architectures to generate sCT from MRI. We introduce the GAN architecture and three most popular GAN-based extensions: least squares-GAN, conditional-GAN, and cycle-GAN.

i) GAN

The adversarial learning strategy was proposed by Goodfellow et al. [81] to generate better sCT images than previous generator-only models. The original way is to simultaneously train two separate neural networks (Fig. 5), the generator G (one of the generator-only models described in i) and Fig. 4) and the discriminator D. These two neural networks form a two-player min–max game where G tries to produce realistic images to fool D while D tries to distinguish between real and synthetic data [81,82]. Compared to generator-only models, GAN

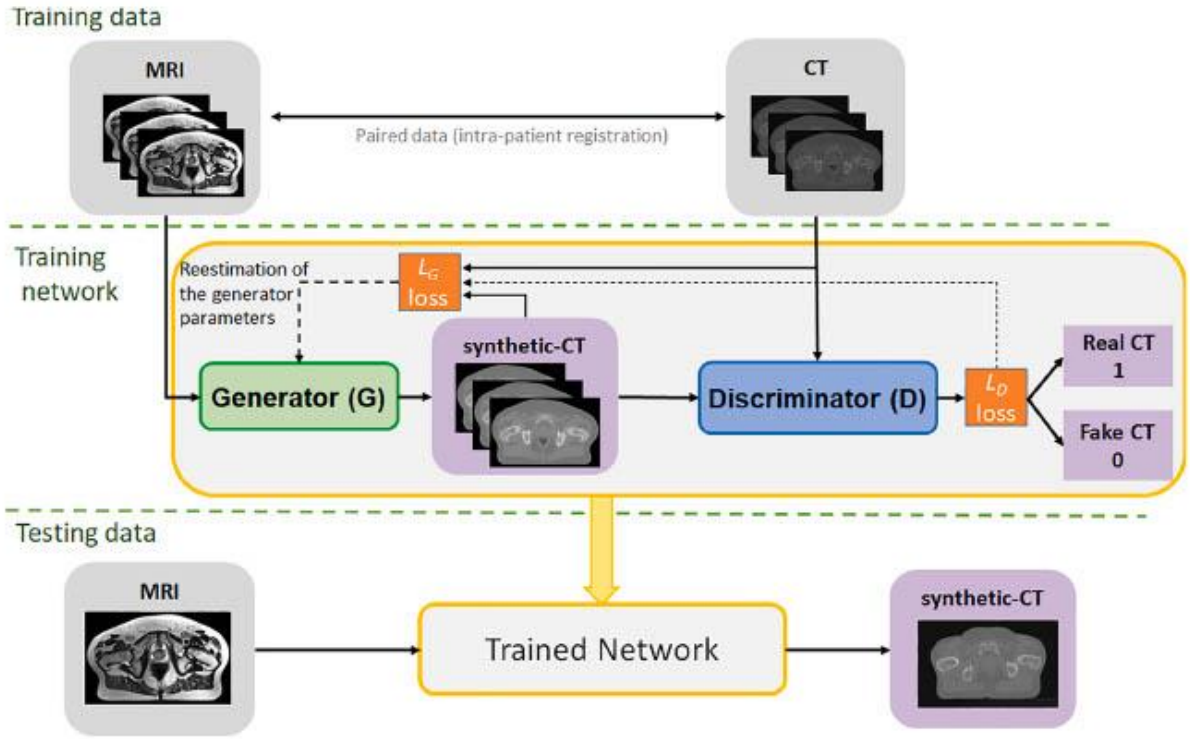

Fig. 5. Generative adversarial network (GAN) architecture. GAN consists of two adversarial CNNs. The first CNN, called the generator (such illustrated in Fig. 3), trained to synthetise images that resemble real images (such as real CT). The second CNN, called discriminator trained to differentiate fake image (synthetic image) from real images (which is considered a binary classification problem). The loss function L_D of the discriminator (called adversarial loss) is generally the binary cross-entropy. G and D are trained alternatively and share the same objective function of adversarial loss. The overall loss function L_G combined the adversarial loss and a voxel-wise loss function (measuring the similarity between the real CT and synthetic-CT voxels).

introduces a data-driven regularizer, the adversarial loss, to ensure that the learned distribution approaches the ground truth.

In the original version [81], the discriminator and generator are implemented as multilayer perceptrons (MLPs) and more recently implemented as CNNs. The architecture of the generator is often the conventional U-Net. Another proposed generator architecture in a GAN is ResNet [62] which is easy to optimize and can gain accuracy from considerably increased depth. The discriminator of the GAN [81] consists of six convolutional layers with different filter sizes but the same kernel sizes and strides, followed by five fully connected layers. ReLU was used as the activation function and a batch normalization layer for the convolutional layers. The dropout layer was added to the fully connected layers, and a sigmoid activation function was used in the last fully connected layer.

The discriminator used in [64] a convolutional "PatchGAN" classifier (markovian discriminator) models high frequency image structure in local patches and only penalizes structure at the scale of image patches.

Using adversarial loss L_D, the classical GAN model can generate high-quality sCT images with less blurry results [29,80] than generator-only models. The discriminator tries to maximize it while the generator tries to minimize it.

In this review, six studies used classical GAN-based architectures to generate sCT from MRI [20,29,62,77,83,84]. The overall loss functions intergrating the adversarial loss function L_D and evaluating sCT and the original CTs used in these GANs are :

- L2-norm alone [20,84],
- perceptual loss [20,83] and the multiscale perceptual loss [20].

The adversarial loss function L_D of the discriminator used in these GANs was generally the binary cross-entropy [29].

Perceptual regularization, used by Largent et al. [20], helps to prevent images over-smoothing and loss of structure details. The perceptual loss functions are based on high-level features extracted from pre-trained VGG network (7th VGG16 in [20]).

As shown by several studies [29,62,85], (1) the adversarial network prevents the generated images from blurring and better preserve details, especially edge features; (2) the accuracy of sCT within the bone region is increased; and (3) the discriminator detects patch features in both real and fake images, mitigating misregistration problem caused by an imperfect alignment between multi-parametric MRI and CT. General convergence in GANs is heavily dependent on hyperparameter tuning to avoid vanishing [86] or exploding gradients, and they are prone to mode collapse. To tackle the training instability of GANs, a plethora of

extensions and subclasses have been proposed.

ii) Least Squares-GAN (LS-GAN)

Most GANs use the binary cross-entropy as the discriminator loss function. However, this cross-entropy loss function leads to the saturation problem in GANs learning (the well-known problem of vanishing gradients [86]). Least square loss function strongly penalized the fake samples away from decision boundary and improve the stability of learning process. Mao et al. [87] adopted the least-squares loss function for the discriminator and showed that minimizing the objective function of LS-GAN minimizes the Pearson χ^2 divergence [88]. Emami et al. [62] replaced the negative log-likelihood objective with a least square loss function (L2 loss), which was more stable during training and generated better sCT quality.

iii) Conditional-GAN (cGAN)

Since the original GAN allows no explicit control on the actual data generation, Goodfellow et al. [81] proposed the conditional GAN (cGAN) to incorporate additional information such as class labels in the synthesis process. cGAN is an extension of the GAN model in which both the generator and the discriminator are conditioned on some additional information. The sCT image output is conditioned on the MR image input.

Different generator architectures in a cGAN have been proposed, including SE-ResNet [61,62], DenseNet [63], U-Net [56,58,59], Embedded Net [30], and the atrous spatial pyramid pooling (ASPP) method [56]. Fetty et al. [89] evaluated four different generator architectures: SE-ResNet, DenseNet, U-Net, and Embedded Net in a cGAN to generate sCT from T2 MRI. Olberg et al. [56] explored two generators: the conventional U-Net architecture implemented in the Pix2Pix framework [64] and the ASPP method [90,91]. The discriminator of the GAN framework was similar in both implementations.

Twenty studies used a cGAN architecture to generate sCT from MRI [31,33,50,56–59,88,89,92–101]. The overall loss functions L_G intergrating the adversarial loss function L_D and evaluating sCT and real CTs used in these cGANs were as follows:

- adversarial loss function (binary cross entropy) [59,101],
- L1-norm (MAE) [92],
- least squares loss function (L2 loss) [88,101],
- mutual information (MI) [58,59],
- focal regression loss [102] used in [99],
- the combination of adversarial (binary cross-entropy) and L2-norm [56],
- the combination of L1-norm and PatchGAN loss (as proposed by Isola et al. [64]) used in [50,89,93],
- the combination of adversarial (binary cross-entropy) and term derived from the log-likelihood of the Laplace distribution [95],
- the combination of Lp-norm, adversarial and gradient [33],
- the combination of multiscale L1-norm, L1 norm and PatchGAN loss [64] used in [88].

The loss functions L_D of the discriminator evaluating sCT and real CTs used in these cGANs areas follows:

- the mostly used adversarial loss (binary cross entropy) [56,58,59,88,93],
- least squares loss function (L2 loss) [88,94,101],
- L1-norm [101].

The L2-based loss function of the generator can cause image blurring. To alleviate blurriness and improve the prediction accuracy, the L1 norm [46] makes the learning more robust to outliers in the training data, such as noise or other artifacts in the images or due to imperfect matching between MR and CT images. The Markovian Discriminator loss or Patch-GAN loss [64], which can be understood as a form of texture/style loss, effectively models the image as a Markov random field, assuming independence between pixels separated by more than a patch diameter.

Pix2Pix proposed by Isola et al. [64] is a successful cGAN variant for high-resolution image-to-image translation. Pix2Pix model generally uses Unet generator and PatchGAN discriminator. As investigated by Isola et al. [64], the use of a loss function based on L1 alone leads to reasonable but blurred results; while cGAN alone leads to sharp results but introduces image artifacts. The authors showed that training in an adversarial setting together with an L1 norm generated sharp images with few artefacts (tissue-classification errors, especially for bone and air differentiation).

In Hemsley et al. [95], the L1 term in cGAN loss function [64] is replaced by a term derived from the log-likelihood of the Laplace distribution to capture data dependent uncertainty.

To overcome MR/CT registration issues, Kazemifar et al. [58,59] used a generator loss function based on the MI in cGAN. The MI loss allows the cGAN to use unregistered data to generate sCT and seems to accurately distinguish between air and bone regions.

Instead of the usual cross-entropy L_D loss in cGAN, Mao et al. [87] recommend the quadratic version of the least square GAN. Olberg et al. [56] evaluated a Pix2Pix framework with two different generators: the conventional U-net and a proposed generator composed of stacked encoders and decoders separated by dilated convolutions applied to increase rates in parallel to encode large-scale features. The overall loss function was composed of adversarial (sigmoid cross-entropy) and MAE losses.

Twelve studies used a Pix2Pix architecture [31,50,56,88,89,92–94,97,98,101,103]. Most of these Pix2Pix frameworks used only one MRI sequence as input and generated one sCT as output (called single-input single-output, SISO). A variant of Pix2Pix architecture proposed by Sharma et al. [103] is multi-input and multiple-output (MIMO) combining information from all available MRI sequences and synthesizes the missing ones.

One of the main advantages of cGANs is that the networks learn reasonable image-to-image translations even if the training dataset size is small. However, cGANs require coregistered MR-CT image pairs for training except with MI as loss function [58,59].

iv) Cycle-GAN

For image-to-image translations between two modalities, the principles of the cycle-GAN are to extract characteristic features of both modalities and discover the underlying relationship between them [104]. The cycle-GAN involved two GANs: one to generate sCT from MRI and a second to generate synthetic-MRI (sMRI) from sCT (the output of the first GAN). These dual GANs learn simultaneously and a cyclic loss function minimizes the discrepancy between the original CT and the sCT obtained from the chained generators.

Cycle GAN-based framework does not require paired MRI/CT images [80,105]. Wolterink et al. [80] found that training using unpaired images could, in some cases, outperform a GAN-model on paired images.

Eleven studies used a cycle-GAN architecture to generate sCT from MRI [32,33,57,77,80,100,101,105–108]. The overall loss functions L_G intergrating the adversarial loss function L_D and comparing the generated sCT and real CTs used in these cycle-GANs were:

- the combination of adversarial loss (cross-entropy) and L1-norm [33,101],
- the combination of the adversarial loss based on cross-entropy, the cycle consistency loss based on L1-norm, and the structural consistency loss based on L1-MIND [105] (the modality-independent neighborhood descriptor, MIND, introduced in [109]),

- the combination of L2-norm, adversarial loss (binary cross-entropy), the gradient difference loss and cycle consistency loss (based on L1 norm) [80],
- the combination of Lp-norm (mean P distance, MPD), adversarial loss and gradient loss [32,33,77,107].

Loss functions L_D of the discriminator used in cycle-GAN are:

- L2-norm (least squares loss) [62] as proposed in [87,110],
- MAD (L1-norm) [32,77,101,107],
- Lp-norm (MPD) [108].

Since L2-based loss functions tend to generate blurry images and L1-based loss functions may introduce tissue-classification errors, some authors [32,33,77,107,108] used an lp-norm ($p = 1.5$) distance, the MPD (Mean P distance). Using the MPD-based loss term, the authors also integrated an image gradient difference (GD) loss term (proposed in [29]) into the loss function [32,33,77,107,108], to retain sharpness in synthetic images, which maintain zones with strong gradients, such as edges. Cycle-GAN-based methods use MSE loss as distance loss function, which often leads to blurring and over-smoothing.

B. Data for sCT generation from MRI

1. MRI/CT image preprocessing and post-processing

In eighteen studies an MRI bias correction [20,30,32,33,43,44,47,49,78,83,89,92,94,101,105–108] was reported. In [30,32,44,47], intensity inhomogeneity (or non-uniformity) correction was performed in all MR images using the N3 bias field correction algorithm [111,112] to correct the bias field before training or synthesis. In [33,43,78,83,89,92,94,105–108], the authors reported that the intensity inhomogeneity of the MRI was corrected using the N4 bias field correction algorithm.

A 2D or 3D MRI geometry correction provided by the vendor was sometimes reported [49,57,70,105]. We can think that most of MR images had a geometry correction, but that it was not mentioned.

In [30,33,78,83,94], all MR images were normalized using a histogram-based intensity normalization [113] to minimize the inter-patient MR intensity variation. Intensity normalization was also used in [30,32]. In [44], all MR images were then histogram-matched to a randomly chosen template to help standardize image intensities across different patients using the method described by Cox et al. [114]. All MR volumes were normalized by aligning the white matter peak identified by fuzzy C-means in [105]. In [49,101], histogram standardizations performed using vendor-provided software (CLEAR) were applied as provided by the vendor.

In the study by Maspero et al. [93], the voxel intensity of CT was clipped within the interval HU to avoid an excessively large discretization step and the MR images were normalized to their 95% intensity interval over the whole patient. All the images were converted to 8-bits to conform to the Pix2Pix implementation [64]. Before training, the air cavities were filled in CT images and bulk-assigned (−1000 HU) as located in MR images using an automated method.

2. Training data characteristics

Compared to 2D CNN, 3D CNN can better model 3D spatial information (neighborhood information) owing to the use of 3D convolution operations [28] solving the discontinuity problem across slices, which are suffered by 2D CNN. However, the input type to DL models is mainly in 2D because fully 3D networks are much more difficult to train due to a large numbers of trainable parameters and requires exponentially more (GPU) memory and more data [28,44]. With the 2.5 D approach, Dinkla et al. [70] added 3D contextual information while maintaining a manageable number of trainable parameters. Furthermore, discontinuities across slices present in 2D methods, were decreased. Besides, the 2.5D approaches [45,70,71] include average axial, sagittal, and coronal images as input to train the CNN. In 3D (patch-based) CNN [28,32], an input MR image is first partitioned into overlapping patches. For each patch, the CNN is used to predict the corresponding CT patch and all predicted CT patches are merged into a single CT image by averaging the intensities of overlapping CT regions.

Most of the reviewed studies used one MRI sequence as input and generated one sCT as output; an architecture generally called single-input single-output (SISO). Four studies used several MRI sequences as input to generate one sCT in output [50,72,92,103], these architectures referred to as multi-input single-output (MISO) [50,72,92,103] or multi-input multiple-output (MIMO) [103]. Moreover, most studies used training and evaluation data from one MRI device while eight studies used multi-device MRI. One study reported use of MRI data from different centers [96] and two studies [88,89] used data from the Gold Atlas Data set [115]. Five studies used low MR field (0.35 T) as input images [31,33,56,73,89].

3. Training and evaluation of data size

The studies included in this review used several training strategies including k-fold cross-validation, single-fold validation, or leave-one-out. In k-fold cross-validation, the dataset is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are combined to form a training set. The average error across all k trials is then computed. In single-fold validation, the dataset is separated into two sets, the training and testing sets. The leave-one out strategy consists on k-fold cross-validation taken to its logical extreme, with k equal to N, the number of data patients in the set.

Data size is a fundamental challenge for DL approaches. There is no reported minimal or optimal data size for DL training. In the head area, four studies assessed sCT image quality as a function of the number of available images for training, from 15 to 242 patients for Alvares Andres et al. [43], from 5 to 47 patients for Gupta [48], from 34 to 135 patients for Peng et al. [100], and from 1 to 40 patients for Maspero et al. [96]. Better image results were found for higher numbers of available images. A minimum of 10 patients seems to be needed since it has shown similar performance than a training of 20, 30 or 40 patients. One effective way to improve model robustness is to enhance the diversity of the training dataset. Data augmentation is essential to teach the network the desired invariance and robustness properties when only a few training samples are available. One common data augmentation technique [32,44,92] is to apply random translations, rotations, zooms, and elastic deformations and adding low-level random noise to training images.

C. Evaluation metrics

sCT evaluation can be performed in terms of intensity, geometric fidelity, or dose metrics. A sCT evaluation was performed using intensity-based metrics for all reviewed studies and through dose criteria in 63% of the reviewed studies. The metrics used in the reviewed studies are listed in Table 1.

Table 1
Imaging and dose metrics used for the evaluation of synthetic-CT generation from MRI.

	Type of metrics	Metric	Definition	Ideal value
Image evaluation	Intensity-based metrics	ME: mean error	$ME = \frac{1}{N}\sum_{i=1}^{N} pCT_i - CT_i$	0 HU
		MAE: mean absolute error	$MAE = \frac{1}{N}\sum_{i=1}^{N} \lvert pCT_i - CT_i \rvert$	0 HU
		PSNR: peak signal to noise ratio	$PSNR = 10log_{10}\left(\frac{Q2}{MSE}\right)$	Maximum of dB
		SSIM: structural similarity metric	$SSIM = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\delta_{xy} + C_2\right)}{(\mu x^2+\mu y2+C_1)(\delta x2+\delta y2+C_2)}$	1
		MSE: mean square error	$MSE = \frac{1}{N}\sum_{i=1}^{N} (pCT_i - CT_i)2$	0
		RMSE: rot mean square error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (pCT_i - CT_i)2}$	0 HU
		NCC: normalized cross-correlation	$NCC = \frac{1}{N}\sum_{x,y,z} \frac{(I_{CT}(x,y,z) - \mu_{CT})\left(I_{pCT}(x,y,z) - \mu_{pCT}\right)}{\sigma_{CT}\sigma_{pCT}}$	
	Geometric fidelity metrics	DSC: dice score coefficient	$DSC = \frac{2\left(V_{CT} \cap V_{pCT}\right)}{V_{CT} + V_{pCT}}$	1
		HD: Hausdorff distance	$H(pCT, CT_{ref}) = max\left(h(pCT, CT_{ref}), h(CT_{ref}, pCT)\right)$	0 mm
		MASD: mean absolute surface distance	$MASD(A,R) = \frac{d_{ave}(S_A, S_R) + d_{ave}(S_R, S_A)}{2}$	0 mm
Dose evaluation	Dose difference metrics	Voxel-to-voxel dose differences	Difference between the dose distribution computed on the reference CT and on the sCT	0 Gy or 0%
		DVH difference	Dose differences on DVH specific points (D_{max}, D_{70Gy}, etc.), for a given structure	0 Gy or 0%
	Gamma analysis metrics	Mean gamma	Value of the mean gamma	0
		Gamma pass-rate	Percentage of pixels/voxels with a gamma value lower than 1	100%

Abbreviations: N: number of voxels; MSE: Mean square error; Q: range of voxel value of sCT and reference CT; x: reference CT; y: sCT; μ_x: mean value of x; μ_y: mean value of y; δ_x^2: variance of x; δ_y^2: variance of y; C_1 and C_2 are expressed as $(k_1Q)^2$ and $(k_2Q)^2$; I_{CT}: HU value of the reference CT, I_{sCT}: HU value of the sCT, μ_{CT}: mean intensity value of the reference CT, μ_{sCT}: mean intensity value of the sCT, σ_{CT} and σ_{sCT}: standard deviation of the reference CT and sCT; V: volume on CT and sCT; d_{ave}: absolute Euclidean distance; S_A: surface of the automated segmentation volume; S_R: surface of the reference organ delineation.

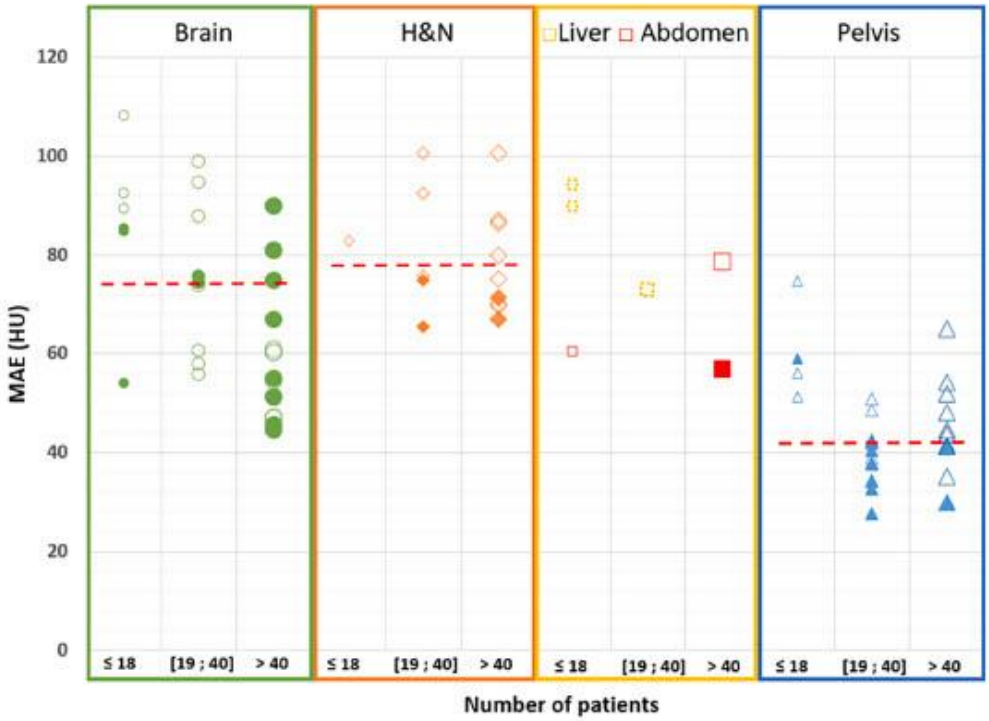

Fig. 6. Mean absolute error (MAE) results for body structure between reference CT and sCT generated with a deep learning method for studies including the brain, H&N, liver, abdomen, and pelvis Each marker represent a study result. Full markers represent generator-only models and empty markers generative models with adversarial. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. Red dotted lines represent the median values. The median values are: 74.2 HU for the brain, 77.9 HU for H&N, and 42.4 HU for the pelvis. The selected values are listed in the Additional tables 1 to 4 (Appendix A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

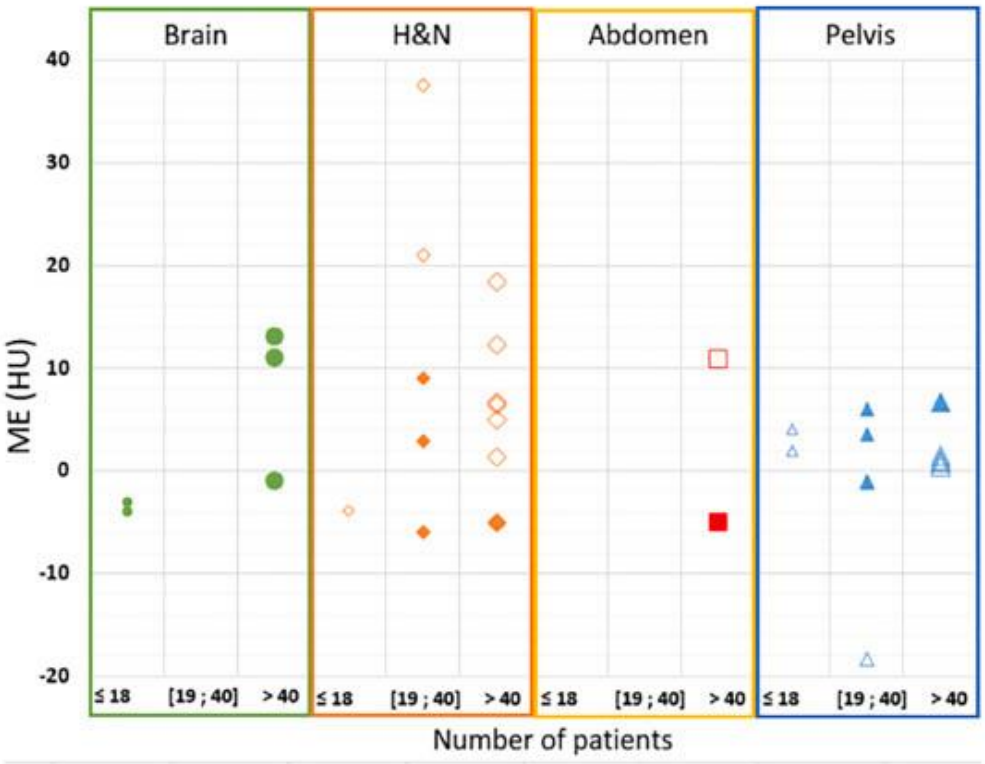

Fig. 7. Mean error (ME) results between reference CT and sCT generated with a deep learning method for studies including the brain, head and neck, abdomen, and pelvis Each marker represent a study result. Full markers represent generator-only models and empty markers generative models with adversarial. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. The selected values are listed in the Additional tables 1 to 4 (Appendix A).

1. Intensity-based evaluation

Only three studies of sCT generation from MRI did not reported MAE values [56,73,98]. Some articles reported MAE in bone or soft tissue while others reported MAE in anatomical structures such as the kidneys, bladder, or rectum [20,45,47,76,88]. Fig. 6 summarizes the MAEs of the studies on brain, H&N, liver, abdomen, and pelvis sCT generation from MRI.

Eighteen studies reported mean error (ME) results. Fig. 7 details the MEs of the studies on brain, H&N, abdomen, and pelvis sCT generation from MRI. For the pelvis, three studies provided ME values for the bladder, rectum and soft tissue [20,47,88]. Some studies have illustrated MAE or ME for one or several slices. Such difference maps allow for qualitative comparisons and spatial analyses.

The peak signal to noise ratio (PSNR) is the simplest and most widely used fidelity measure (full-reference quality metric), which is related to the distortion metric, the MSE. Twenty-two studies on sCT generation from MRI reported PSNR results. Fig. 8 details PSNR results for the brain, H&N, and pelvis sCT generation from MRI studies.

Four studies reported MSE values in the brain and pelvis. Only three studies reported the root MSE (RMSE) value in the brain [98], the breast [56], and the abdomen [50]. Although the MSE is an attractive measure due to its simplicity of calculation, MSE/PSNR can be a poor predictor of visual fidelity in images [116].

More sophisticated measures have been developed to take advantage of the known characteristics of the human visual system. Wang et al. [117] proposed a structural similarity metric (SSIM) to capture the loss of image structure due to variations in lighting (contrast or brightness changes). The SSIM captures image distortion as a combination of three types of distortion: correlation, contrast, and luminance.

2. Geometric fidelity evaluation

Geometric fidelity is based on delineated structures. Nineteen articles reported dice score coefficients (DSCs) between sCT and reference CT for bone, air, or body structures. One study reported DSCs for the bladder and rectum [47]. DSCs were between 0.85 and 0.99 for body and were higher than 0.68 and up to 0.93 for bone structure.

Only two studies reported Hausdorff distance (HD) values for the H&N area [79] and the pelvis [77]. Only one study reported mean absolute surface distance (MASD) values for body, bone, bladder, and rectum volumes [47]. Five studies reported normalized cross-correlation

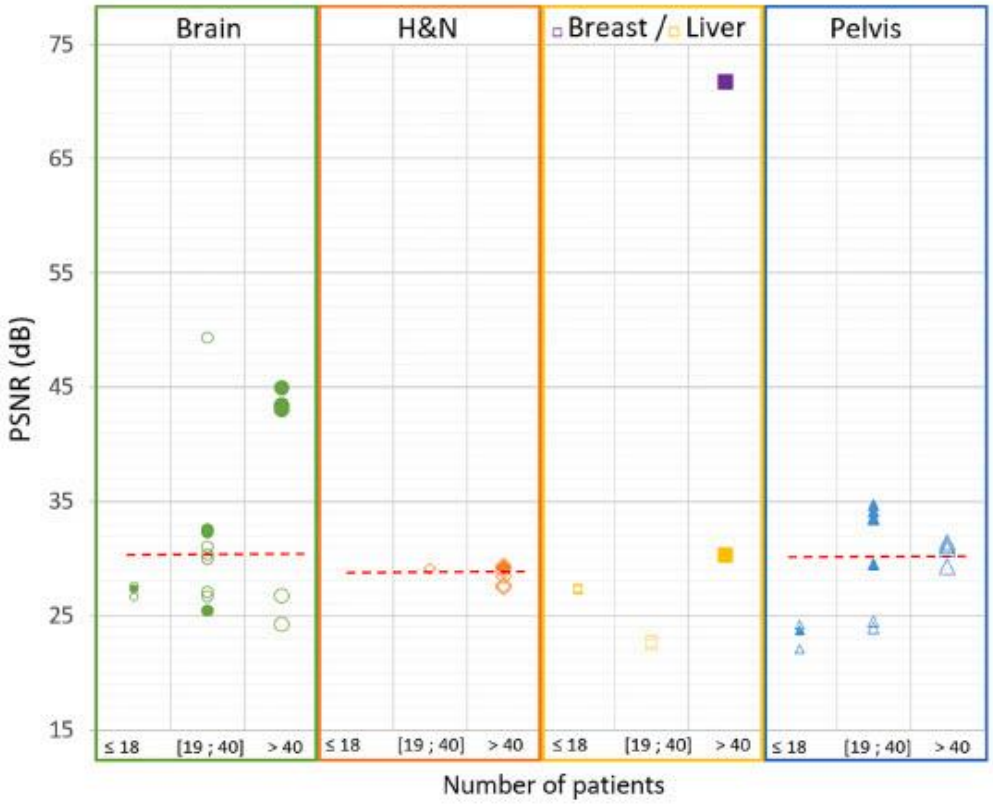

Fig. 8. Peak signal-to-noise ratio (PSNR) results between reference CT and sCT generated with deep learning methods in studies on the brain, head and neck, breast, liver, and pelvis Full markers represent generator-only models and empty markers generative models with adversarial. Results are divided into three categories: studies including less than 18 patients, studies including 19 to 40 patients and studies including more than 40 patients. Red dotted lines represent the median values. The median values are: 30.3 dB for the brain, 28.9 dB for H&N, and 30.2 dB for the pelvis. The selected values are listed in the Additional tables 1 to 4 (Appendix A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(NCC) values in the brain, liver, and pelvis [32,77,106–108].

The penultimate columns of Additional Tables 1, 2, 3, and 4 in Appendix A list the image results of sCT generated from MRI.

3. Dose evaluation

In MRI-only workflows for radiotherapy, a sCT is generated to perform dose calculation. In this context, studies have proposed dosimetric evaluation of dose calculation from sCT with DVH, voxel-to-voxel dose differences or gamma index analysis. Most studies evaluated dose calculation with photon particles, while nine studies investigates sCT dose uncertainties with protons [45,55,59,71,74,77,96,107,108].

v) Dose-volume histogram (DVH)

DVH is a widely used tool in routine clinical radiotherapy. All treatment planning systems allow for the analysis of dose distributions through DVHs. Twenty-two sCT studies reported dose differences at DVH specific points. Eighteen studies reported mean dose differences in selected volume (PTV, CTV, OAR).

vi) Voxel-to-voxel dose difference

The dose difference is defined as the difference between the dose distribution computed on the reference CT and the sCT. The dose difference can be expressed as absolute value (Gy) or relative to the reference dose (%).

Several studies reported mean absolute dose error to express dose uncertainties and mean dose error to express systematic dose uncertainties [47,49,50,70,77,96,101]. Some studies have provided dose differences using dose thresholds such as doses higher than 90% of the prescribed dose, while others have illustrated dose difference maps that allow qualitative and spatial analyses.

vii) Gamma index analysis

Gamma analyses allow spatial analysis (through gamma maps) of dose distributions calculated from sCT compared to those calculated from a reference CT [118]. Gamma analysis can be performed in two or three dimensions. This analysis combines dose and spatial criteria. Several parameters need to be set to perform a gamma analysis, including dose criteria, distance-to-agreement criteria, local or global analysis, and dose threshold. Interpretation and comparison between studies of gamma index results are challenging because they depend on

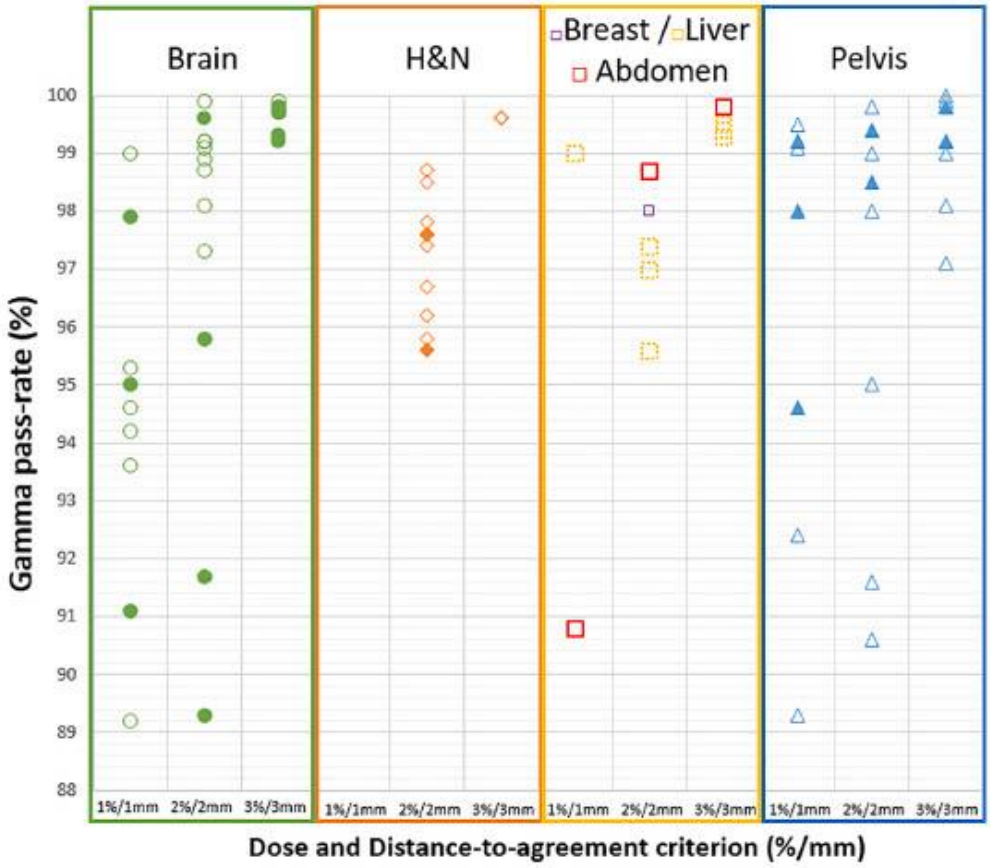

Fig. 9. Gamma pass-rate results between reference CT and sCT dose distributions with deep learning methods for studies including the brain, breast, liver, abdomen, and pelvis Full markers represent generator-only models and empty markers generative models with adversarial. The selected values are listed in the Additional tables 1 to 4 (Appendix A).

the chosen parameters, dose grid size, and voxel resolution [119]. The gamma results can be expressed as gamma pass-rate (percentage of pixels/voxels with a gamma value lower than 1) or mean gamma. Twenty-eight articles reported gamma pass-rate results. Only one study reported mean gamma values in the pelvis [20].

Fig. 9 summarizes the gamma pass-rate results between reference dose distribution and sCT dose distribution for several anatomical localizations. The mean gamma pass-rates were above 89% for all localizations and up to 100%, depending on gamma criteria.

viii) Specific metrics for proton dose calculation

Proton ranges along the beam paths were compared for dose distributions on the reference CT and sCT. In protontherapy, the range of the proton beam strongly depends on the stopping power ratio (SPR) of a given tissue relative to water, which can be determined using the electron density and effective atomic number through the Bethe-Bloch equation. The range is defined as at the 80% distal dose falloff along each beam direction. Several studies have reported the results of range shift or range difference (in mm) per beam [45,55,71,108].

The last columns of Additional Tables 1, 2, 3, and 4 in Appendix A summarize the dose results of the DL sCT generation studies for MRI dose calculation.

D. Image and dose results per anatomical localisation

1. Brain

Twenty-four studies of the brain were performed between 2017 and 2021. Additional Table 1 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies on brain sCT generation from MRI in radiotherapy. T1-weighted (T1w) sequences were mostly used for generating sCTs 88% of the reviewed studies).

For sCT evaluation, all brain studies reported MAEs, which varied from 44 to 129 HU for the whole brain (Fig. 6). For the brain, the MAEs for bone structure were above 174 HU and up to 399 HU in one study. Koike et al. [92] trained a cGAN network with only a T1 sequence or a combination of several sequences (T1w, T2w and FLAIR). The multi-sequence training showed a decrease in MAEs results for the body, soft tissue and bone (between −8 and −33 HU) [92]. Alvarez Andres et al. reported MAE values for CNN and U-Net networks, with higher values for the U-Net network than for the CNN network for the head (an increase of 9 HU) [43]. They also investigated the influence of several sequences (T1, T1-Gd, and T2 FLAIR images) as input in the CNN network. The MAE values were higher with a FLAIR sequence as input in the CNN network than those with T1 sequence (increase of 34 HU). The MAEs were also higher with contrast-enhanced T1-weighted MRI (T1-Gd) than those for T1w MRI (from + 3 to + 32 HU). Only four studies reported ME values [44,45,55,70], which ranged between −4 HU and

13 HU for the whole brain (Fig. 7). The PSNR values were above 24 dB for all brain studies [29,30,32,57,62,72,80,96,97,105] (Fig. 8). The DSCs were above 0.96 for the body, 0.69 for bone, and 0.70 for air structures [42,45,58,70,72]. For SSIM, values varied from 0.63 to 0.94 [57,62,72,96,105]. For NCC, two studies reported values of 0.96 [32,108]. Massa et al. [72] trained models on four different MR sequence: CUBE-FLAIR, T1, T1 post contrast and T2 fatsat. No sequence was statistically better on all the metrics (MAE, PSNR, SSIM, DSC).

Among the 24 brain studies, only 14 reported a dose evaluation [42,43,45,48,55,58,59,70,84,92,96–98,108]. Five studies reported results for proton dose planning [45,55,59,96,108]. All reported DVH mean dose differences were below 2% [42,45,48,59,70,84,92,96–98,108].

For most of the studies, gamma pass-rates were above 89% with the most restrictive criterion (1%/1 mm), and except for one study above 95% for other criteria (Fig. 9). One study [55] reported a mean gamma pass-rate of 89% with 2%/2 mm criteria. With the multi-sequence training, Koike et al. showed an increase in gamma pass-rates (between 0.1% and 1.1%), compared to single sequence training [92].

2. Head and Neck (H&N)

Nine DL sCT generation studies were performed in H&N radiotherapy. Additional Table 2 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies. The MRI sequences used in the H&N sCT studies were T1 and T2. Four studies used the Dixon reconstruction [49,50,79,101].

MAE and ME metrics have been widely reported in the literature. MAEs varied from 65 to 131 HU for the body or head structures (Fig. 6). For the H&N, the MAEs for bone were above 166 HU and up to 357 HU in one study [46]. Qi et al. [50] used multi-sequence input (T1w, T2w, contrast-enhanced T1, and contrast-enhanced T1 Dixon water) images to train a cGAN. The multi-sequence training showed a decrease in MAE for the body, soft tissue, and bone and an increase in PSNR, SSIM, and DSC. They also compared cGAN and U-Net networks. With cGAN and single-sequence input, the MAE, PSNR, and DSC were higher than those obtained using U-Net. For the body, the MEs were mostly around 0 HU, above −6 HU and up to 37 HU (Fig. 7). Five studies reported ME for air, bone, or soft tissue [9,46,49,79,100]. For bone structure, MEs were higher up to 247 HU. PSNR and SSIM were only reported in two studies [50,94]. The PSNR results were approximately 28 dB (Fig. 8). The SSIMs were between 0.78 and 0.92. For bone structure, DSC values were between 0.70 and 0.89 [50,70,71,79,94].

DVH dose difference was performed in nine studies, with a mean difference less than 1.6%. Klages et al. reported a mean dose (D_{mean}) to the parotid glands below 1% and the maximum dose (D_{max}) to the spinal cord below 1.5% [101]. In the only one protontherapy study, the dose differences reached 8% for some OARs [71].

The gamma pass-rates were above 95% for the most restrictive criterion (2%/2mm) and above 98% for the other criterion (3%/3 mm). With the multi-sequence training, Qi et al. showed non-significant gamma pass-rate results [50]. In the same study, they also found higher gamma pass-rates for cGAN than for U-Net architectures.

3. Breast

Two DL sCT generation studies were carried out for breast radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for breast sCT generation from MRI in radiotherapy. These two studies were based on MR images from a low field (0.35 T) MRI device.

Jeon et al. [73] only reported DSC values for two patients. Olberg et al. reported a PSNR of 72 dB, an SSIM of 0.999, and an RMSE of 17 HU [56]. For dose results, they reported gamma pass-rate higher than 98% with 2%/2 mm criteria.

4. Liver

Three DL sCT generation studies were carried out for liver radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for liver sCT generation from MRI in radiotherapy.

Two of the three liver studies used T1 sequence. The MAEs varied from 72 to 94 HU for body structure between studies. Fu et al. [33] compared cGAN and cycle-GAN DLMs for data from three patients. The MAEs were higher for cycle-GAN than for cGAN. The PNSR values were above 22 dB. The NCC values were 0.92 in two liver studies [106,107].

DVH dose difference were calculated in the three liver studies, with the mean differences below 1%. In one study, the dose difference in OARs was less than 0.6% [33]. The dose difference in PTV ($D_{95\%}$) was less than 1.1% [106,107]. The gamma pass-rates were above 90% for the most restrictive criterion (1%/1mm) and above 95% for the other criteria (Fig. 9). In the study by Fu et al. [33], the gamma pass-rates were higher for a cGAN DLM than those for a cycle-GAN DLM.

5. Abdomen

Six DL sCT generation studies were carried out for abdomen radiotherapy. Additional Table 3 (Appendix A) summarizes the DL networks and the image and dose metrics results of the studies for abdomen sCT generation from MRI in radiotherapy.

Acquisitions were performed in breath hold inspiration for two studies on 0.35 T MRI device [31,33].

The MAEs varied from 55 to 94 HU for body structure between abdomen studies. MAEs in lungs were 105 HU in two studies [74,76]. In Florkow et al. [74], PSNR value was 30 dB and DSC values were 0,76 for bone and 0.92 for lungs. Mean dose differences were lower than 1% and gamma pass-rate above 98% in 2%/2mm.

6. Pelvis

Eighteen DL sCT generation studies for the pelvis were performed between 2016 and 2020. Additional Table 4 (Appendix A) summarizes the DL networks and the image and dose metrics results of these studies.

Most MRI sequences in these pelvis studies were T2 sequences. T1 sequences [30,53,78] or Dixon reconstruction [52,54,93] were also used to generate sCT from MRI in the pelvis.

The reported MAEs were 27–65 HU for body structure (Fig. 6) and around 120 HU and up to 250 HU for bone [20,31,47,51,53,78,99]. Fu et al. compared training in 2D and 3D in four patients [78], reporting higher MAEs for 2D than for 3D training (+2–5 HU). Largent et al. compared U-Net and GAN networks with different loss functions [20]. With the L2 loss function, U-Net showed lower MAEs than those for GAN. For all studies, the MEs were generally near to 0 HU for the whole body structure (Fig. 7). One multicenter study reported an ME of −18 HU [88]. For the pelvis, the MEs in the bone were up to 141 HU in one study [31]. The reported PSNRs were between 24 and 34 dB (Fig. 8). Only one study reported SSIM in the pelvic area [75]. Only two studies reported DSC for the body (0.85 and 0.99) [47,77]. The DSCs for bone ranged between 0.70 and 0.93 [47,52,53,75,78]. Only one study reported a DSC for the bladder and rectum of 0.9 [47].

Among the 18 pelvic studies, only nine reported dose evaluations. Liu et al. [107] performed proton dose planning. Most studies reported dose differences below 1.5% for target volumes and OARs. Arabi et al. reported maximum dose differences below 0.5% for the bladder and rectum between 1.1% and 2.9% for the CTV and PTV. Some studies reported a very low dose difference (less than 0.6%) for PTV, bladder,

rectum, and femoral heads [20,51]. However, Liu et al. reported dose differences up to 5% in the rectum and up to 11% in the bladder [77].

The gamma pass-rates were above 89% for the most restrictive criterion and generally above 95% for the other criteria (Fig. 9). In a study using prostate data in training to generate sCT of the rectum and cervix [93], the gamma pass-rates were around 91% for gamma criteria of 2%/2 mm.

Discussion

This article reviewed deep learning methods used to generate sCT from MRI in radiation therapy, and their associated image and dose uncertainties. Two types of DL architectures are widely used; generator-only, and GAN. The most recent DLMs were cGAN and cycle-GAN. A variety of metrics for image evaluation (image intensity and geometric fidelity) has been proposed. The median MAE results were 76 HU for head localization (brain and H&N) and liver, and 42 HU for the pelvic area. Dose evaluations consisted in DVH comparisons, voxel-to-voxel dose differences, or gamma index analyses. The mean dose differences were below 1% in the H&N, liver, breast, and pelvis sCT studies. In brain sCT studies, the mean dose difference was below 2%. For most of the studies, the gamma pass-rates were above 95% (with 2%/2 mm and 3%/3 mm criteria) (Fig. 9).

In radiotherapy, the first sCT generation methods from MRI were bulk density and atlas-based. Other ML methods (non-DLM) have also been investigated, including patch-based or random forest [10,120]. This review focused on DLMs which are the most recent methods with the first study in the pelvic area reported in 2016. Different neural network architectures have been used in the literature with multiple parameters to be set. Compared to other sCT generation methods, DLMs have fast computation times, and do not necessarily require deformable inter-patient registration. sCT generation DLMs have just been commercially available for a clinical use [79,121]. To our knowledge, no open source software is available for sCT generation from MRI with a DLM. Each research team has developed his own DLM with hyper-parameter tuning. This review was not able to identify the most "accurate" DL architecture. Although GAN DLMs are the most recent, for now they do not outperform generator-only DLMs (Figs. 6, 7, 8, and 9). Moreover, we acknowledge that studies are not directly comparable due to the great disparities in input data (imaging protocol, scanner parameters, etc.), training cohort sizes, evaluation cohort sizes, and methods of evaluation. Same data should be used to directly compare the results, such as data in open access from the Gold Atlas Project [115]. Two studies used these data [88,89]. Some studies directly compared several DLMs with the same data (Additional Tables 1, 2, 3, and 4 in Appendix A). Size of patient cohort (training + evaluation) varied according to study and anatomical localization (Additional Tables 1, 2, 3, and 4 in Appendix A). The median number of patients were 45 for the brain, 33 for H&N and 23 for pelvic localization. Studies including few patients (less than 19) did not show the better results (Figs. 6, 7, and 8). But studies with more than 40 patients did not outperform compared with studies with 19 to 40 patients (Fig. 6). Training strategy depends on the number of available data. If you have few data (less than 20 patient data), a leave-one-out strategy is recommended. Image quality of data training is important. Image with artefacts must be removed of the training. A first step of quality image optimization (MR sequence or CT acquisition parameters) can be useful.

Although cycle-GAN or other networks do not require paired data for training step, paired data are required for the evaluation step. For the brain a rigid registration can be sufficient [32] but not for H&N or pelvis area. Few studies used unpaired data in training, even with cycle-GAN [80,105] and with MI as loss function in GAN [58,59]. To have paired data, a deformable registration is needed, with additional uncertainties. Florkow et al. [52] quantified the uncertainties due to MRI-CT registration.

To perform the evaluation, sCTs generated from MRI are compared to reference CT. Even if the time between acquisitions is kept as short as possible, MRI acquisition and reference CT can differ even after non-rigid registration, due to gas volatility and bowel loops displacement in the abdomen, artifacts (teeth, hip prosthesis, fiducials, contrast agent, etc.) or internal movements (bladder and rectum filling) between CT and MR images. Maspero et al. [93] proposed to override gas in the rectum as in the reference CT and performed the imaging evaluation in the intersection volume of the body contours (reference CT ∩ sCT). Cusumano et al. excluded some patients from their studies because of artifacts (artificial implants) or difference of air pocket locations between CT and MR images [31].

Most of studies reported global imaging evaluation metrics, without local or spatial analysis. Hemsley et al. [95] proposed a detailed sCT evaluation with uncertainty heatmaps. Models were proposed to spatially quantify intrinsic and parameter sCT uncertainties [122]. With this method, uncertainty maps are a second output of the DL network [123]. Moreover, an analysis based on image gradient could be performed.

Reviewed articles aimed to show accuracy of sCT generation from MRI compared to a reference CT. In the future, we can imagine an MRI-only workflow without any CT acquisition. In this case, image evaluation metrics without reference need to be developed. Before any use in clinical practice, commissioning and quality assurance process must be implemented [124]. Practical guidelines on the use of MRI for external radiotherapy treatment planning were recently proposed by a multidisciplinary working group of the Institute of Physics and Engineering in Medicine (IPEM) [125]. This document overviews all the aspects of MRI implementation for radiotherapy are described (MR safety, training and education, patient set-up, MRI sequence, MR quality assurance, etc.).

To date, few DL studies have been carried out on abdomen, liver, breast, or H&N radiotherapy. This limited number may be due to the small number of patients undergoing MRI for liver or abdomen radiotherapy compared to brain or prostate radiotherapy. Moreover, standard acquisition of breast MRI is not in radiotherapy treatment position. Number of breast studies are increasing with the availably of MR acquisitions from MRI-linac device. The lack of sCT generation from H&N and abdomen MRI may be due to the complexity of these anatomical localizations with the large part of heterogeneities. MRI in the treatment position can be challenging for H&N acquisitions because specific coils are used [83]. No study has yet investigated lung sCT generation from MRI with DLM. Movement is a huge challenge for MR imaging.

Several MRI sequences have been used to generate sCT from MRI in radiotherapy, with T2w sequences the most common. Some studies used specific reconstruction techniques such as mDixon or FLAIR. The FLAIR sequence is an inversion-recovery sequence. This sequence improves the detection of lesions of the cerebral parenchyma and enables visualization of edemas. It also facilitates the detection of white matter pathologies (softening, demyelination process), which appear as hypersignals.

Three studies investigated the impact of MISO, compared to a SISO [50,92,103]. MISO has the advantage of better tissue description. Koike et al. reported that MISO decreased the MAE and improved gamma pass-rate results compared to SISO [92]. Qi et al. used four sequences individually and combined them as input. The combination of sequences improved the sCT accuracy and robustness [50]. Sharma et al. [103] proposed a MIMO method generalizing to any combination of available and missing MRI sequence.

Moreover, three studies evaluated the impact of generating a sCT from a device other than the one used during training [88,89,96]. Such "multidevice" or "multicenter" impact is a key challenge to a commercial development.

The emergence of linacs combining MRI in the treatment room (MRI-linacs) increase the willingness of MRI-only workflow radiotherapy [126]. Some reviewed studies already used DLM for sCT generation from this device [31,33,56,73]. In this context, dose planning need to consider the presence of magnetic field, with the electron return effect [127,128]. Moreover, on MRI-linac, an MR image is acquired for each

fraction. This image could be used to perform dose monitoring or replanning with the use of a DLM, in the context of MR-guided adaptive radiotherapy [129].

Conclusions

The emergence of DL allows the fast and accurate generation of sCT from MRI in radiotherapy. In the literature, a variety of DLMs have been applied, mainly for brain and pelvis cancer, and also for H&N and liver. Each DL study has showed particularities in terms of hyperparameters or loss functions. Different MRI sequences are used depending on the anatomical location. Many metrics are used for image (voxel intensity and geometric fidelity) evaluation of the generated sCT. The MAE results were around 76 HU for head localization (brain and H&N) and liver, and 40 HU for pelvis. Dosimetric evaluation showed uncertainties below 2% for brain radiotherapy and lower than 1% for H&N, liver, abdomen, and pelvic areas. A better sCT quality was obtained with multiple inputs compared to single input of a DLM. Key challenges of the sCT generation for MRI in radiotherapy with DLMs is the standardization of sCT evaluation, and multicentrer impact.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2021.07.027.

References

[1] Pathmanathan AU, McNair HA, Schmidt MA, Brand DH, Delacroix L, Eccles CL, et al. Comparison of prostate delineation on multimodality imaging for MR-guided radiotherapy. BJR 2019;92:20180948. https://doi.org/10.1259/bjr.20180948.

[2] Kerkmeijer LGW, Maspero M, Meijer GJ, van der Voort van Zyp JRN, de Boer HCJ, van den Berg CAT. Magnetic Resonance Imaging only Workflow for Radiotherapy Simulation and Planning in Prostate Cancer. Clin. Oncol. 2018;30:692–701. 10.1016/j.clon.2018.08.009.

[3] Jonsson J, Nyholm T, Söderkvist K. The rationale for MR-only treatment planning for external radiotherapy. Clin Transl Radiat Oncol 2019;18:60–5. https://doi.org/10.1016/j.ctro.2019.03.005.

[4] Largent A, Barateau A, Nunes J-C, Lafond C, Greer PB, Dowling JA, et al. Pseudo-CT Generation for MRI-Only Radiation Therapy Treatment Planning: Comparison Among Patch-Based, Atlas-Based, and Bulk Density Methods. Int J Radiat Oncol Biol Phys 2019;103:479–90. https://doi.org/10.1016/j.ijrobp.2018.10.002.

[5] Dowling JA, Sun J, Pichler P, Rivest-Hénault D, Ghose S, Richardson H, 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 2015;93:1144–53. https://doi.org/10.1016/j.ijrobp.2015.08.045.

[6] Cusumano D, Placidi L, Teodoli S, Boldrini L, Greco F, Longo S, et al. On the accuracy of bulk synthetic CT for MR-guided online adaptive radiotherapy. Radiol Med 2020;125:157–64. https://doi.org/10.1007/s11547-019-01090-0.

[7] Choi JH, Lee D, O'Connor L, Chalup S, Welsh JS, Dowling J, et al. Bulk Anatomical Density Based Dose Calculation for Patient-Specific Quality Assurance of MRI-Only Prostate Radiotherapy. Front Oncol 2019;9. https://doi.org/10.3389/fonc.2019.00997.

[8] Kemppainen R, Suilamo S, Ranta I, Pesola M, Halkola A, Eufemio A, et al. Assessment of dosimetric and positioning accuracy of a magnetic resonance imaging-only solution for external beam radiotherapy of pelvic anatomy. Phys Imag Radiat Oncol 2019;11:1–8. https://doi.org/10.1016/j.phro.2019.06.001.

[9] Chen S, Quan H, Qin A, Yee S, Yan D. MR image-based synthetic CT for IMRT prostate treatment planning and CBCT image-guided localization. J Appl Clin Med Phys 2016;17:236–45. https://doi.org/10.1120/jacmp.v17i3.6065.

[10] Huynh T, Gao Y, Kang J, Wang L, Zhang P, Lian J, et al. Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model. IEEE Trans Med Imaging 2016;35:174–83. https://doi.org/10.1109/TMI.2015.2461533.

[11] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[12] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med 2018;98:126–46. https://doi.org/10.1016/j.compbiomed.2018.05.018.

[13] Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys 2019;46:e1–36. https://doi.org/10.1002/mp.13264.

[14] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. BJR 2019;92:20190001. https://doi.org/10.1259/bjr.20190001.

[15] Shen C, Nguyen D, Zhou Z, Jiang SB, Dong B, Jia X. An introduction to deep learning in medical physics: advantages, potential, and challenges. Phys Med Biol 2020;65:05TR01. 10.1088/1361-6560/ab6f51.

[16] Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner M-I. Deep Learning: A Review for the Radiation Oncologist. Front Oncol 2019;9. 10.3389/fonc.2019.00977.

[17] Feng M, Valdes G, Dixit N, Solberg TD. Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs. Front Oncol 2018;8. https://doi.org/10.3389/fonc.2018.00110.

[18] Bibault J-E, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett 2016;382:110–7. https://doi.org/10.1016/j.canlet.2016.05.033.

[19] Thompson RF, Valdes G, Fuller CD, Carpenter CM, Morin O, Aneja S, et al. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? Radiother Oncol 2018;129:421–6. https://doi.org/10.1016/j.radonc.2018.05.030.

[20] Largent A, Barateau A, Nunes J-C, Mylona E, Castelli J, Lafond C, et al. Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning. Int J Radiat Oncol Biol Phys 2019;105:1137–50. https://doi.org/10.1016/j.ijrobp.2019.08.049.

[21] Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. Radiat Oncol 2017;12:28. https://doi.org/10.1186/s13014-016-0747-y.

[22] Johnstone E, Wyatt JJ, Henry AM, Short SC, Sebag-Montefiore D, Murray L, et al. Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging-Only Radiation Therapy. Int J Radiat Oncol Biol Phys 2018;100:199–217. https://doi.org/10.1016/j.ijrobp.2017.08.043.

[23] Bird D, Henry AM, Sebag-Montefiore D, Buckley DL, Al-Qaisieh B, Speight R. A Systematic Review of the Clinical Implementation of Pelvic Magnetic Resonance Imaging (MR)-Only Planning for External Beam Radiation Therapy. Int J Radiat Oncol Biol Phys 2019. https://doi.org/10.1016/j.ijrobp.2019.06.2530.

[24] Owrangi AM, Greer PB, Glide-Hurst CK. MRI-only treatment planning: benefits and challenges. Phys Med Biol 2018. https://doi.org/10.1088/1361-6560/aaca4.

[25] Wafa B, Moussaoui A. A review on methods to estimate a CT from MRI data in the context of MRI-alone RT. Mèd Technol J 2018;2:150–78. 10.26415/2572-004X-vol2iss1p150-178.

[26] Wang T, Lei Y, Fu Y, Wynne JF, Curran WJ, Liu T, et al. A review on medical imaging synthesis using deep learning and its clinical applications. J Appl Clin Med Phys 2020. https://doi.org/10.1002/acm2.13121.

[27] Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning-based synthetic-CT generation in radiotherapy and PET: a review. ArXiv:210202734 [Physics] 2021.

[28] Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks. In: Carneiro G, Mateus D, Peter L, Bradley A, Tavares JMRS, Belagiannis V, et al., editors. Deep Learning and Data Labeling for Medical Applications, vol. 10008, Cham: Springer International Publishing; 2016, p. 170–8. 10.1007/978-3-319-46976-8_18.

[29] Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, et al. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, vol. 10435, Cham: Springer International Publishing; 2017, p. 417–25. 10.1007/978-3-319-66179-7_48.

[30] Xiang L, Wang Q, Nie D, Zhang L, Jin X, Qiao Y, et al. Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. Med Image Anal 2018;47:31–44. https://doi.org/10.1016/j.media.2018.03.011.

[31] Cusumano D, Lenkowicz J, Votta C, Boldrini L, Placidi L, Catucci F, et al. A deep learning approach to generate synthetic CT in low field MR-guided adaptive radiotherapy for abdominal and pelvic cases. Radiother Oncol 2020. https://doi.org/10.1016/j.radonc.2020.10.018.

[32] Lei Y, Harms J, Wang T, Liu Y, Shu H-K, Jani AB, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. Med Phys 2019;46:3565–81. https://doi.org/10.1002/mp.13617.

[33] Fu J, Singhrao K, Cao M, Yu V, Santhanam AP, Yang Y, et al. Generation of abdominal synthetic CTs from 0.35T MR images using generative adversarial networks for MR-only liver radiotherapy. Biomed Phys Eng Express 2020;6:015033. 10.1088/2057-1976/ab6e1f.

[34] Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for Medical Image Analysis. Artif Intell Med 2020.

[35] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[36] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: Image traits, technology

trends, case studies with progress highlights, and future promises. ArXiv: 200809104 [Cs, Eess] 2020.
[37] Shen C, Nguyen D, Zhou Z, Jiang SB, Dong B, Jia X. An introduction to deep learning in medical physics: advantages, potential, and challenges. Phys Med Biol 2020. https://doi.org/10.1088/1361-6560/ab6f51.
[38] Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. In: ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning; 2010. p. 807–14.
[39] Maas AL, Hannun AY, Rectifier Ng AY. Nonlinearities Improve Neural Network Acoustic Models 2013.
[40] Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). ICLR 2016.
[41] Ioffe S, Szegedy C. Batch Normalization Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning -; 2015. p. 448–56.
[42] Liu F, Yadav P, Baschnagel AM, McMillan AB. MR-based treatment planning in radiation therapy using a deep learning approach. J Appl Clin Med Phys 2019;20: 105–14. https://doi.org/10.1002/acm2.12554.
[43] Andres EA, Fidon L, Vakalopoulou M, Lerousseau M, Carré A, Sun R, et al. Dosimetry-driven quality measure of brain pseudo Computed Tomography generated from deep learning for MRI-only radiotherapy treatment planning. Int J Radiat Oncol Biol Phys 2020:S0360301620311305. 10.1016/j.ijrobp.2020.05.006.
[44] Han X. MR-based synthetic CT generation using a deep convolutional neural network method. Med Phys 2017;44:1408–19. https://doi.org/10.1002/mp.12155.
[45] Spadea MF, Pileggi G, Zaffino P, Salome P, Catana C, Izquierdo-Garcia D, et al. Deep Convolution Neural Network (DCNN) Multiplane Approach to Synthetic CT Generation From MR images—Application in Brain Proton Therapy. Int J Radiat Oncol Biol Phys 2019;105:495–503. https://doi.org/10.1016/j.ijrobp.2019.06.2535.
[46] Wang Y, Liu C, Zhang X, Deng W. Synthetic CT Generation Based on T2 Weighted MRI of Nasopharyngeal Carcinoma (NPC) Using a Deep Convolutional Neural Network (DCNN). Front Oncol 2019;9. 10.3389/fonc.2019.01333.
[47] Arabi H, Dowling JA, Burgos N, Han X, Greer PB, Koutsouvelis N, et al. Comparative study of algorithms for synthetic CT generation from MRI : Consequences for MRI -guided radiation planning in the pelvic region. Med Phys 2018;45:5218–33. https://doi.org/10.1002/mp.13187.
[48] Gupta D, Kim M, Vineberg KA, Balter JM. Generation of Synthetic CT Images From MRI for Treatment Planning and Patient Positioning Using a 3-Channel U-Net Trained on Sagittal Images. Front Oncol 2019;9. https://doi.org/10.3389/fonc.2019.00964.
[49] Dinkla AM, Florkow MC, Maspero M, Savenije MHF, Zijlstra F, Doornaert PAH, et al. Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network. Med Phys 2019;46:4095–104. https://doi.org/10.1002/mp.13663.
[50] Qi M, Li Y, Wu A, Jia Q, Li B, Sun W, et al. Multi-sequence MR image-based synthetic CT generation using a generative adversarial network for head and neck MRI-only radiotherapy. Med Phys 2020;47:1880–94. https://doi.org/10.1002/mp.14075.
[51] Chen S, Qin A, Zhou D, Yan D. Technical Note: U-net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning. Med Phys 2018;45:5659–65. https://doi.org/10.1002/mp.13247.
[52] Florkow MC, Zijlstra F, M.d LGWK, Maspero M, Berg CAT van den, Stralen M van, et al. The impact of MRI-CT registration errors on deep learning-based synthetic CT generation. Medical Imaging 2019: Image Processing, vol. 10949, International Society for Optics and Photonics; 2019, p. 1094938. 10.1117/12.2512747.
[53] Florkow MC, Zijlstra F, Willemsen K, Maspero M, van den Berg CAT, Kerkmeijer LGW, et al. Deep learning–based MR-to-CT synthesis: The influence of varying gradient echo–based MR images as input channels. Magn Reson Med 2020;83:1429–41. https://doi.org/10.1002/mrm.28008.
[54] Stadelmann JV, Schulz H, Heide UA van der, Renisch S. Pseudo-CT image generation from mDixon MRI images using fully convolutional neural networks. Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 10953, International Society for Optics and Photonics; 2019, p. 109530Z. 10.1117/12.2512741.
[55] Neppl S, Landry G, Kurz C, Hansen DC, Hoyle B, Stöcklein S, et al. Evaluation of proton and photon dose distributions recalculated on 2D and 3D Unet-generated pseudoCTs from T1-weighted MR head scans. Acta Oncol 2019;58:1429–34. https://doi.org/10.1080/0284186X.2019.1630754.
[56] Olberg S, Zhang H, Kennedy WR, Chun J, Rodriguez V, Zoberi I, et al. Synthetic CT reconstruction using a deep spatial pyramid convolutional framework for MR-only breast radiotherapy. Med Phys 2019;46:4135–47. https://doi.org/10.1002/mp.13716.
[57] Li W, Li Y, Qin W, Liang X, Xu J, Xiong J, et al. Magnetic resonance image (MRI) synthesis from brain computed tomography (CT) images based on deep learning methods for magnetic resonance (MR)-guided radiotherapy. Quant Imaging Med Surg 2020;10:1223–36. 10.21037/qims-19-885.
[58] Kazemifar S, McGuire S, Timmerman R, Wardak Z, Nguyen D, Park Y, et al. MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. Radiother Oncol 2019;136: 56–63. https://doi.org/10.1016/j.radonc.2019.03.026.
[59] Kazemifar S, Barragán Montero AM, Souris K, Rivas ST, Timmerman R, Park YK, et al. Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors: Dosimetric evaluation of synthetic CT generated with GANs for MRI-only proton therapy treatment planning of brain tumors. J Appl Clin Med Phys 2020;21:1–11. https://doi.org/10.1002/acm2.12856.
[60] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481–95. https://doi.org/10.1109/TPAMI.2016.2644615.
[61] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Comput Vis Pattern Recogn 2015.
[62] Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. Med Phys 2018;45:3627–36. https://doi.org/10.1002/mp.13047.
[63] Huang G, Liu Z, Maaten VD, Weinberger KQ. Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017;2017:2261–9. https://doi.org/10.1109/CVPR.2017.243.
[64] Isola P, Zhu J-Y, Zhou T, Efros AA. In: Image-to-Image Translation with Conditional Adversarial Networks. Honolulu, HI: IEEE; 2017. p. 5967–76. https://doi.org/10.1109/CVPR.2017.632.
[65] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Trans Pattern Anal Mach Intell 2018;42:7132–41. https://doi.org/10.1109/TPAMI.2019.2913372.
[66] Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. Computer Vision and Pattern Recognition; 2017.
[67] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv:150504597 [Cs] 2015.
[68] Wolterink JM, Leiner T, Viergever MA, Išgum I. Dilated Convolutional Neural Networks for Cardiovascular MR Segmentation in Congenital Heart Disease. Reconstruction, Segmentation, and Analysis of Medical Images, RAMBO 2016, HVSMR 2016 Lecture Notes in Computer Science 2017;10129:95–102. 10.1007/978-3-319-52280-7_9.
[69] Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. In: Niethammer M, Styner M, Aylward S, Zhu H, Oguz I, Yap P-T, editors. Information Processing in Medical Imaging. Cham: Springer International Publishing; 2017. p. 348–60.
[70] Dinkla AM, Wolterink JM, Maspero M, Savenije MHF, Verhoeff JJC, Seravalli E, et al. MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network. Int J Radiat Oncol Biol Phys 2018;102:801–12. https://doi.org/10.1016/j.ijrobp.2018.05.058.
[71] Thummerer A, de Jong BA, Zaffino P, Meijers A, Marmitt GG, Seco J, et al. Comparison of the suitability of CBCT- and MR-based synthetic CTs for daily adaptive proton therapy in head and neck patients. Phys Med Biol 2020. https://doi.org/10.1088/1361-6560/abb1d6.
[72] Massa HA, Johnson JM, McMillan AB. Comparison of deep learning synthesis of synthetic CTs using clinical MRI inputs. Phys Med Biol 2020;65:23NT03. 10.1088/1361-6560/abc5cb.
[73] Jeon W, An HJ, Kim J, Park JM, Kim H, Shin KH, et al. Preliminary Application of Synthetic Computed Tomography Image Generation from Magnetic Resonance Image Using Deep-Learning in Breast Cancer Patients. J Radiat Prot Res 2019;44: 149–55. 10.14407/jrpr.2019.44.4.149.
[74] Florkow MC, Guerreiro F, Zijlstra F, Seravalli E, Janssens GO, Maduro JH, et al. Deep learning-enabled MRI-only photon and proton therapy treatment planning for paediatric abdominal tumours. Radiother Oncol 2020;153:220–7. https://doi.org/10.1016/j.radonc.2020.09.056.
[75] Bahrami A, Karimian A, Fatemizadeh E, Arabi H, Zaidi H. A new deep convolutional neural network design with efficient learning capability: Application to CT image synthesis from MRI. Med Phys 2020;47:5158–71. https://doi.org/10.1002/mp.14418.
[76] Liu L, Johansson A, Cao Y, Dow J, Lawrence TS, Balter JM. Abdominal synthetic CT generation from MR Dixon images using a U-net trained with 'semi-synthetic' CT data. Phys Med Biol 2020;65:125001. https://doi.org/10.1088/1361-6560/ab8cd2.
[77] Liu Y, Lei Y, Wang Y, Shafai-Erfani G, Wang T, Tian S, et al. Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning. Phys Med Biol 2019;64:205022. https://doi.org/10.1088/1361-6560/ab41af.
[78] Fu J, Yang Y, Singhrao K, Ruan D, Chu F-I, Low DA, et al. Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. Med Phys 2019;46:3788–98. https://doi.org/10.1002/mp.13672.
[79] Palmér E, Karlsson A, Nordström F, Petruson K, Siversson C, Ljungberg M, et al. Synthetic computed tomography data allows for accurate absorbed dose calculations in a magnetic resonance imaging only workflow for head and neck radiotherapy. Phys Imag Radiat Oncol 2021;17:36–42. https://doi.org/10.1016/j.phro.2020.12.007.
[80] Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, Berg CAT van den, Išgum I. Deep MR to CT Synthesis using Unpaired Data. ArXiv:170801155 [Cs] 2017.
[81] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. ArXiv:14062661 [Cs, Stat] 2014.
[82] Yi X, Walia E, Babyn P. Generative Adversarial Network in Medical Imaging: A Review. Med Image Anal 2019;58:101552. https://doi.org/10.1016/j.media.2019.101552.
[83] Largent A, Marage L, Gicquiau I, Nunes J-C, Reynaert N, Castelli J, et al. Head-and-Neck MRI-only radiotherapy treatment planning: From acquisition in

treatment position to pseudo-CT generation. Cancer/Radiothérapie 2020: S1278321820300615. 10.1016/j.canrad.2020.01.008.
[84] Liu X, Emami H, Nejad-Davarani SP, Morris E, Schultz L, Dong M, et al. Performance of deep learning synthetic CTs for MR-only brain radiation therapy. J Appl Clin Med Phys 2021;22:308–17. https://doi.org/10.1002/acm2.13139.
[85] Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. Comput Vis Pattern Recogn 2017.
[86] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Trans Neural Netw 1994.
[87] Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least Squares Generative Adversarial Networks. IEEE International Conference on Computer Vision 2017:9.
[88] Brou Boni KND, Klein J, Vanquin L, Wagner A, Lacornerie T, Pasquier D, et al. MR to CT synthesis with multicenter data in the pelvic era using a conditional generative adversarial network. Phys Med Biol 2020. https://doi.org/10.1088/1361-6560/ab7633.
[89] Petty L, Löfstedt T, Heilemann G, Purtado H, Nesvacil N, Nyholm T, et al. Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion. Phys Med Biol 2020. https://doi.org/10.1088/1361-6560/ab857b.
[90] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Comput Vis Pattern Recogn 2018.
[91] Chen LC, Papandreou G, Schroff F, Adam F. Rethinking atrous convolution for semantic image segmentation. Comput Vis Pattern Recogn 2017.
[92] Koike Y, Akino Y, Sumida I, Shiomi H, Mizuno H, Yagi M, et al. Feasibility of synthetic computed tomography generated with an adversarial network for multi-sequence magnetic resonance-based brain radiotherapy. J Radiat Res 2020;61:92–103. https://doi.org/10.1093/jrr/rrz063.
[93] Maspero M, Savenije MHF, Dinkla AM, Seevinck PR, Intven MPW, Jurgenliemk-Schulz IM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. Phys Med Biol 2018;63:185001. https://doi.org/10.1088/1361-6560/aada6d.
[94] Tie X, Lam S, Zhang Y, Lee K, Au K, Cai J. Pseudo-CT generation from multi-parametric MRI using a novel multi-channel multi-path conditional generative adversarial network for nasopharyngeal carcinoma patients. Med Phys 2020;47:1750–62. https://doi.org/10.1002/mp.14062.
[95] Hemsley M, Chugh B, Ruschin M, Lee Y, Tseng C-L, Stanisz G, et al. Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, vol. 12261, Cham: Springer International Publishing; 2020, p. 834–44. 10.1007/978-3-030-59710-8_81.
[96] Maspero M, Bentvelzen LG, Savenije MHF, Guerreiro F, Seravalli E, Janssens GO, et al. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. Radiother Oncol 2020;153:197–204. https://doi.org/10.1016/j.radonc.2020.09.029.
[97] Tang B, Wu F, Fu Y, Wang X, Wang P, Orlandini LC, et al. Dosimetric evaluation of synthetic CT image generated using a neural network for MR-only brain radiotherapy. J Appl Clin Med Phys 2021:acm2.13176. 10.1002/acm2.13176.
[98] Bourbonne V, Jaouen V, Hognon C, Boussion N, Lucia F, Pradier O, et al. Dosimetric Validation of a GAN-Based Pseudo-CT Generation for MRI-Only Stereotactic Brain Radiotherapy. Cancers 2021;13:1082. https://doi.org/10.3390/cancers13051082.
[99] Bird D, Nix MG, McCallum H, Teo M, Gilbert A, Casanova N, et al. Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning. Radiother Oncol 2021;156:23–8. https://doi.org/10.1016/j.radonc.2020.11.027.
[100] Peng Y, Chen S, Qin A, Chen M, Gao X, Liu Y, et al. Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. Radiother Oncol 2020;150:217–24. https://doi.org/10.1016/j.radonc.2020.06.049.
[101] Klages P, Benslimane I, Riyahi S, Jiang J, Hunt M, Deasy JO, et al. Comparison of Patch-Based Conditional Generative Adversarial Neural Net Models with Emphasis on Model Robustness for Use in Head and Neck Cases for MR-Only Planning 2020:27. arXiv:1902.00536.
[102] Weber M, Fürst M, Zöllner JM. Automated Focal Loss for Image based Object Detection. IEEE Intelligent Vehicles Symposium (IV) 2020;2020:1423–9. https://doi.org/10.1109/IV47402.2020.9304830.
[103] Sharma A, Hamarneh G. Missing MRI Pulse Sequence Synthesis Using Multi-Modal Generative Adversarial Network. IEEE Trans Med Imaging 2020;39:1170–83. https://doi.org/10.1109/TMI.2019.2945521.
[104] Zhu J-Y, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. IEEE International Conference on Computer Vision 2017.
[105] Yang H, Sun J, Carass A, Zhao C, Lee J, Xu Z, et al. Unpaired Brain MR-to-CT Synthesis Using a Structure-Constrained CycleGAN. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, et al., editors. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham: Springer International Publishing; 2018, p. 174–82. 10.1007/978-3-030-00889-5_20.
[106] Liu Y, Lei Y, Wang T, Kayode O, Tian S, Liu T, et al. MRI-based treatment planning for liver stereotactic body radiotherapy: validation of a deep learning-based synthetic CT generation method. BJR 2019;92:20190067. https://doi.org/10.1259/bjr.20190067.
[107] Liu Y, Lei Y, Wang Y, Wang T, Ren L, Lin L, et al. MRI-based treatment planning for proton radiotherapy: dosimetric validation of a deep learning-based liver synthetic CT generation method. Phys Med Biol 2019;64:145015. https://doi.org/10.1088/1361-6560/ab25bc.
[108] Shafai-Erfani G, Lei Y, Liu Y, Wang Y, Wang T, Zhong J, et al. MRI-Based Proton Treatment Planning for Base of Skull Tumors. Int J Particle Ther 2019;6:12–25. 10.14338/IJPT-19-00062.1.
[109] Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson Fergus V, Brady SM, et al. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Med Image Anal 2012.
[110] Mao X, Li Q, Xie H. Multi-class Generative Adversarial Networks with the L2 Loss Function. Comput Vis Pattern Recogn 2016.
[111] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20. https://doi.org/10.1109/TMI.2010.2046908.
[112] Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data 1998:87–97.
[113] Nyul LG, Udupa JK, Zhang X. New Variants of a Method of MRI Scale Standardization. IEEE Trans Med Imag 2000:143–50.
[114] Cox I, Roy S, Hingorani SL. Dynamic histogram warping of image pairs for constant image brightness. Proc Int Conf Image Proc 1995:366–9.
[115] Nyholm T, Svensson S, Andersson S, Jonsson J, Sohlin M, Gustafsson C, et al. MR and CT data with multiobserver delineations of organs in the pelvic area—Part of the Gold Atlas project. Med Phys 2018;45:1295–300. https://doi.org/10.1002/mp.12748.
[116] Girod B. What's wrong with mean-squared error. Digital Images and Human Vision (A B Watson, Ed) 1993:207–20.
[117] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans Image Process 2004;13:600–12. https://doi.org/10.1109/TIP.2003.819861.
[118] Low DA, Harms WB, Mutic S, Purdy JA. A technique for the quantitative evaluation of dose distributions. Med Phys 1998;25:656–61.
[119] Hussein M, Clark CH, Nisbet A. Challenges in calculation of the gamma index in radiotherapy–towards good practice. Phys Med 2017;36:1–11.
[120] Yang X, Lei Y, Shu H-K, Rossi P, Mao H, Shim H, et al. Pseudo CT estimation from MRI using patch-based random forest. Medical Imaging 2017: Image Processing, vol. 10133, International Society for Optics and Photonics; 2017, p. 101332Q. 10.1117/12.2253936.
[121] Lerner M, Medin J, Jamtheim Gustafsson C, Alkner S, Siversson C, Olsson LE. Clinical validation of a commercially available deep learning software for synthetic CT generation for brain. Radiat Oncol 2021;16:66. https://doi.org/10.1186/s13014-021-01794-6.
[122] Bragman FJS, Tanno R, Eaton-Rosen Z, Li W, Hawkes DJ, Ourselin S, et al. Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning. ArXiv:180606595 [Cs] 2018;11073:3–11. https://doi.org/10.1007/978-3-030-00937-3_1.
[123] Tanno R, Worrall D, Kaden E, Ghosh A, Grussu F, Bizzi A, et al. Uncertainty Quantification in Deep Learning for Safer Neuroimage Enhancement. ArXiv:190713418 [Cs, Eess, Stat] 2019.
[124] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020. https://doi.org/10.1016/j.radonc.2020.09.008.
[125] Speight R, Dubec M, Eccles CL, George B, Henry A, Herbert T, et al. IPEM topical report: guidance on the use of MRI for external beam radiotherapy treatment planning *. Phys Med Biol 2021;66:055025. https://doi.org/10.1088/1361-6560/abdc30.
[126] Kurz C, Buizza G, Landry G, Kamp F, Rabe M, Paganelli C, et al. Medical physics challenges in clinical MR-guided radiotherapy. Radiat Oncol 2020;15:93. https://doi.org/10.1186/s13014-020-01524-4.
[127] Cusumano D, Teodoli S, Greco F, Fidanzio A, Boldrini L, Massaccesi M, et al. Experimental evaluation of the impact of low tesla transverse magnetic field on dose distribution in presence of tissue interfaces. Physica Med 2018;53:80–5. https://doi.org/10.1016/j.ejmp.2018.08.007.
[128] Raaijmakers AJE, Raaymakers BW, Lagendijk JJW. Experimental verification of magnetic field dose effects. Phys Med Biol 2007;52:4283–91. https://doi.org/10.1088/0031-9155/52/14/017.
[129] Otazo R, Lambin P, Pignol J-P, Ladd ME, Schlemmer H-P, Baumann M, et al. MRI-guided Radiation Therapy: An Emerging Paradigm in Adaptive Radiation Oncology. Radiology 2021;298:248–60. https://doi.org/10.1148/radiol.2020202747.

Discussion

Three categories of methods were analysed consecutively to calculate dose distribution on MRI, as summarised in Table 1.2. Figure 1.8 illustrates the results obtained for each family of methods applied to the same MRI of a female pelvis. The atlas and machine learning based approaches involve the utilisation of a database of co-registered CT-MRI images (intra-patient). Inter-patient image registration is also necessary in atlas methods. Bulk density methods, on the other hand, can be directly performed using the patient's MRI. While all the methods utilise standard MRI sequences, ultra-short echo time (UTE) MRI sequences, which can differentiate between air and bone tissue, are also employed in bulk density and machine learning approaches. Density assignment methods typically rely on the average intensity (in HU values) of specific volumes of interest, calculated from a CT scan image database, while Atlas-based methods incorporate intensity information and integrate spatial and shape information for the volumes of interest. Machine learning methods, on the other hand, employ image descriptors that capture texture and contour information in the neighbourhood of each voxel.

Deep learning-based sCT generation methods employ convolutional neural networks (CNNs) as model architectures, but recently transformers have demonstrated great potential in image synthesis [48]. Hybrid networks, combining CNNs and transformers, have been proposed to extract both local texture and global information [49]–[52]. The primary advantage of transformers lies in their ability to better understand contextual information compared to CNNs. However, they do tend to come with a higher computational cost and require larger amounts of data.

Bulk-density and atlas methods have received fewer recent publications, suggesting that they may be more challenging to improve and may struggle to compete with the performance of DLMs. While deep learning methods show great promise, they heavily rely on the quality of intra-patient registration within the learning cohort. Although cycle-GAN architectures have the potential to avoid intra-patient registration [53], [54], recent studies using this architecture still employ registered data to provide better results[54], [55]. Therefore, further investigation is needed in this area.

Analysis of the literature raises several questions regarding the evaluation of improvements of various methods for calculating dose from MRI. One key aspect that has not been adequately addressed is the impact of different irradiation techniques (such as IMRT, VMAT, SBRT, brachytherapy, proton therapy, etc.) on these dose calculation methods. Furthermore, there is variation in MRI sequences and acquisition parameters across studies. Wang et al.[56] compared the effect of different MRI sequences on a deep-learning based sCT generation method (consistent cycle-GAN) for paediatric brain tumour, while Florkow et al. [57] were interested in studying the influence of gradient echo–based contrasts on a 3D patch-based neural network. But these evaluations are conducted on different datasets and are not systematic. It would be valuable to establish multicentre image databases that allow for the

evaluation of the proposed methods on standardised datasets using consistent validation tools. This would facilitate the development of more generalised deep learning models and enable better comparisons between different approaches.

There is also a notable methodological gap in the evaluation of different approaches[11]. Some studies focus solely on comparing CT-scan and sCT images, while others focus only on dosimetric assessment.

All these studies based their evaluation on full reference metrics (i.e comparison with a ground truth), as described in Boulanger et al.[1]. These metrics provide an insight into the overall accuracy of the method but do not allow for the identification of the limitations of the sCT generation approach. Additionally, they cannot be applied in a daily sCT quality assurance process as no reference CT will be available in an MRI-only workflow.

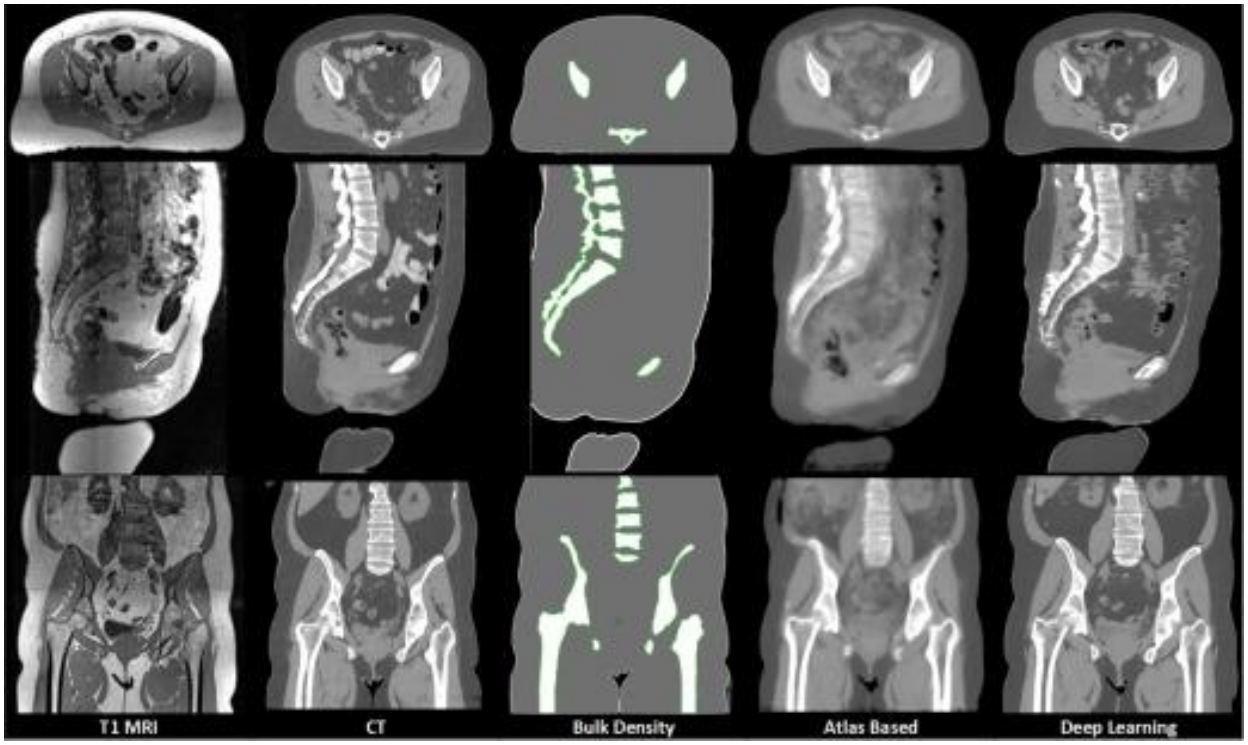

Figure 1.8: Example sCT comparison between 3 generation methods from a T1-weighted MRI of a female pelvis. The columns show: the original MRI, the actual planning CT, a bulk density sCT, a multi-atlas based sCT, and a conditional generative adversarial network (deep learning) based sCT. The rows show the axial, sagittal, and coronal views from the 3D volumes. Figure from the book "Biomedical Image synthesis and simulation", chapter 20 [5].

Table 1.2: Principle, advantages, and drawbacks of each category of dose calculation methods from MRI

Methods	Principle	Advantages	Drawbacks
Bulk-density	Manual or automatic delineation of volumes of interest on the patient's MRI, then assignment of a density value (electronic or physical) to each region.	- Simplicity of the methodology	- Highly operator-dependent (manual delineation) - Time-consuming to calculate (automatic delineation) - Segmentation errors - Homogeneity of tissue - Less accurate dose - Restricted to certain locations
Atlas	Pairwise mapping of a learning database of CT and MRI images to the patient's MRI, followed by CTs fusion.	- Fully automated method - Good accuracy of the calculated dose - Heterogeneity of tissue density - Automatic delineation of volumes of interest - Anatomical genericity	- High computation time - Sensitive to anatomical dissimilarities - Requires intra-patient deformable multimodal registration and inter-patient deformable registration of the training cohort - Uncertainty caused by registration errors - Smoothing of intensities
Machine learning **(Including deep-learning models (DLM))**	Modelling the relationships between the intensities of MRI and CT voxels using machine learning tools, then applying the model to the patient's MRI. The model is established in 2 stages: learning (generating the model) and validation (application to the patient).	- Speed of execution - Good accuracy of estimated dose - No inter-patient registration (except patch-based) - Heterogeneity of tissue density	- Requires intra-patient registration (multimodal) of the training cohort, except for non-supervised DLM - Large amount of data needed (especially for DLMs)

Conclusion

MRI offers better contrast between soft tissues compared to CT imaging, making it a valuable reference imaging modality for treatment planning. This advantage allows for more accurate delineation of specific target volumes and eliminates the need for MRI-CT registration, reducing associated errors. Currently, two treatment strategies incorporating MRI are being explored. The first strategy involves replacing planning CT with planning MRI while carrying out treatment using a standard LINAC. The second strategy involves treatment with an MRI-LINAC machine, where an MRI scan is acquired before each irradiation session. In MRI-LINAC treatments, ART strategies play a central role. It is planned that treatment plans can be adjusted dynamically during each session based on the real-time anatomy, utilising a live ART approach. Dose calculation based on MRI has thus become of interest and is a rapidly advancing field in radiotherapy. The emergence of deep learning has enabled fast and accurate generation of sCT from MRI. In the literature, various DLMs have been applied, primarily for brain and pelvic cancer, as well as for head and neck and liver. Each deep-learning study has demonstrated unique characteristics in terms of hyperparameters and loss functions, and different MRI sequences are utilised depending on the anatomical location. It has been observed that employing multiple inputs in a DLM yields better sCT quality compared to using a single input. One of the key challenges in sCT generation from MRI in radiotherapy using DLMs is addressing the multicentre impact, as well as the standardisation of sCT evaluation.

Indeed, multiple metrics are employed to evaluate the generated sCT, including voxel intensity and geometric fidelity, but no consensus has been established in the scientific community.

State-of-the-art of quality assessment methods

Overview

Quality assessment of images involves the evaluation and measurement of different aspects related to the perceived quality, fidelity, and accuracy of images. It aims to assess the extent to which an image accurately represents the original scene or meets specific criteria or standards. Image quality assessment (IQA) can be approached through both objective and subjective methods.

1- *Subjective methods*

In subjective IQA, human observers are presented with a set of images, which may include reference images and distorted images in double stimulus approaches, or only distorted images in single stimulus approaches. The observers are then asked to rate or rank the quality of the distorted images according to their perceived visual quality. The ratings can be based

on the overall image quality, sharpness, colour, accuracy, and artifact visibility. To ensure reliable and consistent results, subjective assessments are often conducted under controlled conditions. Multiple observers are typically involved to account for individual differences in perception. The results of subjective assessments are often used to validate and refine objective image quality metrics. By incorporating human perception, subjective assessment ensures that the resulting images are not only technically accurate but also visually satisfying to human observers. However, conducting subjective assessments can be time-consuming and resource-intensive, requiring careful planning, coordination, and analysis.

2- *Objective methods*

Image quality assessment metrics are used to quantitatively evaluate the quality of images. These metrics provide objective measures that assess several aspects of image quality, such as sharpness, contrast, noise, and distortion.

Objectives metrics can be classified in 3 categories:

- Full reference metrics involve comparing a generated image with a reference image considered as the ground truth. This comparison can be conducted through voxel-wise difference or measures of distorting noise that affect the quality of image representation, such as peak signal-to-noise ratio (PSNR). Human visual system (HVS) based methods, such as visual information fidelity (VIF)[58] or the structural similarity index (SSIM)[59] and its variations, assess the similarity between the reference image and the distorted image by considering structural information, luminance, and contrast. These metrics provide a measure of perceived quality by mimicking human visual perception.

- Reduced reference IQA metrics rely on comparing specific features or information from a distorted image with a reference image, but they do not require a full reference image for comparison. Instead, they use a reduced set of reference information or features to assess image quality.

- No-reference IQA metrics do not require a reference image for comparison. Instead, they are designed to assess the quality of an image based solely on its own content and characteristics. Sharpness, noise, texture, and structural information within the image are analysed to estimate its quality. Examples of popular no-reference IQA metrics include BRISQUE (Blind/Reduced-Reference Image Quality Evaluator) [60], NIQE (Naturalness Image Quality Evaluator), and PIQE (Perceptual Image Quality Evaluator) [61]. These metrics are trained on a large dataset of images and utilise statistical or machine learning techniques to provide objective quality scores without relying on a reference image. Deep-learning based methods also recently arise [62]–[64].

Objective metrics provide quantitative measures of image quality, allowing comparisons between different images or different processing algorithms. However, it's important to note

that no single metric can fully capture all aspects of human perception, and the choice of metric depends on the specific application and requirements.

IQA in CT

A recently published report by the French Society of Medical Physicists (SFPM) presents the metrics used to assess a CT scan[1]. The classical methods can be divided into four categories:

- Signal and contrast

The signal of each voxel in a CT scan is measured in HU. This scale is defined based on the relationship between the linear absorption coefficient of water (μ_{water}) and the average linear absorption coefficient (μ_X) of the contents within the volume defined by voxel X.

$$HU = 1000 \times \frac{\mu_X - \mu_{water}}{\mu_{water}} \quad (1)$$

The signal measured within a region of interest (ROI) is determined by calculating the average value of the HU for the voxels included in the ROI:

$$signal(ROI) = \overline{HU(ROI)} \quad (2)$$

Finally, the contrast between 2 ROIs is defined as follow:

$$C(ROI_1, ROI_2) = signal(ROI_1) - signal(ROI_2) \quad (3)$$

- Noise

In CT scans, noise arises from both quantum noise, which is associated with the random emission and detection of photons, and electronic noise. The assessment of noise in an image of a homogeneous object involves calculating the standard deviation of the HU within a ROI.

- Spatial resolution

Spatial resolution refers to the minimum distance that can be measured between two structures and is closely associated with the concept of point spread function (PSF). The PSF, also known as the spatial impulse response, is a mathematical function that characterizes the imaging system's response to a point object.

[1] The report is available online:
https://www.sfpm.fr/sites/www.sfpm.fr/files/Bibliotheque/Documents_SFPM/Public/Rapports_SFPM/Rapports_GT/sfpm_2023_41_tdm_metriques.pdf

The Modulation Transfer Function (MTF) is a function that describes the imaging system's capability to preserve contrast as a function of the level of object detail, also known as spatial frequency.

- Detectability

Signal-to-noise ratio (SNR) is the comparison of a specific signal to the background noise. A low SNR value will result in a significant amount of noise that can partially or completely mask the signal, making it challenging to interpret the image.

The contrast-to-noise ratio (CNR) is a measure used to evaluate the difference in average attenuation or signal intensity between a structure of interest and the background, relative to the background noise. The CNR is commonly defined as follows:

$$CNR(structure, background) = \frac{signal(structure) - signal(background)}{signal(background)} \quad (4)$$

A "low" CNR value will manifest in the image as a significant amount of noise that partially or completely obscures the contrast of the lesion, thereby posing challenges for the radiologist in interpreting the image.

These metrics show that contrast and noise are two key features to describe the quality of a CT scan. However, they are not sufficient to assess the quality of a sCT in the context of MRI-only RT.

IQA in synthetic-CT generation for MRI-only radiation therapy

Methods and algorithms developed for generating sCT scans require validation, not only during the development phase of new generation methods but also for the clinical validation of implementing an sCT system in a radiotherapy center prior to routine use. This validation is achieved using full-reference metrics, which involve comparing the generated sCT image to its corresponding planning CT. Discrepancies in terms of HU values compared to a reference CT are measured using metrics such as mean absolute error (MAE) and mean error (ME). This difference can be assessed at a voxel level, resulting in 3D error maps. Additionally, mean square error (MSE), root mean square error (RMSE), and PSNR[65] are computed. The MSE is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(HU_{sCT}(i) - HU_{CT}(i))^2 \quad (5)$$

With N the number of voxels in the image, $HU_{sCT}(i)$ the intensity in HU of the i^{th} voxel in the sCT and $HU_{CT}(i)$ the intensity in HU of the i^{th} in the reference CT. While this metric provides

insight into the dispersion of the error, the PSNR, defined as follows, measures the level of noise corruption in the image:

$$PSNR = 10log_{10}\left(\frac{Q^2}{MSE}\right) \quad (6)$$

With Q the dynamic of the image.

Perception-based metrics like the visual information fidelity (VIF), the structural similarity index (SSIM) (eq. 7) [5], [12] and multi-scale SSIM [52] are also commonly employed and focus on the structure, contrast, and luminance of an image.

$$SSIM = \frac{(2\mu_{CT}\mu_{sCT} + C_1)(2\sigma_{CTsCT} + C_2)}{(\mu_{CT}^2 + \mu_{sCT}^2 + C_1)(\sigma_{CT}^2 + \sigma_{sCT}^2 + C_2)} \quad (7)$$

Here, C_1 and C_2 are two variables to stabilize the division with weak denominator, μ_{CT} and μ_{sCT} represent respectively the mean value of the reference CT and the sCT, σ_{CT}^2 and σ_{sCT}^2 their variance and σ_{CTsCT} the covariance of the CT and the sCT.

The assessment of geometric fidelity for automatically segmented structures, such as bones and body contours, involves metrics like the Dice similarity coefficient (DSC), Hausdorff distance, and mean absolute surface distance (MASD). These metrics provide valuable information about the accuracy of the image contours.

Dose accuracy is evaluated using full-reference metrics that compare the dose calculations obtained from the sCT with those derived from the reference CT. Many studies in the sCT generation literature focus on dosimetric endpoints, including gamma analysis and dose-volume histogram (DVH) metric [66]. DVH is a widely used tool in radiation therapy routine. Gamma analysis allows for the spatial analysis of dose distributions obtained from the sCT and reference CT by utilizing gamma maps. It can be performed in two or three dimensions, incorporating dose and spatial criteria. Various parameters must be set for a gamma analysis, including dose criteria, distance-to-agreement criteria, local or global analysis, and dose threshold.

Figure 1.9 summarises the different metrics commonly used by order of complexity. They also have been described in Boulanger et al.[1].

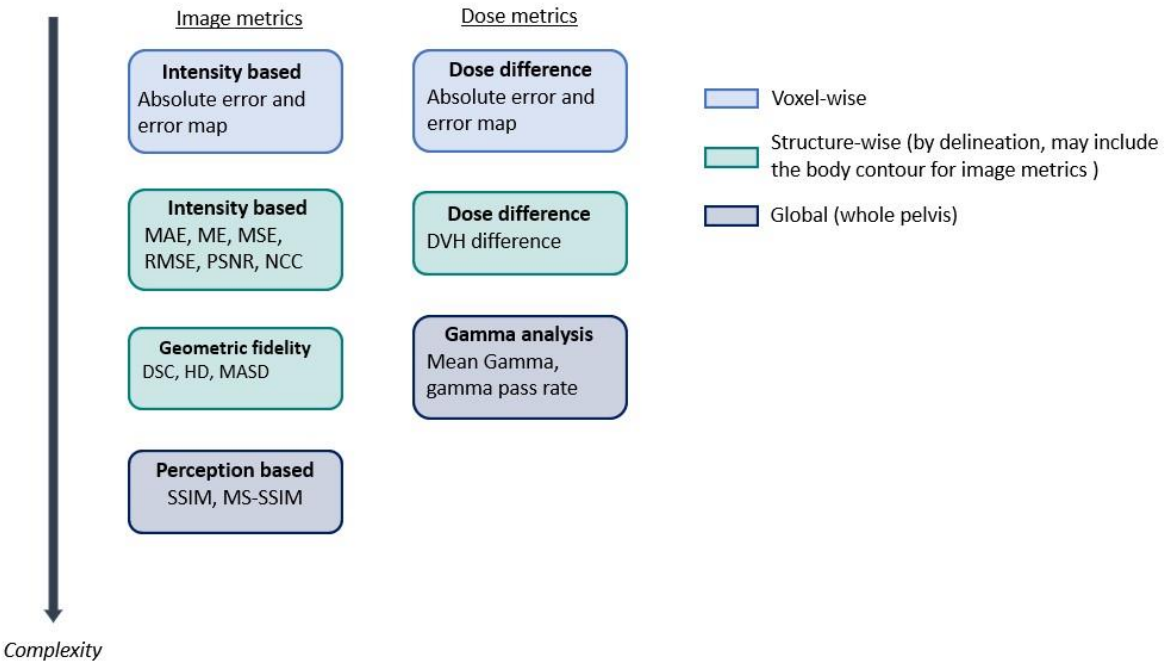

Figure 1.9: Imaging and dose metrics used for the evaluation of synthetic-CT.
These diverse metrics provide a comprehensive assessment of the image quality, ranging from voxel-level analysis to overall body evaluation, including delineated organs. Abbreviations: MAE: Mean absolute error, ME: mean error, MSE: Me an square error, RMSE: Root mean square error, PSNR: Peak signal-to-noise-ratio, NMI: Normalized mutual information, NCC: normalized cross-correlation, DSC: Dice score coefficient, HD: Hausdorff distance, MASD: Mean absolute surface distance, SSIM: structural similarity, MS-SSIM: Multi-scale SSIM, VIF: Visual information fidelity, DVH: Dose-volume histogram.

Patient specific sCT QA

In practice, patient specific sCT assessment can be used for validation (offline or online) or to validate sCT generation methods as part of a clinical evaluation stage. In the literature, CBCTs were used to assess patient-specific sCTs generated from MRI by comparing dose distributions and CT number accuracy in both images [67]–[69]. For sCT used in an adaptive proton therapy workflow, a range probing approach has been proposed by Oria et al.[70]. Film and 3D gel dosimetry were also investigated. It relies on printing 3D case-specific phantoms[71] . Choi et al.[72] proposed a strategy for assessing sCT by comparing the resulting dose distribution with the dose distribution obtained from a bulk- density image.
Probabilistic estimation of errors in sCT at a voxel level has also been explored in previous studies. Van Harten et al.[73] proposed a method to obtain a voxel-wise uncertainty map by analysing the discrepancies between sCT volume reconstructions on different axes (axial, sagittal, and coronal) using a DLM trained on 2D data for each axis. DLM also allows for the computation of epistemic (model-dependent) and aleatoric (data-dependent) uncertainties[30], [74], [75]. Johansson et al. [76], introduced a method estimating the probability of error in sCT generated from a Gaussian mixture model. These methods provide

3D maps of the probability of errors but are developed for specific models and provide estimates without generalisability.

Conclusion

Several metrics exist for assessing image quality, and in the context of sCT quality assurance, the ones currently utilised are full-reference metrics. However, they pose a limitation as they require a planning CT as a ground truth, which may not be available in an MRI-only workflow. The methods proposed to tackle this issue in the literature are defined for specific sCT generation methods or specific application (i.e ART with CBCT or proton therapy).
In contrast, existing no-reference IQA metrics rely on extensive datasets for model training and provide a global evaluation of image quality. However, before these models can be effectively implemented in clinical workflow, they would need to be trained on large multicentre datasets.
It is important to note that an sCT image does not need to be perfect to enable safe treatment planning in MRI-only radiation therapy. Therefore, it is crucial to establish acceptance criteria for errors in generated sCT and develop a standardised QA method for sCT validation that can be universally applied.

References

[1] M. Boulanger *et al.*, "Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review," *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[2] T. Nyholm, M. Nyberg, M. G. Karlsson, and M. Karlsson, "Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments," *Radiation Oncology*, vol. 4, no. 1, Nov. 2009, doi: 10.1186/1748-717X-4-54.

[3] J. Jalil ur Rehman *et al.*, "Intensity modulated radiation therapy: A review of current practice and future outlooks," *J Radiat Res Appl Sci*, vol. 11, no. 4, pp. 361–367, 2018, doi: https://doi.org/10.1016/j.jrras.2018.07.006.

[4] G. Noël, D. Antoni, I. Barillot, and B. Chauvet, "Délinéation des organes à risque et contraintes dosimétriques," *Cancer/Radiothérapie*, vol. 20, pp. S36–S60, 2016, doi: https://doi.org/10.1016/j.canrad.2016.07.032.

[5] J. Dowling *et al.*, *Image synthesis for MRI-only radiotherapy treatment planning*. 2022. doi: 10.1016/B978-0-12-824349-7.00027-X.

[6] J. Ng *et al.*, "MRI-LINAC: A transformative technology in radiation oncology," *Frontiers in Oncology*, vol. 13. Frontiers Media S.A., Jan. 27, 2023. doi: 10.3389/fonc.2023.1117874.

[7] D. Yan, “Adaptive Radiotherapy: Merging Principle Into Clinical Practice,” *Semin Radiat Oncol*, vol. 20, no. 2, pp. 79–83, 2010, doi: https://doi.org/10.1016/j.semradonc.2009.11.001.

[8] P. J. Keall *et al.*, “Review of Real-Time 3-Dimensional Image Guided Radiation Therapy on Standard-Equipped Cancer Radiation Therapy Systems: Are We at the Tipping Point for the Era of Real-Time Radiation Therapy?,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 102, no. 4, pp. 922–931, 2018, doi: https://doi.org/10.1016/j.ijrobp.2018.04.016.

[9] J. Bertholet *et al.*, “Real-time intrafraction motion monitoring in external beam radiotherapy,” *Phys Med Biol*, vol. 64, no. 15, p. 15TR01, Aug. 2019, doi: 10.1088/1361-6560/ab2ba8.

[10] M. P. P. N. Y. P. M. L.-C. A. Gervaise F. Esperabe-Vignau, “Évaluation des connaissances des prescripteurs de scanner en matière de radioprotection des patients,” *J Radiol*, vol. 1845, no. 7, pp. 621–749, 2011, doi: http://dx.doi.org/10.1016/j.jradio.2011.03.023.

[11] E. Johnstone *et al.*, “Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 100, no. 1, pp. 199–217, Jan. 2018, doi: 10.1016/j.ijrobp.2017.08.043.

[12] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” *Med Phys*, 2021, doi: 10.1002/mp.15150.

[13] J. Lambert *et al.*, “MRI-guided prostate radiation therapy planning: Investigation of dosimetric accuracy of MRI-based dose planning,” *Radiotherapy and Oncology*, vol. 98, no. 3, pp. 330–334, Mar. 2011, doi: 10.1016/j.radonc.2011.01.012.

[14] M. Köhler, T. Vaara, M. van Grootel, R. M. Hoogeveen, R. Kemppainen, and S. Renisch, “MR-only simulation for radiotherapy planning White paper: Philips MRCAT for prostate dose calculations using only MRI data,” 2015.

[15] J. Kim *et al.*, “Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy,” *Radiation Oncology*, vol. 10, no. 1, 2015, doi: 10.1186/s13014-015-0549-7.

[16] J. Sjölund, D. Forsberg, M. Andersson, and H. Knutsson, “Generating patient specific pseudo-CT of the head from MR using atlas-based regression,” *Phys Med Biol*, vol. 60, no. 2, p. 825, Jan. 2015, doi: 10.1088/0031-9155/60/2/825.

[17] J. A. Dowling *et al.*, “An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy,” *Int J Radiat Oncol Biol Phys*, vol. 83, no. 1, 2012, doi: 10.1016/j.ijrobp.2011.11.056.

[18] J. Uh, T. E. Merchant, Y. Li, X. Li, and C. Hua, “MRI-based treatment planning with pseudo CT generated through atlas registration,” *Med Phys*, vol. 41, no. 5, p. 51711, 2014, doi: https://doi.org/10.1118/1.4873315.

[19] B. Demol, C. Boydev, J. Korhonen, and N. Reynaert, “Dosimetric characterization of MRI-only treatment planning for brain tumors in atlas-based pseudo-CT images generated from

standard T 1-weighted MR images," *Med Phys*, vol. 43, no. 12, pp. 6557–6568, Dec. 2016, doi: 10.1118/1.4967480.

[20] J. A. Dowling *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*, vol. 93, no. 5, pp. 1144–1153, 2015, doi: 10.1016/j.ijrobp.2015.08.045.

[21] M. J. and G. F. and V. C. and M. M. and M. J. and K. A.-C. and P. S. and A. D. and A. S. R. and H. B. F. and O. S. Burgos Ninon and Cardoso, "Robust CT Synthesis for Radiotherapy Planning: Application to the Head and Neck Region," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, J. and W. W. M. and F. A. Navab Nassir and Hornegger, Ed., Cham: Springer International Publishing, 2015, pp. 476–484.

[22] C. Wang, M. Chao, L. Lee, and L. Xing, "MRI-based Treatment Planning with Electron Density Information Mapped from CT Images: A Preliminary Study," *Technol Cancer Res Treat*, vol. 7, no. 5, pp. 341–347, 2008, doi: 10.1177/153303460800700501.

[23] D. Andreasen, K. Van Leemput, and J. M. Edmund, "A patch-based pseudo-CT approach for MRI-only radiotherapy in the pelvis," *Med Phys*, vol. 43, no. 8Part1, pp. 4742–4752, 2016, doi: https://doi.org/10.1118/1.4958676.

[24] W. Boukellouz and A. Moussaoui, "Magnetic resonance-driven pseudo CT image using patch-based multi-modal feature extraction and ensemble learning with stacked generalisation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 8, pp. 999–1007, 2021, doi: https://doi.org/10.1016/j.jksuci.2019.06.002.

[25] A. M. Dinkla *et al.*, "Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network," *Med Phys*, vol. 46, no. 9, pp. 4095–4104, 2019, doi: 10.1002/mp.13663.

[26] A. Largent *et al.*, "Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning," *Int J Radiat Oncol Biol Phys*, vol. 105, no. 5, pp. 1137–1150, 2019, doi: 10.1016/j.ijrobp.2019.08.049.

[27] Y. Lei *et al.*, "MRI-based synthetic CT generation using semantic random forest with iterative refinement," *Phys Med Biol*, vol. 64, no. 8, 2019, doi: 10.1088/1361-6560/ab0b66.

[28] T. Huynh *et al.*, "Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model," *IEEE Trans Med Imaging*, vol. 35, no. 1, pp. 174–183, 2016, doi: 10.1109/TMI.2015.2461533.

[29] D. Andreasen, K. Van Leemput, R. H. Hansen, J. A. L. Andersen, and J. M. Edmund, "Patch-based generation of a pseudo CT from conventional MRI sequences for MRI-only radiotherapy of the brain," *Med Phys*, vol. 42, no. 4, pp. 1596–1605, 2015, doi: https://doi.org/10.1118/1.4914158.

[30] M. Hemsley *et al.*, "Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions," 2020, pp. 834–844. doi: 10.1007/978-3-030-59710-8_81.

[31] D. Bird *et al.*, “Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning,” *Radiotherapy and Oncology*, vol. 156, pp. 23–28, 2021, doi: 10.1016/j.radonc.2020.11.027.

[32] H. Emami, M. Dong, S. P. Nejad-Davarani, and C. K. Glide-Hurst, “Generating synthetic CTs from magnetic resonance images using generative adversarial networks,” *Med Phys*, vol. 45, no. 8, pp. 3627–3636, 2018, doi: 10.1002/mp.13047.

[33] X. Han, “MR-based synthetic CT generation using a deep convolutional neural network method:,” *Med Phys*, vol. 44, no. 4, pp. 1408–1419, 2017, doi: 10.1002/mp.12155.

[34] Y. Lei *et al.*, “MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks,” *Med Phys*, vol. 46, no. 8, pp. 3565–3581, 2019, doi: 10.1002/mp.13617.

[35] M. Lerner, J. Medin, C. Jamtheim Gustafsson, S. Alkner, C. Siversson, and L. E. Olsson, “Clinical validation of a commercially available deep learning software for synthetic CT generation for brain,” *Radiation Oncology*, vol. 16, no. 1, pp. 1–11, 2021, doi: 10.1186/s13014-021-01794-6.

[36] Y. Zhao *et al.*, “Compensation cycle consistent generative adversarial networks (Comp-GAN) for synthetic CT generation from MR scans with truncated anatomy,” *Med Phys*, Feb. 2023, doi: 10.1002/mp.16246.

[37] B. Tang *et al.*, “Dosimetric evaluation of synthetic CT image generated using a neural network for MR-only brain radiotherapy,” *J Appl Clin Med Phys*, vol. 22, no. 3, pp. 55–62, 2021, doi: 10.1002/acm2.13176.

[38] J. Fu *et al.*, “Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging,” *Med Phys*, vol. 46, no. 9, pp. 3788–3798, 2019, doi: 10.1002/mp.13672.

[39] Y. Li, S. Xu, Y. Lu, and Z. Qi, “CT synthesis from MRI with an improved multi-scale learning network,” *Front Phys*, vol. 11, Jan. 2023, doi: 10.3389/fphy.2023.1088899.

[40] A. Largent *et al.*, “Pseudo-CT Generation for MRI-Only Radiation Therapy Treatment Planning: Comparison Among Patch-Based, Atlas-Based, and Bulk Density Methods,” *Int J Radiat Oncol Biol Phys*, vol. 103, no. 2, pp. 479–490, 2019, doi: 10.1016/j.ijrobp.2018.10.002.

[41] N. Tyagi *et al.*, “Dosimetric and workflow evaluation of first commercial synthetic CT software for clinical use in pelvis,” *Phys Med Biol*, vol. 62, no. 8, pp. 2961–2975, 2017, doi: 10.1088/1361-6560/aa5452.

[42] L. M. O’Connor *et al.*, “Optimisation and validation of an integrated magnetic resonance imaging-only radiotherapy planning solution,” *Phys Imaging Radiat Oncol*, vol. 20, pp. 34–39, 2021, doi: https://doi.org/10.1016/j.phro.2021.10.001.

[43] S. J. Hoogcarspel, J. M. Van der Velden, J. J. W. Lagendijk, M. van Vulpen, and B. 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*, vol. 59, no. 23, p. 7383, Nov. 2014, doi: 10.1088/0031-9155/59/23/7383.

[44] M. Kapanen and M. 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 (Madr)*, vol. 52, no. 3, pp. 612–618, 2013, doi: 10.3109/0284186X.2012.692883.

[45] J. Korhonen, M. Kapanen, J. Keyriläinen, T. Seppälä, and M. 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*, vol. 41, no. 1, p. 11704, 2014, doi: https://doi.org/10.1118/1.4842575.

[46] C. Wachinger, M. Brennan, G. C. Sharp, and P. Golland, “Efficient Descriptor-Based Segmentation of Parotid Glands With Nonlocal Means,” *IEEE Trans Biomed Eng*, vol. 64, no. 7, pp. 1492–1502, 2017, doi: 10.1109/TBME.2016.2603119.

[47] M. C. Florkow *et al.*, “The impact of MRI-CT registration errors on deep learning-based synthetic CT generation,” SPIE-Intl Soc Optical Eng, 2019, p. 116. doi: 10.1117/12.2512747.

[48] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[49] O. Dalmaz, M. Yurt, and T. Çukur, “ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis,” *IEEE Trans Med Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022, doi: 10.1109/TMI.2022.3167808.

[50] B. Zhao *et al.*, “CT synthesis from MR in the pelvic area using Residual Transformer Conditional GAN,” *Computerized Medical Imaging and Graphics*, vol. 103, p. 102150, 2023, doi: https://doi.org/10.1016/j.compmedimag.2022.102150.

[51] X. Li, K. Shang, G. Wang, and M. D. Butala, “DDMM-Synth: A Denoising Diffusion Model for Cross-modal Medical Image Synthesis with Sparse-view Measurement Embedding,” 2023.

[52] S. Pan *et al.*, “Synthetic CT Generation from MRI using 3D Transformer-based Denoising Diffusion Model,” 2023.

[53] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 2242–2251, 2017, doi: 10.1109/ICCV.2017.244.

[54] H. Yang *et al.*, “Unpaired Brain MR-to-CT Synthesis using a Structure-Constrained CycleGAN,” 2018, [Online]. Available: http://arxiv.org/abs/1809.04536

[55] A. Jabbarpour, S. R. Mahdavi, A. Vafaei Sadr, G. Esmaili, I. Shiri, and H. Zaidi, “Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and CycleGAN: Dosimetric assessment for 3D conformal radiotherapy,” *Comput Biol Med*, vol. 143, Apr. 2022, doi: 10.1016/j.compbiomed.2022.105277.

[56] C. Wang *et al.*, “Toward MR-only proton therapy planning for pediatric brain tumors: Synthesis of relative proton stopping power images with multiple sequence MRI and

development of an online quality assurance tool," *Med Phys*, vol. 49, no. 3, pp. 1559–1570, Mar. 2022, doi: 10.1002/mp.15479.

[57] M. C. Florkow *et al.*, "Deep learning–based MR-to-CT synthesis: The influence of varying gradient echo–based MR images as input channels," *Magn Reson Med*, vol. 83, no. 4, pp. 1429–1441, Apr. 2020, doi: 10.1002/mrm.28008.

[58] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006, doi: 10.1109/TIP.2005.859378.

[59] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[60] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011, pp. 723–727. doi: 10.1109/ACSSC.2011.6190099.

[61] N. Venkatanath, D. Praneeth, B. H. Maruthi Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 21st National Conference on Communications, NCC 2015*, Institute of Electrical and Electronics Engineers Inc., Apr. 2015. doi: 10.1109/NCC.2015.7084843.

[62] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans Neural Netw Learn Syst*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015, doi: 10.1109/TNNLS.2014.2336852.

[63] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the Use of Deep Learning for Blind Image Quality Assessment," Feb. 2016, doi: 10.1007/s11760-017-1166-8.

[64] Q. Gao *et al.*, "Combined global and local information for blind CT image quality assessment via deep learning," SPIE-Intl Soc Optical Eng, Mar. 2020, p. 39. doi: 10.1117/12.2548953.

[65] G. Wang, Y. Zhang, X. Ye, and X. Mou, "Image quality assessment," in *Machine Learning for Tomographic Imaging*, in 2053-2563. IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

[66] R. Kemppainen *et al.*, "Assessment of dosimetric and positioning accuracy of a magnetic resonance imaging-only solution for external beam radiotherapy of pelvic anatomy," *Phys Imaging Radiat Oncol*, vol. 11, pp. 1–8, Jul. 2019, doi: 10.1016/j.phro.2019.06.001.

[67] E. Palmér, E. Persson, P. Ambolt, C. Gustafsson, A. Gunnlaugsson, and L. E. Olsson, "Cone beam CT for QA of synthetic CT in MRI only for prostate patients," *J Appl Clin Med Phys*, vol. 19, no. 6, pp. 44–52, Nov. 2018, doi: 10.1002/acm2.12429.

[68] J. J. Wyatt, R. A. Pearson, C. P. Walker, R. L. Brooks, K. Pilling, and H. M. McCallum, "Cone beam computed tomography for dose calculation quality assurance for magnetic resonance-only radiotherapy," *Phys Imaging Radiat Oncol*, vol. 17, pp. 71–76, Jan. 2021, doi: 10.1016/j.phro.2021.01.005.

[69] S. Irmak, L. Zimmermann, D. Georg, P. Kuess, and W. Lechner, “Cone beam CT based validation of neural network generated synthetic CTs for radiotherapy in the head region,” *Med Phys*, vol. 48, no. 8, pp. 4560–4571, Aug. 2021, doi: 10.1002/mp.14987.

[70] C. Seller Oria *et al.*, “Range probing as a quality control tool for CBCT-based synthetic CTs: In vivo application for head and neck cancer patients,” *Med Phys*, vol. 48, no. 8, pp. 4498–4505, Aug. 2021, doi: 10.1002/mp.15020.

[71] S. Neppl *et al.*, “Measurement-based range evaluation for quality assurance of CBCT-based dose calculations in adaptive proton therapy,” *Med Phys*, vol. 48, no. 8, pp. 4148–4159, Aug. 2021, doi: 10.1002/mp.14995.

[72] J. H. Choi *et al.*, “Bulk Anatomical Density Based Dose Calculation for Patient-Specific Quality Assurance of MRI-Only Prostate Radiotherapy,” *Front Oncol*, vol. 9, 2019, doi: 10.3389/fonc.2019.00997.

[73] L. D. van Harten, J. M. Wolterink, J. J. C. Verhoeff, and I. Išgum, “Automatic online quality control of synthetic CTs,” SPIE-Intl Soc Optical Eng, Mar. 2020, p. 57. doi: 10.1117/12.2549286.

[74] M. Abdar *et al.*, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” 2020, [Online]. Available: http://arxiv.org/abs/2011.06225

[75] C. A. T. van den Berg and E. F. Meliadò, “Uncertainty Assessment for Deep Learning Radiotherapy Applications,” *Seminars in Radiation Oncology*. W.B. Saunders, Oct. 01, 2022. doi: 10.1016/j.semradonc.2022.06.001.

[76] A. Johansson and M. Karlsson, “Voxel-wise uncertainty in CT substitute derived from MRI,” vol. 39, no. June, pp. 3283–3290, 2012.

Chapter 2: Aims of the thesis

Numerous methods for generating synthetic CTs have been developed, and recent advancements in deep learning have facilitated the production of accurate results. However, for the systematic use of MRI-based dose planning in clinical routine, the issue of quality control for the generated images still needs to be addressed.

The main objectives of this thesis are as follows:

- To identify the limitations and shortcomings of the synthetic CT generation methods through statistical evaluation. This will provide a better understanding of their capabilities and constraints.

- To quantify the impact of errors in intensity on dose distribution. By measuring the effects of these errors, the goal is to assess their significance and potential impact on treatment.

- To develop a framework for evaluating the quality of each patient specific sCT. This assessment will ensure that the resulting images meet the required standards and are acceptable for treatment planning purposes.

The research conducted in this thesis aims to address these objectives and will contribute to the development of recommended best practices for inclusion in a clinical protocol. The objectives and the content of the various chapters comprising the thesis are presented in Figure 2.1, providing a comprehensive overview of the study's scope and structure.

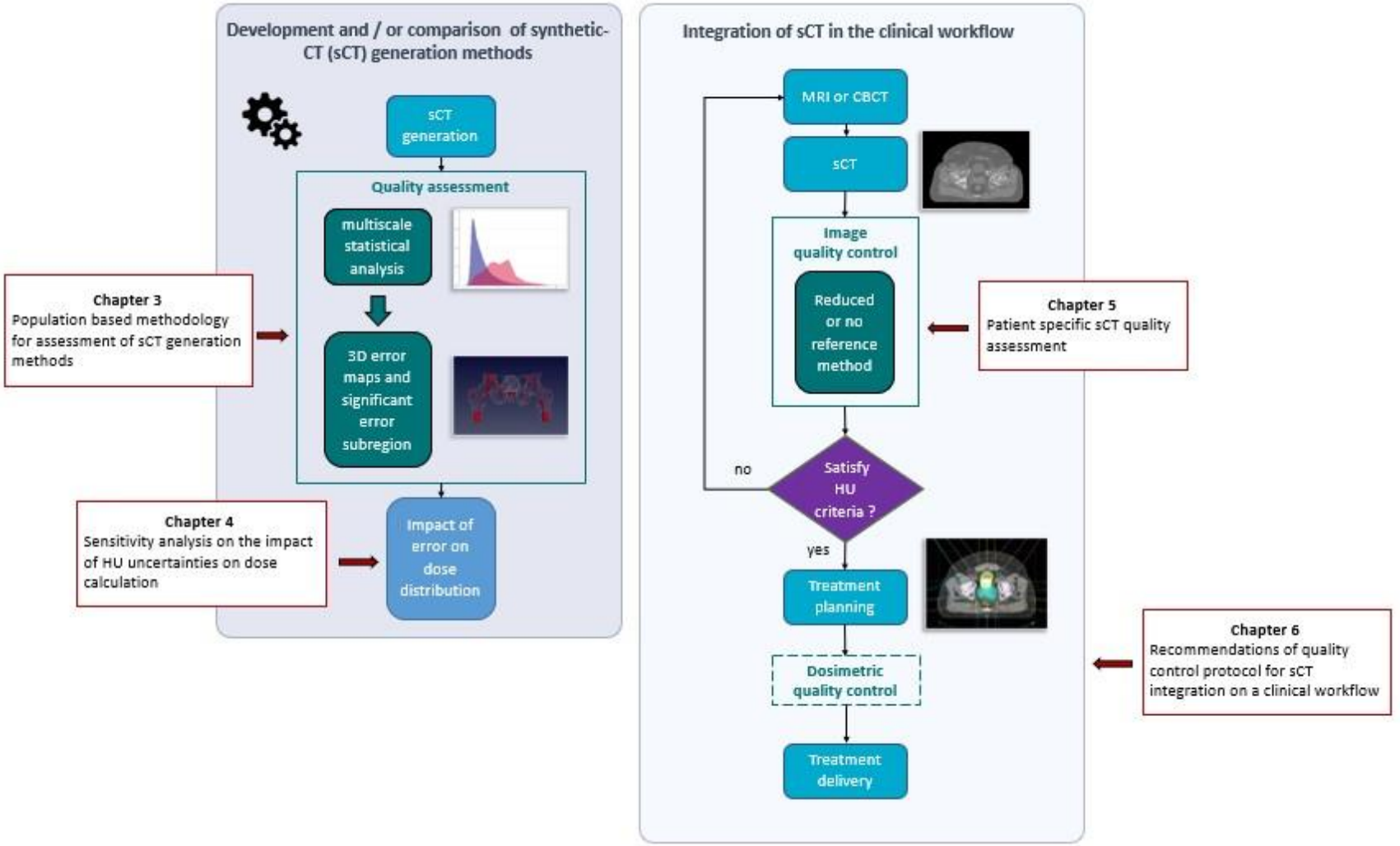

Figure 2.1: Contribution of the thesis. The first part of the thesis (chapters 3 and 4) focuses on the analysis of the limitations of synthetic CT generation methods and how the errors in Hounsfield Units (HU) will impact the treatment calculation. The chapter 5 presents non-reference methods for patient-specific sCT quality assessment. Chapter6 proposed recommendations for quality control steps to integrate daily use of sCT into the clinical routine.

Chapter 3: Quality assurance for MRI-only radiation therapy: A voxel-wise population-based methodology for image and dose assessment of synthetic-CT generation methods

This chapter introduces a methodology for comprehensive assessment of synthetic-CT generation methods at the voxel level, encompassing both image quality and dose accuracy. This evaluation may be used prior to incorporating a specific sCT generation approach into a clinical workflow to ensure its robustness and reliability. By following this methodology, a better understanding of the capabilities and limitations of these methods can also be achieved.
The content of this chapter has been published in Frontiers in Oncology in 2022.

"Quality assurance for MRI-only radiation therapy: a voxel-wise population-based methodology for image and dose assessment of synthetic-CT generation methods"
Hilda Chourak, Anaïs Barateau, Safaa Tahri, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Adrien Boue-Rafle, Mathias Perazzi, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta (Frontiers in Oncology, 2022)

Preliminary versions of this work were presented at international and national conferences.
"Voxel-Wise Analysis for Spatial Characterisation of Pseudo-CT Errors in MRI-Only Radiotherapy Planning"
Hilda Chourak, Anaïs Barateau, Eugenia Mylona, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - ISBI 2021 (poster)
"Spatial Characterization of errors in pseudo-CT generation for MRI-only radiotherapy"
Hilda Chourak, Anaïs Barateau, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - ESTRO 2021 (poster)
"Caractérisation spatiale d'erreurs de pseudo-CT pour la planification de dose à partir d'IRM"
Hilda Chourak, Anaïs Barateau, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - SFPM 2021 (oral presentation)

Abstract

Quality assurance (QA) of synthetic-CT (sCT) is crucial for safe clinical transfer to an MRI-only radiotherapy planning workflow. The aim of this work is to propose a population-based process assessing local errors in the generation of sCTs, and their impact on dose distribution.

For the analysis to be anatomically meaningful, a customised inter-patient registration method brought the population data to the same coordinate system. Then, the voxel-based process was applied on two sCT generation methods: a bulk-density method (BDM) and a Generative Adversarial Network (GAN). CT and MRI pairs of 39 patients treated by radiotherapy for prostate cancer were used for sCT generation, and 26 of them with delineated structures were selected for the analysis. Voxel-wise errors in sCT compared to CT were assessed for image intensities and dose calculation, and a population-based statistical test was applied to identify regions where discrepancies were significant. Cumulative histograms of mean absolute dose error per volume of tissue were computed to give a quantitative indication of the error for each generation method.
Accurate inter-patient registration was achieved, with mean Dice scores higher than 0.91 for all organs. The proposed method produces 3D maps that precisely show the location of the major discrepancies for both sCT generation, highlighting the heterogeneity of image and dose errors for sCT generation methods from MRI across the pelvic anatomy. Hence, this method provides additional information that will assist with both sCT development and quality control for MRI-based planning radiotherapy.

Introduction

Magnetic resonance imaging (MRI) is becoming increasingly integrated into clinical radiotherapy (RT) planning and monitoring. MRI guided RT is motivated by the superior soft tissue contrast compared to CT and the non-ionizing modality. However, MRI does not provide information on electron density of tissue, essential for radiotherapy dose calculation. To overcome this issue, several approaches to generate synthetic CT (sCT) in Hounsfield Units (HU) from a specific MRI have been developed[1], [2]. These include: bulk density[3], [4], atlas-based[5], machine-learning models, such as patch-based methods with feature extraction[6], and more recently deep-learning models (DLMs)[6]–[12].
Currently, sCT image quality assessment is based on global metrics that measure discrepancies between reference CT and the corresponding sCT[12], [13]. The most common are intensity-based[14] metrics, like mean absolute error (MAE), mean error (ME) , mean squared error (MSE) and peak signal to-noise ratio (PSNR). Structural similarity (SSIM)[15], [16] is also often computed. These metrics have been reported at a global level, restricted to a single value describing agreement within the body contour of the patient, or within an organ[12]. Regarding dosimetric evaluation, dose distributions obtained from sCT are assessed by comparing dose-volume histogram (DVH) and gamma analysis[17]–[20] to the ground truth (dose distribution from reference CT).
DVHs are volume-based statistics that are not relatable to spatial locations; while gamma are spatial distributions, they are usually condensed to a single pass-rate metric and gamma scores are difficult to interpret clinically. For sCT evaluations each patient is usually assessed

in isolation and results are then combined. However, it has been reported that errors might appear heterogeneously distributed across different tissue densities[6], [16], [21]–[24].
Assessing the spatial distribution of errors at a population level may help to identify their origin as well as clinical impact and may subsequently improve the accuracy of sCT generation methods. It can also be useful to compare and select sCT generation methods, and to a large extent it may lead to the introduction of quality control protocols within the MRI-based RT planning workflow.
Voxel-wise population analysis can provide powerful tools to assess clinical impacts of image and dose difference across individuals[25], [26]. However, their application requires an accurate non-rigid registration of a whole population to a single coordinate system, and the implementation of voxel-wise statistical tests. Previous preliminary work has demonstrated the feasibility of this method, but the analysis methods were limited in clinical scope [27].
The aim of this paper was to propose a multiscale strategy to assess accuracy of sCT generation methods, starting with a standard error evaluation in the whole pelvis, followed by assessment of organ errors and finally by the implementation of a voxel-wise workflow.
The whole scan population was brought to the same coordinate system via a customized non-rigid registration method. Two different sCT generation approaches were chosen as examples to illustrate the methodology: a bulk-density method (BDM) and a deep-learning method, based upon a generative adversarial network (GAN) architecture [6], [28]. Then a comprehensive population based statistical analysis is performed, including a permutation test adapted to non-parametric paired data and the evaluation of the error dispersion at a voxel-wise scale for each method. The presented methodology provides not only a population spatial quantification of sCT image value and dose errors, but it also allows comparison across different sCT generation approaches using the same dataset.

Materials and Methods

Data

A cohort of 39 patients with prostate cancer aged 58 to 78 years were used to generate sCT scans. For each patient, a CT scan was acquired on a GE LightSpeed RT or a Toshiba Aquilion, (256 x 256 x 128 matrix with a voxel size of 1.17 mm x 1.17 mm x 2.5 mm or 2.0 mm) and a T2-weighted MRI was acquired on a Siemens Skyra 3T in the treatment position (resolution of 1.6 mm x 1.6 mm x 1.6 mm). Each CT was resampled and registered to the corresponding MRI via a symmetric rigid registration followed by a structure-guided non-rigid method[29], [30] to rectify the main anatomical variations due to the delay between both acquisitions.
MRI were then pre-processed to correct non-uniformity [31] with the Insight Toolkit Library (ITK).
As some organs' delineation, crucial for the interpatient-registration, were incomplete, the voxel-wise analysis was performed on the 26 patients with bones, prostate, bladder, and

rectum delineated on MRI by 2 physicians. The rectal length started at 2 cm below the clinical target volume (CTV). Two clinical target volumes (CTVs) were defined: CTV1 including prostate and seminal vesicles, and CTV2 corresponding to the prostate only.

Workflow

The proposed workflow is presented in Figure 3.1. It includes the generation of sCTs using two methods (BDM and GAN) and dose computation. Then, sCTs and dose distributions followed a standard evaluation in the native space. Finally, an accurate customized organ-driven non-rigid algorithm was applied to bring all the data to the same coordinate system, where voxel-wise analysis was performed.

sCT generation methods

Bulk-density method (BDM)

Bulk density methods have application to quality assurance of sCT scans [4] and are also employed in this work to demonstrate that differences between scan quality for different sCT method can be determined with our workflow. sCTs were obtained by assigning Hounsfield Units (HU) values to the patient's soft tissue, bones and air segmented from MRI. For bone segmentation, automatic tools from Varian Eclipse were used on CT. This contour was then rigidly aligned to the MRI scan and contours were manually adjusted by a research radiation therapist [31]. The volume of air resulted from thresholds in the inner part of the rectum delineated on MRI. The soft tissue area corresponds to the subtraction of bones and air from the body contour. A water equivalent density (0 HU) was assigned to the soft tissue [3], [32]. For bones and air, the densities allocated were respectively 350 HU and -450 HU, which are the mean CT values of the cohort in the corresponding segmented regions[28].

Generative Adversarial Network (GAN)

The GAN architecture used in this study to generate sCT is fully described in Largent et al.[6]. The generator was a U-Net inspired by Han et al.[33], with L2 norm as loss function:

$$L_G(I, C) = \left|\left|C - G(I)\right|\right|_2^2 \quad (1)$$

Where I corresponds to the MRI intensity, $G(I)$ to the generated sCT and C to the reference CT.

The discriminator was a PatchGAN, using binary cross entropy as loss function:

$$L_D(G(I), C) = -\sum_{i=1}^{n} C_i \log(G(I)_i) + (1 - C_i) \log(1 - G(I)_i) \quad (2)$$

$G(I)$ is the sCT produced by the generator from the target MRI, C the corresponding reference CT and n the number of voxels in C.$L_G(I, C)$ and $L_D(G(I), C)$ were combined to create the adversarial loss. Axial 2D CT and MRI slices were used to train the model, and a three-fold

cross validation was applied. The training cohort comprised 26 patient data and the validation cohort of size 13.

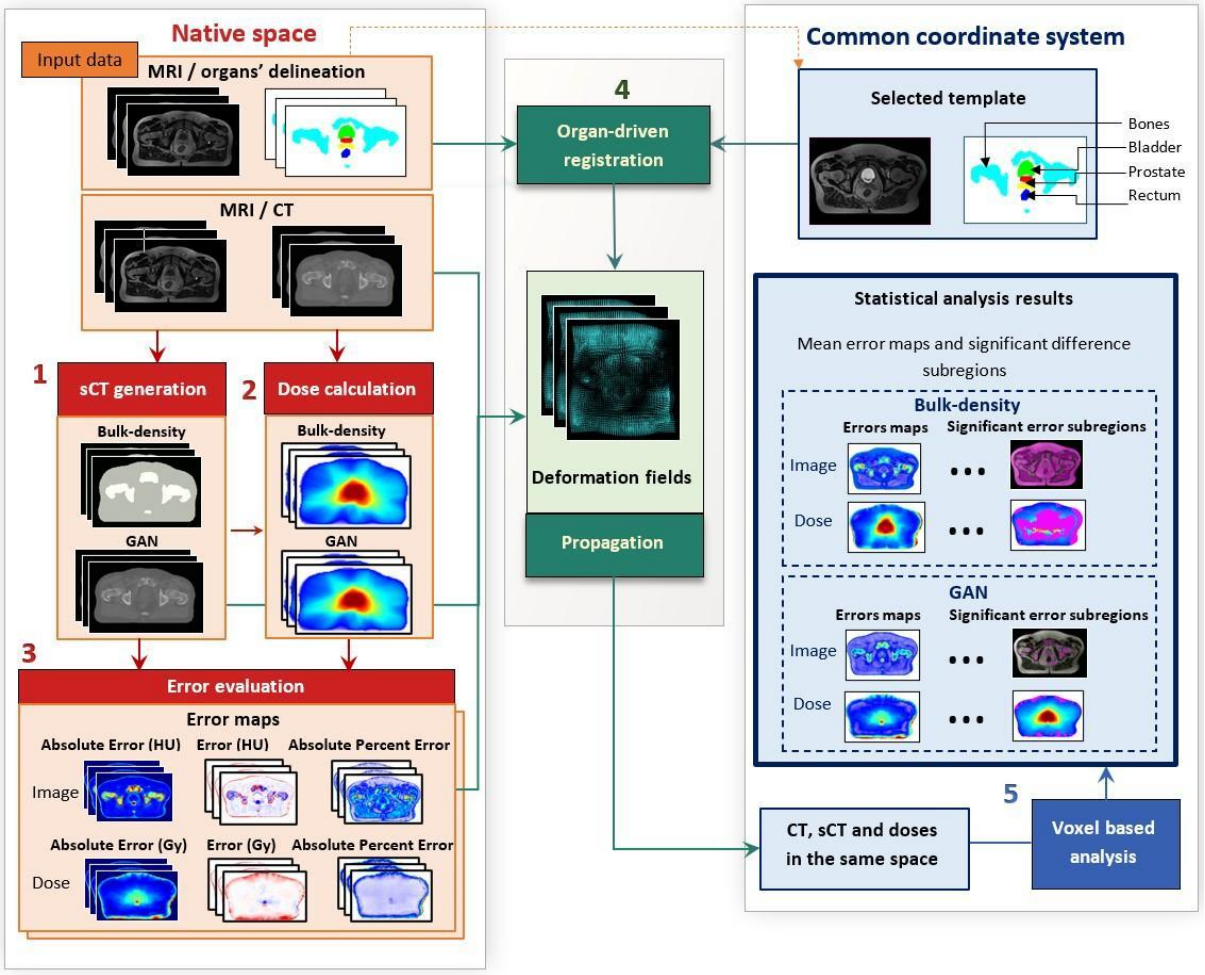

Figure 3.1: Workflow of voxel-wise population-based analysis
This workflow comprises 5 steps: (1) synthetic-CT (sCT) generation with a bulk-density and a Generative Adversarial Network (GAN) method, (2) dose calculation and (3) error evaluation of images and doses in the native space of each patient. This evaluation includes the computation of Absolute Error, Error, and the Absolute Percent Error. The non-rigid registration step (4) resulted in deformation fields, allowing for propagation of the whole data to a common coordinate system. Once all data were in the same anatomical space, statistical analysis was performed (5), producing 3D error maps for each sCT generation method and highlighting significant difference subregions for both image and dose distributions.

Dose calculation in native space

Volumetric modulated arc therapy (VMAT) was planned on reference CT images with the Pinnacle v.9.10 (Philips) treatment planning system (TPS) using the collapsed cone convolution algorithm and a dose grid resolution of 3 mm. For all patients, a sequential treatment was delivered with a total dose of 50 Gy to the CTV1, followed by a boost of 28 Gy

in the CTV2, both at 2 Gy per fraction. The beam parameters used to compute the dose on the reference CT were used to calculate dose on the sCT.

Image and dose error evaluation in native space

The accuracy of the sCT generation in HU and in Gy was first assessed in the native space, to reduce bias induced by the inter-patient non-rigid registration.
Absolute error (AE), error (E) and absolute percent error (APE) were computed by comparing corresponding CT and sCT pairs at a voxel level, producing 3D error maps for each patient.
The global quality of sCT was evaluated with respect to the patient's structures (prostate, rectum and bladder) and the whole pelvis by computing the mean absolute error (MAE), the mean error (ME) and the mean absolute percent error (MAPE) in these regions from the previous maps.

$$AE(i) = |X_{CT}(i) - X_{sCT}(i)| \quad (3a)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} AE(i) \quad (3b)$$

$$E(i) = X_{CT}(i) - X_{sCT}(i) \quad (4a)$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} E(i) \quad (4b)$$

$$APE(i) = \left|\frac{X_{CT}(i) - X_{sCT}(i)}{X_{CT}(i)}\right| \quad (5a)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} APE(i) \quad (5b)$$

with n the number of voxels, $X_{CT}(i)$ and $X_{sCT}(i)$ the intensities of the i^{th}voxel in, respectively, the reference and the generated image, in HU for image evaluation or in Gy for dose evaluation.
The closer to zero the AE, the E, the APE, and so their respective means, the more accurate is the prediction.

Organ-driven registration

First, an individual MRI scan from the cohort was selected as a template (exemplar) by considering the median volumes of bladder, rectum, and prostate. Then, a customized organ-driven registration, based upon previously proposed methods [25], [34] was performed with an overall optimized alignment across the organs.

Input images for the registration were a combination of MR images and structural descriptions (SD) of the delineated organs obtained as follows:

- Euclidean distances to the surface were computed for all structures[35].
- For the rectum, a scalar field was generated by applying the Laplacian equation inside the volume[36]. The Laplacian field provided a normalised distance map to the central path of the organ.
- For the prostate, the Laplacian was also computed with respect to its barycentre.

Finally, scalar fields of all structures were merged into a global structural description of the organs and combined to the MRI (Figure 3.2). Afterwards, all the structures were rigidly aligned using the Elastix toolbox (translation). From bones to bladder, each structure requires a different level of deformation. To handle this high variability, non-rigid registration based on diffeomorphic demons[37] with 4 levels of resolution was successively applied to: i) the bladder, ii) the whole pelvis, iii) the prostate, iv) the rectum, v) the bones.

The demons algorithm uses Gaussian regularisation, which involves smoothing the deformation field. The sigma of the Gaussian filter was set to 1, and the number of iterations for the 4 levels of resolution were: i) 300, 300, 200, 20 for the bladder contour, ii) 200, 200, 100,0 for the whole pelvis, iii) 200, 200, 150, 5 for the prostate SD , iv) 100, 100, 100,5 for the rectum SD, v) 100, 100, 150,50 for the bones SD.

For the bladder, a b-spline transform using the Elastix toolbox was also performed on SD prior to the demons registration (step i)).

Each step resulted in deformation fields: 3D vectors defined at each voxel and providing the appropriate transformation. The resulting 3D deformation fields were combined and applied to delineated structures, reference CTs, sCTs, dose planning and error maps to propagate all the data from their native spaces to a common coordinate system (CCS). After the propagation of CT in the CCS the bones, including the femoral heads, were split between spongy and cortical and separately registered to preserve their inner structure composition. This final transformation was then applied to sCT, dose and error maps.

For the propagation of CT in the CCS to be meaningful, each CT-MRI patient pair had to be properly co-registered prior to the inter-patient registration. This step-by-step approach can accommodate the high anatomical inter-individual variability, and facilitates the propagation of delineated structures, including the registered reference CTs, sCTs, dose distributions and the error maps from their native spaces to a common coordinate system (CCS).

As a visual indicator of the performance of this process, a checkerboard of the template MRI with the mean population MRI in the CCS, and a checkerboard of the template CT with the mean population CT in the CCS are presented in Figure 3.3. The probability maps, also in Figure 3.3, allow visualization of the discrepancies between the delineated organs contours following registration.

Table 3.2 summarises the volumes of the delineated organs prior and after the registration process. The Dice similarity coefficient (DSC) between the template structures, $V_{t_{MRI}}$, and the corresponding deformed delineated organ, V_{MRI}, was also used for validation.

$$DSC = \frac{2\left(V_{t_{MRI}} \cap V_{MRI}\right)}{V_{t_{MRI}} + V_{MRI}} \quad (6)$$

For the voxel-based population analysis to be meaningful, only accurately registered data were included (DSC > 0.85 for all the segmented organs). The 26 cases passed this criterion.

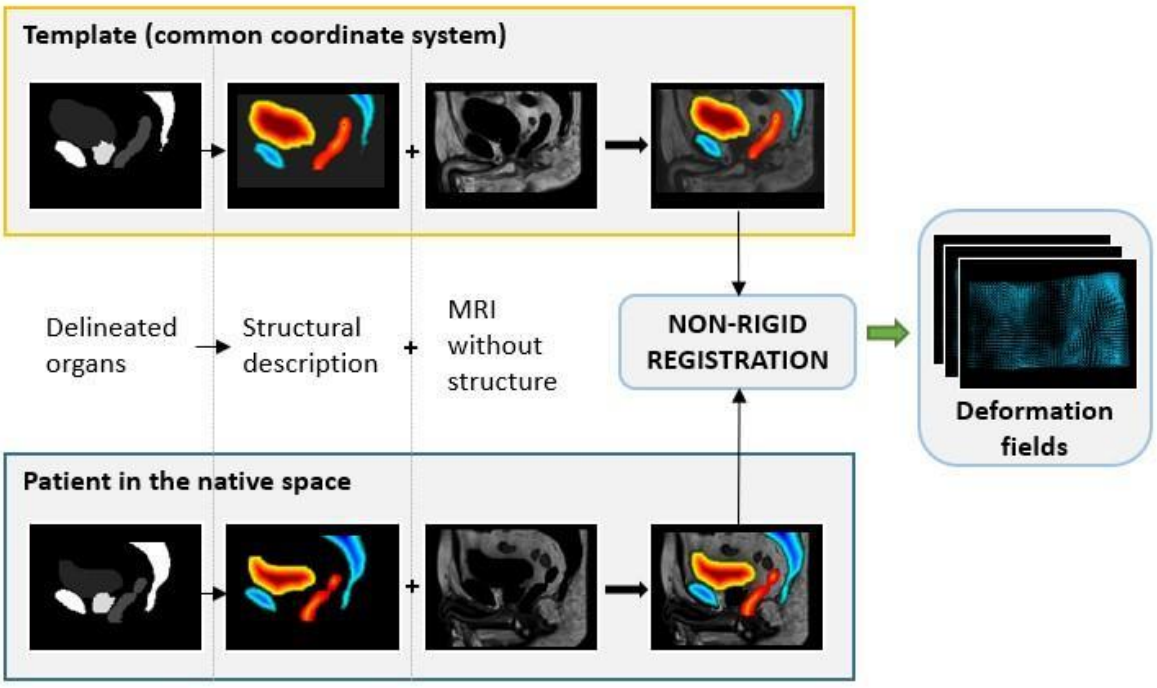

Figure 3.2: Pre-processing step for the non-rigid registration process.
After organ delineation, a structural description was performed by computing the Euclidean distances to the surface and the Laplacian equation. This was finally combined to MR images to obtain the deformation fields used to bring all the data from their native space to the common coordinate system (CCS).

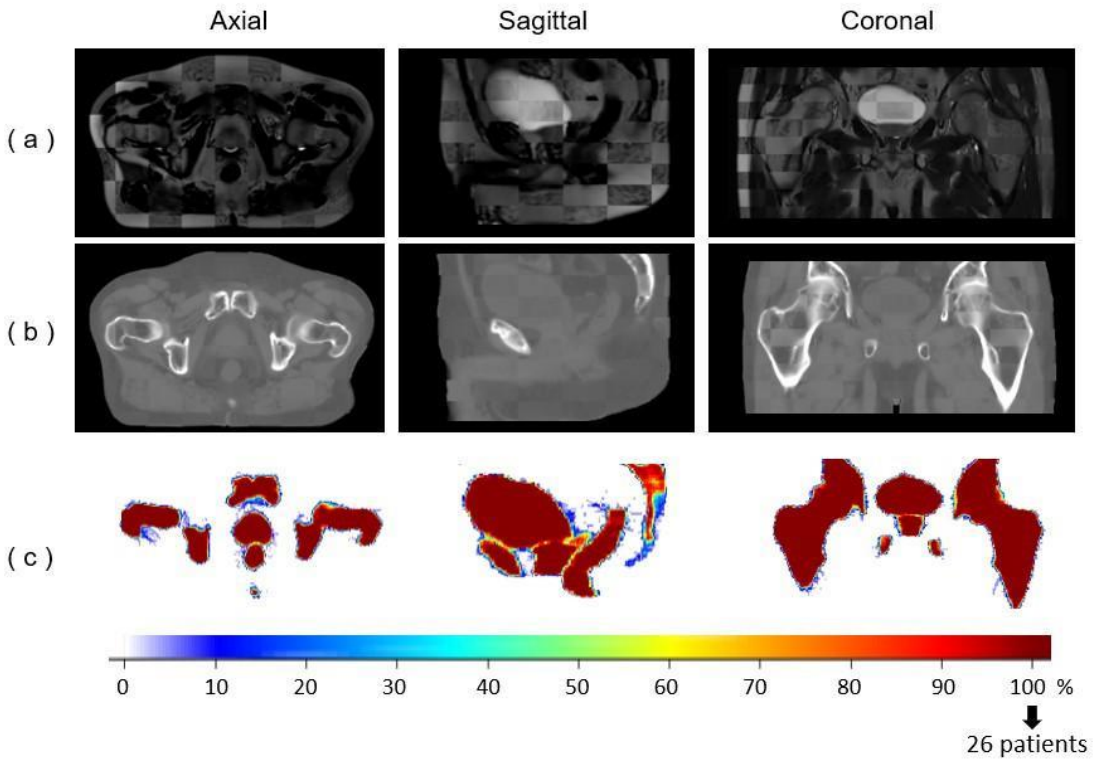

Figure 3.3: Visual quality control of the interpatient registration.
Checkerboard comparison of (a) the template MRI with the mean of all the population MRIs registered in the common coordinate system (CCS) and (b) the template CT with the mean population CTs in the CCS. Probability maps are presented in (c). It is the result of the overlapping of all the delineated structures in the same space to estimate the precision of the registration. In blue, few structures are overlaid (poor quality of registration). In red, all the patient structures correspond to the same anatomical location (100%, perfect registration).

Voxel-wise analysis in CCS

Image and dose mean error maps computation

Once all data were in the CCS, voxel-wise MAE (vMAE), ME (vME) and MAPE (vMAPE) maps for images and dose distributions were obtained by averaging the voxel errors data across the cohort. The v represents that these data are now voxel specific and hence spatial, i.e they are not averaged across a particular patient's voxels, they are found by considering all the patient cohort values for a particular voxel i.
So, in the CCS errors are defined as follow:

$$vMAE(i) = \frac{1}{p}\sum_{j=1}^{p} |X_{CT}(i,j) - X_{sCT}(i,j)| \quad (7)$$

$$vME(i) = \frac{1}{p}\sum_{j=1}^{p} X_{CT}(i,j) - X_{sCT}(i,j) \tag{8}$$

$$vMAPE(i) = \frac{1}{p}\sum_{j=1}^{p} \left|\frac{X_{CT}(i,j) - X_{sCT}(i,j)}{X_{CT}(i,j)}\right| \tag{9}$$

$vMAE(i)$ is the mean absolute error, $vME(i)$ the mean error and $vMAPE(i)$ the mean absolute percent error for a voxel i. $X_{CT}(i,j)$ and $X_{sCT}(i,j)$ represent the values, in HU for the image assessment or in Gy for the dose assessment, of the reference CT and the sCT, for the i^{th}voxel of the j^{th} image of the population, and p the total number of patients in the population.

The template scan body contour was applied to these images to focus on the region of interest and discard slight body contour variation due to registration. Then, the relative standard deviation of the absolute error (RSDAE), also known as coefficient of variation, was used for the evaluation of the dispersion of the prediction error at a voxel-wise scale.

$$RSD_{AE}(i) = \frac{\sqrt{\sum_{j=0}^{p}\left(AE(i,j) - vMAE(i)\right)^2}}{vMAE(i)} \tag{10}$$
$$\text{with } AE(i,j) = |X_{CT}(i,j) - X_{sCT}(i,j)|$$

So, for each voxel i , the lower is the RSDAE, the higher is the probability to have an absolute error close to the $vMAE(i)$ value. Figure 3.4 and Figure 3.5 illustrate the results, respectively for image and dose assessment.

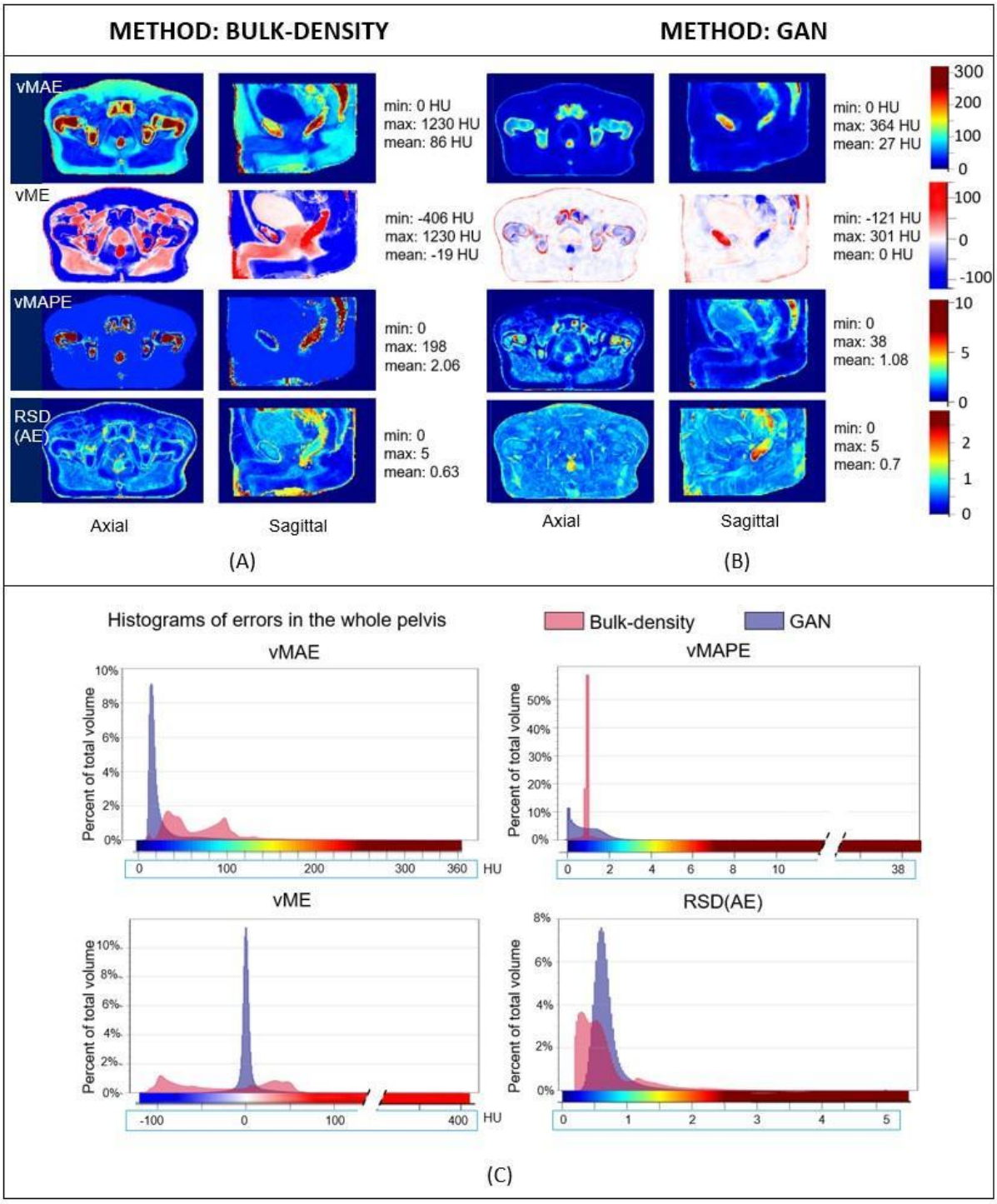

Figure 3.4: HU error maps in the common coordinate system.
Axial and sagittal views of voxel-wise mean absolute error (vMAE), mean error (vME) and mean absolute percent error (vMAPE) maps in the same anatomical space and the corresponding histograms (3) for sCT generated with (1) bulk-density and (2) GAN method. The relative standard deviation of the absolute error (RSD(AE)) is also illustrated. Colours scales of error maps were associated to histograms.

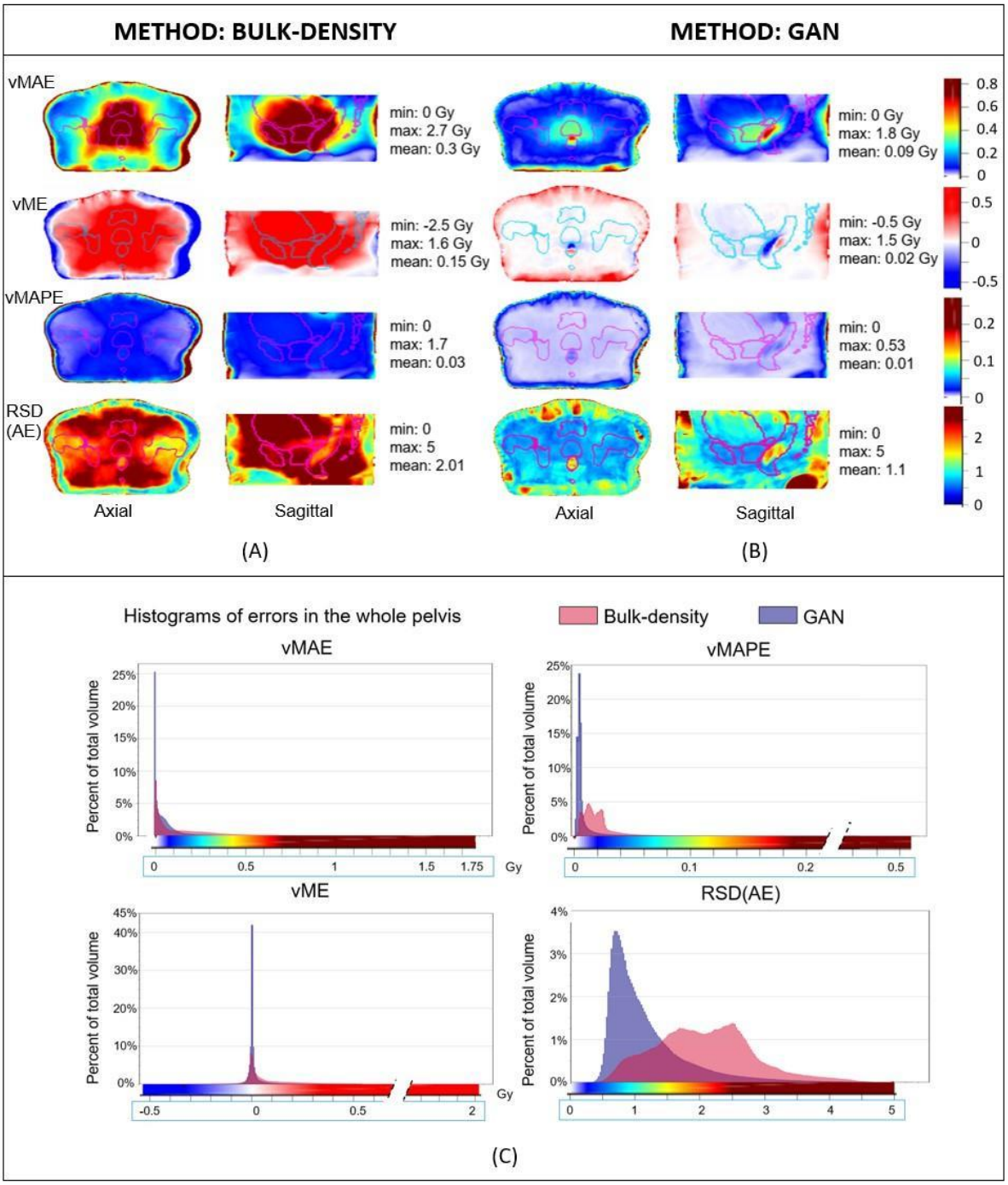

Figure 3.5: Mean dose error maps in the common coordinate system.
Axial and sagittal views of voxel-wise mean absolute error (vMAE), mean error (vME) and mean absolute percent error (vMAPE) maps in the same anatomical space and the corresponding histograms (3) for dose computed from sCT generated with (1) bulk-density and (2) GAN method. The relative standard deviation of absolute error (RSD(AE)) is also illustrated. Contours of delineated organs of the template were overlaid on each image, and colour scales of error maps were associated to histograms.

Permutation test

To complete this study, voxel-wise paired permutation tests proposed by Konietschke et al.[38] were performed for each method with the R software package for nonparametric multiple comparisons[39]. This statistical approach is an adaptation of the Student's test for non-parametric paired data and includes permutation tests. The hypothesis in this study was that the intensity in Hounsfield units, or the dose in Gy, of the generated sCT scans were identical to the value of the reference scans (Figure 3.6).

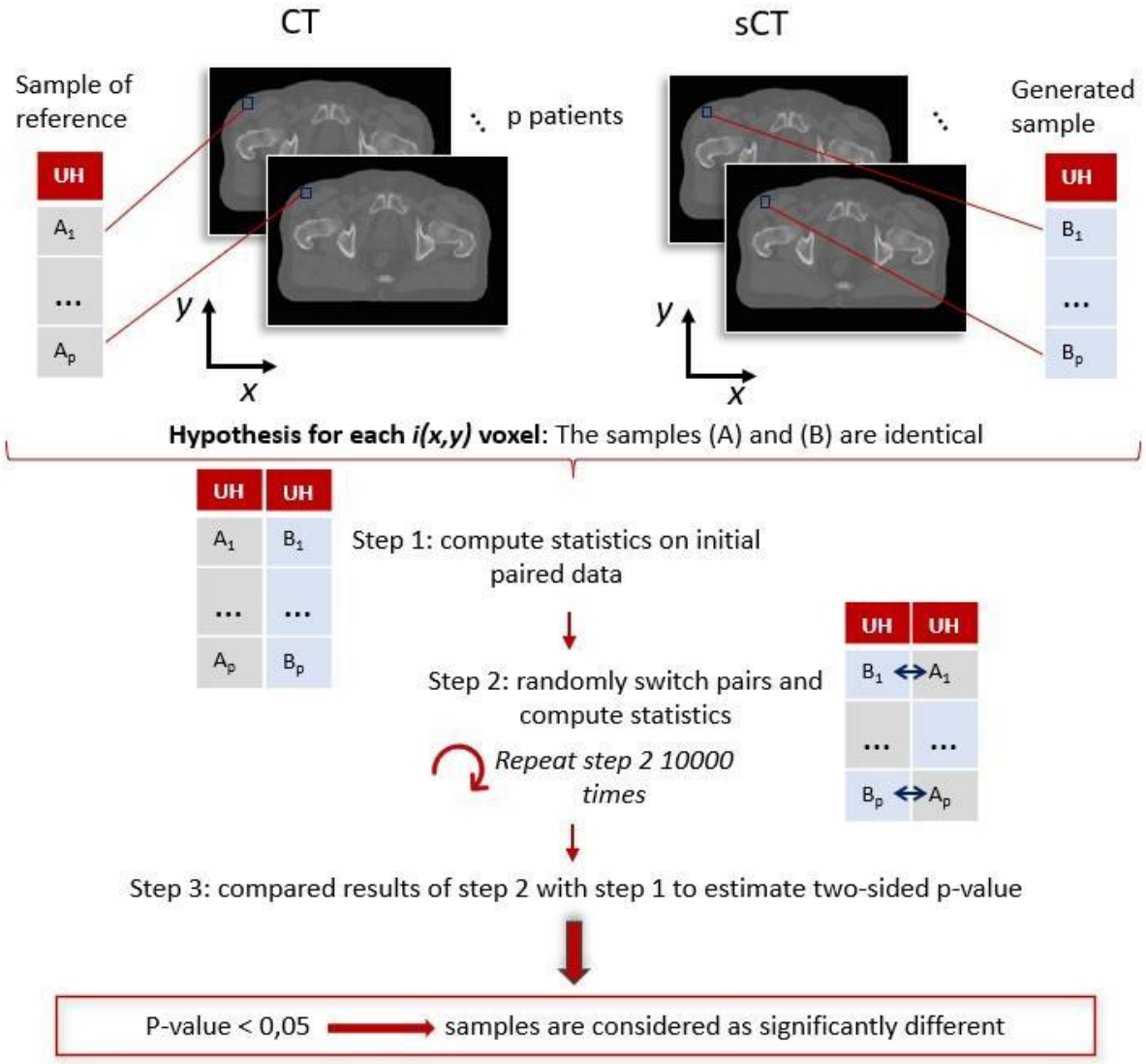

Figure 3.6: Paired permutation test general workflow: example for the image evaluation using Hounsfield Units.
To each voxel coordinate (x,y) correspond paired data (A1, B1), …, (Ap, Bp). These pairs were used to determine if the generated (B) and the reference (A) samples were identical or not following the procedure proposed by Konietschke et al[38]. A p-value(x,y) is obtained for each voxel, highlighting the regions where the differences are significant. The same process was applied on dose distributions.

Two paired lists of values were determined for each voxel and compared.

Multiple comparisons may lead to type I errors, namely the false positive rate. So, to limit these errors, 10000 random permutations were utilized to estimate the p-value.
The procedure to estimate the p-value followed these steps:

- Computation of the statistics[38] on the initial data: $U = (\ U_1, \dots,\ U_p\ )$, with $U_1 = (X_{CT}(1),\ X_{sCT}(1)\ )$ the paired values for patient 1, and p the total number of patients in the population
- Computation of the statistics on randomly permuted data defined as $U_{perm} = (\ U_{perm1}, \dots,\ U_{permp}\ )$, with $U_{perm1} = \{\big((X_{CT}(1),\ X_{sCT}(1)\big), \big((X_{sCT}(1),\ X_{CT}(1)\big)\}$ the two possible paired values for patient 1. This step was repeated 10 000 times
- Comparison of the results obtained with the swapped data U_{perm} and the one obtained in the first step to estimate the p-value [38].

This test resulted in 3D maps, where a voxel i corresponds to the probability that the initial hypothesis was true for the ith voxel of the generated sCTs. Regions of significant differences (p-value < 0.05) between CTs and sCTs on the one hand, and between dose plans calculated on CTs and sCTs on the other, were generated. These volumes, referred to as Error Sub-Regions (ESR), are illustrated in Figure 3.7.

Mean absolute dose error – volume histogram

This cumulative histogram is a quantitative tool, allowing for assessment of absolute error in the dose calculations on the sCT and CT scans with respect to the volumes of tissue. It was built in the same way as dose volume histograms (DVH) and computed from the vMAE map in the CCS. The regions of interest for this evaluation were bladder, rectum, prostate and pelvis. To focus on the region of the dose distribution, the pelvic region was cropped to within 2 cm above and 2 cm below the rectum, according to the superior to inferior axis.
Two criteria for evaluation were selected: V0.5Gy and V1Gy, which correspond respectively to the total volume with an absolute error greater than or equal to 0.5 Gy and 1 Gy.

Dosimetric endpoints

- *Gamma analysis*

Dose plans were propagated to the CCS and combined, resulting in mean reference CT dose and mean dose for each sCT generation method. Thus, a spatial dose evaluation was conducted comparing mean dose distributions with a 3D gamma analysis (local, 1%/1mm, dose threshold 10%) using VeriSoft software. The gamma pass-rate, corresponding to the percentage of voxels with gamma inferior to 1, and mean gamma were reported, additionally to gamma maps in the axial plan.

- *DVH criteria*

Absolute differences between dosimetric values calculated on the reference CT propagated in the CCS and those calculated using sCT generated from BDM and GAN were determined. The contours used were the bladder, rectum and prostate of the template in the CCS. Table 3.4 presents the average differences of mean dose, D2%, D50% and D95% for each method, with Dx % representing the dose in x% of the volume of interest.

Results

Image and dose error evaluation in native space

Table 3.1 depicts the results of the evaluation in the native space for both bulk-density and GAN methods. The BDM presented higher MAE, MAPE, and ME than the deep-learning based approach. The worst MAE scores for both methods were in the bone regions (244.4 HU for BDM, 124.3 HU for the GAN). This structure also had higher mean CT number and standard deviation (342 HU ± 317 HU).

Table 3.1: Error evaluation performed in the native space for sCT generation methods. Global scores for the whole pelvis and per organ are presented. Mean absolute error (MAE), mean absolute percentage error (MAPE) and mean error (ME) were computed between reference CT and sCT (image results in HU) and between dose distribution calculated from these images (dose results in Gy). Reference CT number and mean dose in each anatomical region are also indicated.
Table 3.3: Error evaluation performed in the native space for sCT generation methods.

			IMAGE (HU)			DOSE (Gy)		
			GAN	**BULK-DENSITY**	**Mean CT number**	**GAN**	**BULK-DENSITY**	**Mean dose**
Global	PELVIS	MAE	**33.9 ± 7.6**	96.4 ± 16.5	18 ± 184	**0.06 ± 0.02**	0.2 ± 0.36	8.9 ± 13.4
		MAPE	**1.3 ± 0.6**	2.3 ± 0.8		**0.1 ± 0.03**	0.12 ± 0.04	
		ME	**3.4 ± 15.6**	-10.4 ± 24.3		**0.11 ± 0.05**	0.19 ± 0.27	
Organ-wise	BONES	MAE	**124.3 ± 22.4**	244.4 ± 29.8	342 ± 317	**0.06 ± 0.03**	0.24 ±0.46	14.6 ± 15.1
		MAPE	**1.3 ± 0.8**	3.9 ± 1.8		**0.04 ± 0.02**	0.06 ± 0.03	
		ME	23.9 ± 45.7	**20.4 ± 62.3**		**0.03 ± 0.08**	0.24 ± 0.47	
	BLADDER	MAE	18.2 ± 4.9	**17.1 ± 5.8**	4 ± 19	**0.11 ± 0.1**	0.72 ± 1.88	25.8 ± 22.7
		MAPE	2.2 ± 1.2	**1.1 ± 0.1**		**0.01 ±0.01**	0.02 ± 0.05	
		ME	4.9 ± 12.0	**4.9 ± 12.9**		**-0.02 ± 0.15**	0.69 ± 1.89	
	RECTUM	MAE	**67.1 ± 66.6**	140.9 ± 71.8	-13 ± 135	**0.23 ± 0.23**	0.79 ±1.62	36.7 ± 19.2
		MAPE	**2.1 ± 1.2**	6.8 ± 6.2		**0.01 ± 0.0**	0.02 ±0.05	
		ME	**-16.3 ± 77.6**	98.2 ± 82.9		**-0.04 ± 0.18**	0.58 ± 1.68	
	PROSTATE	MAE	**17.6 ± 3.8**	34.2 ± 8.5	29 ± 24	**0.34 ± 0.2**	1.46 ±3.54	78.7 ± 0.8
		MAPE	1.2 ± 1.0	**1.0 ± 0.0**		**0.0 ±0.0**	0.02 ± 0.05	
		ME	**3.7 ± 11.3**	30.7 ± 11.6		**-0.04 ± 0.38**	1.3 ± 3.61	

Regarding dose calculation, MAE reached 1.46 Gy, equivalent to 1.85% of the expected dose, in the prostate for the BDM and 0.34 Gy for the GAN. For each method, MAPE was similar for the prostate, rectum and bladder (around 0.02 for BDM and 0.01 for GAN), and superior in bones (0.06 and 0.04). Standard deviation for all error types and all delineated organs were larger for BDM compared to GAN.

Registration

The customized non-rigid registration process accurately brought the 26 patients of the cohort in the same anatomical space, as shown by the average dice score of 0.98 ± 0.01 for the body contour, 0.93 ± 0.01 for the bones, 0.96 ± 0.01 for the bladder, 0.91 ± 0.02 for the rectum and 0.91 ± 0.02 for the prostate. The mean volume, in cubic centimeters, of each delineated structure ended close to the volume of the template's organs in the CCS (Table 3.2) confirming the efficiency of the method.
The accuracy of the registration inside the body is also illustrated visually in Figure 3.3.

Voxel-based error maps

Image assessment

Figure 3.4 depicts the vMAE, vME and vMAPE error maps computed in the CCS for both BDM and GAN methods. RSDAE map, representing the dispersion of the absolute error distribution at each voxel considering the overall cohort, are also included. It illustrates the voxel-wise quality assessment of sCT generated for each method. Histograms of these 3D error maps are presented in this figure, which allows comparison of the accuracy of both methods. Difference in intensity up to 250 HU in the rectum and more than 500 HU in cortical bones were found for the BDM. An underestimation (in red, Figure 3.4) of more than 200 HU in the cortical bones, and around 140 HU in the rectum were observed in sCT generated from BDM, as well as an overestimation (in blue, negative values) of 200 HU in spongy bones. For the GAN, the highest vMAE was found in bones (around 100 HU, and up to 220 HU in denser regions). The vMAE reached 200 HU in a small specific region within the rectum, close to the prostate and seminal vesicles. According to the vME map, the GAN approach led to an overestimation (in blue, Figure 3.4) in the previously described location in the rectum, with a score equal to -85 HU, and in spongy bones (-40 HU). An underestimation of 110 HU in cortical bones (in red, Figure 3.4). Errors highlighted with the vMAPE were in spongy bones and in the rectum for both methods, also in the contour of the bladder for the GAN. The vMAPE histogram for the BDM has a narrow distribution around 1in soft tissue, as computing the MAPE in this area, where sCT value is equal to 0 HU, results in dividing the reference CT value by itself. Though the RSDAE were more than to 1.5 and to 2 respectively for the BDM and the GAN in the rectum, the highest values were not at same location.

Figure 3.7 presents significant ESRs, in red, overlaid on the mean MR images in the CCS and on the mean dose distribution. Most of the HU values predicted with the BDM were significantly different from the reference CT HU values, except in an important part of the bladder and in tissue interfaces. According to the studentized permutation test result, ESRs were preferentially located in cortical bones, skin, a part of the prostate, and regions scattered around the bladder and the rectum for the sCT obtained with the DLM.

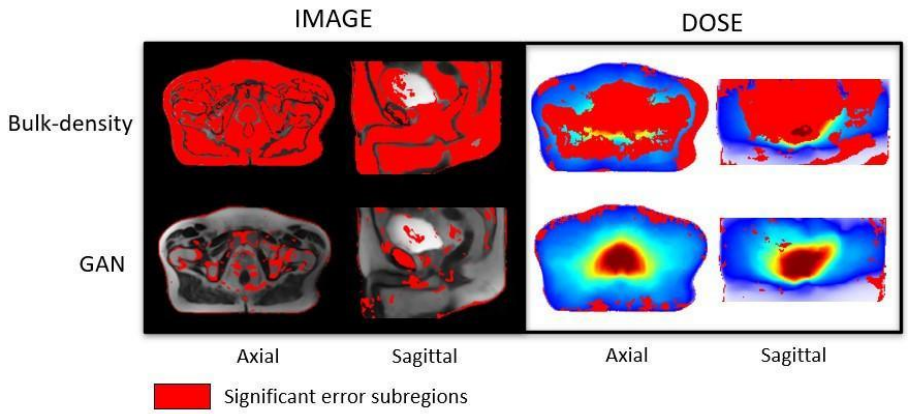

Figure 3.7: Studentized paired permutation test results
Significant error subregions brought out by the Konietschke's paired permutation test, in red, overlaid on mean MR images in the common coordinate system (CCS) for HU values (left) and overlaid on the mean dose plans in the CCS for Gy values (right). This statistical test produced p-value maps. Differences of intensities (HU) in one hand, and dose (Gy) in the other hand, were considered as significant for p-value < 0.05.

Table 3.2: Volume of delineated structure in cm3 prior and after the non-rigid registration.
These data are presented regarding the volume of the template in the common coordinate system (CCS).

	VOLUME IN NATIVE SPACE (cm^3)				*REGISTERED VOLUME (cm^3)*				*TEMPLATE IN CCS (cm^3)*
	mean	***std***	***min***	***max***	***mean***	***std***	***min***	***max***	
BODY	*14362*	*2092*	*10608*	*18300*	*15392*	*261*	*14363*	*15812*	*15374*
BLADDER	*274*	*142*	*113*	*633*	*243*	*3*	*237*	*251*	*246*
BONES	*1259*	*205*	*908*	*1817*	*1082*	*36*	*1031*	*1183*	*1076*
PROSTATE	*40*	*19*	*16*	*82*	*33*	*1*	*31*	*37*	*34*
RECTUM	*66*	*29*	*25*	*133*	*36*	*1*	*34*	*37*	*36*

Dose assessment

Figure 3.5 illustrates the dose differences for the whole population data. As for the image assessment, the resulting maps allowed to evaluate and compare locally resulting the dose

calculation of both sCT generation methods. For the BDM, vMAE in the organs at risk increased up to 1.7 Gy, just near the prostate. The most predominant absolute errors for the GAN appeared in the rectum with differences up to 0.75 Gy, and the first centimetre of the body contour. In the prostate the vMAE was around 0.3 Gy. The vME reached 0.4 Gy on the body contour for the DLM. The vMAPE confirmed the error on the body contour but not in the rectum for both approaches. RSDAE highlighted the same area in the rectum than vMAE and vME maps (RSDAE > 1.5). The higher the delivered dose, the higher was the error observed, with an underestimation of the dose distribution of 1.3 Gy in the prostate for the BDM. As for image analysis, dose error maps histograms appeared wider than for GAN (Figure 3.5).

According to Figure 3.7, a major part of the dose plans computed from the BDM were considered as significantly different from the ground truth. For those calculated from sCT generated with GAN, ESR were localized surrounding the body, mainly on the skin and until 3 cm inside the body.

Mean absolute dose error per volume

Figure 3.8 presents the comparison of the two sCT methods by showing the absolute dose difference (Gy) per percentage of tissue volume. This metric reveals a larger error for BDM than GAN, regardless of the organ considered. No volume reached 1Gy of dose difference for the GAN sCT (Table 3.3).

Table 3.3: Percent of tissue volume with a mean absolute error (MAE) reaching 0.5 Gy ($V_{0.5\ Gy}$) and 1 Gy ($V_{1\ Gy}$) for both sCT generation methods. The mean of voxel values of the vMAE map in the common coordinate system was computed in the whole pelvis and in the template's structures (bladder, rectum, and prostate).

	BULK-DENSITY				GAN			
	PELVIS	**BLADDER**	**RECTUM**	**PROSTATE**	**PELVIS**	**BLADDER**	**RECTUM**	**PROSTATE**
$V_{0.5\ Gy}$	16.58%	77.03%	80.93%	100%	1.10%	0%	10.03%	0%
$V_{1\ Gy}$	3.63%	16.48%	31.85%	100%	0.08%	0%	0%	0%

Dosimetric endpoints

Results of 3D gamma analysis (criteria: local, 1%/ 1mm, low dose threshold = 10%) performed on mean dose volume in the CCS are presented Figure 3.. This allows for local comparison of gamma maps of each sCT generation methods.

In Table 3.4, dosimetric criteria assessment shows an absolute difference superior to 1 Gy in the prostate for the BDM, while the GAN results are around 0.33 Gy in this location.

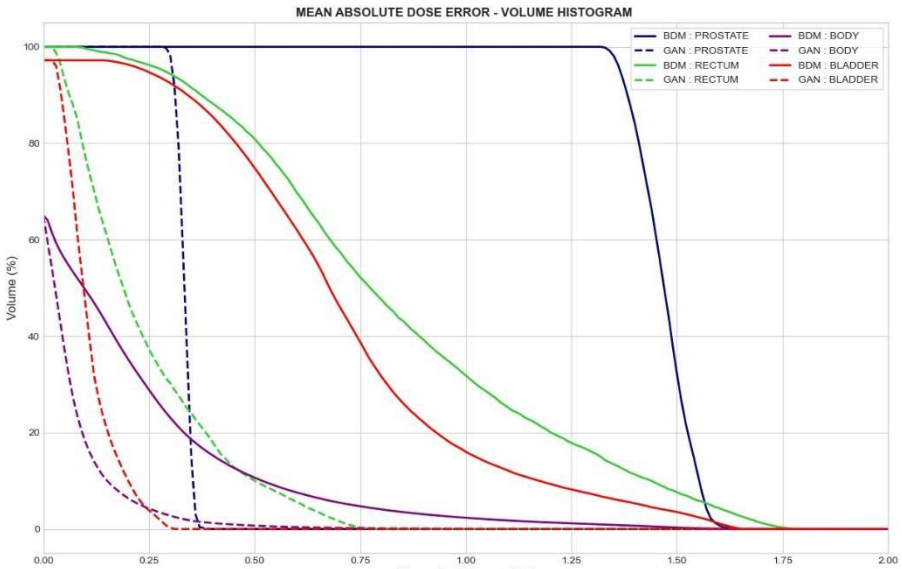

Figure 3.8: Mean absoulte dose error - volume histogram
Mean absolute difference between dose computed from the reference CT and dose computed from the synthetic CT generated with bulk-density method (BDM, continuous line) and GAN (dotted line) for a specific volume of delineated structures. Each colour represents a tissue volume.

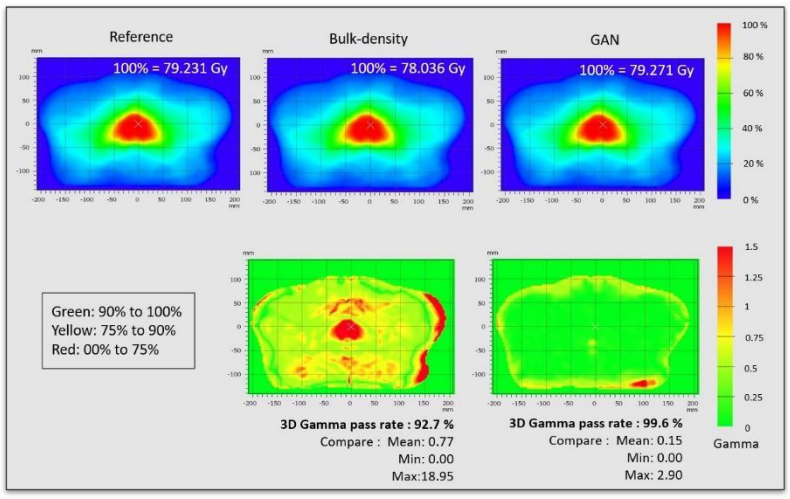

Figure 3.9: Dose distributions and gamma maps
Dose distributions were propagated to the common coordinate system (CCS) and combined, resulting in mean reference CT dose, mean dose for sCT generated from bulk-density and mean dose for sCT generated from GAN method. These dose distributions were used to calculate the gamma pass rate (criteria: 3D, local, 1%/ 1mm, low dose threshold = 10%).

Table 3.4: Absolute difference of dosimetric criteria computed for both bulk-density and GAN methods using the template contours in the common coordinate system (CCS).
Absolute difference of the dose means, D2%, D50%, and D95% computed between the reference CT and the synthetic CTs in the rectum, bladder and prostate. Dx% represents the dose in x% of the volume of interest.

	BULK-DENSITY			GAN		
	BLADDER	**RECTUM**	**PROSTATE**	**BLADDER**	**RECTUM**	**PROSTATE**
mean dose absolute difference (Gy) ± std	0.71 ±1.90	0.69 ± 1.85	1.45 ± 3.62	0.10 ± 0.10	0.16 ± 0.12	0.33 ± 0.21
D2% absolute difference (Gy) ± std	1.59 ± 3.66	1.44 ± 3.28	1.41 ± 3.58	0.27 ± 0.19	0.50 ± 0.62	0.33 ± 0.22
D50% absolute difference (Gy) ± std	0.67 ± 1.88	0.66 ± 1.68	1.43 ± 3.63	0.09 ± 0.09	0.18 ± 0.20	0.33 ± 0.21
D95% absolute difference (Gy) ± std	0.24 ± 0.83	0.22 ± 0.56	1.49 ± 3.63	0.03 ± 0.05	0.04 ± 0.04	0.32 ± 0.22

Discussion

This study proposed a methodology based on voxel-wise population analysis to assess the local errors in sCT generation approaches and their impact on the dose distribution. It also allows comparison of performance of several sCT generation methods. The full evaluation process was applied on two sCT generation methods, allowing for the examination of heterogeneity of errors in HU but also in 3D dose distributions across the pelvis.

The presented methodology relies on the accuracy of the interindividual non-rigid registration step, as for all voxel-based approaches[40]. Registration methods have been developed in morphometry studies [41]–[43] Previous studies in the pelvic area included structural descriptions of the bladder and prostate only[25], rectum only[34], or were combined to CT[44]. The voxel-wise statistical analysis performed here includes a novel integration of bones, with a step dedicated to the preservation of their inner structure. The combination of these structural descriptions with MR images is also original in this context and achieved a precise registration of the whole pelvis as it offers superior contrast in soft tissue. With the Demons algorithm for deformable registration, the amount of deformation is limited by the deformation field smoothing at each iteration, which heaps avoid large and unnatural displacement. The algorithm is quite robust to breaking down, however this is possible if the anatomy or modality is very different, particularly if the rigid registration step has failed prior to the Demons algorithm.

The same pelvic MRI data used in this study had been successfully evaluated in previous work which has relied on the same registration method (for example, “Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences” Dowling et al. 2015, and “A Multi-center Prospective Study for Implementation of an MRI-Only Prostate Treatment Planning Workflow” Greer et al. 2019).

While the reported DSC highlight the structural similarities, these are also robust indicators for when the analysis would break down. Major displacement of the organs leading to non-realistic deformation within the body during the registration will impact the DSC of the contours and can provide a good quality assurance step to ensure the registration has not failed. The mean DSC of 0.98 for the body contours indicates that the registration on this dataset appears to be accurate.

This method permits to map organs, images, and doses in a single coordinate system. Comparison by voxel is thus anatomically meaningful for both images and doses.

vMAE, vME, vMAPE and RSDAE 3D maps were produced, showing the distribution of mean error across the pelvis for a whole population. The error maps histograms are a quantitative tool to compare the chosen methods. As vMAE map values appear to be correlated to the reference intensity (the most important errors are in cortical bones, where the mean HU value is the highest), the relative difference, vMAPE, was also computed as a measure of prediction accuracy. The purpose of vME maps is to determine if the prediction tends to be systematically superior or inferior to the reference, and the RSDAE, also known as coefficient of variation, can be interpreted as uncertainties maps of each method [45]. RSDAE gives an insight into regions where HU prediction is trustworthy or not. Therefore, each 3D map computed in this study illustrated complementary information on errors produced in both sCT and dose distributions.

To define if the errors were significant across the anatomy in the CCS, a voxel-wise statistical test was applied on images and on dose distributions. The permutation test proposed by Konietschke et al.[38] was used to cope with the multiple comparison problem and is appropriate for paired and non-parametric data. Other permutation tests, such as Chen’s[46] used in Chourak et al.[27], does not appear suitable in our approach as it does not compare each CT to its corresponding sCT.

The two evaluated methods were BDM and DLM using GAN. BDM is an historical approach for MRI-only radiation planning and was the first integrated in a commercialized device (MRCAT, Philips[47]). The BDM also have application to quality assurance of sCT scans[4]. This approach is simple and does not involve registration, but it lacks accuracy as it does not take tissue heterogeneity into account. The BDM presented in this paper was chosen as an illustration of the proposed methodology, but it has been shown that more accurate methods exist[3], [47]–[49].

Although several sCT generation methods have been proposed in the literature, recent studies head towards deep-learning strategies[12], [50] DLMs such as GAN trained with

paired-data rely on intra-patient registration precision[51]. Multimodal registration of the input data and training is time-consuming, but generated sCT are in general more accurate[6], [20].

According to the RSDAE map, the GAN was more consistent in HU prediction and resulted in more reliable dose planning. For both methods, important MAEs and MEs arose in the rectum, near the prostate. This area corresponded to a high RSDAE regarding other structures and a high MAPE, expressing the lack of accuracy of both methods in this location. Besides, the error did not stand out as significant with the studentized permutation test for GAN. This wide error might be due to the change in patients' anatomy between CT and MRI acquisition, but is not necessarily related to an incorrect prediction of the HU. Another possibility is that the change in patients' anatomy disrupted the training phase for the GAN.

BDM statistically lacked accuracy for HU prediction and dose calculation. For the GAN HU values, significant differences were observed in cortical bones, especially in the femoral heads, but no significant consequence appeared in the dose distribution.

Although HU prediction accuracy is important, sCT generation needs to be reliable for dose planning. Dosimetric assessment is thus crucial, and is usually based on DVH, which is an organ-based metric, and gamma analysis. The gamma was computed in the CCS, allowing for the extraction of local values across the population. The location of dose discrepancies is clearly visible, with gamma superior to 1 in the prostate for the BDM (Figure 3.). Gamma results allow a spatial dose analysis of the sCT generation method for chosen criteria (1%/1mm in this study).

Recent studies in sCT generation involve deep-learning for different anatomical locations. Nevertheless, artificial intelligence (AI) is not yet fully trusted for clinical use and key points to assess AI solutions in radiology are raised[52]. Critical questions for performance and validation are related to robustness to input variability, training data and potential sources of bias identified by developers. As the GAN was trained with paired CT and MRI, the multi-modal registration accuracy directly impacts the quality of sCT[51]. In addition, uncertainties inherent to deep learning models[53] also generate misprediction.

These uncertainties may produce errors in sCT HU values, and so may impact dose computation.

The population-based strategy presented in this paper offers the possibility to define at a voxel level the capability of a method to be accurate across a cohort of patients, having variable tissue density and anatomy, in HU and on the resulting dose distribution. It gives an insight on the reliability of sCT generation, where usually the assessment is limited to global or organ-wise assessment [1], [54], [55].

A limitation of the registration process might be the accuracy of the contours. Inter-observer delineation for bladder, prostate and rectum on a similar dataset appeared to be close in a previous study[31]. However, the experts may have been more experienced than the physicians who segmented the data for this project. Nevertheless, relations between HU errors and their impact on dose computations are yet to be investigated. In silico models with

simulated HU errors in specific tissue followed by dose computation could help to determine the acceptable level of error in sCT that will not affect the dose.
Overall, voxel-wise analysis brought out significant differences which did not show up with the global scores and allowed the assessment of both HU prediction and dose distribution. This process identified locations where the sCT were more prone to errors. This will provide a way forward for translation to a clinical radiotherapy practice. However, the analysis accuracy highly depends on the quality of the interpatient registration. As misregistration can remain, dissociating registration error to those inherent to the generation methods is an issue of interest and this is yet to be fully explored.
Even if the sCT generation method appeared to be accurate, there is no guarantee that each new sCT will be reliable for dose calculation, especially for a patient anatomically different from the training cohort or if the MR image presents artefacts, is acquired with a different sequence or device.
The implemented voxel-based analysis workflow depends on interpatient registration accuracy: mismatch between structures will lead to biased results. Moreover, the statistical test presented in this paper is time-consuming, as simulation studies show that at least 10 000 random permutation are needed for each voxel for an adequate p-value estimation[38]. Furthermore, type I error may remain in the ESR.
This methodology is a tool for assessing and comparing sCT generation methods and illustrate inhomogeneities. But more research is required to go further in quality assurance process. Part of our future work is to investigate the ability to assess a single sCT, without reference, before its use for dose calculation.
This study focused on the male pelvic area considering prostate cancer irradiation, however the methodology can be applied to any other anatomical location provided accurate registration is achieved.

Conclusion

The proposed voxel-wise population-based workflow resulted in 3D error maps for sCT generation from MRI. This methodology relies on a robust organ-driven non-rigid registration which brings all the patients to the same anatomical space. The assessment of HU and dose distributions calculated from sCT accuracy followed a multi-scale strategy, whereby errors were computed for the whole pelvis, followed by the organs and finally at a voxel level, allowing for spatial characterization of the differences across the methods. This analysis was completed with a quantitative assessment via error map histograms comparison and the mean absolute dose error per volume histogram to compare different sCT generation methods. Thus, this workflow will be useful in comparison and localization of errors in sCT generation method and provides a way forward to sCT quality control within the MRI-based planning RT.

References

[1] E. Johnstone *et al.*, "Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 100, no. 1. 2018. doi: 10.1016/j.ijrobp.2017.08.043.

[2] D. Bird, A. M. Henry, D. Sebag-Montefiore, D. L. Buckley, B. Al-Qaisieh, and R. Speight, "A Systematic Review of the Clinical Implementation of Pelvic Magnetic Resonance Imaging–Only Planning for External Beam Radiation Therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 105, no. 3. Elsevier Inc., pp. 479–492, Nov. 01, 2019. doi: 10.1016/j.ijrobp.2019.06.2530.

[3] J. Kim *et al.*, "Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy," *Radiation Oncology*, vol. 10, no. 1, 2015, doi: 10.1186/s13014-015-0549-7.

[4] J. H. Choi *et al.*, "Bulk Anatomical Density Based Dose Calculation for Patient-Specific Quality Assurance of MRI-Only Prostate Radiotherapy," *Frontiers in Oncology*, vol. 9, 2019, doi: 10.3389/fonc.2019.00997.

[5] J. A. Dowling *et al.*, "An atlas-based electron density mapping method for magnetic resonance imaging (MRI)-alone treatment planning and adaptive MRI-based prostate radiation therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 83, no. 1, 2012, doi: 10.1016/j.ijrobp.2011.11.056.

[6] A. Largent *et al.*, "Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning," *International Journal of Radiation Oncology Biology Physics*, vol. 105, no. 5, pp. 1137–1150, 2019, doi: 10.1016/j.ijrobp.2019.08.049.

[7] J. Fu *et al.*, "Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging," *Medical Physics*, vol. 46, no. 9, pp. 3788–3798, 2019, doi: 10.1002/mp.13672.

[8] X. Tie, S. K. Lam, Y. Zhang, K. H. Lee, K. H. Au, and J. Cai, "Pseudo-CT generation from multi-parametric MRI using a novel multi-channel multi-path conditional generative adversarial network for nasopharyngeal carcinoma patients," *Medical Physics*, vol. 47, no. 4, pp. 1750–1762, 2020, doi: 10.1002/mp.14062.

[9] D. Bird *et al.*, "Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning," *Radiotherapy and Oncology*, vol. 156, pp. 23–28, 2021, doi: 10.1016/j.radonc.2020.11.027.

[10] H. Yang *et al.*, "Unpaired Brain MR-to-CT Synthesis using a Structure-Constrained CycleGAN," 2018.

[11] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” *Medical Physics*, 2021, doi: 10.1002/mp.15150.

[12] M. Boulanger *et al.*, “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[13] Y. Liu *et al.*, “Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning,” *Physics in Medicine and Biology*, vol. 64, no. 20, 2019, doi: 10.1088/1361-6560/ab41af.

[14] G. Wang, Y. Zhang, X. Ye, and X. Mou, “Image quality assessment,” in *Machine Learning for Tomographic Imaging*, IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[16] A. Bahrami, A. Karimian, and H. Arabi, “Comparison of different deep learning architectures for synthetic CT generation from MR images,” *Physica Medica*, vol. 90, no. August, pp. 99–107, 2021, doi: 10.1016/j.ejmp.2021.09.006.

[17] A. M. Dinkla *et al.*, “Dosimetric evaluation of synthetic CT for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network,” *Medical Physics*, vol. 46, no. 9, pp. 4095–4104, 2019, doi: 10.1002/mp.13663.

[18] B. Tang *et al.*, “Dosimetric evaluation of synthetic CT image generated using a neural network for MR-only brain radiotherapy,” *Journal of Applied Clinical Medical Physics*, vol. 22, no. 3, pp. 55–62, 2021, doi: 10.1002/acm2.13176.

[19] H. Wang, H. Chandarana, K. T. Block, T. Vahle, M. Fenchel, and I. J. Das, “Dosimetric evaluation of synthetic CT for magnetic resonance-only based radiotherapy planning of lung cancer,” *Radiation Oncology*, vol. 12, no. 1, 2017, doi: 10.1186/s13014-017-0845-5.

[20] H. Arabi *et al.*, “Comparative study of algorithms for synthetic CT generation from MRI: Consequences for MRI-guided radiation planning in the pelvic region,” *Medical Physics*, vol. 45, no. 11, pp. 5218–5233, 2018, doi: 10.1002/mp.13187.

[21] M. Hemsley *et al.*, “Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions,” 2020, pp. 834–844. doi: 10.1007/978-3-030-59710-8_81.

[22] M. F. Spadea *et al.*, “Deep Convolution Neural Network (DCNN) Multiplane Approach to Synthetic CT Generation From MR images—Application in Brain Proton Therapy,” *International Journal of Radiation Oncology Biology Physics*, vol. 105, no. 3, pp. 495–503, 2019, doi: 10.1016/j.ijrobp.2019.06.2535.

[23] D. Cusumano *et al.*, “A deep learning approach to generate synthetic CT in low field MR-guided adaptive radiotherapy for abdominal and pelvic cases,” *Radiotherapy and Oncology*, vol. 153, pp. 205–212, 2020, doi: 10.1016/j.radonc.2020.10.018.

[24] K. N. D. Brou Boni *et al.*, "MR to CT synthesis with multicenter data in the pelvic area using a conditional generative adversarial network," *Physics in Medicine and Biology*, vol. 65, no. 7, 2020, doi: 10.1088/1361-6560/ab7633.

[25] E. Mylona *et al.*, "Voxel-Based Analysis for Identification of Urethrovesical Subregions Predicting Urinary Toxicity After Prostate Cancer Radiation Therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 104, no. 2, pp. 343–354, 2019, doi: 10.1016/j.ijrobp.2019.01.088.

[26] R. N. Finnegan *et al.*, "A statistical, voxelised model of prostate cancer for biologically optimised radiotherapy," *Physics and Imaging in Radiation Oncology*, vol. 21, pp. 136–145, Jan. 2022, doi: 10.1016/j.phro.2022.02.011.

[27] H. Chourak *et al.*, "Voxel-Wise Analysis for Spatial Characterisation of Pseudo-CT Errors in MRI-Only Radiotherapy Planning," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 395–399. doi: 10.1109/ISBI48211.2021.9433800.

[28] A. Largent *et al.*, "Pseudo-CT Generation for MRI-Only Radiation Therapy Treatment Planning: Comparison Among Patch-Based, Atlas-Based, and Bulk Density Methods," *International Journal of Radiation Oncology Biology Physics*, vol. 103, no. 2, pp. 479–490, 2019, doi: 10.1016/j.ijrobp.2018.10.002.

[29] D. Rivest-Hénault, P. Greer, jurgen frip, and J. Dowling, *Structure-Guided Nonrigid Registration of CT–MR Pelvis Scans with Large Deformations in MR-Based Image Guided Radiation Therapy*. 2014. doi: 10.1007/978-3-319-05666-1_9.

[30] D. Rivest-Hénault, N. Dowson, P. B. Greer, J. Fripp, and J. A. Dowling, "Robust inverse-consistent affine CT-MR registration in MRI-assisted and MRI-alone prostate radiation therapy," *Medical Image Analysis*, vol. 23, no. 1, pp. 56–69, 2015, doi: 10.1016/j.media.2015.04.014.

[31] J. A. Dowling *et al.*, "Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences," *International Journal of Radiation Oncology Biology Physics*, vol. 93, no. 5, pp. 1144–1153, 2015, doi: 10.1016/j.ijrobp.2015.08.045.

[32] Y. K. Lee *et al.*, "Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone", doi: 10.1016/S0.

[33] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method:," *Medical Physics*, vol. 44, no. 4, pp. 1408–1419, 2017, doi: 10.1002/mp.12155.

[34] G. Dréan, O. Acosta, C. Lafond, A. Simon, R. De Crevoisier, and P. Haigron, "Interindividual registration and dose mapping for voxelwise population analysis of rectal toxicity in prostate cancer radiotherapy," *Medical Physics*, vol. 43, no. 6, pp. 2721–2730, 2016, doi: 10.1118/1.4948501.

[35] P.-E. Danielsson, "Euclidean distance mapping," *Computer Graphics and Image Processing*, vol. 14, no. 3, pp. 227–248, 1980, doi: https://doi.org/10.1016/0146-664X(80)90054-4.

[36] S. E. Jones, B. R. Buchbinder, and I. Aharon, “Three-dimensional mapping of cortical thickness using Laplace’s equation,” *Human Brain Mapping*, vol. 11, no. 1, 2000, doi: 10.1002/1097-0193(200009)11:1<12::AID-HBM20>3.0.CO;2-K.

[37] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: efficient non-parametric image registration.,” *NeuroImage*, vol. 45, no. 1 Suppl, 2009, doi: 10.1016/j.neuroimage.2008.10.040.

[38] F. Konietschke and M. Pauly, “A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data,” *Electron J Stat*, vol. 6, pp. 1358–1372, 2012, doi: 10.1214/12-EJS714.

[39] F. Konietschke, M. Placzek, F. Schaarschmidt, and L. A. Hothorn, “nparcomp: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals,” 2015.

[40] G. Palma, S. Monti, and L. Cella, “Voxel-based analysis in radiation oncology: A methodological cookbook,” *Physica Medica*, vol. 69. Associazione Italiana di Fisica Medica, pp. 192–204, 2020. doi: 10.1016/j.ejmp.2019.12.013.

[41] J. Ashburner and K. J. Friston, “Voxel-based morphometry - The methods,” *NeuroImage*, vol. 11, no. 6 I, pp. 805–821, 2000, doi: 10.1006/nimg.2000.0582.

[42] L. Shi *et al.*, “Radiation-induced gray matter atrophy in patients with nasopharyngeal carcinoma after intensity modulated radiotherapy: A MRI magnetic resonance imaging voxel-based morphometry study,” *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 9, pp. 902–909, 2018, doi: 10.21037/qims.2018.10.09.

[43] A. A. Joshi, R. M. Leahy, R. D. Badawi, and A. J. Chaudhari, “Registration-based morphometry for shape analysis of the bones of the human wrist,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 416–426, Feb. 2016, doi: 10.1109/TMI.2015.2476817.

[44] G. Dréan *et al.*, “MRI to CT Prostate Registration for Improved Targeting in Cancer External Beam Radiotherapy,” *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 370–373, 2017, doi: 10.1109/JBHI.2016.2581881.

[45] T. Higaki, K. Akita, and K. Katoh, “Coefficient of variation as an image-intensity metric for cytoskeleton bundling,” *Scientific Reports*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-79136-x.

[46] C. Chen, M. Witte, W. Heemsbergen, and M. Van Herk, “METHODOLOGY Open Access Multiple comparisons permutation test for image based data mining in radiotherapy,” 2013.

[47] N. Tyagi *et al.*, “Dosimetric and workflow evaluation of first commercial synthetic CT software for clinical use in pelvis,” *Physics in Medicine and Biology*, vol. 62, no. 8, pp. 2961–2975, 2017, doi: 10.1088/1361-6560/aa5452.

[48] J. H. Choi *et al.*, “Synthetic CT generation using MRI with deep learning: How does the selection of input images affect the resulting synthetic CT?,” *Medical Image Analysis*, vol. 9, no. 1, pp. 0–30, 2019, doi: 10.1016/j.snb.2007.07.003.

[49] K. Eilertsen, L. Nilsen Tor Arne Vestad, O. Geier, and A. Skretting, “A simulation of MRI based dose calculations on the basis of radiotherapy planning CT images,” *Acta Oncologica*, vol. 47, no. 7, pp. 1294–1302, 2008, doi: 10.1080/02841860802256426.

[50] M. Lerner, J. Medin, C. Jamtheim Gustafsson, S. Alkner, C. Siversson, and L. E. Olsson, “Clinical validation of a commercially available deep learning software for synthetic CT generation for brain,” *Radiation Oncology*, vol. 16, no. 1, pp. 1–11, 2021, doi: 10.1186/s13014-021-01794-6.

[51] M. C. Florkow *et al.*, “The impact of MRI-CT registration errors on deep learning-based synthetic CT generation,” 2019, p. 116. doi: 10.1117/12.2512747.

[52] P. Omoumi *et al.*, “To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines),” *European Radiology*, 2021, doi: 10.1007/s00330-020-07684-x.

[53] M. Abdar *et al.*, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” 2020.

[54] J. M. Edmund and T. Nyholm, “A review of substitute CT generation for MRI-only radiation therapy,” *Radiation Oncology*, vol. 12, no. 1, p. 28, Dec. 2017, doi: 10.1186/s13014-016-0747-y.

[55] J. A. Dowling and J. Korhonen, “MR-Only Methodology,” in *MRI for Radiotherapy*, Cham: Springer International Publishing, 2019, pp. 131–151. doi: 10.1007/978-3-030-14442-5_9.

Chapter 4: Determination of acceptable Hounsfield Units uncertainties via a sensitivity analysis for an accurate dose calculation in the context of prostate MRI-only radiotherapy

The previous chapter focused on detecting significant errors subregions of sCT generation methods, examining both image and dose aspects. However, these assessments were conducted independently, and their specific impact on the dose remains unclear. If a method tends to fail in a specific region, what would be the consequences for the treatment?

This chapter presents a sensitivity analysis as a valuable tool to address this question. The analysis explores the correlation between intensity changes in different structures to identify regions where mispredictions will have the most substantial impact on the dose in the target volume. Furthermore, the study investigates the influence of error volume on the dose at the isocenter, considering three criteria: size, location relative to the target volume, and the intensity change within the error volume.

The findings of this study have been accepted for publication, after minor revisions, in the Physical and Engineering Sciences in Medicine journal in August 2023.

"Determination of acceptable Hounsfield Units uncertainties via a sensitivity analysis for an accurate dose calculation in the context of prostate MRI-only radiotherapy"
H. Chourak, A. Barateau, P. Greer, C. Lafond, J-C Nunes, R. de Crevoisier, J. Dowling, O. Acosta (PESM, 2023)

Preliminary work of this study has been presented at the AUS MRinRT conference in 2022.
"MRI-only radiation therapy for prostate cancer: exploration of the impact of synthetic-CT uncertainties on dose calculation"
Hilda Chourak, Dowling Jason, Peter Greer, Anais Barateau, Safaa Tahri, Renaud de Crevoisier, Jean-Claude Nunes, Oscar Acosta - AUS MrinRT 2022 (oral presentation)

Abstract

Radiation therapy is moving from CT based to MRI guided planning, particularly for soft tissue anatomy. An important requirement of this new workflow is the generation of synthetic-CT (sCT) from MRI to enable treatment dose calculations. Automatic methods to determine the acceptable range of CT Hounsfield Unit (HU) uncertainties to avoid dose distribution errors is thus a key step toward safe MRI-only radiotherapy. This work has analysed the effects of controlled errors introduced in CT scans on the calculated radiation dose for prostate cancer patients. Spearman correlation coefficient has been computed, and a global sensitivity analysis performed following the Morris screening method. This allows the classification of different error factors according to their impact on the dose at the isocentre. sCT HU estimation errors in the bladder appeared to be the least influential factor, and sCT quality assessment should not only focus on organs surrounding the radiation target, as errors in other soft tissue may significantly impact the dose in the target volume. This methodology links dose and intensity-based metrics, and is the first step to define a threshold of acceptability of HU uncertainties for accurate dose planning.

Keywords: Sensitivity analysis**;** quality assurance; synthetic-CT; MRI-only radiotherapy; prostate cancer.

Introduction

External beam radiation therapy (EBRT) involves the application of high-energy x-ray beams from multiple directions, depositing energy (dose) within a tumour to destroy cancer cells. EBRT is a well-established treatment modality for localised prostate cancer. Until recently, treatment has traditionally been planned based on Computed Tomography (CT), with Magnetic Resonance Imaging (MRI) also acquired for diagnostic information. For prostate cancer, MRI has added significant value to EBRT due to its superior soft tissue contrast which results in the improved accuracy of manual labelling of the target volume (the prostate gland) and nearby organs at risk (bladder, rectum, bones). This improved accuracy may reduce the risk of toxicity in healthy tissue[1], [2].

The deployment of MRI-only radiotherapy (RT) provides greater efficiency and accuracy in the clinical workflow by bypassing the MR to planning CT registration step and removes the need for an extra CT scan. This justifies the increasing worldwide deployment of dedicated MRI scanners and MRI-linear accelerator (MRI-linac) hybrid machines for treatment delivery, the latter also allows for better patient positioning and tumour targeting[3]. However, MRI does not provide information on the electron density of tissues, which is necessary for dose calculation. Synthetic-Computed Tomography (sCT) generation is thus a critical component of MRI-only RT workflows.

Currently, sCT images are assessed against a ground truth CT in two ways: image and dose[4]. The first method involves a comparison of Hounsfield Units (HU)[5], [6]. The most commonly used metrics are full reference intensity-based and include mean absolute error (MAE), mean error (ME) and peak signal-to-noise ratio (PSNR). Perception-based models like the structural similarity (SSIM) may also be assessed[7], [8], and more specifically the multiscale SSIM (MS-SSIM)[9]. These metrics result in global or organ-wise values, but local errors such as air incorrectly included within an organ may have an impact on treatment delivery and may not be identified with a global metric. For sCT in the pelvic area, the HU uncertainties are typically observed in the cortical bone and rectum when air pockets are present[10], [11].

The quality of sCT images is also assessed by the dose accuracy. For the different EBRT treatment techniques such as intensity-modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT), the beams cross several healthy tissues before reaching the target. Errors in these beams' trajectories will have consequences on the dose distribution in the target. Most of the sCT generation literature describe dosimetric endpoints such as gamma analysis and dose-volume histograms (DVH) metrics[12]. These measures give an insight of the overall dose distribution accuracy on the sCT. A previous study proposed a voxel-wise statistical analysis strategy to locally assess sCT generation approaches in image and dose domains[13], but no correlation was made between both. Choi et al.[14] investigated the correlation between image metrics as a global value (computed within the body contour) and dose accuracy in the target volume and proposed a water equivalent depth method as a metric. However, no information was given on the origin of dosimetric errors. Generated images must be sufficiently correct to ensure accurate dose planning in the tumour area. So, determining the origin of local erroneous dose will allow focusing on the most meaningful HU error and provide thresholds of HU uncertainties acceptability.

The aim of this study is to investigate the correlation between localised HU errors and dose at the centre of the target volume, here the prostate. To do so, a sensitivity analysis (SA) was performed, by applying the Morris screening method[15]. An SA is designed to quantify the effect of parameters on the output[16]; in this study, the effect of HU error on the dose distribution at the isocentre (centre of the prostate).

Several SA methods exist and can be classified in two types: local and global. Local methods allow for the examination of the model at a specific point in the input space. Most of these approaches induce a low computational cost. However, they do not give an indication of interactions between parameters or on the linearity of their effects. Global methods measure the sensitivity in several points in the input space and highlight the type of effect and the possibility of interactions[17]. SA has previously been applied to assess the ability of quality assurance protocols to detect events affecting MRI in RT[18], or to evaluate the sensitivity of electron dose calculation with respect to stopping power and transport coefficients[19].

In this study, a global one-at-a-time (OAT) approach, the Morris screening method, has been chosen to identify the impact of uncertainties in synthetic-CT on the isodose. The Morris method has previously demonstrated its ability to simplify models predicting biochemical

recurrence after radiotherapy[20] by discarding parameters with a low impact on the output. Applying this methodology to sCT for MRI-only RT is the first step in the definition of thresholds of acceptability of HU errors in sCT for safe MRI-only RT practice.

Material and methods

Two experiments have been conducted to determine the errors in sCT that are more likely to affect the dose at the isocentre. First, the errors have been assessed in terms of HU number, volume, and location by adding an artefact in the reference CTs. Spearman correlation coefficient (SCC) between error features (intensity, volume, location) and dose at the isocentre were computed. While the SCC indicates if the different features have a monotonic impact on the dose, the SA will help to classify the features according to their influence on the output and give information on the linearity and or interaction between factors.
In a second phase, we focused on the impact of errors in specific anatomical location by changing the mean intensity in the bladder, rectum, bones, prostate and in the remaining soft tissues.

Dataset

Data of 39 patients with localised prostate cancer aged 58 to 78 years were used in this study. Ethics approval for the study protocol was obtained from the local area health ethics committee, and informed consent was obtained from all patients. For each patient, a CT scan was acquired on a GE LightSpeed RT or a Toshiba Aquilion, (256 x 256 x 128 matrix with a voxel size of 1.17 mm x 1.17 mm x 2.5 mm or 2.0 mm). Bones, bladder, rectum, and prostate were manually delineated by experts.

Sensitivity analysis: Morris screening method

The Morris screening method is a randomised OAT global SA. The parameters are modified individually, and cover a K-dimensional cube, with K representing the number of factors (Figure 4.1).
Feature values were generated using the Sensitivity R package[21] and were randomly assigned to efficiently cover the K-dimensional space. Elementary effects (EE) given by (1) are calculated to assess the effect of the X_i factor variation on the output. The model is evaluated $N = R \times (K + 1)$ times for each j patient, with R the number of repetitions, i.e the number of EE computed per factor. It offers an insight of the influence of parameters $X = [X_1, \dots, X_i, \dots, X_K]$ on the model $Y = f_j(X)$ with a moderate computational cost. This approach also provides information on the type of impact (linear / non-linear, monotonic or not) and on the interaction between the factors assessed[17].

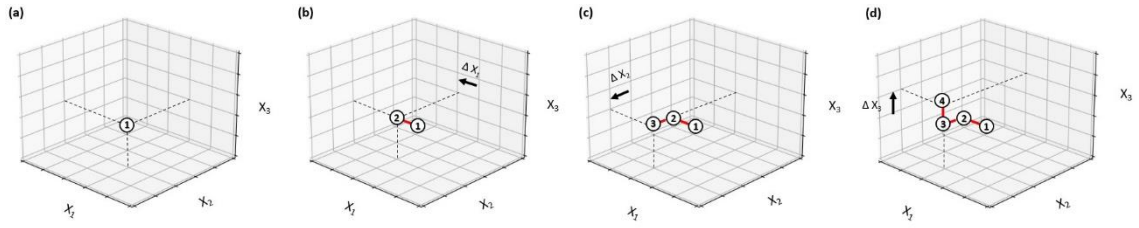

Figure 4.1: Example of a trajectory for the evaluation of the influence of K = 3 factors. First, one point is randomly selected in the 3-dimensional space (a). Then, three other points are created by changing one parameter value at a time (b, c and d).

$$EE_{i,j} = \frac{f_j(X_1, \dots, X_i + \Delta_i, \dots, X_K) - f_j(X_1, \dots, X_i, \dots, X_K)}{\Delta_i} \quad (1)$$

Δ_i is the discrete variation of the parameter.

For each factor and each patient, the mean $\mu_{i,j}$ (2) of EE, the standard deviation $\sigma_{i,j}$ (3), and the mean of the absolute values of the EE $\mu *_{i,j}$ (4) are computed to summarise the EE and thus estimate the global sensitivity in the output space[22]. $\mu *_{i,j}$ is used to solve the effect of opposite signs for non-monotonic functions.

$$\mu_{i,j} = \frac{\sum_{r=1}^{R} EE_{i,j}^{r}}{R} \quad (2)$$

$$\sigma_{i,j} = \frac{\sum_{r=1}^{R} \left(EE_{i,j}^{r} - \mu_{i,j}\right)}{R} \quad (3)$$

$$\mu *_{i,j} = \frac{\sum_{r=1}^{R} \left|EE_{i,j}^{r}\right|}{R} \quad (4)$$

To illustrate the impact of the parameters on the output, the Euclidean distance of each point to the origin $(\mu * 0, \sigma = 0)$ $D_i = \sqrt{\mu *_{i,j}{}^2 + \sigma_{i,j}{}^2}$ has been calculated[23].
Low $\mu *$ and σ indicate an insignificant impact for a chosen factor, and high $\mu *$ and/or σ stand for significant impact. High value of σ compare to $\mu *$indicates a factor involved in interaction with others factors or whose effect is non-linear (Figure 4.2).
In this study, the Morris screening approach aimed to emphasise the impact of localised HU errors on dose calculation, according to:

- descriptive characteristics of the error (intensity, size and location),
- mean intensity within the organs.

These two approaches are described in the experiment's sections below.

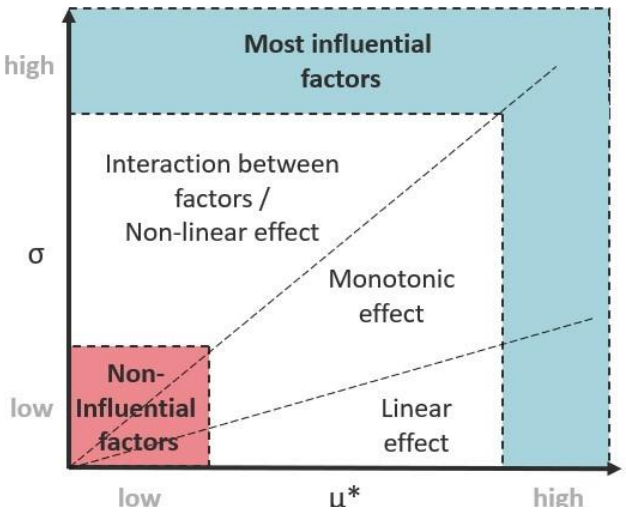

Figure 4.2: Classification of parameters according to the mean of the absolute elementary effects (μ*) and their dispersion (σ).

Experiment 1

The first experiment aimed to assess the impact of error according to 3 factors: intensity, size and location. To achieve this, an artefact with various combinations of these three parameters has been added to the 39 planning CTs. The artefact was built as follows:

- HU variation, from -250 HU to + 250 HU.
- Distance to the isocentre, from 0 to 100 mm. The artefact displacement followed one of the beams' axis.
- Diameter of the artefact, from 2 mm to 50 mm.

The model has been evaluated 200 times for each patient: $N = R \times (K + 1)$,with $R = 50$ repetitions, and $K = 3$ factors (intensity, distance, size), resulting in 7800 simulations.

The Spearman correlation coefficient (SCC) has also been computed in this experiment. This is a nonparametric measure of statistical dependence of ranking between two variables.
An SCC close to -1 or 1 denotes a strong correlation, while an SCC close to 0 illustrates a weak relationship.
To compute the SCC for each error features, the following parameters have been defined:

- For the effect of changes in HU, the step was set to 25 HU. The diameter of the artefact was fixed to 50 mm and its centre aligned to the isocentre, allowing for complete coverage of the target and ensuring a homogeneous distribution of the dose within the error volume.
- For the effect of distance, the step was set to 10 mm, with an error fixed at +200 HU and a size of 50 mm. The displacement followed a beam axis, minimizing the impact of the dose on the result. (For the error to have consequences on the dose at the isocentre, it must be encountered by one of the beams delivering the treatment).

- For the effect of size, the error was fixed at +200 HU and located at 30 mm from the isocentre. This location corresponds approximately to the rectum, where high HU variation can be observed due to the difficulty of predicting air pockets.

2145 images were generated to compute the SCC.

Experiment 2

Errors in sCT are more likely to be evaluated in terms of mean HU error within the body or per organs[24]–[27]. So, in this experiment, mean intensity changes in the following locations have been applied in order to assess their potential impact on the dose:

- Bladder (from -100 HU to +100 HU),
- Rectum (from -1000 HU to +200 HU),
- Bones (from -500 HU to +500 HU),
- Prostate (from -100 HU to +100 HU),
- Remaining soft tissue (from -100 HU to +100 HU).

Remaining soft tissue volumes are generated by subtraction of bone, bladder, prostate and rectum volumes from the body contour. The model was evaluated 240 times for each patient $R = 40$ repetitions, and $K = 5$factors), resulting in 9360 simulations. Higher threshold has been defined for bone and rectum, according to the difficulty for a sCT generation method to predict HU in these locations. Especially for the rectum, where the presence of gas (-1000 HU) is uncertain.

Dose planning

IMRT with 7 beams (photons of 6 MV) was planned for 39 fractions (2 Gy per fraction) on reference CT images using a dose grid resolution of 3x3x3 mm with MatRad[28], an open-source software for radiation treatment planning developed for research purposes[29]–[31]. The beam parameters used to compute the dose on the CT were then copied to calculate the dose on each modified CTs. Figure 4.3 presents examples of modified CT and their corresponding dose used in this study.

Results

Experiment 1

The relationship between the 3 error features and the isodose appear to be monotonic, with an SCC of -0.99 for the intensity variation, -0.95 for the size and 0.73 for the distance. As shown in Figure 4.4 , an overestimation of HU will reduce the dose distributed in the target, while an under estimation will result in a higher dose at the isocentre. Also, there is an important interaction between the size of the volume error and the beams delivering the

dose. As the amount of this volume within the beam increases there are greater impacts on the treatment. An artefact with a diameter of 30mm will decrease the dose in the target of 0.5 Gy in average. As the error is fixed at +200HU to assess the impact of the size, the dose distribution will decrease in this graph.

Regarding the distance, the closer is the volume from the isocentre, the more important is the impact of the error in this location. For all of the patient cohort, when the distance to the isocentre reaches 40 mm, the impact of the artefact starts to be constant, without reaching the prescribed dose (78Gy). This might be explained by the variation of the dose going through the volume of error.

The SCC gives an insight of the effect of each parameter on the dose distribution, but this covers only a few possible combinations of factors compared to the Morris screening method.

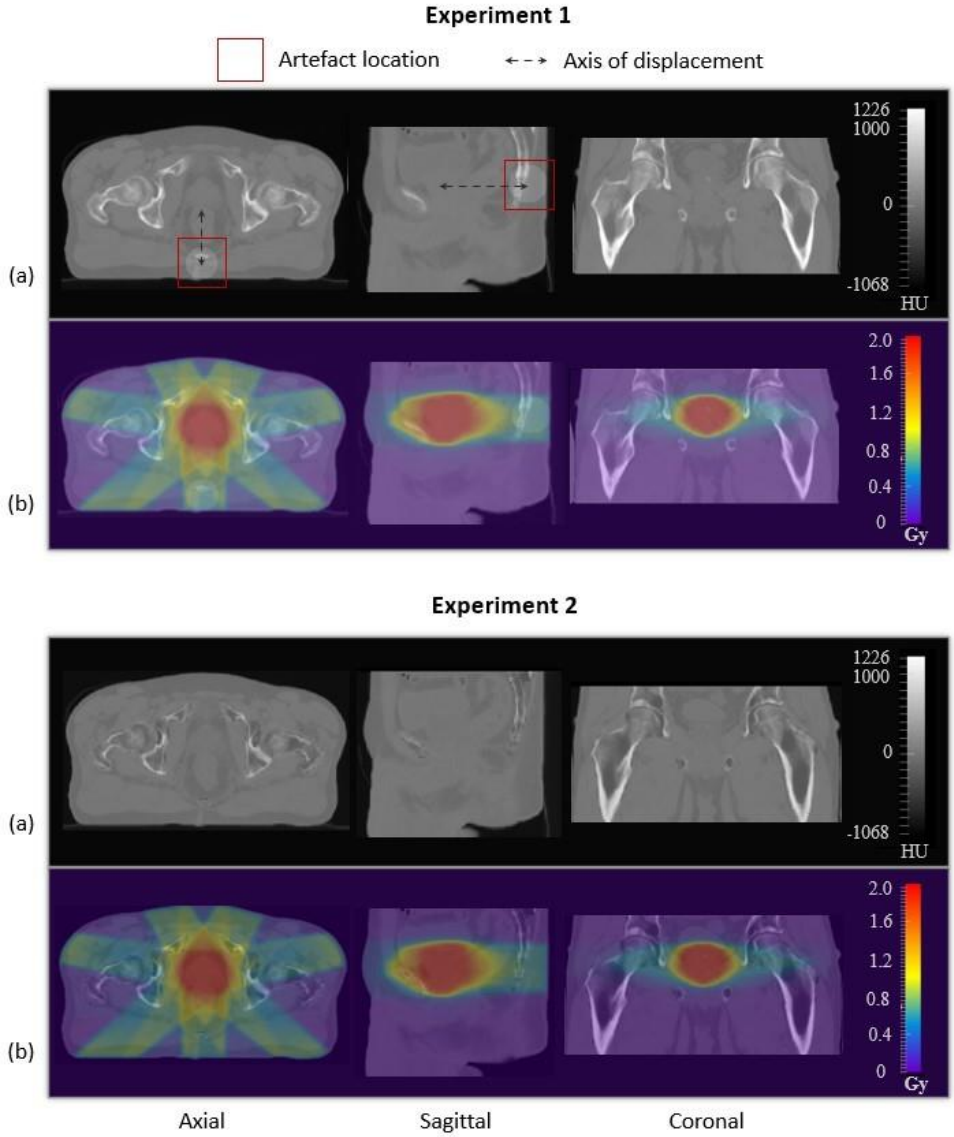

Figure 4.3: Example of a Modified CT (a) and its corresponding dose distribution overlaid (b) for both experiments. For Experiment 1, the error volume is visible in the red square, and the dotted arrow represent its axis of displacement. A modified image of the same patient has been randomly selected to illustrate Experiment 2. Here, 100 HU were added in the bladder, 66.6 HU in the rectum, 55.5 HU in the prostate, and 78 HU in the remaining soft tissue. 146.5 HU were subtracted in the bones.

Figure 4.5 (a) presents the results of the SA. σ is superior to $\mu*$ for all the factors assessed: their effect on the dose distribution at the isocentre is thus non-linear / non-monotonic and/or they interact with each other. This figure also shows that intensity and size are the two most impactful parameters. This statement is confirmed by Figure 4.5 (b), as the Euclidean distance to the origin of the graph is an indication of the influence of a factor on the output. Indeed, it shows that on average, the intensity and the size have both a similar impact on the output.

Experiment 2

Figure 4.6 shows that change in bladder and prostate intensities do not imply significant change in the dose at the isocentre. The dose appeared to be more sensitive to errors in the bones and rectum. The standard deviation of $\mu*$ and σ are more important for these anatomical locations, so change of intensity had a less constant impact across the patient cohort than for the bladder or prostate.

Errors in the remaining soft tissue are the most impactful.

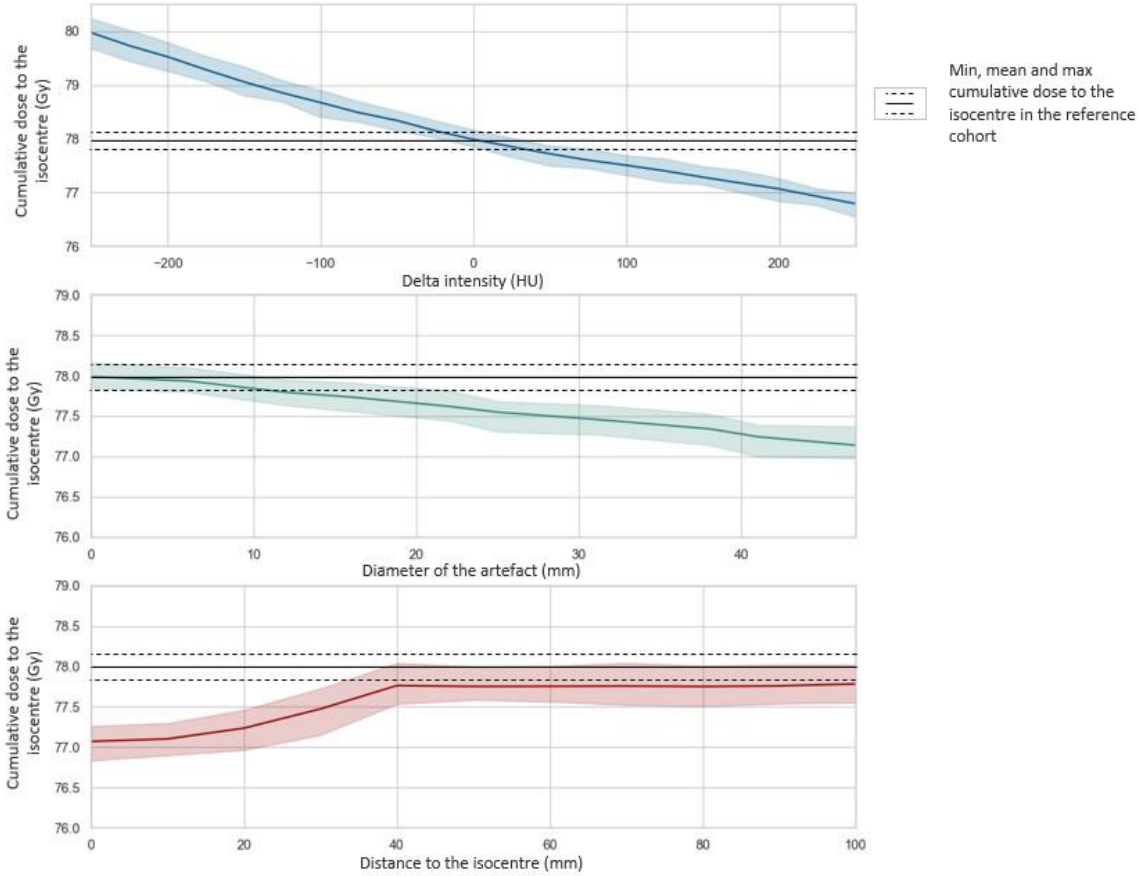

Figure 4.4: Experiment 1: Impact of the error (in terms of intensity in blue, size in green and distance in red) on the dose distribution at the isocentre.

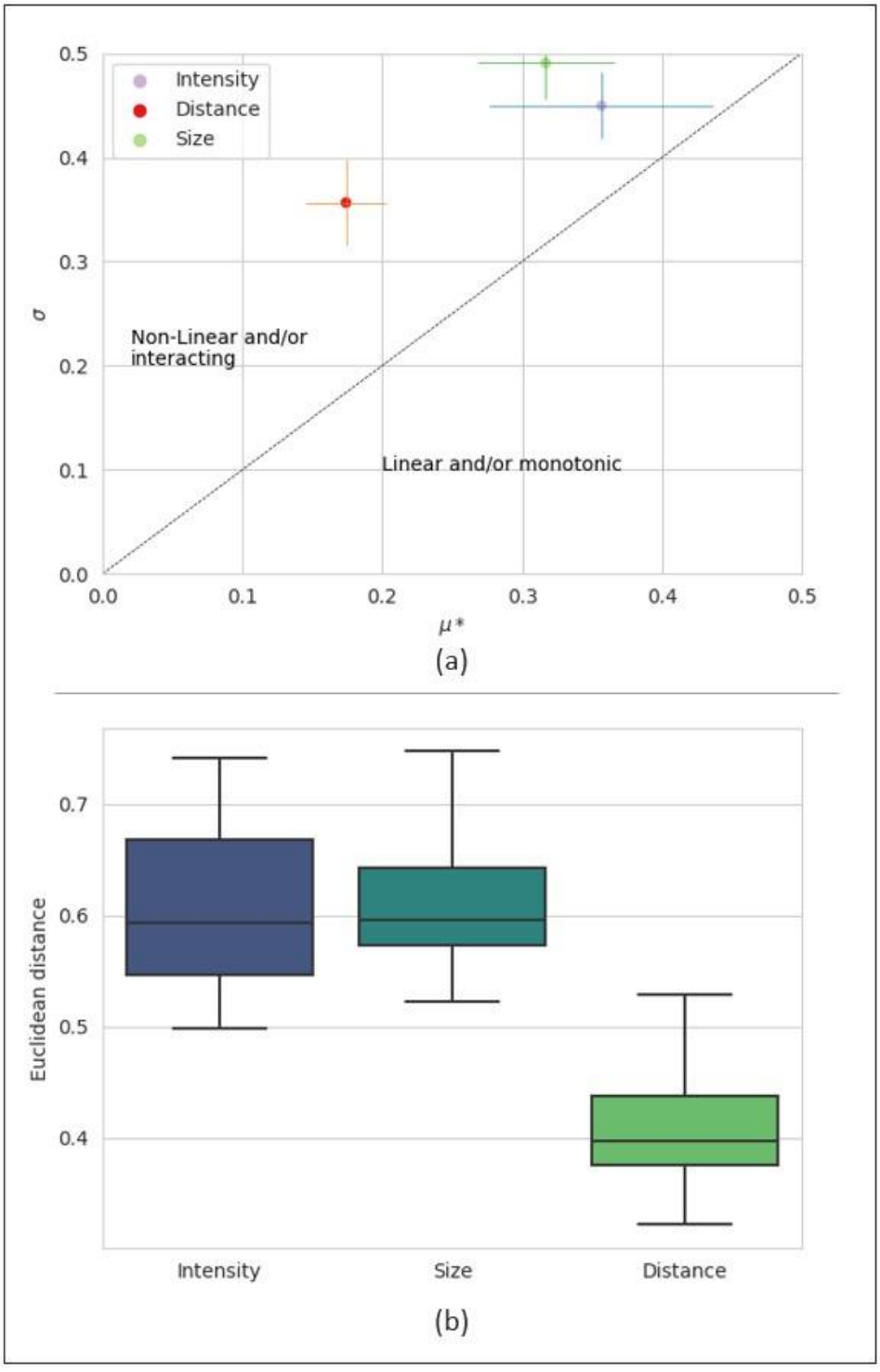

Figure 4.5: Morris screening results for the Experiment 1. (a) Mean of (μ*i,j,σi,j) for each factor. The bars correspond to the standard deviation of μ*i,j and σi,j across the patient cohort. (b) Euclidean distance of each point (μ*i,j,σi,j) to the origin of the graph σ = f(μ∗) in descending order of importance.

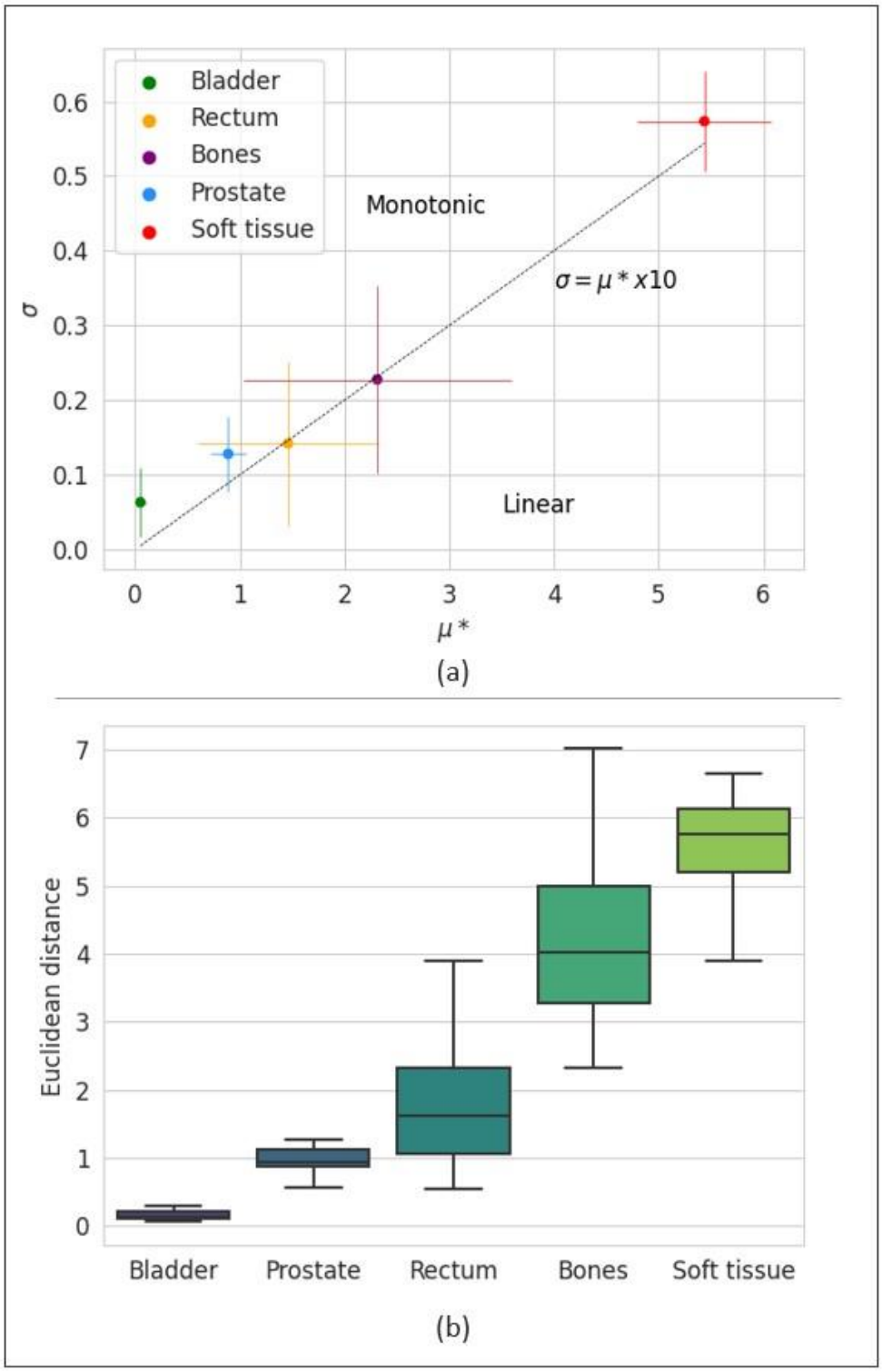

Figure 4.6: Morris screening results for Experiment 2. (a) Mean of (μ*i,j,σi,j) for each factor. The bars correspond to the standard deviation of μ*i,j and σi,j across the patient cohort. (b) Euclidean distance of each point (μ*i,j,σi,j) to the origin of the graph σ = f(μ∗) in descending order of importance.

Discussion

As different sCT generation methods will produce different and inhomogeneous HU uncertainties across the patient's body[11], [13], [26], two experiments have been performed. The first one highlighted the sensitivity of the dose to changes in intensity and size of the volume of error. According to Figure 4.4 , the three features assessed had a monotonic effect on the output, and the Morris screening analysis demonstrated that the three parameters interact with each other.

The second experiment presented the result of the SA on organ-wise error. The variation of HU has been applied homogeneously across each structure to be consistent with the way the methods are usually assessed in the literature (mean error within the organs and the body contours).

Sensitivity to errors in the bones and rectum is less consistent across the patient cohort (Figure 4.6) compared to errors in the bladder and prostate. This might be due to higher variability of size and HU in these structures, with the presence or absence of rectal gas, and different densities in the bone structure (cortical and spongy bones, with a variability in density across the population due to age and body mass[32]). The size of the bones, varying with size of the individual, would also depend on patient weight, where for a thin person an error in bone would have more influence as there is less soft tissue. In this experiment, unlike in the first one, the impact of the different parameters assessed tended to be linear. This might be explained by the consistency of the size and distance of each structure assessed, so the only changing factor is the variation of HU.

Some studies evaluated the dosimetric impact of HU to density curve variation. For example, in a previous study, Thomas et al.[33] reported a dosimetric error of 1.0% for a difference of 8.0% in bone electron density. Notable HU variations affect the accuracy of dose calculation [34], [35]. In case of HU to density curve error, the whole CT image is impacted for a given tissue. In this study, we focused on specific local area.

An absolute threshold of acceptability cannot be universally defined since it depends on each specific sCT generation method and treatment scenario. Therefore, it is recommended to apply this methodology to each clinical centre's specific data. The obtained results are specific to the dose calculation algorithm, the number of beams crossed by the volumes and the amount of dose distributed by each of them. In this study, we assessed the effect of errors on IMRT dose plans, but other treatment techniques may be used in the clinic like VMAT, and stereotactic body radiation therapy (SBRT)[36], may result in different dose distributions across the body and thus will have a strong impact on the results. For particle therapy, the dose in the normal tissue outside the target volume is reduced[37], and the dosimetric impact due to misprediction in HU are likely to be larger. Different results are thus expected for proton and carbon ion therapy. Future work will investigate the other treatment techniques, with a more significant link with the sCT generation.

We focused on the dose at the isocentre, but changes in HU also have consequences on dose distributions in the organs at risk (bladder, rectum, femoral heads), leading to toxicity and inconvenient secondary effects such as chronic bladder inflammation. Therefore, future work will also explore the local influence of HU modification, using dose-volume histogram differences in each specific location.

In the pelvic area, the anatomy of the patient is subject to change due to variation of the bladder and rectal filling for example, which may have consequences on the accuracy of the treatment delivery[38], [39]. The method proposed in this paper could also be used to determine an acceptance criteria of organs motion during the treatment.

The methodology presented in this study can be adapted to each specific generation method, once the location of HU uncertainties has been identified, and the treatment plans defined. Deep-learning based sCT generation methods tend to be the most common[40], and more effective models should to be developed in the future. Aleatoric (data dependant) and epistemic (model dependant) uncertainties are specific to machine-learning models and can be assessed[41]–[43]. Including the impact of these uncertainties on the dose distribution during the learning process might be a way to create more clinically valid image generation.

Conclusion

A sensitivity analysis was performed, allowing for determining the less influential HU errors on the dose distribution at the isocentre. sCT assessment should not only focus on delineated contours, and sparse error in the body contours should not be neglected. This study confirms the necessity to locally assess each sCT prior to its use in a clinical workflow, particularly in steep dose gradient areas.

The main contribution of this paper is to provide a bridge between intensity-based metrics and dose, which are often used independently to assess the quality of sCT for EBRT. This approach can be used to generate clinical thresholds, and potentially model constraints, for both training and validation of sCT generation methods. The study is the first step in the definition of threshold of uncertainty acceptability in sCT to ensure accurate MRI-only RT.

References

[1] A. M. E. Bruynzeel *et al.*, “A Prospective Single-Arm Phase 2 Study of Stereotactic Magnetic Resonance Guided Adaptive Radiation Therapy for Prostate Cancer: Early Toxicity Results,” *Int J Radiat Oncol Biol Phys*, vol. 105, no. 5, pp. 1086–1094, Dec. 2019, doi: 10.1016/j.ijrobp.2019.08.007.

[2] A. U. Kishan *et al.*, “Magnetic resonance imaging-guided versus computed tomography-guided stereotactic body radiotherapy for prostate cancer (MIRAGE): Interim analysis of a phase III randomized trial.,” *Journal of Clinical Oncology*, vol. 40, no. 6_suppl, p. 255, Feb. 2022, doi: 10.1200/JCO.2022.40.6_suppl.255.

[3] J. Ng *et al.*, “MRI-LINAC: A transformative technology in radiation oncology,” *Front Oncol*, vol. 13, Jan. 2023, doi: 10.3389/fonc.2023.1117874.

[4] E. Johnstone *et al.*, “Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy,” *International Journal of Radiation Oncology Biology Physics*, vol. 100, no. 1. 2018. doi: 10.1016/j.ijrobp.2017.08.043.

[5] G. Wang, Y. Zhang, X. Ye, and X. Mou, “Image quality assessment,” in *Machine Learning for Tomographic Imaging*, IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

[6] L. S. Chow and R. Paramesran, “Review of medical image quality assessment,” *Biomedical Signal Processing and Control*, vol. 27. Elsevier Ltd, pp. 145–154, May 01, 2016. doi: 10.1016/j.bspc.2016.02.006.

[7] J. Dowling *et al.*, *Image synthesis for MRI-only radiotherapy treatment planning*. 2022. doi: 10.1016/B978-0-12-824349-7.00027-X.

[8] M. Boulanger *et al.*, “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[9] Y. Li, S. Xu, Y. Lu, and Z. Qi, “CT synthesis from MRI with an improved multi-scale learning network,” *Front Phys*, vol. 11, Jan. 2023, doi: 10.3389/fphy.2023.1088899.

[10] A. Largent *et al.*, “Pseudo-CT Generation for MRI-Only Radiation Therapy Treatment Planning: Comparison Among Patch-Based, Atlas-Based, and Bulk Density Methods,” *Int J Radiat Oncol Biol Phys*, vol. 103, no. 2, pp. 479–490, 2019, doi: 10.1016/j.ijrobp.2018.10.002.

[11] H. Chourak *et al.*, “Voxel-Wise Analysis for Spatial Characterisation of Pseudo-CT Errors in MRI-Only Radiotherapy Planning,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 395–399. doi: 10.1109/ISBI48211.2021.9433800.

[12] R. Kemppainen *et al.*, “Assessment of dosimetric and positioning accuracy of a magnetic resonance imaging-only solution for external beam radiotherapy of pelvic anatomy,” *Phys Imaging Radiat Oncol*, vol. 11, pp. 1–8, Jul. 2019, doi: 10.1016/j.phro.2019.06.001.

[13] H. Chourak *et al.*, “Quality assurance for MRI-only radiation therapy: A voxel-wise population-based methodology for image and dose assessment of synthetic CT generation methods,” *Front Oncol*, vol. 12, Oct. 2022, doi: 10.3389/fonc.2022.968689.

[14] J. Hyuk Choi *et al.*, “Investigation of a water equivalent depth method for dosimetric accuracy evaluation of synthetic CT,” *Physica Medica*, vol. 105, Jan. 2023, doi: 10.1016/j.ejmp.2022.11.011.

[15] M. D. Morris, “Factorial Sampling Plans for Preliminary Computational Experiments,” *Technometrics*, vol. 33, no. 2, pp. 161–174, May 1991, doi: 10.1080/00401706.1991.10484804.

[16] D. G. Cacuci, Mihaela. Ionescu-Bujor, and I. Michael. Navon, *Sensitivity and uncertainty analysis.* Chapman & Hall/CRC Press, 2003.

[17] G. Qian and A. Mahdi, "Sensitivity analysis methods in the biomedical sciences," *Mathematical Biosciences*, vol. 323. Elsevier Inc., May 01, 2020. doi: 10.1016/j.mbs.2020.108306.

[18] M. Adjeiwaah, A. Garpebring, and T. Nyholm, "Sensitivity analysis of different quality assurance methods for magnetic resonance imaging in radiotherapy," *Phys Imaging Radiat Oncol*, vol. 13, pp. 21–27, Jan. 2020, doi: 10.1016/j.phro.2020.03.001.

[19] R. C. Barnard, M. Frank, and K. Krycki, "Sensitivity analysis for dose deposition in radiotherapy via a Fokker-Planck model," *Mathematical Medicine and Biology*, vol. 34, no. 1, pp. 109–123, Mar. 2017, doi: 10.1093/imammb/dqv039.

[20] C. Sosa-Marrero *et al.*, "Towards a Reduced in Silico Model Predicting Biochemical Recurrence after Radiotherapy in Prostate Cancer," *IEEE Trans Biomed Eng*, vol. 68, no. 9, pp. 2718–2729, Sep. 2021, doi: 10.1109/TBME.2021.3052345.

[21] A. Bertrand Iooss *et al.*, "Package 'sensitivity' Title Global Sensitivity Analysis of Model Outputs," 2022.

[22] F. Campolongo, A. Saltelli, and J. Cariboni, "From screening to quantitative sensitivity analysis. A unified approach," *Comput Phys Commun*, vol. 182, no. 4, pp. 978–988, Apr. 2011, doi: 10.1016/j.cpc.2010.12.039.

[23] D. Ojeda *et al.*, "Sensitivity analysis and parameter estimation of a coronary circulation model for triple-vessel disease," *IEEE Trans Biomed Eng*, vol. 61, no. 4, pp. 1208–1219, 2014, doi: 10.1109/TBME.2013.2296971.

[24] B. Zhao *et al.*, "CT synthesis from MR in the pelvic area using Residual Transformer Conditional GAN," *Computerized Medical Imaging and Graphics*, vol. 103, Jan. 2023, doi: 10.1016/j.compmedimag.2022.102150.

[25] S. Tahri *et al.*, "A high-performance method of deep learning for prostate MR-only radiotherapy planning using an optimized Pix2Pix architecture," *Physica Medica*, vol. 103, pp. 108–118, Nov. 2022, doi: 10.1016/j.ejmp.2022.10.003.

[26] A. Largent *et al.*, "Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning," *Int J Radiat Oncol Biol Phys*, vol. 105, no. 5, pp. 1137–1150, 2019, doi: 10.1016/j.ijrobp.2019.08.049.

[27] Y. Liu *et al.*, "Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning," *Phys Med Biol*, vol. 64, no. 20, 2019, doi: 10.1088/1361-6560/ab41af.

[28] H. P. Wieser *et al.*, "Development of the open-source dose calculation and optimization toolkit matRad," *Med Phys*, vol. 44, no. 6, pp. 2556–2568, Jun. 2017, doi: 10.1002/mp.12251.

[29] M. MacFarlane, D. A. Hoover, E. Wong, J. J. Battista, and J. Z. Chen, “Technical Note: A fast inverse direct aperture optimization algorithm for volumetric-modulated arc therapy,” *Med Phys*, vol. 47, no. 4, pp. 1558–1565, Apr. 2020, doi: 10.1002/mp.14074.

[30] R. Kamal *et al.*, “Efficiency of a novel non-monotonic segmented leaf sequence delivery of Varian MLC for non-split IMRT fields,” *Reports of Practical Oncology and Radiotherapy*, vol. 25, no. 5, pp. 801–807, Sep. 2020, doi: 10.1016/j.rpor.2020.07.005.

[31] E. J. Her *et al.*, “Voxel-level biological optimisation of prostate IMRT using patient-specific tumour location and clonogen density derived from mpMRI,” *Radiation Oncology*, vol. 15, no. 1, Jul. 2020, doi: 10.1186/s13014-020-01568-6.

[32] R. Nuti, G. Martini, and C. Gennari, “Age-related changes of whole skeleton and body composition in healthy men,” *Calcif Tissue Int*, vol. 57, no. 5, pp. 336–339, 1995, doi: 10.1007/BF00302068.

[33] S. J. Thomas, “Relative electron density calibration of CT scanners for radiotherapy treatment planning.,” *Br J Radiol*, vol. 72, no. 860, pp. 781–786, 1999, doi: 10.1259/bjr.72.860.10624344.

[34] B. Zurl, R. Tiefling, P. Winkler, P. Kindl, and K. S. Kapp, “Hounsfield units variations: Impact on CT-density based conversion tables and their effects on dose distribution,” *Strahlentherapie und Onkologie*, vol. 190, no. 1, pp. 88–93, Jan. 2014, doi: 10.1007/s00066-013-0464-5.

[35] A. T. Davis, A. L. Palmer, and A. Nisbet, “Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review,” *British Journal of Radiology*, vol. 90, no. 1076. British Institute of Radiology, 2017. doi: 10.1259/bjr.20160406.

[36] E. T. and E. R. J. Podder Tarun K. and Fredman, “Advances in Radiotherapy for Prostate Cancer Treatment,” in *Molecular & Diagnostic Imaging in Prostate Cancer: Clinical Applications and Treatment Strategies*, H. Schatten, Ed. Cham: Springer International Publishing, 2018, pp. 31–47. doi: 10.1007/978-3-319-99286-0_2.

[37] S. J. Dowdell, P. E. Metcalfe, J. E. Morales, M. Jackson, and A. B. Rosenfeld, “A comparison of proton therapy and IMRT treatment plans for prostate radiotherapy,” 2008.

[38] Z. Chen, Z. Yang, J. Wang, and W. Hu, “Dosimetric impact of different bladder and rectum filling during prostate cancer radiotherapy,” *Radiation Oncology*, vol. 11, no. 1, Aug. 2016, doi: 10.1186/s13014-016-0681-z.

[39] Y. Xiong *et al.*, “Assessment of intrafractional prostate motion and its dosimetric impact in MRI-guided online adaptive radiotherapy with gating,” *Strahlentherapie und Onkologie*, 2022, doi: 10.1007/s00066-022-02005-1.

[40] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” *Med Phys*, 2021, doi: 10.1002/mp.15150.

[41] M. Abdar *et al.*, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” 2020, [Online]. Available: http://arxiv.org/abs/2011.06225

[42] M. Hemsley *et al.*, “Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions,” 2020, pp. 834–844. doi: 10.1007/978-3-030-59710-8_81.

[43] C. A. T. van den Berg and E. F. Meliadò, “Uncertainty Assessment for Deep Learning Radiotherapy Applications,” *Seminars in Radiation Oncology*. W.B. Saunders, Oct. 01, 2022. doi: 10.1016/j.semradonc.2022.06.001.

Statements and Declarations

Funding

This work was partially supported by region Bretagne (France) through ARED scholarship program and a PhD scholarship Grant from e-health Research Centre-CSIRO (Australia). The authors have no relevant financial or non-financial interest to disclose.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

The original idea was conceived by Oscar Acosta and Hilda Chourak. All authors contributed to the study design. Medical physics expertise was provided by Anais Barateau, Peter Greer, Caroline Lafond and Renaud de Crevoisier. The first draft of the manuscript was written by Hilda Chourak then improved by Jason Dowling, Anais Barateau and Jean-Claude Nunes. All authors commented on previous version of the manuscript. All authors read and approved the final manuscript.

Ethics approval

Ethics approval for the study protocol was obtained from Hunter New England Human Research Ethics Committee (reference number 05/09/14/3.06). All patients provided their written informed consent, including consent to participate and consent to publish**.**

Chapter 5: Patient specific synthetic-CT quality assessment without reference

In this chapter, two methods are proposed for assessing the quality of patient-specific synthetic-CT when no planning CT is available as a ground truth. The first approach involves creating an atlas using a cohort of CT scans and detecting outliers to highlight significant error subregions. A voxel-wise Shapiro-Wilk test allowed us to conclude that the density of tissue follows a normal distribution (this hypothesis had been rejected in the rectum and in areas where the body contour was misregistered). However, this method can be time-consuming, and the accuracy of the results depends on the precision of the registration.

To address these issues, a radiomics-based approach is proposed. This involves selecting significant radiomics features from a cohort of CT scans. These features are then computed on the new image, and a score is assigned based on their distance from the values obtained on the reference images.

The first section has been published and presented in the Digital Image Computing: Techniques and Applications (DICTA) conference.
"Local quality assessment of patient specific synthetic-CT via voxel-wise analysis"
Hilda Chourak, Anaïs Barateau, Jean-Claude Nunes, Peter Greer, Safaa Tahri, Caroline Lafond, Renaud de Crevoisier, Jason Dowling, Oscar Acosta – DICTA 2022 (oral presentation)

The second section is ongoing work and need further investigation before being submitted to a journal.

1- Local quality assessment of patient specific synthetic-CT via voxel-wise analysis

Abstract

Synthetic-Computed Tomography (sCT) generation is a critical component of Magnetic Resonance Imaging (MRI)only radiation therapy workflows. The sCT computed from MRI is generally assessed by measuring Hounsfield Units (HU) discrepancies with a reference CT. The aim of this work was to propose a process for the blind assessment of local errors in generated sCTs where a reference CT is unavailable, allowing for safe MRI-only radiation therapy treatment planning. A personalised inter-patient registration method was applied to align a cohort of reference CTs into the same coordinate system. This process resulted in probability maps for each segmented organ, a mean CT image and a standard deviation map. These data were propagated to the anatomical space for each sCT, allowing for out of distribution intensities to be detected at a voxel level by computing local z-scores. Probability maps of organs were used to weight the resulting z-scores, reducing the bias induced by the registration around structures. Two sCT generation methods were chosen as examples to illustrate this methodology: an atlas-based method (ABM) and a deep-learning approach based on a Generative Adversarial Network (GAN) architecture. 39 patients treated with external beam radiotherapy for prostate cancer, with co-registered CT and MR pairs, were used for sCT generation. 26 of these patients were selected as reference CT, and sCT of the remaining 13 patients were assessed. Accurate inter-individual registration was achieved, with mean Dice scores higher than 0.91 for all organs. The average volume of error represented 0.29% of the image for the ABM, 0.37% for the GAN. The proposed methodology produced 3D volumes which identify significant local sCT errors. Depending on their size and location, these errors could lead to inaccurate tissue density computation during radiation therapy. This work provides an automated QA method aimed at preventing incorrect radiation dose delivery to patients.

Introduction

Radiation therapy is a well established, cost-effective treatment which has an evidence-based indication for 48% of cancer patients [1]. Most treatment planning is based on Computed Tomography (CT) imaging. Magnetic Resonance Imaging (MRI) is a non-ionizing modality. providing improved soft tissue contrast than CT, which leads to more accurate tumour and organ delineation in radiation therapy treatment planning. This justifies the increasing worldwide deployment of dedicated MRI scanners and MRI-linear accelerator hybrid machines for treatment delivery. However, unlike CT scans, MRI does not provide information on electron density of tissue, crucial for dose calculation. Therefore, several approaches to

generate synthetic-CT (sCT) from MRI have been developed to allow for MRI-only radiation therapy [2], [3], including: bulk density [4], [5], atlas-based [6], machinelearning models, such as patch-based methods with feature extraction [7], and more recently deep-learning models [7]– [13]. sCT image quality is currently assessed via global metrics, which measure discrepancies between a reference CT and the corresponding sCT [13], [14]. The most commonly used are intensity-based metrics [15], such as mean absolute error (MAE), mean error (ME), mean absolute percent error (MAPE), mean squared error (MSE) and peak signal-to-noise ratio (PSNR). Structural similarity (SSIM) [16] and visual information fidelity (VIF) [17] may also computed. These metrics are reported at a global level: either restricted to a single value describing agreement within the body contour of the patient, within a class of tissue, or within contoured organ boundaries [13]. A limitation of all of these metrics is that they are based on comparison to a ground truth CT scan, which means they are only useful for validation prior to clinical deployment. Once deployed in an MRI-only radiotherapy clinical practice, quality assurance becomes very important as errors may appear sporadically distributed across different tissue densities [7], [18], and there will not be a reference to assess the sCT (this is referred to as blind quality in this paper). As Hounsfield Units (HU) intensities are correlated to tissue density, the inaccurate prediction of these values may lead to error in dose calculation. Thus, it is crucial to localise and determine the volume of error in HU prediction for each new sCT. Previous studies on blind CT quality assessment have focused on noise detection [19]. Choi et al [5] investigated the use of a bulk-density map generated from MRI as reference to assess sCT, but this does not consider the contrast of intensities in soft tissue. Voxel-wise analysis has proved to be efficient in the assessment of the clinical impacts of image and dose difference across individuals [20], [21]. However, their application requires an accurate non-rigid registration of a whole population to a single coordinate system, and the implementation of voxel-wise statistical tests [22]. Previous work gave an insight of the feasibility of this method, but the analysis was based upon comparison of generated data to their ground truth [23].

The aim of this paper is to propose a strategy to assess the quality of patient specific sCT at a voxel-level to ensure safe MRI-only treatment planning. The first step was the offline computation of an atlas from a cohort of reference CTs: these data were registered in the same anatomical space following an adapted non-rigid registration process. The second step for online processing of a new sCT involves extracting mean and standard deviation maps from this atlas, which are propagated in the sCT space. This allows the computation of a z-score map to highlight local outliers. As the inter-patient registration in the atlas construction may induce bias in the analysis, probability maps extracted from the organ delineation of the atlas were used to assign weights to z-score according to their location.

The output of the blind QA method includes 3D volumes showing predicted areas of significant errors in HU values. Two different sCT generation approaches were chosen as examples to demonstrate the efficiency of this methodology: an atlas-based method (ABM) [6] and a deep-learning method, based upon a generative adversarial network (GAN) architecture [7].

Material et methods

A. Treatment planning for radiation therapy

The aim of the treatment planning phase in external beam radiation therapy is to define the optimal beam settings to maximise the dose that will be delivered in the tumour while minimising it in the surrounding organs. The standard treatment planning workflow is presented Figure 5.1.1. A planning CT is acquired, giving electron density information. This step is crucial as the dose delivered depend on the density of tissue crossed by the beams. Then, this image is registered to the diagnosis MRI. The MRI aim for tumour and organs delineation. During the beam's and dose computation step, these contours are used to control the dose that will be delivered in the target and the healthy tissue. Once the optimum settings are determined, a radiation oncologist validate the treatment. For prostate cancer, the treatment will then be delivered in 30 to 40 fractions, at a rate of 5 fractions a week.

One of the advantages of MRI-only radiation therapy is the simplification of the planning phase for the physicist and physician involved in the process, but also for the patient.

B. Data

Retrospective data from 39 patients with localised prostate cancer aged 58 to 78 years were used in this work. For each patient, a CT scan was acquired on a GE LightSpeed RT or a

Toshiba Aquilion, (256 x 256 x 128 matrix with a voxel size of 1.17 mm x 1.17 mm x 2.5 mm or 2.0 mm) and a T2-weighted MRI was acquired on a Siemens Skyra 3T in the treatment position (resolution of 1.6 mm x 1.6 mm x 1.6 mm). Each CT was resampled and registered to the corresponding MRI via a symmetric rigid registration followed by a structure-guided non-rigid method [24], [25] to rectify the anatomical variations due to the delay between both acquisitions. Non-uniformity of MRI was then corrected [6] with the Insight Toolkit Library (ITK). Organ delineation (labelling) was performed on the MRI by three experienced observers. These organs include the bladder, prostate, rectum, bones, and body contour (Figure 5.1.2).

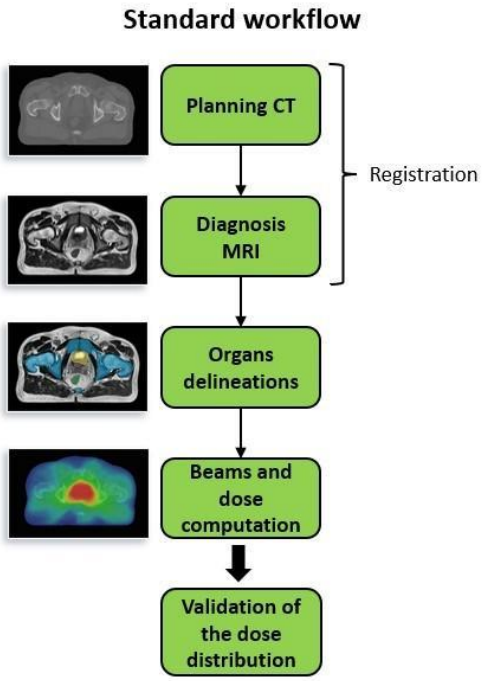

Figure 5.1.1: Standard workflow for treatment planning in external beam radiation therapy.

C. synthetic-CT generation methods

1. **Atlas based method:** The Atlas-based method (ABM) [6] used rigid followed by non-rigid registration of a set of 3D MRI-CT atlases to a target MRI. sCT voxel intensities were then propagated from the co-registered CTs according to a patch based local intensity match between the target MRI and the multiple MRIs composing the atlas.
2. **Generative adversarial network (GAN):** The GAN architecture chosen for this study to generate sCT was composed of a U-Net for the generator, and a PatchGAN for the discriminator. Axial 2D CT and MRI slices, were used to train the model. A three-fold cross validation was used to validate the model: 26 patient data composed the training cohort, and 13 patient data composed the validation set. The architecture of the model is described in Largent et al. [7].

D. Workflow

The voxel-based analysis workflow process is presented in Figure 5.1.3.

a) Structure guided non-rigid registration*:* In radiation therapy, target (the prostate in this study) and organs at risk (bladder, rectum, bones) delineation is a crucial step in the treatment planing process. Thus, delineations are systematically achieved by the radiation oncologist, allowing for the use of these contours in the registration process. All the reference data were registered following the personalized organ driven non-rigid registration method described below. First, a representative study patient was selected as a template (this patient was selected based on their similarity to the median volume of

body, prostate, rectum and bladder). Following this, a customised organ-driven registration, based upon previously proposed methods [20], [23] was performed. As MRI provides better contrast in soft tissues compared to CT, the registration process has a combination of MR intensities and structural descriptions (SD) of the structure contours (bones, prostate, bladder, rectum) obtained in a pre-processing step. The structural description was obtained as follows: - Euclidean distances to the surface were computed for all structures [26].

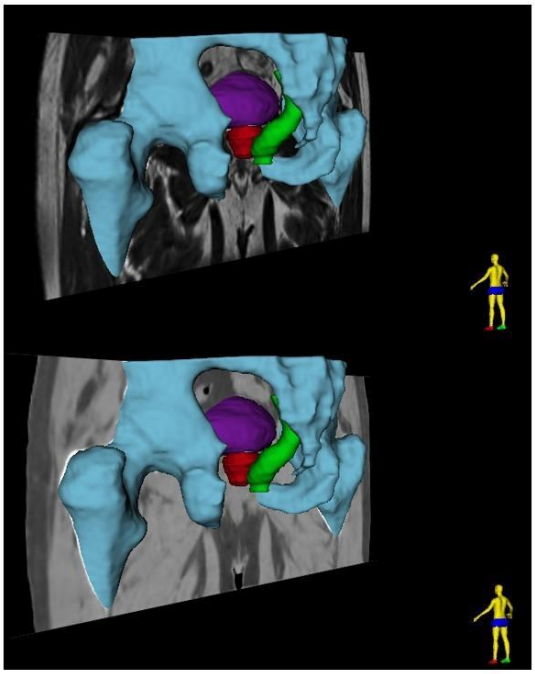

Figure 5.1.2: MRI of one patient of the cohort (top) and its co-registered CT (bottom) with 3D models of delineated main organs in the pelvic area: bones, bladder, prostate and rectum respectively in blue, purple, red and green.

A scalar field was generated by applying the Laplacian equation inside the rectum [27], resulting in a normalised distance map to the central path of the organ. - The Laplacian was also computed for the prostate with respect to its barycentre.

All the input images were aligned using the Elastix toolbox (translation). Then, non-rigid registration based on diffeomorphic demons [28] was successively applied to: i) the bladder SD, ii) the whole pelvis, iii) the prostate SD, iv) the rectum SD, v) the bones SD. This resulted in a 3D vectors fields, allowing for the propagation of CTs in the template space. Bones were divided between spongy and cortical and separately registered, to preserve the composition of inner structure. This step-by step process allows for fine-tunes, providing accurate registration of both contours and inner structure of all organs. The non-rigid registration was validated by computing the Dice Similarity Coefficient (DSC) between the template structures, $V_{t_{MRI}}$ and the corresponding deformed delineated organ, V_{MRI}:

$$DSC = \frac{2(V_{t_{MRI}} \cap V_{MRI})}{V_{t_{MRI}} + V_{MRI}} \quad (1)$$

b) Reference atlas data: Once all the reference CT are in the common coordinate system, the voxel-wise mean and standard deviation of the intensities' distribution were computed.

c) Probability maps*:* The probability map resulted from the superposition of the reference CT organs' delineation in the common coordinate system (CCS). This allowed for the visualisation and estimation of the discrepancies between the delineated organ contours following registration, and provides an indication of the probability of a voxel's inclusion within an organ.

E. sCT assessment

Mean HU values, standard deviation and probability maps of structures were propagated to each sCT anatomical space for the voxel-wise analysis. We assume that, at each location, the tissue density across the cohort follows a normal distribution. To detect outliers in the sCT, the z-score, also called standard score, was computed. It is an indication of the probability of the value to be part of the reference distribution, and was calculated at each voxel i as follows:

$$z(i) = \frac{sCT(i) - \mu(i)}{\sigma(i)} \times \omega(i) \quad (2)$$

Where $sCT(i)$ is the HU value of the i^{th} voxel in the synthetic image. $\mu(i)$ and $\sigma(i)$ are respectively the mean intensity and the standard deviation at this location in the CT atlas. The inter-patient registration may lead to bias in the reference data. Thus, to reduce the impact of mis-registration, $\omega(i)$, the weight corresponding to the normalised probability map value at the i^{th} voxel, were multiplied on the z-scores. All voxel values outside the 95% confidence interval in the resulting 3D map, i.e all values superior to 2 or inferior to -2, were considered as outliers.

A conventional image quality assessment was proceeded to highlight the relevance of the method. Thus, mean absolute error (MAE), mean error (ME) and mean absolute percent error (MAPE) were computed as follow:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |HU_{CT}(i) - HU_{sCT}(i)| \quad (3)$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} HU_{CT}(i) - HU_{sCT}(i) \quad (4)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{HU_{CT}(i) - HU_{sCT}(i)}{HU_{CT}(i)}\right| \quad (5)$$

with n the total number of voxels, $HU_{CT}(i)$ and $HU_{sCT}(i)$ the intensities of the i^{th} voxel in, respectively, the reference and the generated image. The closer to zero these values are, the more accurate is the prediction. These metrics were applied to assess HU errors in the whole pelvis, by organ and in the volume of outliers.

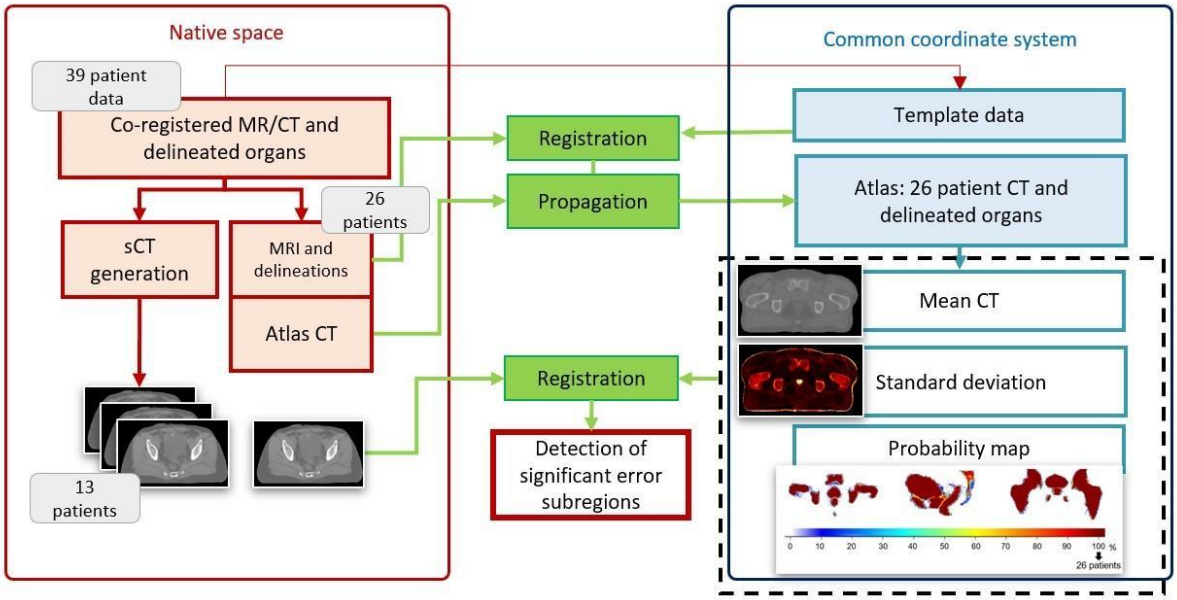

Figure 5.1.3: Workflow of voxel-based analysis. The input dataset was composed of 39 images, and sCT were generated for all of them. 26 patient CTs were selected as a reference cohort, allowing for blind sCT quality assessment for the 13 remaining patients. The 26 patient data were non-rigidly registered into a common coordinate system (CCS), resulting in the reference atlas. Probability maps of the delineated organs, mean reference CT and standard deviation of CT in the CCS were then computed at a voxel-level. These data are then registered to each sCT anatomical space, providing detection of significant error subregions.

Results and discussion

The personalised non-rigid registration method accurately brought the 26 patients of the reference cohort in the same anatomical space, as demonstrated by the average DSC scores of 0.98 ± 0,01 for the body contour, 0.93 ± 0,01 for the bones, 0.96 ± 0,01 for the bladder, 0.91 ± 0,02 for the rectum and 0.91 ± 0,02 for the prostate. Previous work has supported the use of the organ structural description [20], and its combination to MRI for better preservation of the different soft tissue class within the body [23]. The reference atlas for the assessment was thus considered reliable, allowing for the voxel-wise analysis to be proceeded [22]. On this sample of 13 sCT assessed, ABM and GAN obtained close MAE, ME and MAPE

results, as shown in Table 5.1.1. Despite the small volume represented by the outliers, these presented a MAE up to 155.37 HU for the ABM and 135.82 for the GAN, and a MAPE of 4.03 for the ABM, against 2.44 for the GAN. MAE values appeared to be correlated to the reference intensity (the most important errors are in cortical bones, where the mean HU value is the highest) and illustrates how far from the ground truth the mean HU prediction is. The MAPE, which correspond to the relative difference, was computed as a measure of prediction accuracy. The ME determined if the prediction tends to be systematically superior (negative results) or inferior to the reference. These metrics offer complementary information and can be computed at a voxel level, and should not be used alone for global assessment. For example, a ME close to zero in a volume do not imply an accurate prediction, as this volume might include both high positive and negative scores. And, as full reference metrics, they need a CT ground truth, making them useful to assess the efficiency of a sCT generation method during the development phase, but unsuitable in a clinical workflow.

The average volume of the outliers identified within the sCT scans represented 0.29% ± 0.26 of the image for the ABM, 0.37% ± 0.35 for the GAN (Table 5.1.1). Figure 5.1.4 illustrates the regions where the out of distribution HU values were detected in the worst case on one hand (patient 3), and the best case (patient 5) on the other. Few errors appeared in the prostate and in the organs at risk (bladder, rectum, femoral heads), except for one patient where the voxel-wise approach has detected significant errors in the bladder for both sCT generation method, especially for the GAN, as shown in Figure 5.1.4. This might be explained by a high anatomical variation compared to the training cohort for the GAN. The ABM performed better in this location, for all the sCT assessed (Table 5.1.1Table 5.1.1).

A similar process was used in Wang et al. [29] on pediatric brain data. However, the registration errors were not considered and the threshold of a difference of 100 HU compared to the mean chosen in this study may not be valuable regarding the variability of tissue densities within the pelvic area (from -1000 HU for the air to 1500 HU for dense bones) and for each patient.

Applying a weight to the z score map according to the probability of mis-registration of each structure reduce the bias linked to the inter-patient registration, but error may still remain, especially when assessing sCT for a patient with high anatomical variation compared to the reference cohort. So, part of our future work is to separate errors due to the inter-patient non-rigid registration from errors inherent to the generation method. Deep-learning can help to estimate local error registration [30]. Unfortunately, this study is limited by the number of data available. Thus, using 26 patients as reference atlas might not be enough to be representative of the population concerned by prostate cancer.

Table 5.1.1: Mean absolute error (MAE), mean error (ME) and mean absolute percent error (MAPE) obtained for the 13 sCT generated with the atlas-based method (ABM) on one hand, and the GAN on the other. These scores were computed in the whole pelvis (body), by organ and within the volume of outliers.

		ABM			GAN		
		mean		std	mean		std
BODY	MAE (HU)	37.89	±	8.65	32.35	±	7.70
	MAPE	1.53	±	0.73	1.33	±	0.69
	ME (HU)	-7.06	±	14.47	-3.53	±	13.87
BLADDER	MAE (HU)	17.30	±	8.46	20.84	±	14.25
	MAPE	2.42	±	1.01	3.12	±	1.13
	ME (HU)	8.49	±	11.35	6.98	±	19.94
RECTUM	MAE (HU)	76.54	±	59.97	68.83	±	66.85
	MAPE	2.12	±	1.09	1.84	±	1.32
	ME (HU)	-14.40	±	78.81	-38.19	±	69.57
PROSTATE	MAE (HU)	23.94	±	4.96	19.08	±	5.80
	MAPE	1.62	±	1.26	1.64	±	1.59
	ME	-4.05	±	13.29	-3.34	±	14.05
BONES	MAE (HU)	127.29	±	27.58	122.29	±	21.39
	MAPE	1.58	±	1.21	1.34	±	1.03
	ME (HU)	23.90	±	48.18	24.90	±	38.23
VOL. OF OUTLIERS	% body vol.	0.29	±	0.26	0.37	±	0.35
	MAE (HU)	155.37	±	30.56	135.82	±	29.72
	MAPE	4.03	±	2.29	2.44	±	1.39
	ME (HU)	25.49	±	56.02	-10.84	±	62.20

In addition, a potential uncertainty while using an atlas for sCT assessment in the pelvic area is the air in the digestive system, as there is no consistent state from one patient to another, and even for the same patient over time. One way to correct this issue might be to include a step to compare the volume of air in the delineated rectum in the input MRI with its resulting sCT. A limitation of the proposed approach in this work may be the computation time involved with the non-rigid registration process (20 min for each sCT). However, this is generally not an issue in treatment planning (as there are usually a number of days between an initial planning

scan and treatment delivery), and faster registration methods could be investigated if clinically required.

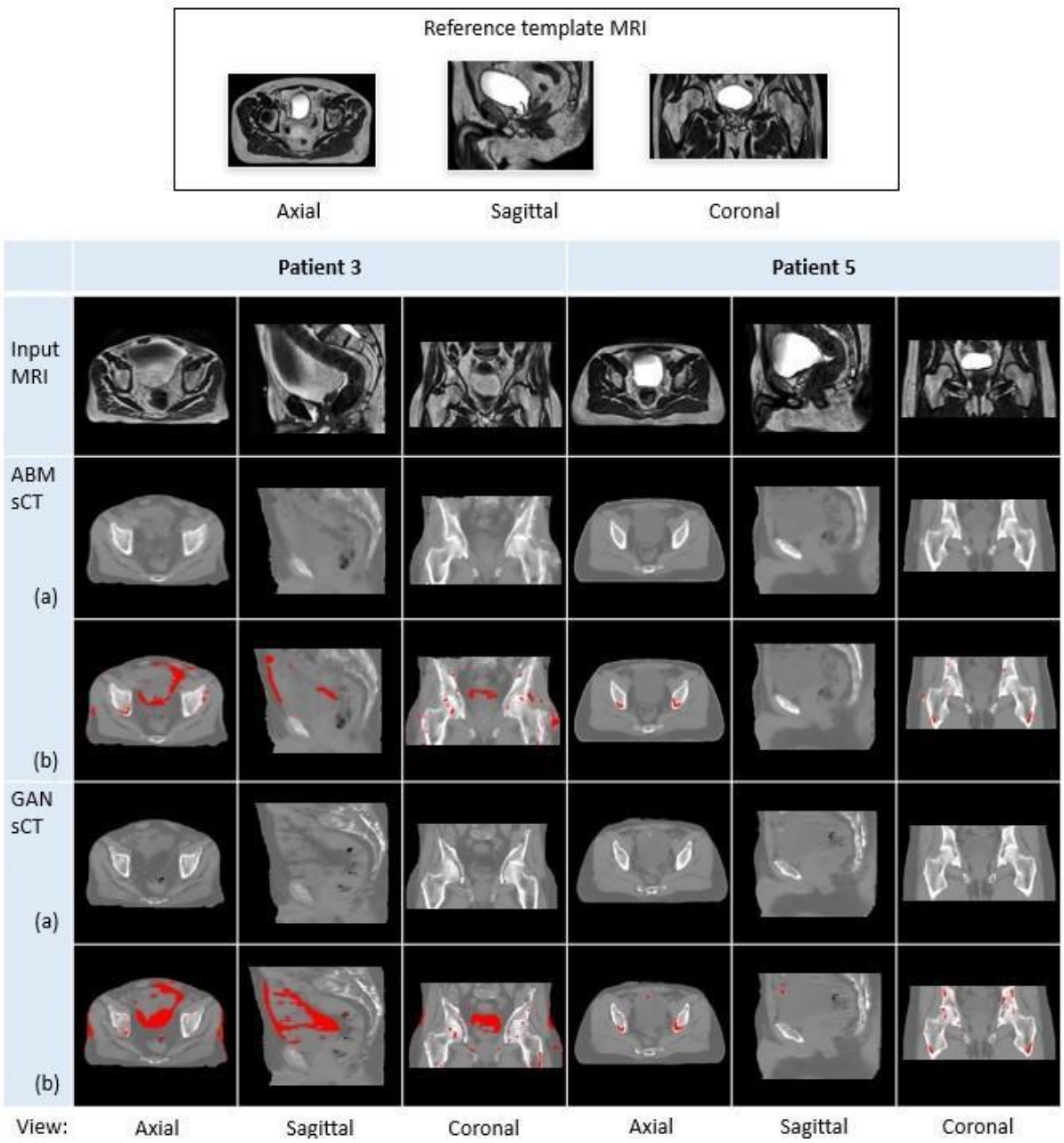

Figure 5.1.4: Axial, sagittal and coronal views of the MRI used as template to create the reference atlas followed by two patients MRI and (a) their resulting synthetic-CT (sCT) generated from the atlas-based method (ABM) and the GAN method. (b) represents these sCT with the significant outlier (error) volume overlaid in red

Integrating this workflow in a treatment planning system will allow for manual correction of the density by the radiation oncologist in detected volume of outliers before dose calculation. This methodology can also be used for offline validation or to validate sCT generation methods as part of a clinical evaluation stage. This study presents results for the male pelvis

(prostate cancer radiation therapy), but the method can be applied to other anatomical locations.

Conclusion

In this paper, we have presented a voxel-wise analysis method based on an efficient non-rigid registration process. The step-by-step approach has been shown to be robust to high inter-individual anatomical variability. The method results in a 3D volume, highlighting regions where the estimated HU in the generated sCT is significantly different from an atlas of previously acquired reference CT data. The proposed methodology has been shown to be capable of detecting local errors in sCT generated from MRI, which is an important contribution towards safe MRI based radiation treatment planning.

Compliance with ethical standard

Ethics approval for the study protocol was obtained from Hunter New England Human Research Ethics Committee. (reference number 05/09/14/3.06). All patients provided their written informed consent.

Acknowledgment

This work was partially funded by region Bretagne (France) through ARED scholarship program, and a PhD scholarship Grant from e-health Research Centre- CSIRO (Australia).

References

[1] M. B. Barton *et al.*, "Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012," *Radiotherapy and Oncology*, vol. 112, no. 1, pp. 140–144, Jul. 2014, doi: 10.1016/j.radonc.2014.03.024.

[2] E. Johnstone *et al.*, "Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 100, no. 1. 2018. doi: 10.1016/j.ijrobp.2017.08.043.

[3] D. Bird, A. M. Henry, D. Sebag-Montefiore, D. L. Buckley, B. Al-Qaisieh, and R. Speight, "A Systematic Review of the Clinical Implementation of Pelvic Magnetic Resonance Imaging–Only Planning for External Beam Radiation Therapy," *International Journal of Radiation Oncology Biology Physics*, vol. 105, no. 3. Elsevier Inc., pp. 479–492, Nov. 01, 2019. doi: 10.1016/j.ijrobp.2019.06.2530.

[4] J. Kim *et al.*, "Dosimetric evaluation of synthetic CT relative to bulk density assignment-based magnetic resonance-only approaches for prostate radiotherapy," *Radiation Oncology*, vol. 10, no. 1, 2015, doi: 10.1186/s13014-015-0549-7.

[5] J. H. Choi *et al.*, "Bulk Anatomical Density Based Dose Calculation for Patient-Specific Quality Assurance of MRI-Only Prostate Radiotherapy," *Front Oncol*, vol. 9, 2019, doi: 10.3389/fonc.2019.00997.

[6] J. A. Dowling *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*, vol. 93, no. 5, pp. 1144–1153, 2015, doi: 10.1016/j.ijrobp.2015.08.045.

[7] A. Largent *et al.*, "Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning," *Int J Radiat Oncol Biol Phys*, vol. 105, no. 5, pp. 1137–1150, 2019, doi: 10.1016/j.ijrobp.2019.08.049.

[8] J. Fu *et al.*, "Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging," *Med Phys*, vol. 46, no. 9, pp. 3788–3798, 2019, doi: 10.1002/mp.13672.

[9] X. Tie, S. K. Lam, Y. Zhang, K. H. Lee, K. H. Au, and J. Cai, "Pseudo-CT generation from multi-parametric MRI using a novel multi-channel multi-path conditional generative adversarial network for nasopharyngeal carcinoma patients," *Med Phys*, vol. 47, no. 4, pp. 1750–1762, 2020, doi: 10.1002/mp.14062.

[10] D. Bird *et al.*, "Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning," *Radiotherapy and Oncology*, vol. 156, pp. 23–28, 2021, doi: 10.1016/j.radonc.2020.11.027.

[11] H. Yang *et al.*, "Unpaired Brain MR-to-CT Synthesis using a Structure-Constrained CycleGAN," 2018, [Online]. Available: http://arxiv.org/abs/1809.04536

[12] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, "Deep learning based synthetic-CT generation in radiotherapy and PET: A review," *Med Phys*, 2021, doi: 10.1002/mp.15150.

[13] M. Boulanger *et al.*, "Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review," *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[14] Y. Liu *et al.*, "Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning," *Phys Med Biol*, vol. 64, no. 20, 2019, doi: 10.1088/1361-6560/ab41af.

[15] G. Wang, Y. Zhang, X. Ye, and X. Mou, "Image quality assessment," in *Machine Learning for Tomographic Imaging*, in 2053-2563. IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[17] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006, doi: 10.1109/TIP.2005.859378.

[18] A. Bahrami, A. Karimian, and H. Arabi, “Comparison of different deep learning architectures for synthetic CT generation from MR images,” *Physica Medica*, vol. 90, no. August, pp. 99–107, 2021, doi: 10.1016/j.ejmp.2021.09.006.

[19] A. P. Reeves, Y. Xie, and S. Liu, “Automated image quality assessment for chest CT scans:,” *Med Phys*, vol. 45, no. 2, pp. 561–578, Feb. 2018, doi: 10.1002/mp.12729.

[20] E. Mylona *et al.*, “Voxel-Based Analysis for Identification of Urethrovesical Subregions Predicting Urinary Toxicity After Prostate Cancer Radiation Therapy,” *Int J Radiat Oncol Biol Phys*, vol. 104, no. 2, pp. 343–354, 2019, doi: 10.1016/j.ijrobp.2019.01.088.

[21] R. N. Finnegan *et al.*, “A statistical, voxelised model of prostate cancer for biologically optimised radiotherapy,” *Phys Imaging Radiat Oncol*, vol. 21, pp. 136–145, Jan. 2022, doi: 10.1016/j.phro.2022.02.011.

[22] G. Palma, S. Monti, and L. Cella, “Voxel-based analysis in radiation oncology: A methodological cookbook,” *Physica Medica*, vol. 69. Associazione Italiana di Fisica Medica, pp. 192–204, 2020. doi: 10.1016/j.ejmp.2019.12.013.

[23] H. Chourak *et al.*, “Voxel-Wise Analysis for Spatial Characterisation of Pseudo-CT Errors in MRI-Only Radiotherapy Planning,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 395–399. doi: 10.1109/ISBI48211.2021.9433800.

[24] D. Rivest-Hénault, P. Greer, jurgen frip, and J. Dowling, *Structure-Guided Nonrigid Registration of CT–MR Pelvis Scans with Large Deformations in MR-Based Image Guided Radiation Therapy*. 2014. doi: 10.1007/978-3-319-05666-1_9.

[25] D. Rivest-Hénault, N. Dowson, P. B. Greer, J. Fripp, and J. A. Dowling, “Robust inverse-consistent affine CT-MR registration in MRI-assisted and MRI-alone prostate radiation therapy,” *Med Image Anal*, vol. 23, no. 1, pp. 56–69, 2015, doi: 10.1016/j.media.2015.04.014.

[26] P.-E. Danielsson, “Euclidean distance mapping,” *Computer Graphics and Image Processing*, vol. 14, no. 3, pp. 227–248, 1980, doi: https://doi.org/10.1016/0146-664X(80)90054-4.

[27] S. E. Jones, B. R. Buchbinder, and I. Aharon, “Three-dimensional mapping of cortical thickness using Laplace’s equation,” *Hum Brain Mapp*, vol. 11, no. 1, 2000, doi: 10.1002/1097-0193(200009)11:1<12::AID-HBM20>3.0.CO;2-K.

[28] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: efficient non-parametric image registration.,” *Neuroimage*, vol. 45, no. 1 Suppl, 2009, doi: 10.1016/j.neuroimage.2008.10.040.

[29] C. Wang *et al.*, “Toward MR-only proton therapy planning for pediatric brain tumors: Synthesis of relative proton stopping power images with multiple sequence MRI and development of an online quality assurance tool,” *Med Phys*, vol. 49, no. 3, pp. 1559–1570, Mar. 2022, doi: 10.1002/mp.15479.

[30] K. A. J. Eppenhof and J. P. W. Pluim, “Supervised local error estimation for nonlinear image registration using convolutional neural networks,” in *Medical Imaging 2017: Image Processing*, SPIE, Feb. 2017, p. 101331U. doi: 10.1117/12.2253859.

2- Radiomic features selection for quality assessment of patient specific synthetic-CT

Abstract

The MRI-only workflow has gained popularity in external beam radiation therapy for improving efficiency and accuracy in treatment planning. It relies on generating synthetic-CT (sCT) from MRI for dose calculation, but assessing the accuracy of these images without reference remains challenging. Current methods that rely on the comparison to a reference CT as a ground truth are inadequate for MRI-only workflows. This study aims to investigate a protocol for assessing patient-specific sCTs in radiation therapy (RT) workflow using radiomics analysis. The study involved 39 patients with localized prostate cancer. sCTs were generated using four methods: multi-atlas, bulk-density, patch-based, and a Generative adversarial network (GAN). Radiomic features extracted from CT scans and sCTs were used to select the most significant features through a random forest algorithm train to classify the images according to their gamma pass-rate. These features were then compared to expected values from the reference CT cohort, assigning a quality score based on out-of-distribution values. Radiomics analysis is shown to be an efficient method for quality control of patient-specific sCTs in MRI-only RT. However, further investigations are necessary before deploying this approach in clinical settings. Nevertheless, the findings provide valuable insights for evaluating the accuracy of predicted Hounsfield Units (HU), enhancing dose calculation and treatment outcomes in MRI-only RT.

Introduction

Radiation therapy (RT) is a well-established treatment with evidence-based indications for 48% of cancer patients [1]. The standard RT workflow relies on two imaging modalities: computed tomography (CT) for dose calculation based on electron density information, and magnetic resonance imaging (MRI) for better soft tissue contrast, enabling more accurate target delineation [2] and minimising the risk of toxicity in healthy tissue [3]. To define the treatment plan, traditionally both CT and MRI images are co-registered. However, the MR-CT registration step introduces uncertainties, with reported calculations of up to 2 mm for prostate cancer patients [4].

In order to enhance efficiency and accuracy in the clinical workflow, MRI-only RT has gained popularity, eliminating the need for CT scans and relying solely on MRI. This has led to the widespread deployment of dedicated MRI scanners and MRI-linear accelerator (MRI-LINAC) hybrid machines for treatment delivery. The advantage of MRI-LINAC is its ability to accommodate adaptive RT, which considers daily internal anatomical changes and recalculates the dose distribution prior to each session. However, as MRI does not provide

electron density information, the generation of synthetic-CT (sCT) becomes crucial for MRI-only RT [5]–[7].

The use of sCT in clinical practice faces a significant limitation: the lack of robust approaches to evaluate the accuracy of generated images and ensure the correctness of predicted Hounsfield Units (HU) for accurate dose calculation. Currently, the assessment of sCT involves the use of full reference intensity-based metrics, i.e. requiring a reference CT, such as mean absolute error (MAE), mean error (ME), and peak signal-to-noise ratio (PSNR) [8]. Perception-based metrics like the structural similarity index measure (SSIM) [9], [10] and the multiscale SSIM are also commonly employed. However, these metrics rely on a simulation CT as the ground truth and the accuracy of the registration between the reference CT and the MRI used for sCT generation. Consequently, these metrics are inadequate for quality control in an MRI-only clinical workflow. To address this challenge and enable patient-specific sCT assessment, a previous study proposed the use of an atlas of reference CT to identify local out-of-distribution HU numbers in sCT [11]. This approach, based on voxel-wise analysis, aimed to highlight discrepancies between the predicted and expected HU values. However, it is necessary to note that the reliability of the results depends on the robustness of the interpatient registration process used to generate the atlas prior to statistical analysis.

The purpose of this study is to establish a simple and reliable protocol based on the selection of significant image features from radiomics, allowing for patient-specific sCT assessment before its use in clinical practice.

Radiomics analysis has been successfully employed in RT to improve diagnosis, assess treatment response [12]–[14], classify errors [15], and detect errors [16] in intensity-modulated RT (IMRT) quality assurance (QA). However, to our knowledge, its application for patient-specific sCT QA has not been explored. In this paper, quantitative image features have been extracted from a cohort of CT scans and sCT generated from four previously published methods: an atlas-based method [17], a bulk-density method, a patch-based method [18], and a Generative Adversarial Network (GAN) [19]. A random forest (RF) algorithm is then employed to select the most significant image features according to the gamma pass-rate (GPR).

Material and methods

The workflow of the study is presented Figure 5.2.1. First, radiomic features were computed on a cohort of reference CTs and on sCTs generated through four different methods: an atlas-based method, a bulk-density method, a patch-based method, and a GAN. Volumetric modulated arc therapy (VMAT) was planned on reference CT images (treatment planning system Pinnacle v.9.10, Philips) using the collapsed cone convolution algorithm and a dose grid resolution of 3 mm. For all patients, a sequential treatment was delivered with a total dose of 78 Gy in the clinical target volume (CTV). The same beam parameters were used to compute the dose on the sCT. Key features were selected using an RF algorithm train to

classify images based on their GPR (”excellent” for GPR ≥ 99.9, ”correct” for 99.9 > GPR ≥ 98, ”insufficient” for GPR < 98). GPR were computed using MatRad [20], an open-source software for radiation treatment planning developed for research purposes. The selected features were then used to define a score, computed according to the percent of selected features with values falling within the range of expected values according to the cohort of reference CT.

Dataset

A total of 39 patients with localised prostate cancer, aged 58 to 78 years, were included in this study. For each patient, a CT scan was acquired using a GE LightSpeed RT large-bore scanner or a Toshiba Aquilion scanner, with a matrix size of 256 x 256 x 128 and a voxel size of 1.17 mm x 1.17 mm x 2.5 mm or 2.0 mm. Additionally, a T2-weighted MRI was obtained in treatment position using a Siemens Skyra 3T scanner, with a voxel size of 1.6 mm x 1.6 mm x 1.6 mm. To correct the non-uniformity of the MRI images, the N4 bias field correction algorithm from the Insight Toolkit Library (ITK) was employed [17]. Afterwards, the CT scans were resampled and registered to their corresponding MRIs to account for anatomical variations caused by the time gap between acquisitions. This registration process involved an inverse-consistent affine registration [21], followed by a non-rigid registration [22].

Synthetic-CT generation

Multi-Atlas The multi-atlas technique was originally published by Dowling et al. [17]. It involves non-rigid registrations of MRI-CT atlases that have been co-registered with a target MRI. A fusion step is then performed by assessing local similarities between the training atlas and the target MRI. The local weighting of the registered CT atlases in the corresponding areas is used to create each voxel in the sCT.

Bulk-density sCTs were obtained by assigning HU values to the patient’s soft tissue, bones and air. The volume of air resulted from thresholds in the inner part of the rectum delineated on MRI. The soft tissue area corresponds to the subtraction of bones and air from the body contour. A water equivalent density (0 HU) was assigned to the soft tissue, and densities allocated to bones and air were 350 HU and -450 HU, corresponding to the mean CT values of the cohort in the corresponding segmented regions [18].

Patch-based This approach involves inter-patient registration, feature extraction from MRIs and patch partitioning. The sCT is generated within a matching of multiple patches to the target MRI [18].

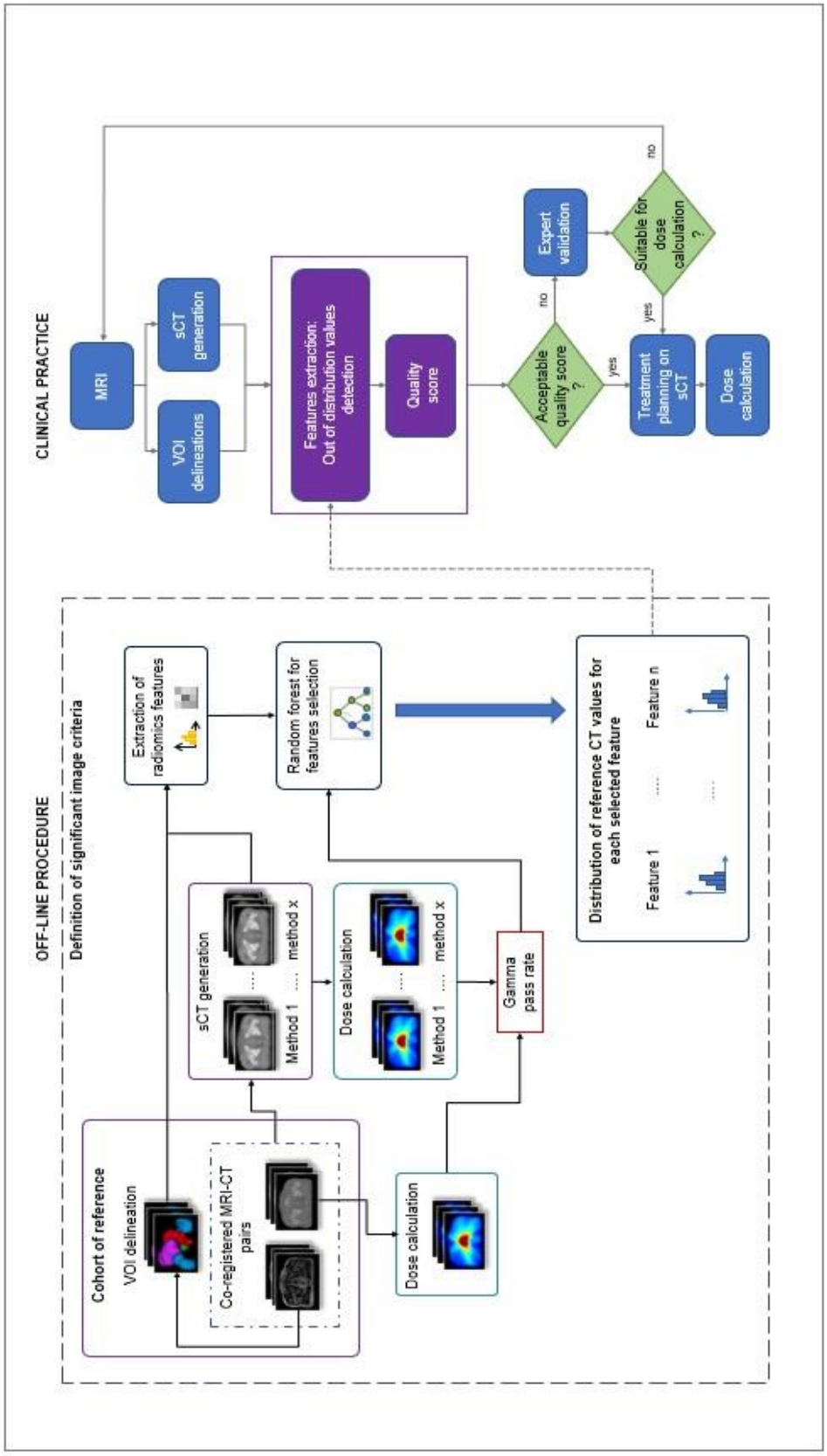

Figure 5.2.1: Workflow of the study: The offline procedure generates distributions of reference values for the *n* most significant features correlated with the gamma pass-rate. These distributions will then be employed in clinical practice to identify anomalies in patient-specific synthetic-CT (sCT) prior to dose calculation. If the resulting score is unsatisfactory, we recommend a subjective quality evaluation by an expert before deciding whether to restart the generation process. This can be repeated for each volume of interest (VOI) delineated for an organ-wise assessment.

GAN The GAN architecture used in this study for sCT generation is fully described in Largent et al. [18]. The generator employed a U-Net inspired by Han et al. [22], with the L2 norm used as the loss function:

$$L_G(I, C) = \|C - G(I)\|_2^2 \quad (1)$$

Here (equation (1)), I corresponds to the MRI intensity, $G(I)$ represents the generated sCT, and C denotes the reference CT. The discriminator utilised a PatchGAN, with binary cross-entropy as the adversarial loss function:

$$L_D(G(I), C) = -\sum_{i=1}^{n} C_i \log(G(I)_i) + (1 - C_i) \log(1 - G(I)_i) \quad (2)$$

In the equation (2), $G(I)$ refers to the sCT produced by the generator from the target MRI,C represents the corresponding reference CT, and n the number of voxels in C. The global loss was created by combining $L_G(I,C)$ and $L_D(G(I),C)$. The model was trained using axial two-dimensional CT and MRI slices, and three-fold cross-validation was applied.

Test data

To test the ability of the algorithm to detect errors, 10 reference CT were randomly selected. Expert delineated MRI contours were used to modified HU values within the bladder, CTV, rectum, bones, and the remaining soft tissue. Random HU shifts were applied:

- From -100 HU to +100 HU in the bladder,
- From -1000 HU to +200 HU in the rectum,
- From -500 HU to +500 HU in the bones,
- From -100 HU to +100 HU in the CTV,
- From -100 HU to +100 HU in the remaining soft tissue.

Remaining soft tissue volumes are generated by subtraction of bone, bladder, CTV and rectum volumes from the body contour. Higher threshold has been defined for bone and rectum, according to the difficulty for a sCT generation method to predict HU in these locations. Especially for the rectum, where the presence of air pocket is uncertain.

A spherical artefact has been added to 10 other randomly selected CT. The size, intensity and location has been randomly assigned to this error volume:

- Intensity from -250 HU to + 250 HU.
- Distance from 0 to 100 mm to the isocentre.

- Size from 2 mm to 50 mm of diameter.

Examples of images used for validation are presented figure 5.2.2.

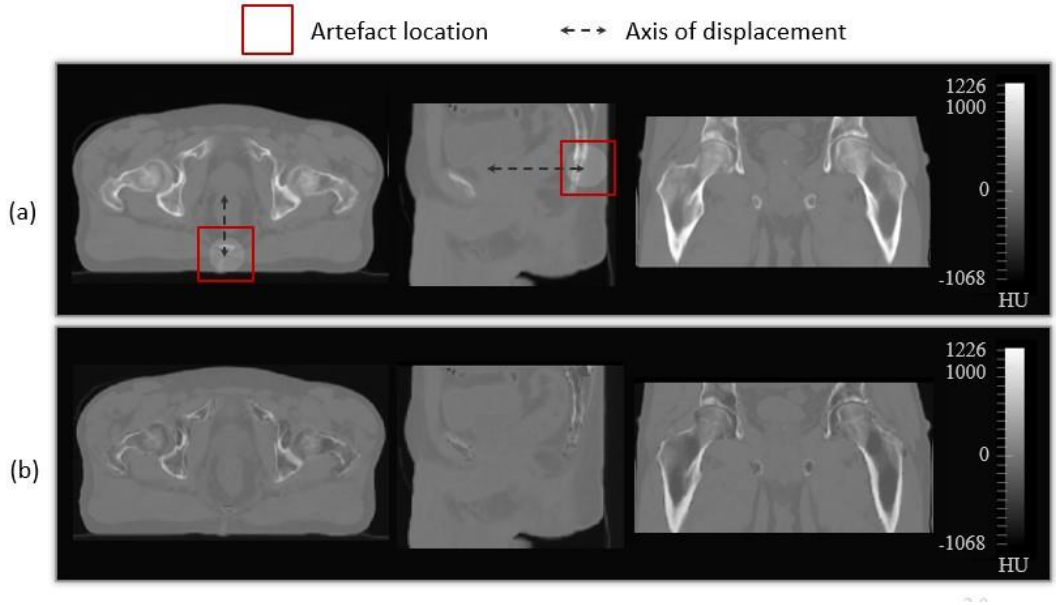

Figure 5.2.2: Example of modification applied on CT for validation, with (a) the presence of an artefact and (b) random HU modification in delineated organs.

Radiomic feature extraction and selection

Images features were extracted from CT scans and sCT using the PyRadiomics Python package [24]. The package includes a class for first-order statistics (18 features), a class for shape descriptors (26 features), and 75 textures features classified in five classes: grey level co-occurrence matrix, grey level run length matrix, grey level size zone matrix, neighbouring grey tone difference matrix, and grey level dependence matrix. In this study, a total of 1275 features were extracted from body contour. These features encompassed first-order statistics and textures from the original image, as well as wavelets and Laplacian of Gaussian (LoG) filters with sigma values of 1 mm, 2 mm, 3 mm, 4 mm, and 5 mm. Morphological features were deemed irrelevant and discarded.

In the feature selection approach, the first step involved eliminating highly correlated features by discarding those with a pairwise Pearson correlation coefficient ≥ 0.85. Then, an RF algorithm was used to select the most significant features as descriptors. The algorithm aimed to classify images according to their GPR (“Excellent” for GPR ≥ 99.9, ”Correct” for 99.9 < GPR ≥ 98, ”Insufficient” for GPR > 98). GPR were computed using MatRad [20], an open-source software for radiation treatment planning developed for research purposes. The criteria to compute the GPR were 1% dose difference and 1mm distance to agreement.

The RF model was implemented using the scikit-learn package [25]. The hyperparameters of the RF model were fine-tuned using a randomised search parameter optimisation process [26]. Then a recursive feature elimination was performed in a cross-validation loop to find the optimal number of features (RFECV). The selected features were computed from the image to assess. Points falling outside the range of the reference cohort were considered outliers, given the limited size of the dataset consisting of real patient data used in this study. A quality score was assigned based on the percentage of out-of-distribution values. Consequently, the results were classified into three categories: “poor” for a score < 40%, “good” for a score between 40% and 80%, and “excellent” for a score ≥ 80%.

Random forest model training and evaluation

3-fold stratified cross-validation was conducted for model training and evaluation. The training/validation dataset included 136 images, representing 70% of the initial dataset, and the test set 59 (10 “Excellent GPR”, 37 “Good GPR”, 12 “Insufficient GPR”).

To evaluate the efficiency of the trained RF model, several performance metrics were computed on the test dataset, including precision (eq.3), sensitivity (eq.4), F1-score (eq.5) and accuracy (eq.6). The results are presented in Table 5.2.1, followed by the confusion matrix, Table 5.2.1.

Precision measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It provides insight into the model’s ability to avoid false positives.

$$precision = \frac{tp}{tp + fp} \tag{3}$$

tp stands for the number of true positive, and fp the number of false positive.

Sensitivity, also known as recall or true positive rate, calculates the proportion of actual positive instances that are correctly predicted by the model. It indicates the model’s effectiveness in identifying true positives.

$$sensitivity = \frac{tp}{tp + fn} \tag{4}$$

with fn the number of false negative.

The F1-score combines precision and sensitivity into a single metric and provides a balanced measure of the model’s performance. It considers both false positives and false negatives, making it useful when there is an imbalance between positive and negative instances.

$$F1 = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{5}$$

Accuracy determines the overall correctness of the model by calculating the proportion of correctly predicted instances out of the total instances. It is a common metric to assess the overall performance of a model.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (6)$$

with tn the number of true negative.

These metrics provide valuable insights into different aspects of the model's performance and help evaluate its effectiveness in making accurate predictions.

Results

RF models were found to accurately distinguish between the first two categories of GPR, but tend to struggle in predicting GPR < 98, as indicated in Table 5.2.1, which displays the performance of the model for each class. The confusion matrix (Table 5.2.2) reveals that the probability for the prediction to be accurate for low GPR is only 58%. In 42% of cases, the GPR is classified as "good" instead of "insufficient". This could potentially be attributed to the lack of poor GPR in the dataset (Figure 5.2.3).

Regarding the evaluation, Table 5.2.3 demonstrates that both random HU errors in VOI and the presence of artefacts have been successfully identified by the score based on the selected radiomics features. However, no correlation can be established regarding the impact of this change on the dose calculation.

Table 5.2.1: Random forest model evaluation.

Gamma pass-rate class	Precision	Sensitivity	F1-score	Support
Excellent	1.00	1.00	1.00	10
Good	0.86	0.81	0.83	37
insufficient	0.50	0.58	0.54	12
Accuracy : 0.80				

Table 5.2.2: Confusion matrix

	GPR "Excellent"	GPR "Good"	GPR "Insufficient"
GPR "Excellent"	1.0	0.0	0.0
GPR "Good"	0.0	0.81	0.19
GPR "Insufficient"	0.0	0.42	0.58

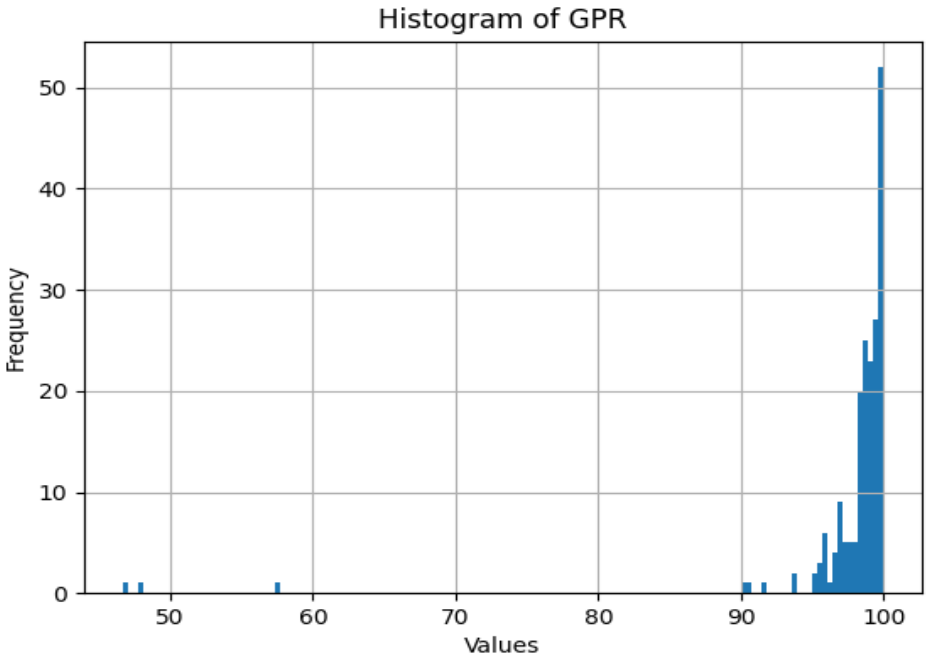

Figure 5.2.3: Histogram of the gamma pass rate (GPR) of the data used to train the Random Forest Classifier.

Table 5.2.3: Scores were obtained using the selected features and applied on the validation dataset. Experiment 1 refers to the data where a random HU variation has been applied, while experiment 2 refers to the presence of a spheric volume of error in the image.

CT score	Experiment 1	Experiment 2
100	73.91	86.96
100	67.39	86.96
100	71.73	86.96
100	91.30	86.96
100	82.61	80.43
100	60.87	89.13
100	78.26	93.48
100	63.04	84.78
100	73.91	84.78
100	78.26	86.96

Discussion

The main objective of this study was to explore the use of radiomics features to introduce a reliable and efficient tool for patient-specific sCT quality control in MRI-only RT. This approach enables daily assessment of synthetic CTs without requiring a simulation CT as a reference.

Probabilistic estimation of errors in sCT at a voxel level has been explored in previous studies. For instance, Van Harten et al. [27] proposed a method to obtain a voxel-wise uncertainty map by analysing the discrepancies between sCT volume reconstructions on different axes (axial, sagittal, and coronal) using a deep-learning model trained on 2D data for each axis. Deep-learning models also allow for the computation of epistemic (model-dependent) and aleatoric (data-dependent) uncertainties [28], [29]. Another method, introduced by Johansson et al.[30], involves estimating the probability of error in sCT generated from a Gaussian mixture model. These methods provide 3D maps of the probability of errors. However, it is important to note that these approaches are developed for specific models and provide estimates without generalisability. In contrast, the method presented in this paper aims to overcome these limitations and provide a more generalisable solution for error detection in sCT. Furthermore, our model enables the direct comparison of CT features, not solely based on voxel-wise intensity changes.

To select the most informative data, features were chosen based on their ability to impact the GPR within the entire pelvis. These features were mainly extracted using wavelet decomposition or a LoG filter with small kernel sizes applied to the original images. Wavelet transforms have proven to be a robust technique for extracting biomarkers in radiomics [31]. They have demonstrated efficiency in predicting the tumour type of early-stage lung nodules in CT[32], assessing the treatment response of gastric carcinoma to low-dose rate radiotherapy [33], detecting liver cirrhosis[34], and evaluating the neoadjuvant chemotherapy response in breast cancer using MRI [35]. However, since wavelet-derived features primarily capture local variations, they might be too restrictive to serve as the sole criteria for rejecting an sCT. Therefore, subjective image quality assessment remains necessary, as not having a good score in a specific location doesn't necessarily lead to significant consequences in dose calculation.

Another advantage of the proposed approach is its potential use as a reduced-reference metric when the patient's simulation CT is available, allowing for a direct comparison of feature values between the two images in the context of image-guided RT workflow, for example. This eliminates uncertainties associated with the intrapatient data registration step, required when using full reference metrics. In such cases, an acceptability threshold needs to be defined.

The method described can also be extended to sCT generated from cone-beam computed tomography (CBCT) [10], as its applicability is not limited to a specific imaging modality. CBCT is widely employed for patient positioning and monitoring during various stages of treatment

delivery. However, CBCT image reconstruction is prone to artifacts, so, the use of daily CBCT for online plan adaptation has been limited. By converting CBCT to CT, accurate dose computation could be achieved, thereby enhancing the quality of image-guided adaptive RT (IGART). Our protocol can be integrated into the IGART workflow, evaluating the accuracy of sCT for precise treatment delivery.

The major limitation of this study is the small size of the dataset and the limited representation of low GPR instances. The results are therefore not statistically robust enough to definitively draw conclusions about the effectiveness of this approach. Furthermore, the model developed in this study is specifically designed for detecting discrepancies within certain sCT generation methods and CT images from the same center. To enhance the generalizability of this approach, it would be beneficial to apply it to multi-center data. Adapting this methodology to a different monocenter dataset would necessitate repeating the entire procedure, considering factors such as varying field of view and image resolution, as these factors can impact the extracted features. While this method enables patient-specific sCT quality assurance uses GPR for clinically relevant feature selection, the selected features may be overly restrictive.

The key improvement of this approach lies in the robustness of the protocol. Indeed, the process of extracting and selecting radiomic features is very sensitive to the scanner and to the quality of the segmentation of the region of interest. Several repetitions with contour variations and tests on different datasets should be carried out to guarantee the robustness and reliability of the selected features.

It is crucial to explore this methodology on a larger dataset for both training and validation purposes, and to assess the correlation between detected anomalies and their impact on the dose distribution at different scales. Part of the future work will thus involve creating a more diverse validation set and applying the methodology per VOI.

Conclusion

The proposed radiomics-based workflow for patient-specific QA involves selecting quantitative image features to assess the similarity between a sCT and a reference CT cohort. This approach utilises optimised random-forest models, ensuring both performance and ease of implementation. The resulting score effectively highlights significant discrepancies without relying on a specific simulation CT as ground truth, enabling rapid assessment of new sCT images. Radiomics analysis proves to be an efficient tool for quality control of patient-specific sCTs in an RT workflow. However, further investigation is needed prior to use this approach in clinical settings. Nevertheless, this workflow may serve as the initial step in a QA procedure aimed at facilitating safe MRI-only RT, given its efficiency in detecting artefacts and shift of HU within specific organs.

Compliance with ethical standards

Ethics approval for the study protocol was obtained from Hunter New England Human Research Ethics Committee (reference number 05/09/14/3.06).All patients provided their written informed consent.

Acknowledgements

This work was partially supported by region Bretagne (France) through ARED scholarship program and a PhD scholarship Grant from e-health Research Centre-CSIRO (Australia). The authors have no relevant financial or non-financial interest to disclose.

References

[1] M. B. Barton et al., “Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012,” Radiotherapy and Oncology, vol. 112, no. 1, pp. 140–144, Jul. 2014, doi: 10.1016/j.radonc.2014.03.024.

[2] B. Hentschel, W. Oehler, D. Straußß, A. Ulrich, and A. Malich, “Definition of the CTV prostate in CT and MRI by using CT-MRI image fusion in IMRT planning for prostate cancer,” Strahlentherapie und Onkologie, vol. 187, no. 3, pp. 183–190, Mar. 2011, doi: 10.1007/s00066-010-2179-1.

[3] A. M. E. Bruynzeel et al., “A Prospective Single-Arm Phase 2 Study of Stereotactic Magnetic Resonance Guided Adaptive Radiation Therapy for Prostate Cancer: Early Toxicity Results,” Int J Radiat Oncol Biol Phys, vol. 105, no. 5, pp. 1086–1094, Dec. 2019, doi: 10.1016/j.ijrobp.2019.08.007.

[4] T. Nyholm, M. Nyberg, M. G. Karlsson, and M. Karlsson, “Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments,” Radiation Oncology, vol. 4, no. 1, Nov. 2009, doi: 10.1186/1748-717X-4-54.

[5] J. M. Edmund and T. Nyholm, “A review of substitute CT generation for MRI-only radiation therapy,” Radiation Oncology, vol. 12, no. 1, p. 28, Dec. 2017, doi: 10.1186/s13014-016-0747-y.

[6] E. Johnstone et al., “Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy,” International Journal of Radiation Oncology Biology Physics, vol. 100, no. 1. 2018. doi: 10.1016/j.ijrobp.2017.08.043.

[7] M. Boulanger et al., “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” Physica Medica, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021. doi: 10.1016/j.ejmp.2021.07.027.

[8] G. Wang, Y. Zhang, X. Ye, and X. Mou, “Image quality assessment,” in Machine Learning for Tomographic Imaging, in 2053-2563. IOP Publishing, 2019, pp. 9–1 to 9–30. doi: 10.1088/978-0-7503-2216-4ch9.

[9] J. Dowling et al., Image synthesis for MRI-only radiotherapy treatment planning. 2022. doi: 10.1016/B978-0-12-824349-7.00027-X.

[10] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, “Deep learning based synthetic-CT generation in radiotherapy and PET: A review,” Med Phys, 2021, doi: 10.1002/mp.15150.

[11] H. Chourak et al., “Local quality assessment of patient specific synthetic-CT via voxel-wise analysis,” in 2022 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/DICTA56598.2022.10034622.

[12] S. Fan et al., “CT-Based Radiomics Signature: A Potential Biomarker for Predicting Postoperative Recurrence Risk in Stage II Colorectal Cancer,” Front Oncol, vol. 11, Mar. 2021, doi: 10.3389/fonc.2021.644933.

[13] Y. M. Zhang, G. Z. Gong, Q. T. Qiu, Y. W. Han, H. M. Lu, and Y. Yin, “Radiomics for Diagnosis and Radiotherapy of Nasopharyngeal Carcinoma,” Frontiers in Oncology, vol. 11. 2022. doi: 10.3389/fonc.2021.767134.

[14] H. Huang, F. Xu, Q. Chen, H. Hu, F. Qi, and J. Zhao, “The value of CT-based radiomics nomogram in differential diagnosis of different histological types of gastric cancer,” Phys Eng Sci Med, vol. 45, no. 4, pp. 1063–1071, Dec. 2022, doi: 10.1007/s13246-022-01170-y.

[15] C. Ma et al., “The structural similarity index for IMRT quality assurance: radiomics-based error classification,” Med Phys, vol. 48, no. 1, pp. 80–93, Jan. 2021, doi: 10.1002/mp.14559.

[16] L. S. Wootton, M. J. Nyflot, W. A. Chaovalitwongse, and E. Ford, “Error Detection in Intensity-Modulated Radiation Therapy Quality Assurance Using Radiomic Analysis of Gamma Distributions,” Int J Radiat Oncol Biol Phys, vol. 102, no. 1, pp. 219–228, Sep. 2018, doi: 10.1016/j.ijrobp.2018.05.033.

[17] J. A. Dowling 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, vol. 93, no. 5, pp. 1144–1153, 2015, doi: 10.1016/j.ijrobp.2015.08.045.

[18] A. Largent et al., “Pseudo-CT Generation for MRI-Only Radiation Therapy Treatment Planning: Comparison Among Patch-Based, Atlas-Based, and Bulk Density Methods,” Int J Radiat Oncol Biol Phys, vol. 103, no. 2, pp. 479–490, 2019, doi: 10.1016/j.ijrobp.2018.10.002.

[19] A. Largent et al., “Comparison of Deep Learning-Based and Patch-Based Methods for Pseudo-CT Generation in MRI-Based Prostate Dose Planning,” Int J Radiat Oncol Biol Phys, vol. 105, no. 5, pp. 1137–1150, 2019, doi: 10.1016/j.ijrobp.2019.08.049.

[20] H. P. Wieser et al., “Development of the open-source dose calculation and optimization toolkit matRad,” Med Phys, vol. 44, no. 6, pp. 2556–2568, Jun. 2017, doi: 10.1002/mp.12251.

[21] D. Rivest-Hénault, N. Dowson, P. B. Greer, J. Fripp, and J. A. Dowling, “Robust inverse-consistent affine CT-MR registration in MRI-assisted and MRI-alone prostate radiation therapy,” Med Image Anal, vol. 23, no. 1, pp. 56–69, 2015, doi: 10.1016/j.media.2015.04.014.

[22] D. Rivest-Hénault, P. Greer, jurgen frip, and J. Dowling, Structure-Guided Nonrigid Registration of CT–MR Pelvis Scans with Large Deformations in MR-Based Image Guided Radiation Therapy. 2014. doi: 10.1007/978-3-319-05666-1_9.

[23] X. Han, “MR-based synthetic CT generation using a deep convolutional neural network method:,” Med Phys, vol. 44, no. 4, pp. 1408–1419, 2017, doi: 10.1002/mp.12155.

[24] J. J. M. Van Griethuysen et al., “Computational radiomics system to decode the radiographic phenotype,” Cancer Res, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.

[25] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[26] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” 2012. [Online]. Available: http://scikit-learn.sourceforge.net.

[27] L. D. van Harten, J. M. Wolterink, J. J. C. Verhoeff, and I. Išgum, “Automatic online quality control of synthetic CTs,” SPIE-Intl Soc Optical Eng, Mar. 2020, p. 57. doi: 10.1117/12.2549286.

[28] M. Abdar et al., “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” 2020, [Online]. Available: http://arxiv.org/abs/2011.06225

[29] C. A. T. van den Berg and E. F. Meliadò, “Uncertainty Assessment for Deep Learning Radiotherapy Applications,” Seminars in Radiation Oncology. W.B. Saunders, Oct. 01, 2022. doi: 10.1016/j.semradonc.2022.06.001.

[30] A. Johansson and M. Karlsson, “Voxel-wise uncertainty in CT substitute derived from MRI,” vol. 39, no. June, pp. 3283–3290, 2012.

[31] F. Prinzi, C. Militello, V. Conti, and S. Vitabile, “Impact of Wavelet Kernels on Predictive Capability of Radiomic Features: A Case Study on COVID-19 Chest X-ray Images,” J Imaging, vol. 9, no. 2, Feb. 2023, doi: 10.3390/jimaging9020032.

[32] R. Jing et al., “A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules,” Sci Rep, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-01470-5.

[33] Z. Hou et al., “Radiomic analysis using contrast-enhanced CT: Predict treatment response to pulsed low dose rate radiotherapy in gastric carcinoma with abdominal cavity metastasis,” Quant Imaging Med Surg, vol. 8, no. 4, pp. 410–420, May 2018, doi: 10.21037/qims.2018.05.01.

[34] K. Kotowski et al., “Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features,” Comput Biol Med, vol. 152, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106378.

[35] J. Zhou et al., “Predicting the response to neoadjuvant chemotherapy for breast cancer: Wavelet transforming radiomics in MRI,” BMC Cancer, vol. 20, no. 1, Feb. 2020, doi: 10.1186/s12885-020-6523-2.

Chapter 6: AI-based synthetic CT quality control guidelines for accurate MRI-only radiation therapy

This chapter presents methods and metrics used for assessing sCT, along with recommendations for quality control steps to integrate daily use of sCT into the clinical routine. This chapter is the beginning of a manuscript draft. It is still in its early stages and requires discussion with experts in image processing, medical physicists, and clinicians to establish a consensus ensuring the accuracy of MRI-only RT.

Introduction

Radiotherapy (RT) is a highly effective treatment option for cancer patients, with approximately 40% of patients undergoing RT during their cancer care[1]. In the conventional RT workflow, CT imaging is used for treatment planning and dose calculation, while MRI provides superior soft tissue contrast for accurate tumour identification during planning and treatment delivery. MRI also offer better soft tissue visualization and enable more precise delineation of organs at risk. MR-only RT workflow has thus emerged[2]. It can be implemented using a standard linear accelerator (LINAC) or an MR-LINAC—a device combining an MRI and a LINAC in the same room. In the standard workflow, there are challenges related to dose calculation based on MRI and the registration between planning MRI and daily images (such as 3D CBCT or 2D images). The MR-LINAC offers the advantage of daily adaptation of the initial treatment plan based on the day's anatomy captured by MR images. However, a major drawback of MRI for MR-only RT, with or without MR-LINAC, is its inability to provide information on tissue densities, which is crucial for accurate dose calculation. To overcome this limitation, several methods have been developed to generate synthetic CTs (sCTs) that allow for the use of MRI in treatment planning. These approaches, mainly based on deep-learning models (DLMs)[3] nowadays, have demonstrated high accuracy and robustness.

Despite the effectiveness of DLMs in predicting Hounsfield Units (HU) values from MRI sequences, challenges remain in evaluating the image quality of the resulting sCTs[4]. The current literature assesses sCT quality by comparing them to their corresponding planning CTs. However, with the adoption of MRI-only RT treatment planning, these CT scans will be no longer available. Alternative methods for patient-specific quality assurance (QA) without CT have been proposed, such as using cone-beam CT (CBCT) to evaluate patient-specific sCTs generated from MRI[5]–[7]. However, incorporating CBCT acquisition in adaptive RT with MRI-LINAC is unlikely. To our knowledge, there are no widely accepted practices or standards for ongoing QA of sCTs derived from MRI. Therefore, establishing a standardized quality control protocol for patient-specific sCTs obtained on a daily basis becomes imperative.

The aim of this chapter was to provide guidelines for sCT QA, ensuring reliable treatment delivery in MRI-only RT. Qualitative and quantitative approaches, with and without reference CT, for daily patient specific QA have been considered.

Qualitative evaluation

Different scenario can be encountered in RT (Figure 6.1), for both the planning phase and the treatment delivery. For accurate dose delivery, the treatment plan is adjusted to the anatomy of the day by aligning the treatment plan to a CBCT or MRI acquired prior to delivering the treatment. The use of sCT during each phase allows for not having to do registration and thus should provide a more precise targeting of the tumour volume.

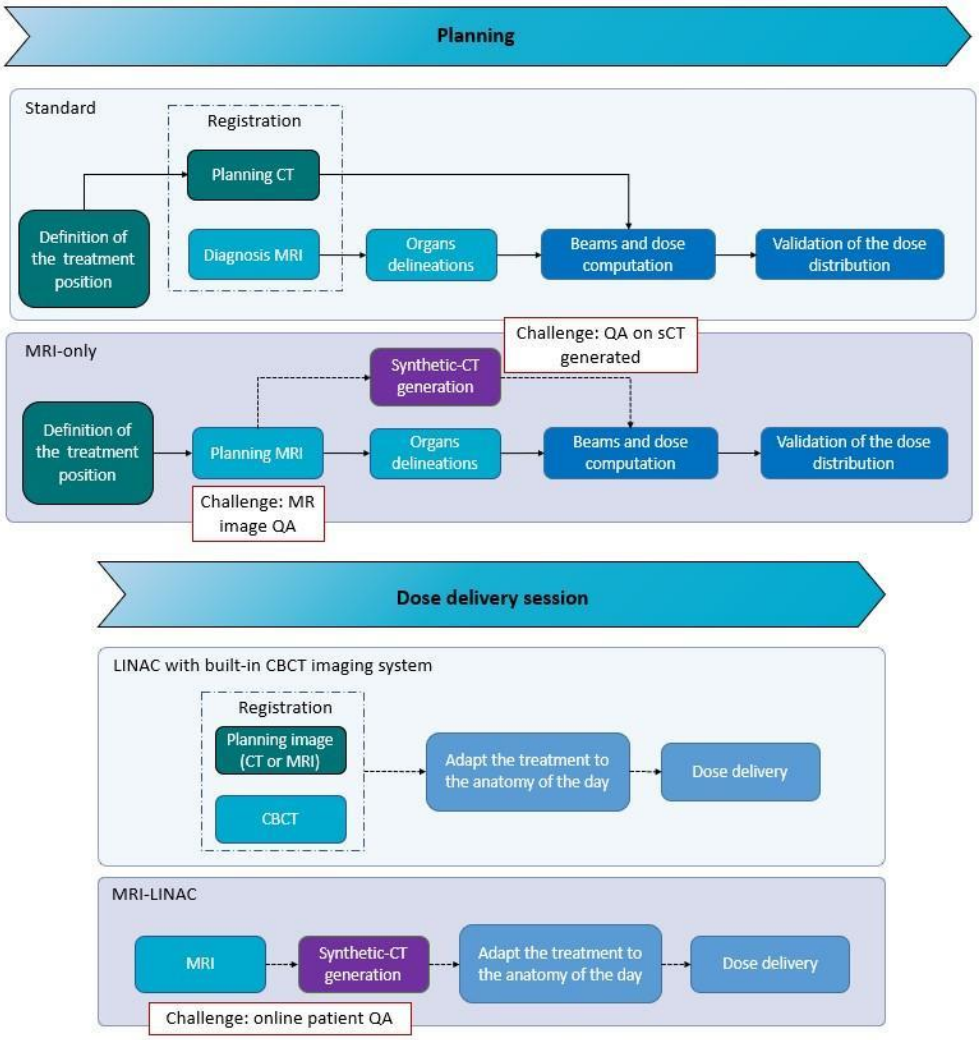

Figure 6.1: Standard and MRI-online adaptive radiation therapy (RT) workflows. The planning phase in standard RT involves the acquisition of a planning CT for dose calculation, and a diagnosis MRI for accurate tumour and organs at risk delineation. Both images are then non-rigidly registered before calculating the dose. In MRI-only RT planning, a planning MRI is acquired. This image is then used to generate a synthetic-CT (sCT), allowing for dose calculation. Traditionally in ART, the planning CT is registered to a CBCT to adapt the treatment to the anatomy of the day. With the MR-LINAC, a sCT can be generated from the MRI and the treatment plan can be transposed on this image.

But, the sCT is an image generated, and not acquired, it is thus important to check its accuracy each time prior to use it.

The type of sequence and the overall quality of the MRI used to generate the sCT will have an impact on the resulting image. The quality control process thus begins with the assessment of the MRI, then qualitative and quantitative metrics can be used to assess the sCT. In figure 6.2, we propose a decision tree from the MRI acquisition to the treatment delivery.

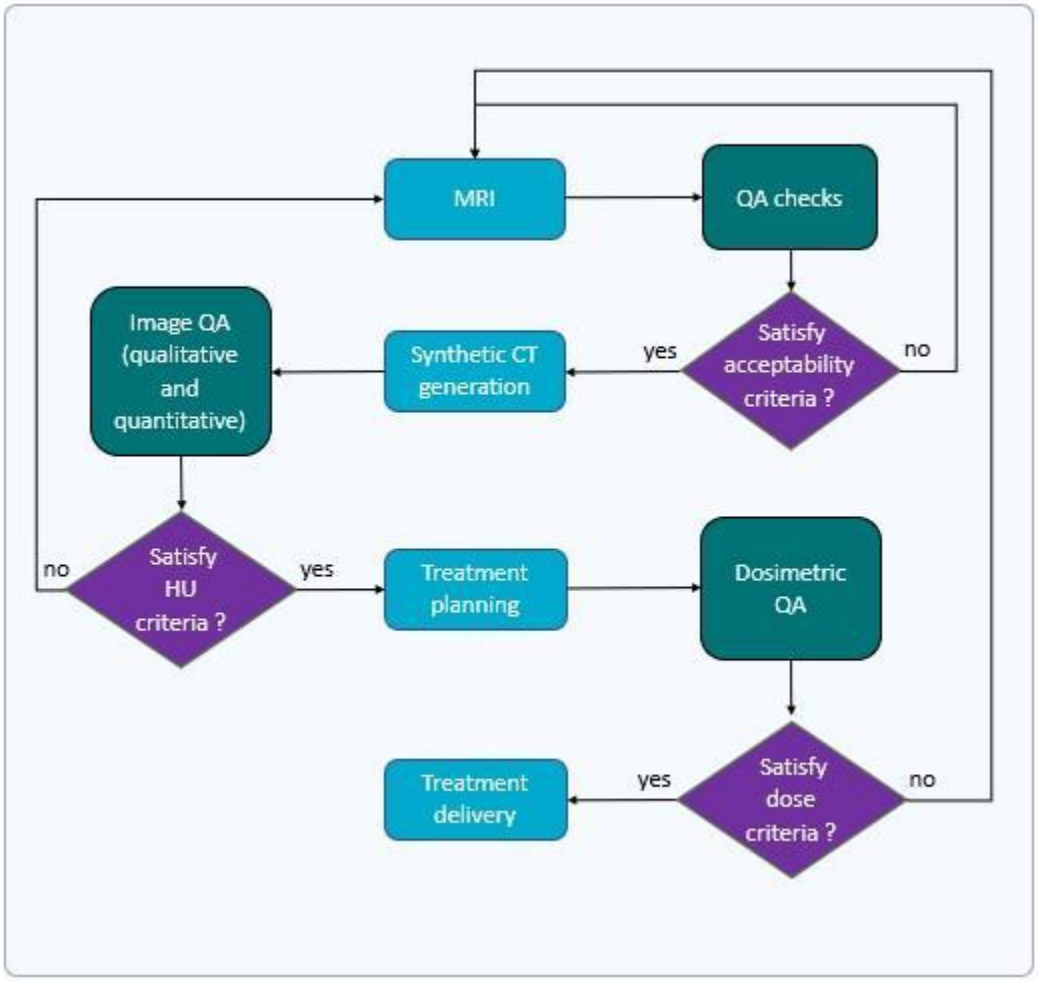

Figure 6.2: Quality assurance (QA) steps for MRI-based treatment planning

MR images QA

Automatic verification of acquisition parameters by reading the DICOM files will allow for detection of errors in: sequence type acquisition, image orientation, application of gradient distortion correction, spatial resolution (pixel size and slice thickness) and detection of gap. The QA of the MRI need to be completed by the overall visual inspection by radiotherapist and medical physicists.

Also, DLMs are trained using specific MRI sequences. When generating daily sCTs, it's essential to use the same sequence for optimal accuracy.
Table 6.1 presents the key elements to check in order to determine if the MRI scan is suitable for sCT generation.

Table 6.1: Quality assurance checks of MRI for RT dose calculation.
LFOV stands for large field of view, SNR for signal to noise ratio, OAR for organ at risk. This checklist has been established according to the suggestions made by Speight et al.[1] and Dowling et al.[8]. Some of these elements can be checked with the DICOM tags. These elements are parts of QA checks used in the HIPSTER[9] and NINJA[10] clinical trials.

Check list	Details
Distortion correction	3D distortion correction has been activated for the LFOV scan. Check distortion corrections for other scans.
Voxel spacing and gap	Ideally like CT slice thickness. Should be less than 2mm and near isotropic. There shouldn't be any slice gap.
3D versus 2D acquisition	3D acquisitions should be used to provide high resolution isotropic imaging. 2D acquisition may be used if it offers good soft tissue contrast.
FOV position	The centre of the FOV should be positioned over the anatomy of interest (FOV and magnet isocentre are aligned for better geometric fidelity)
Gradient non-linearity correction	2D distortion correction should be applied as a minimum. 3D distortion correction should be applied if available. Some systems allow 3D distortion correction to be applied to 2D multi-slice datasets.
Contrast agent	May be used to highlight targets or OARs. The decision whether to use contrast agent should follow discussion with a radiologist.
Fiducial marker visibility	Verify that the fiducial markers are clearly visible on MRI and distinguishable from calcifications.
Size	The skin must be in the FOV and the image must be large enough for insertion of a virtual couch (required for treatment planning system (TPS))
MRI sequence	The sequence must be the same than the data used to train the DLM.
Anatomy	The image is the correct anatomy, and the orientation is correct

Intensity distribution	The intensity distribution is within the same range of the intensities of MRI used to train the DLM.

sCT QA

Once the quality of the acquired MRI scan is deemed suitable for sCT generation and image guidance, the image can be produced and assessed.

Table 6.2 lists the elements to check before using a sCT for treatment planning. If the image respects all the conditions below, complementary quantitative QA can be applied.

Table 6.2: sCT qualitative assessment check list. (Dowling et al.[8]). These elements are parts of QA checks used in the HIPSTER[9] and NINJA[10] clinical trials.

Check list	Details
Image transfer	Confirm that the correct sCT has been assigned to the patient and confirm that sCT is correctly oriented.
Image integrity	Visually inspect the entire sCT volume for any missing tissue or major artefacts. These differences may not affect dose calculation but should be noted.
Field of view	Ensure that the sCT has sufficient field-of-view to cover all relevant anatomy, skin contour, and sufficient extension for dose calculation.
Body contour	Check if the sCT body contour match the MRI body contour (require segmentation).
HU to electron density conversion	Check that the correct calibration curve has been applied to the sCT.

Quantitative evaluation of sCT

Most of the methods used for the evaluation of sCT are also used in general image QA. Figure 6.3 shows the methods employed for sCT QA by order of complexity. These metrics can be classified in 3 categories:

- Full reference, i.e compare the generated image to a ground truth. For voxel-wise comparison, the two images must be registered.
- Reduced reference. In this case, only features of a ground truth image are compared.
- No reference. These are used when no ground truth is available for direct comparison.

The choice of metrics will then depend on the availability of a planning CT (ground truth).

With reference CT

sCT can be generated to adapt the treatment to the anatomy of the day, but the acquisition of a CT for the planning phase may remain. In this scenario, the planning CT can be used as a reference to assess the quality of the sCT.

The most commonly used in the literature are the mean absolute error (MAE), the mean error (ME), the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM)[11]. The mean absolute percent error (MAPE) is the relative error, indicating the precision of the prediction regarding the reference. The absolute difference, absolute percent difference and the difference of intensities between the sCT and the planning CT can be computed either within a contour (body contour or organs), or at a voxel level resulting respectively in absolute error (AE), absolute percent error (APE), and error (E) maps. The PSNR is an indication of the distorting noise in the generated image, while the SSIM is a measure of the difference in luminance, contrast, and structure. The visual information fidelity (VIF) is an alternative measure of the global similarity of the images. All these metrics give complementary information. However, the sCT and the planning CT need to be registered to compute these metrics.

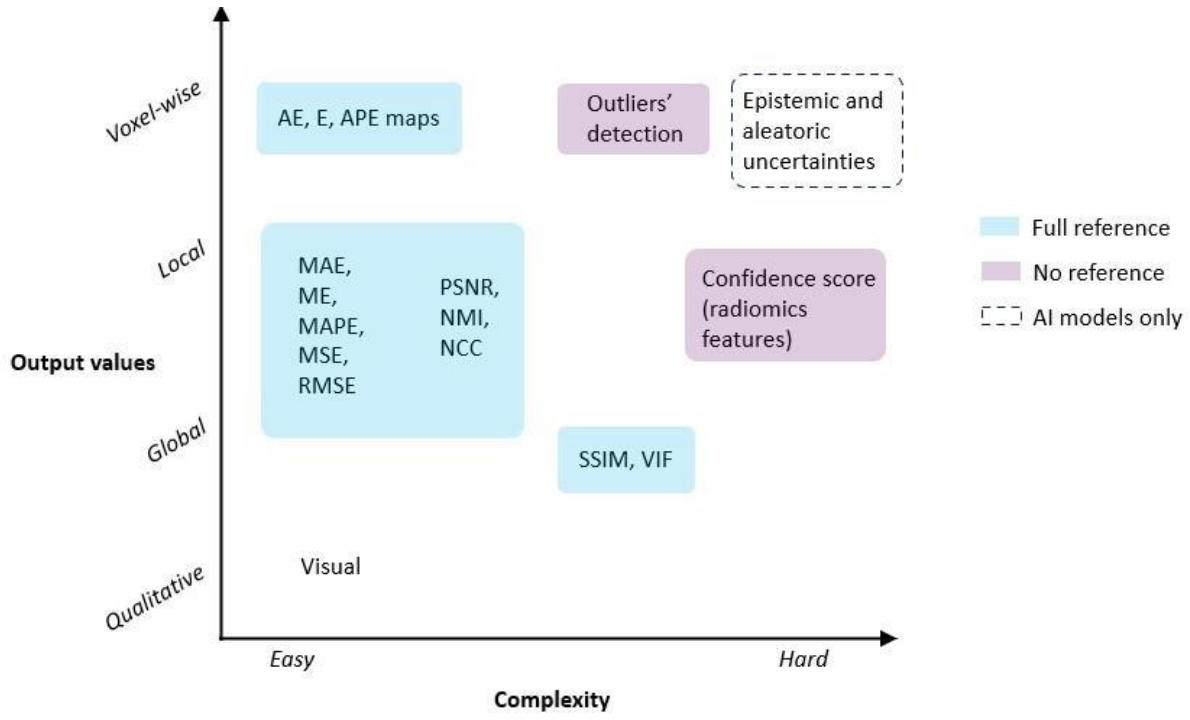

Figure 6.3: Methods for image quality assessment of synthetic CT by order of complexity. This diagram presents metrics comparing directly the sCT to a planning CT (full reference metrics, in blue) and methods to assess the sCT when no planning CT is available (no reference, in purple). The no-reference metrics in this context are based on statistics derived from retrospective CT scans. The voxel intensities (outliers' detection)

or key radiomics features (confidence score) are compared to those obtained from a reference CT cohort, representing the range of expected values.
These quantitative metrics can be computed at a voxel level, organ-wise (local) or within the whole body contour (global). Aleatoric and epistemic allow for the probability of having a misprediction of Hounsfield Unit value for each voxel and are specific to deep-learning, or Artificial Intelligence (AI) – based models. The acronyms are details at the end of the chapter.

Without reference CT

In the scenario of MRI-only RT, i.e when no reference CT is available, the generated image can be assessed by statistical analysis of HU values compared to a cohort of CT. A voxel-wise outliers' detection method was previously proposed by Chourak et al.[12]. This method relies on the registration of the sCT to an atlas of reference. A confidence score according to the values of specific image features can also be computed, as proposed in the 2nd section of chapter 5.

DLM allows for fast sCT generation. One approach to reducing uncertainties is generating multiple sCTs from the same MRI scan and measuring the voxel-level differences in Hounsfield Units (HU). However, if the error stems from a bias in the model, the same misprediction could be present in all the generated images, making it less discernible through multiple sCT generations alone. The idea is thus to compare images obtained from two different methods[13]. In a clinical setting, the sCT derived from a commercial device can be compared to an in-house method that has demonstrated its effectiveness using center-specific data.

Predictive uncertainty maps

Epistemic and aleatoric uncertainties are specific to deep-learning model[14]–[16].

Epistemic uncertainty could be reduced with a better dataset and refers to the uncertainty of a model. It can be divided as :

- Structural uncertainty : related to the architecture of the model, i.e. is it the best choice for the task?
- Parametric uncertainty : related to the estimation of the parameters of the model.

The measure of these uncertainties can be interpreted as the standard deviation of the output according to variations of the model. Monte Carlo dropout, which consists of deactivating random neurons in the network during the learning phase, is one way to estimate the epistemic uncertainty.

Aleatoric, also known as data or intrinsic uncertainty, refers to the inherent property of the data distribution and is thus irreducible[17]. It can be divided as :

- Homoscedastic uncertainty : constant across all input data.
- Heteroscedastic uncertainty : varies across the input data.

Using probabilistic models, i.e. models incorporating probabilistic layers, like Bayesian Neural Networks (BNNs), enables to capture both aleatoric and epistemic uncertainties.

Another effective approach is the use of ensemble models[18], where multiple models with the same architecture are trained with different initialisations or on diverse data subsets. The variance in predictions across the ensemble provides an indication of aleatoric uncertainty. Computing these uncertainties results in 3D maps, providing insight into the probability of an accurately predicted HU value at a voxel level. This gives an indication of the areas where the generation method may lack accuracy.

Recommendations

Table 6.3: Threshold of acceptability for quantitative evaluation of sCT generated from MRI with deep-learning based methods.

	Methods	Recommendations
With reference CT	MAE per organ and within the body contour	(Particular case for air cavities where the presence of air must be compared with the MRI to correspond to the anatomy of the day) Within 10% of accuracy[19].
	Geometric integrity	≤ 1mm for structure within 10cm radial distance of isocentre, ≤ 2mm if outside of this perimeter[20].
Without reference CT	Statistics on retrospective CT (voxel-wise or confidence score)	Within the 95% confidence interval or within the minimal and maximal values if the sample size of the reference CT cohort is < 100.
	Double sCT computation	Absolute HU difference < 50 HU in each non-bony structure.

Discussion

This paper proposes guidelines for sCT QA to ensure reliable treatment delivery in MRI-only RT. Both qualitative and quantitative approaches were considered. Full-reference metrics, which involve comparing the sCT to a planning CT, as well as no-reference methods, have been presented to provide QA solutions for different RT scenarios involving sCT from MRI. These methods allow for the assessment of the generated image at different scales, from the overall images to the organs and voxel level. Ideally, automatic tools for sCT QA in MRI-LINAC workflow (online ART) should be integrated in the device. Low computation time and easy interpretation of the results are key elements for fast and efficient quality control.

The results of images generated from DLM are highly dependent on the training database. For non-'in-house' models, harmonisation of scanners and pre-processing is beneficial but may not be sufficient. Multicentre studies have shown the positive impact of mixing data from different hospitals on the results, compared to training on a dataset from one centre and testing on another[21], allowing for more generalisable models. This also allows training a model to produce sCT from MRI of different field strengths (0.35T to 3T, for example). Additionally, training the model with a large variability in the dataset provides a better chance to produce good quality images for patients with low representation in the population.

For “in-house” models trained on monocentre data, the MRI used to generate the sCT must be consistent with the type of sequences and the field of view of the MRI used to train the model. If change appears in practice, the model will have to be re-evaluated. For commercial software, a test on a known image dataset should be conducted after each update. Changes in the population over time might also cause the model to degrade[2]. In future work, setting up a protocol to predict locally the potential errors in the sCT by assessing the MRI will permit to save time and resources.

This chapter presents suggestions for sCT QA focusing on the accuracy of HU prediction. However, the finality is the RT dose calculation and the treatment delivery. Once a sCT is deemed of sufficient quality, a dosimetric QA is necessary. In case of the existence of a planning CT, the dose calculation of the day can be compared to the initial DVH. Dosimetric QA if no ground truth is available is a main concern and this need to be addressed. In the International Commission on Radiation Units and Measurements report published in 2022, Keall et al.[19] suggest respecting an accuracy of 10% in CT number as 20% variation in HU may result in a systematic dose error of 1.5%[22]. The treatment to determine this threshold have been determined using a 6 MV photon beam, but proton therapy is more sensitive to HU variation. Therefore, these recommendations must be adjusted according to the prescribed treatment and anatomical location.

The location of the errors needs also to be identified, as the influence of an error is correlated to its interaction within the beams (area where the dose is homogenous or strong gradient). Impact of HU errors on dose calculation should also be investigated to better determine threshold of acceptability.

Due to the complexity of the treatment planning and delivery workflow in RT, it is important to note that not only the quality of images should be assessed; the contours and the dose calculation need to be controlled prior to treatment delivery. The whole quality control pipeline needs to be applied each time an sCT is used, because, unlike images acquired from a device, the sCT is an image resulting from a computational process and thus errors in its generation can occur at any time.

[2] https://www.fda.gov/media/160125/download

Conclusion

Practical QA for clinical use of sCT in MRI-only RT has been proposed in this chapter, taking into consideration the existence, or not, of a planning CT. These QA criteria aim to be implemented in clinics and future clinical trials.

Metrics acronyms

AE	Absolute error
APE	Absolute percent error
DVH	Dose-volume histogram
E	Error
MAE	Mean absolute error
MAPE	Mean absolute percent error
ME	Mean error
MSE	Mean square error
NCC	Normalized cross-correlation
NMI	Normalized mutual information
PSNR	Peak signal-to-noise ratio
RMSE	Root mean square error
SSIM	Structural similarity metric
VIF	Visual information fidelity

References

[1] R. Speight *et al.*, “IPEM topical report: Guidance on the use of MRI for external beam radiotherapy treatment planning∗,” *Phys. Med. Biol.*, vol. 66, no. 5, Mar. 2021, doi: 10.1088/1361-6560/abdc30.

[2] M. A. Schmidt and G. S. Payne, “Radiotherapy planning using MRI,” *Physics in Medicine and Biology*, vol. 60, no. 22. Institute of Physics Publishing, pp. R323–R361, Oct. 28, 2015, doi: 10.1088/0031-9155/60/22/R323.

[3] M. Boulanger *et al.*, “Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review,” *Physica Medica*, vol. 89. Associazione Italiana di Fisica Medica, pp. 265–281, Sep. 01, 2021, doi: 10.1016/j.ejmp.2021.07.027.

[4] M. Claessens *et al.*, “Quality Assurance for AI-Based Applications in Radiation Therapy,” *Seminars in Radiation Oncology*, vol. 32, no. 4. W.B. Saunders, pp. 421–431, Oct. 01, 2022, doi: 10.1016/j.semradonc.2022.06.011.

[5] E. Palmér, E. Persson, P. Ambolt, C. Gustafsson, A. Gunnlaugsson, and L. E. Olsson,

"Cone beam CT for QA of synthetic CT in MRI only for prostate patients," *J. Appl. Clin. Med. Phys.*, vol. 19, no. 6, pp. 44–52, Nov. 2018, doi: 10.1002/acm2.12429.

[6] J. J. Wyatt, R. A. Pearson, C. P. Walker, R. L. Brooks, K. Pilling, and H. M. McCallum, "Cone beam computed tomography for dose calculation quality assurance for magnetic resonance-only radiotherapy," *Phys. Imaging Radiat. Oncol.*, vol. 17, pp. 71–76, Jan. 2021, doi: 10.1016/j.phro.2021.01.005.

[7] S. Irmak, L. Zimmermann, D. Georg, P. Kuess, and W. Lechner, "Cone beam CT based validation of neural network generated synthetic CTs for radiotherapy in the head region," *Med. Phys.*, vol. 48, no. 8, pp. 4560–4571, Aug. 2021, doi: 10.1002/mp.14987.

[8] J. Dowling *et al.*, *Image synthesis for MRI-only radiotherapy treatment planning*. 2022.

[9] P. Greer *et al.*, "A multi-center prospective study for implementation of an MRI-only prostate treatment planning workflow," *Front. Oncol.*, vol. 9, no. AUG, 2019, doi: 10.3389/fonc.2019.00826.

[10] "TROG 18.01 The NINJA Clinical Trial: Novel Integration of New prostate radiation schedules with adJuvant Androgen deprivation."

[11] G. Wang, Y. Zhang, X. Ye, and X. Mou, "Image quality assessment," in *Machine Learning for Tomographic Imaging*, IOP Publishing, 2019, pp. 9–1 to 9–30.

[12] H. Chourak *et al.*, "Local quality assessment of patient specific synthetic-CT via voxel-wise analysis," 2022, doi: 10.1109/DICTA56598.2022.10034622.

[13] R. Dal Bello *et al.*, "Patient-specific quality assurance strategies for synthetic computed tomography in magnetic resonance-only radiotherapy of the abdomen," *Phys. Imaging Radiat. Oncol.*, vol. 27, p. 100464, Jul. 2023, doi: 10.1016/j.phro.2023.100464.

[14] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021, doi: 10.1007/s10994-021-05946-3.

[15] R. Barbano, S. Arridge, B. Jin, and R. Tanno, "Uncertainty quantification in medical image synthesis," in *Biomedical Image Synthesis and Simulation: Methods and Applications*, Elsevier, 2022, pp. 601–641.

[16] C. A. T. van den Berg and E. F. Meliadò, "Uncertainty Assessment for Deep Learning Radiotherapy Applications," *Seminars in Radiation Oncology*. W.B. Saunders, Oct. 01, 2022, doi: 10.1016/j.semradonc.2022.06.001.

[17] M. Abdar *et al.*, "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges," *Mach. Learn.*, vol. 11073 LNCS, no. 3, p. 116, 2019, doi: 10.1098/rsta.2015.0203.

[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[19] P. J. Keall *et al.*, "ICRU REPORT 97: MRI-Guided Radiation Therapy Using MRI-Linear

Accelerators," *J. ICRU*, vol. 22, no. 1, pp. 1–100, Dec. 2022, doi: 10.1177/14736691221141950.

[20] L. N. Baldwin, K. Wachowicz, S. D. Thomas, R. Rivest, and B. Gino Fallone, "Characterization, prediction, and correction of geometric distortion in 3 T MR images," *Med. Phys.*, vol. 34, no. 2, pp. 388–399, 2007, doi: 10.1118/1.2402331.

[21] B. Texier *et al.*, "Computed tomography synthesis from magnetic resonance imaging using cycle Generative Adversarial Networks with multicenter learning," *Phys. Imaging Radiat. Oncol.*, vol. 28, Oct. 2023, doi: 10.1016/j.phro.2023.100511.

[22] B. Zurl, R. Tiefling, P. Winkler, P. Kindl, and K. S. Kapp, "Hounsfield units variations: Impact on CT-density based conversion tables and their effects on dose distribution," *Strahlentherapie und Onkol.*, vol. 190, no. 1, pp. 88–93, Jan. 2014, doi: 10.1007/s00066-013-0464-5.

Conclusion: overview, contributions, limitations, and perspectives

Overview

The generation of sCT from MRI is an area of significant interest due to its potential to streamline patient care and enhance the accuracy of treatment delivery in EBRT. Recent advancements in deep-learning models have demonstrated promising precision and robustness in sCT generation. However, it should be noted that these approaches have been trained and evaluated on monocentre datasets of varying sizes. Therefore, their effectiveness and consistency on MRIs from different manufacturers and across the diverse range of anatomies encountered in clinical practice remain uncertain.

Existing literature typically presents results based on body contours or organ-wise evaluations, where an assessment of the overall quality of the generated sCTs is provided. However, this lacks information about the specific locations of mispredictions. In this thesis, a methodology has been proposed to identify voxel-level locations, encompassing both image and dose calculations, where the methods tend to be less accurate. To address the impact of localized HU errors on dose calculation, sensitivity analysis based on the Morris screening method is employed to determine whether a volume of error would have an impact on the dose at the isocentre. This analysis can also assist in establishing acceptable thresholds in terms of size or HU discrepancies based on their respective locations.

Once a generation method is deemed robust and accurate enough to be integrated in a clinical workflow, the assessment of patient-specific sCTs becomes a crucial aspect. However, in the existing literature, the metrics used for assessment either require a planning CT as a reference (full reference metrics) or involve the use of a CBCT, which is not applicable in an MRI-only radiation therapy workflow. To address this challenge, an atlas based approach is proposed using a cohort of CT scans to assist with identifying outliers at a voxel level in daily sCTs. Alternatively, radiomics features can be computed on a cohort of CT scans, and these values can be used to assign an acceptability score to a new sCT.

While an sCT does not need to be perfect in terms of image quality, it must possess sufficient quality to enable accurate dose calculation. Therefore, the selection of features correlated to a dosimetric endpoint, such as the gamma pass-rate, holds significance in the context of EBRT.

Main contributions and limitations

This thesis proposes a methodology to identify the limitations of sCT generation methods through statistical evaluation, aiming to provide a better understanding of their strengths and limitations. To achieve this, the patient cohort data has been brought into the same

anatomical space using a robust non-rigid registration method. The suggested hybrid registration approach utilizes intensity and contour-based methods to combine the MRI images to the structural description of its delineated organs, resulting in improved alignment of soft tissues, contours, and inner structures. Statistical analysis is then conducted on both images and doses, to subsequently generate 3D error maps that highlight regions where mispredictions of HU tend to be significant. It is important to note that this methodology has a major limitation: the bias induced by the interpatient registration process, as well as the intra-patient registration process (where CT and MRI are registered for each patient before generating the sCT).

The sensitivity analysis presented in this study follows the Morris screening method. It highlights the structures in which a change in HU values will have the most significant impact on the dose in the centre of the prostate. Specifically, the bones and the soft tissues (excluding the bladder, rectum, and prostate) were found to have the highest impact. This can be attributed to the number of beams that pass through these structures.

By assessing the impact of size, location, and changes in HU of an artifact on the dose in the target volume, the proposed methodology allows us to conclude that the distance to the isocentre is the least influential factor. The impact of changes in intensity shows the most variability across the patient cohort and is equally significant to the size factor. These findings provide valuable insights for determining thresholds of acceptability for uncertainties in HU. However, it is important to note that the approach should be adapted based on factors specific to each sCT generation method, such as location, error intensity rank, size, or any other relevant factor for this method. It is essential to interpret these conclusions with caution as they are specific to the dose calculation algorithm, the number of beams crossed by the volumes and the amount of dose delivered in each of them. In this study, we assessed the effect of errors on IMRT dose plans. The results cannot be generalized to other treatment techniques such as VMAT, SBRT, or proton therapy. Furthermore, another limitation of this study is that we focused solely on the impact in the centre of the target volume. Changes in HU may also have consequences on the dose distribution in organs at risk, leading to toxicity and potentially side effects such as bladder inflammation.

To address the challenge of assessing patient-specific sCT without a reference, two strategies have been developed. The first strategy involves creating an atlas from a cohort of reference CTs. This atlas is then registered to the new sCT, and a statistical analysis is conducted to detect outliers at a voxel level by comparing the HU values of the new sCT to the distribution of HU values from the cohort of reference CTs. To mitigate the bias induced by the interpatient registration process, a weight has been applied according to the probability of mis-registration of each structure. This method highlights regions of significant discrepancies. However, it is time-consuming and relies on the accuracy of interpatient registration, making it potentially inappropriate for patients with significantly different anatomies compared to the atlas.

To avoid registration uncertainties, the second approach utilizes radiomics-based feature calculation. These features are computed on a cohort of generated and reference data. An optimized random forest algorithm is then employed to select the most meaningful features for regarding their dosimetric outcomes. These selected features are then computed on each new sCT, and based on their deviation from the expected range of values, a score is assigned. This process can be repeated for several volumes of interest, including the body contour, bladder, rectum, clinical target volume, and heads of femurs, enabling assessment at both a global and organ-wise level. While this method has a low computation time and may allow for online QA, the selected features can present vendor dependency. Additionally, the field of view and the accuracy of the contours also impact the results.

A major limitation of this work is the data we worked with. As all the experiments were performed with a unique dataset of 39 patients with localized prostate cancer, the results presented cannot be considered generalizable but give an insight into the performance of the model assessed. Also, while the results are not generalizable, the methods proposed are and can be adapted to different generation approaches and anatomical locations.

The knowledge acquired during my thesis on sCT generation and the significance of QA have led to suggested guidelines for integrating quality control measures at various stages of the clinical workflow. However, this work is still in its early stages and requires discussion with experts in image processing, medical physicists, and clinicians to establish a consensus on metrics and best practices that ensure the accuracy of MRI-only RT.

The contributions of this thesis to sCT QA can be applied to sCT generated from CBCT, as they are not dependent on the modality used for image generation. Additionally, these contributions can be extended to other anatomical locations.

Perspectives

Using a large multicentre dataset to train deep-learning models for generating sCTs from MRI can enhance the generalizability and robustness of these methods. The availability of such a database to the public would also facilitate meaningful comparisons among published methods. Currently, each publication compares its results with other models based on their own dataset, and each method optimizes hyperparameters to achieve the best performance on their specific dataset, which complicates the implementation of these methods in different centres. Creating such a database poses several challenges, including the requirement for paired images where both a CT and an MRI are acquired from the same patient. Additionally, it necessitates international collaboration and should aim to represent diverse anatomies, socio-economical levels, ethnicities, and age groups to ensure fairness of the models[1]. This database will also prove valuable for the statistical approaches investigated in this study, as the patient data used were acquired at a single centre and consisted of only 39 CT-MRI pairs, potentially limiting its representativeness of the prostate cancer population.

To address the limitations associated with uncertainties arising from the non-rigid registration process in the previously presented methodologies, one potential solution is to differentiate between errors attributed to registration and those intrinsic to the method itself. This may be achieved by adapting mixed-effects models, initially developed for longitudinal data analysis[2], or by applying deep-learning approach[3]. Incorporating complementary metrics, such as Hausdorff distance and negative mean square difference, to assess the accuracy of the registration process would provide a more comprehensive understanding of potential biases introduced.

The sensitivity analysis can be enhanced by investigating the local influence of modified HU values, incorporating differences in dose-volume histograms for each volume of interest, including the target volume as well as the organs at risk (bladder, rectum, and heads of femur). Future work will also explore other treatment techniques. It would be valuable to model the impact of each error factor based on its interaction with the beams in different treatment modalities.

In the pelvic area, the patient's anatomy is susceptible to changes caused by variations in bladder and rectal filling for example. These changes may have implications for the accuracy of treatment delivery [4], [5]. The presented methodology in this study can be adapted to each specific sCT generation method, once the locations of HU uncertainties have been identified and treatment plans have been defined. Deep-learning-based sCT generation methods are commonly used, and the development of more effective models should be pursued in the future. Aleatoric uncertainties (data-dependent) and epistemic uncertainties (model-dependent) are specific to machine-learning models and can be assessed [6]. Considering the impact of these uncertainties on the dose distribution during the learning process could lead to the creation of more clinically valid image generation. The method proposed in this paper has also the potential to determine acceptance criteria for organ motion during treatment.

The voxel-wise approach for patient-specific sCT QA is time-consuming and impractical for online use. Therefore, exploring faster registration methods can be beneficial. The use of this methodology for offline validation or as part of the clinical evaluation stage to validate sCT generation methods can also be explored.

The radiomics feature-based methodology is more efficient and does not require interpatient registration. However, it's important to note that the study was conducted on a single-centre dataset. If applying this methodology to a different dataset, the entire procedure would need to be repeated, considering factors such as different field of view, image resolution, and variations in the feature extraction process. Additionally, when bladder injection is involved in the CT procedure for the training phase, it can affect the extracted features. The reproducibility of radiomic features is also influenced by various factors such as image

acquisition settings, reconstruction algorithms, and the software used for feature extraction [7]. A part of future work will involve investigating the reproducibility of selected features and testing their effectiveness as an assessment tool on sCT generated from data outside the training centre. Furthermore, exploring the benefits of using a multicentre database to train the random forest during the feature selection phase will be valuable. To achieve generalizability, a protocol for normalizing the data in terms of field of view and image resolution needs to be defined prior to using this approach. Additionally, an assessment of the contours used to delineate the volume of interest is necessary.

Conclusion

In conclusion, DLM have demonstrated great potential in sCT generation for MRI-only EBRT. Research has already proposed DLMs that achieve a high level of accuracy. However, the integration of DLM-based sCT into clinical practice faces two main obstacles:

- The need for training cohorts in reference centres or access to multicentre databases
- The requirement to analyse the limitations of these methods and assess the quality of the generated images before their clinical use.

Generation of sCT for MRI-based radiotherapy RT shows promise in reducing toxicity and improving local control. For this approach to be effective, the generated images must meet quality standards in terms of both visual representation and electron densities specific to each patient. Nevertheless, clinical trials are necessary to demonstrate the clinical benefits of this approach, and international consensus for sCT QA need to be established.

References

[1] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, “Addressing fairness in artificial intelligence for medical imaging,” *Nat Commun*, vol. 13, no. 1, p. 4581, 2022, doi: 10.1038/s41467-022-32186-3.

[2] J. L. Bernal-Rusiel, D. N. Greve, M. Reuter, B. Fischl, and M. R. Sabuncu, “Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models,” *Neuroimage*, vol. 66, pp. 249–260, 2013, doi: https://doi.org/10.1016/j.neuroimage.2012.10.065.

[3] K. A. J. Eppenhof and J. P. W. Pluim, “Supervised local error estimation for nonlinear image registration using convolutional neural networks,” in *Medical Imaging 2017: Image Processing*, SPIE, Feb. 2017, p. 101331U. doi: 10.1117/12.2253859.

[4] Y. Xiong *et al.*, “Assessment of intrafractional prostate motion and its dosimetric impact in MRI-guided online adaptive radiotherapy with gating,” *Strahlentherapie und Onkologie*, 2022, doi: 10.1007/s00066-022-02005-1.

[5] Z. Chen, Z. Yang, J. Wang, and W. Hu, “Dosimetric impact of different bladder and rectum filling during prostate cancer radiotherapy,” *Radiation Oncology*, vol. 11, no. 1, Aug. 2016, doi: 10.1186/s13014-016-0681-z.

[6] M. Abdar *et al.*, “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges,” 2020, [Online]. Available: http://arxiv.org/abs/2011.06225

[7] A. Traverso, L. Wee, A. Dekker, and R. Gillies, “Repeatability and Reproducibility of Radiomic Features: A Systematic Review,” *Int J Radiat Oncol Biol Phys*, vol. 102, no. 4, pp. 1143–1158, Nov. 2018, doi: 10.1016/j.ijrobp.2018.05.053.

Scientific valorisation

Journal papers published

"*Determination of acceptable Hounsfield Units uncertainties via a sensitivity analysis for an accurate dose calculation in the context of prostate MRI-only radiotherapy*"
H. Chourak, A. Barateau, P. Greer, C. Lafond, J-C Nunes, R. de Crevoisier, J. Dowling, O. Acosta (Physical and Engineering Sciences in Medicine, 2023)

"Quality assurance for MRI-only radiation therapy: a voxel-wise population-based methodology for image and dose assessment of synthetic-CT generation methods"
Hilda Chourak, Anaïs Barateau, Safaa Tahri, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Adrien Boue-Rafle, Mathias Perazzi, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta (Frontiers in Oncology, 2022)

"A high-performance method of deep learning for prostate MR-only radiotherapy planning using an optimized Pix2Pix architecture",
S. Tahri, A. Barateau, C. Cadin, **H. Chourak**, S. Ribault, F. Nozahic, O. Acosta, J.A. Dowling, P.B. Greer, A. Largent, C. Lafond, R. De Crevoisier, J.C. Nunes (Physica Medica, 2022)

Book chapter: "*Image synthesis for MRI-only radiotherapy treatment planning"*
Jason Dowling, Laura O'Connor, Oscar Acosta, Parnesh Raniga, Renaud de Crevoisier, Jean-Claude Nunes, Anais Barateau, **Hilda Chourak**, Jae Hyuk Choi, Peter Greer (In book: Biomedical Image Synthesis and Simulation, 2022)

"Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review"
M. Boulanger, Jean-Claude Nunes, **H. Chourak**, A. Largent, S. Tahri, O. Acosta, R. De Crevoisier, C. Lafond, A. Barateau (Physica Medica, 2021)

Invitation for oral communication

"Image synthesis for external beam radiation therapy"
Conferences on Artificial Intelligence for medical image synthesis, University of Medellin, Colombia (2022)

Communication in conference with lecture committee

"Voxel-Wise Analysis for Spatial Characterisation of Pseudo-CT Errors in MRI-Only Radiotherapy Planning"
Hilda Chourak, Anaïs Barateau, Eugenia Mylona, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - ISBI 2021 (poster presentation)

"*Spatial Characterization of errors in pseudo-CT generation for MRI-only radiotherapy*"
Hilda Chourak, Anaïs Barateau, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - ESTRO 2021 (poster presentation)

"Caractérisation spatiale d'erreurs de pseudo-CT pour la planification de dose à partir d'IRM"
Hilda Chourak, Anaïs Barateau, Capucine Cadin, Caroline Lafond, Jean-Claude Nunes, Peter Greer, Jason Dowling, Renaud de Crevoisier, Oscar Acosta - SFPM 2021 (oral presentation)

"Local quality assessment of patient specific synthetic-CT via voxel-wise analysis"
Hilda Chourak, Anaïs Barateau, Jean-Claude Nunes, Peter Greer, Safaa Tahri, Caroline Lafond, Renaud de Crevoisier, Jason Dowling, Oscar Acosta – DICTA 2022 (oral presentation)

"MRI-only radiation therapy for prostate cancer: exploration of the impact of synthetic-CT uncertainties on dose calculation"
Hilda Chourak, Dowling Jason, Peter Greer, Anais Barateau, Safaa Tahri, Renaud de Crevoisier, Jean-Claude Nunes, Oscar Acosta - AUS MrinRT 2022 (oral presentation)

"MRI-only radiation therapy for prostate cancer: a sensitivity analysis of the impact of synthetic-CT uncertainties on dose calculation"
Hilda Chourak, Jason Dowling, Peter Greer, Anais Barateau, Caroline Lafond, Renaud de Crevoisier, Jean-Claude Nunes, Oscar Acosta - ISBI 2023 (poster presentation)

"When to use Augmentation - Variability Insufficient for Cortical Thickness Estimation Improvement"
Filip Rusak, Rodrigo Santa Cruz, **Hilda Chourak**, Elliot Smith, Jurgen Fripp, Clinton Fookes, Pierrick Bourgeat, Andrew P. Bradley – ISBI 2023 (poster presentation)

COLLEGE DOCTORAL BRETAGNE | MATHS, TELECOMS INFORMATIQUE, SIGNAL SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Evaluation de la qualité de CT synthétiques générés à partir d'IRM pour la planification de traitement en radiothérapie externe

Mots clés : CT synthétiques, contrôle qualité, radiothérapie externe

Résumé : La radiothérapie (RT) externe repose sur deux modalités d'imagerie: la tomodensitométrie (scanner CT) pour le calcul de la dose basée sur la densité électronique des tissus, et l'imagerie par résonance magnétique (IRM) permettant une segmentation plus précise de la tumeur et des organes a risque grâce au meilleur contraste que cette modalité offre. L' IRM seule pour la planification de dose a gagné en popularité rendant le CT de planification obsolète. Cependant, cette modalité ne fournit pas d'informations sur la densité électronique des tissus. La génération de CT synthétique (sCT) est donc essentielle à la planification du traitement à partir d'IRM. Plusieurs méthodes ont été développées, et les progrès récents en apprentissage profond permettent d'obtenir des images précises. Mais le contrôle qualité d'un sCT pour une utilisation systématique en routine clinique n'est pas trivial. Les principaux objectifs de cette thèse sont: 1) Identifier les limites des méthodes de génération de sCT via une analyse statistique. 2) Quantifier l'impact des erreurs d'intensité sur la distribution de dose. L'objectif est d'évaluer l'importance de ces erreurs et leurs potentielles impact sur le traitement. 3) Proposer un protocole permettant l'évaluation de chaque sCT, afin de s'assurer que les images générées répondent aux normes requises et sont acceptables pour la planification du traitement.

Title: Synthetic-CT quality assessment for MRI-only based treatment planning in radiation therapy

Keywords: synthetic-CT, quality assessment, radiation therapy

Abstract: The standard external beam radiation therapy (EBRT) workflow relies on two imaging modalities: computed tomography (CT) for dose calculation based on electron density information, and magnetic resonance imaging (MRI) for better soft tissue contrast, enabling more accurate target delineation and minimising the risk of toxicity in healthy tissue. To define the treatment plan, both CT and MRI images are co-registered, inducing uncertainties. MRI- only RT has thus gained popularity by eliminating the need for CT scans. But, as MRI does not provide electron density information, the generation of synthetic-CT (sCT) is essential for MRI-only RT. Several methods have been developed, and recent advancements in deep learning have facilitated the production of more accurate results. However, for the systematic use of MRI-based dose planning in a clinical setting, the issue of quality control for the sCT still needs to be addressed. The main objectives of this thesis are: 1) To identify the limitations of the sCT generation methods through statistical analysis. 2) To quantify the impact of Hounsfield Unit errors on dose distribution. By measuring the effects of these errors, the goal is to assess their significance and potential implications in treatment. 3) To develop a patient specific sCT quality assessment framework, to ensure that generated sCTs meet required standards and are acceptable for treatment planning purposes.