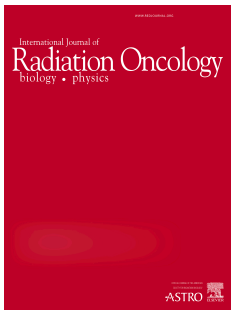

Dosimetry-driven quality measure of brain pseudo Computed Tomography generated from deep learning for MRI-only radiotherapy treatment planning

Emilie Alvarez Andres, Lucas Fidon, Maria Vakalopoulou, Marvin Lerousseau, Alexandre Carré, Roger Sun, Guillaume Klausner, Samy Ammari, Nathan Benzazon, Sylvain Reuzé, Théo Estienne, Stéphane Niyoteka, Enzo Battistella, Angéla Rouyar, Georges Noël, Anne Beaudre, Frédéric Dhermain, Eric Deutsch, Nikos Paragios, Charlotte Robert

PII: S0360-3016(20)31130-5

DOI: https://doi.org/10.1016/j.ijrobp.2020.05.006

Reference: ROB 26340

To appear in: *International Journal of Radiation Oncology • Biology • Physics*

Received Date: 15 November 2019

Revised Date: 21 April 2020

Accepted Date: 5 May 2020

Please cite this article as: Andres EA, Fidon L, Vakalopoulou M, Lerousseau M, Carré A, Sun R, Klausner G, Ammari S, Benzazon N, Reuzé S, Estienne T, Niyoteka S, Battistella E, Rouyar A, Noël G, Beaudre A, Dhermain F, Deutsch E, Paragios N, Robert C, Dosimetry-driven quality measure of brain pseudo Computed Tomography generated from deep learning for MRI-only radiotherapy treatment planning, *International Journal of Radiation Oncology • Biology • Physics* (2020), doi: https://doi.org/10.1016/j.ijrobp.2020.05.006.

Dosimetry-driven quality measure of brain pseudo Computed Tomography generated from deep learning for MRI-only radiotherapy treatment planning

Emilie Alvarez Andres[#1,2,3], Lucas Fidon[#2,4], Maria Vakalopoulou[4], Marvin Lerousseau[1,3,4], Alexandre Carré[1,3], Roger Sun[1,3,4,5], Guillaume Klausner[1,5], Samy Ammari[6], Nathan Benzazon[1,3], Sylvain Reuzé[1,3], Théo Estienne[1,3,4], Stéphane Niyoteka[1,3,] Enzo Battistella[1,3,4], Angéla Rouyar[1,3], Georges Noël[7], Anne Beaudre[3], Frédéric Dhermain[5], Eric Deutsch[1,5], Nikos Paragios[2], Charlotte Robert[1,3]

[1]U1030 Molecular Radiotherapy, Paris-Sud University - Gustave Roussy - Inserm - Paris-Saclay University, Villejuif, France
[2]TheraPanacea, Paris, France
[3]Department of Medical Physics, Gustave Roussy - Paris-Saclay University, Villejuif, France
[4]MICS Laboratory, CentraleSupélec, Paris-Saclay University, 91190, Gif-sur-Yvette, France
[5]Department of Radiotherapy, Gustave Roussy - Paris-Saclay University, Villejuif, France
[6]Department of Radiology, Gustave Roussy - Paris-Saclay University, Villejuif, France
[7]Department of Radiotherapy, Paul Strauss Institute, Strasbourg, France
#: Equal contribution

Corresponding author:
Charlotte Robert
Department of Medical Physics
114 street Edouard Vaillant
94800 Villejuif, France
Tel : +33(0)142115606
CH.ROBERT@gustaveroussy.fr

Authors responsible for the statistical analyses:
Emilie Alvarez Andres
Department of Medical Physics
114 street Edouard Vaillant
94800 Villejuif, France
Tel : +33(0)142115154
emilie.alvarez-andres@gustaveroussy.fr

Roger Sun
Department of Medical Physics
114 street Edouard Vaillant
94800 Villejuif, France
Tel : +33(0)142115050
roger.sun@gustaveroussy.fr

Conflicts of Interest:
Emilie Alvarez Andres reports grants from TheraPanacea, during the conduct of the study.
Sylvain Reuzé is a full-time employee of GE Healthcare since December 2018, outside of the submitted work.
Eric Deutsch reports grants and personal fees from Roche Genentech, grants from Servier, grants from Astrazeneca, grants and personal fees from Merck Serono, grants from BMS, grants from MSD, outside the submitted work.
Nikos Paragios is CEO and Founder of TheraPanacea, during the conduct of the study.

Funding:
No funding to declare.

Research data:
Research data are not available at this time.

Short title: Deep learning for brain pseudo CT generation

Abstract

Purpose: This study aims at evaluating the impact of key parameters on the pseudo Computed Tomography (pCT) quality generated from Magnetic Resonance Imaging (MRI) with a 3D convolutional neural network (CNN).

Methods: 402 brain tumor cases were retrieved yielding to associations of 182 Computed Tomography (CT)/T1 weighted MRI (T1), 180 CT/contrast enhanced T1 weighted MRI (T1-Gd) and 40 CT/T1/T1-Gd. A 3D CNN was used to map T1 or T1-Gd into CT and evaluate the importance of different components. First, the training set size influence on the testing set accuracy was assessed. Moreover, we evaluated the MR sequence impact, using T1 only and T1-Gd only cohorts. Then, we investigated four MRI standardization approaches, namely histogram-based (HB), zero-mean/unit-variance (ZMUV), White Stripe (WS) and no standardization (NS) based on training, validation and testing cohorts composed of 242, 81 and 79 patients cases respectively, as well as a bias field correction influence. Finally, two networks, namely HighResNet and 3D UNet, were compared to evaluate the architecture impact on the pCT quality. The Mean Absolute Error (MAE), gamma indices and dose volume histograms were used as evaluation metrics.

Results: Generating models using all the available cases for training led to higher pCT quality. The T1 and T1-Gd models indicated maximum differences in gamma indices means of 0.07 percent point. The MAE obtained with WS was 78 Hounsfield Units (HU) +/-22HU, which slightly outperformed HB, ZMUV and NS ($p<0.0001$). Regarding the network architectures, 3%/3mm gamma indices of 99.83%+/-0.19% and 99.74%+/-0.24% were obtained for HighResNet and 3D UNet respectively.

Conclusion: Our best pCT were generated using more than 200 samples in the training dataset, while training with T1 only and T1-Gd only did not significantly affect the performance. Regardless of the preprocessing applied, the dosimetry quality remained equivalent and relevant for a potential use in clinical practice.

Introduction

Magnetic Resonance Imaging (MRI) has become prevalent in radiotherapy planning due to its excellent soft tissue contrast compared to Computed Tomography (CT). During a brain tumor radiotherapy process, MRI and CT play a key role in indicating areas of interest and estimating the dosimetry respectively. Yet, dealing with multiple imaging modalities requires to co-register them, leading to errors up to 2mm (1), and target volumes margins increase.

To address this limitation, numerous approaches have been developed to generate a pseudo Computed Tomography (pCT) from MRI (2,3). First, the bulk density approach (4,5) assigns specific Electron Densities (ED) to pre-segmented MRI relying however on the labeling quality. Second, the multi-atlas method constitutes a multiple "atlases" database representing co-registered pairs of CT and MRI acquired from different patients. The incoming MRI is first aligned to the atlases MRI through a deformable registration. The resulting deformation fields are then applied to the atlases CT which are combined to generate the pCT (6,7). Due to the computational complexity of deformable registrations, the multi-atlas approach is time-consuming. To mitigate these limitations, Deep Learning (DL) methods (8–10) have been recently introduced, reporting promising results (11,12). Compared to the other approaches, DL-based methods efficiently exploit large databases to learn a direct mapping from MRI to CT. A deep Convolutional Neural Network (CNN) consists in a composition of convolutional filters and simple non-linear functions organized in layers. The parameters of the CNN are learned using pairs of MRI/CT training data via empirical risk minimization and stochastic gradient descent. DL-based methods benefit from highly efficient Graphical Processing Unit (GPU) implementations which reduce the inference time of the pCT of several orders of magnitude compared to atlas-based methods. Based on a NVIDIA Titan X GPU, Han et al. (13) reported durations of 9 seconds and 10 minutes for the DL and atlas-based approaches respectively.

However, there is still no consensus regarding: 1) the optimal training set size, 2) the best-suited MR sequence, 3) the optimal MR standardization preprocessing, 4) the use of an

inhomogeneity correction and 5) the best suited network architecture (Table A1). Additionally, there is no discussion about the approach to evaluate the generated pCT.

Indeed, training datasets sizes ranging from 15 (14) to 77 patients (12) have been reported, raising the issue of the minimal number of training patients required to ensure a satisfying generalization to unseen examples. Moreover, most of the studies used either T1-weighted MRI (T1) or contrast enhanced T1-weighted MRI (T1-Gd). However, the benefit of using a contrast agent in terms of pCT quality is still unclear. Additionally, only few studies applied MRI intensities standardization as preprocessing. Yet, it can improve the pCT quality (15). A similar question concerns the bias field correction, as only Han et al. (13) applied it. Finally, several CNN architectures have been used in the literature, such as HighResNet (16,17) and UNet (13) for instance, without systematically comparing them.

An additional aspect which it is not explicitly discussed in these works is the influence of these parameters on a dosimetry-based pCT evaluation. Numerous studies report their performances using peak signal-to-noise ratio or Mean Absolute Error (MAE) metrics (13,18,19) which can possibly be irrelevant to the real clinical scenario.

This study aims at evaluating the impact of significant parameters, namely the training dataset size, the input MR sequence, the standardization strategy, the application of an inhomogeneity correction and the network architecture, on the computed pCT accuracy and the associated clinical dosimetry. The pCT evaluation is based on both the MAE based on the intensities and ED, and clinical criteria, namely 1%/1mm, 2%/2mm and 3%/3mm gamma indices and differences in Dose Volume Histograms (DVH) of the Planning Target Volume (PTV).

Methods and Materials

Images acquisition and preprocessing

402 institutional patients treated between 2006 and 2017 for brain tumors were included in this retrospective study. For all of the patients, the delay between the planning CT and T1 or T1-Gd MR acquisitions did not exceed eight days. The dataset was composed of 182 CT/T1, 180 CT/T1-Gd and 40 CT/T1/T1-Gd paired images.

All the CT were acquired with a Sensation Open scanner (Siemens Healthineers, Erlangen, Germany) using a 120kVp tube voltage. The slice thickness was equal to 1mm, 2mm, 3mm and 5mm for 3, 45, 353 and 1 patients cases respectively. The native X and Y voxel sizes were included in [0.50mm; 0.70mm], [0.70mm; 0.90mm] and [0.90mm; 1.10mm] for 208, 76 and 118 patients respectively.

The MRI were all acquired with GE Healthcare devices (GE Healthcare, Milwaukee, Wisconsin, USA). Two patients cases' MR sequences were from external institutes and were acquired on two different 1.5T devices: Optima MR360 and Discovery MR450. The remaining MRI were institutional images, acquired on a 3T Discovery MR750w (224 patient cases), a 1.5T Optima MR450w (9 patient cases) and a 1.5T Signa Excite (167 patient cases). Only 3D axial T1-weighted images with or without a gadolinium injection were used. Initial slice thicknesses were included in [1mm; 1.2mm], [1.4mm; 2mm], [3mm; 3.2mm] and equal to 5mm for 234, 10, 157, 1 patients respectively. Regarding the native X and Y voxel sizes, they were included in [0.44mm; 0.50mm], [0.50mm; 0.58mm] and equal to 0.94mm for 325, 73 and 4 patients respectively.

For each patient, the CT was first rigidly registered to the T1 or T1-Gd images using the Drop library.[1] Then, the images were linearly resampled to a 1mm×1mm×1mm voxel size, before harmonizing the volumes to 300x300x242 voxels. Both the MRI intensities and the CT Hounsfield Units (HU) were clipped, to 0.1 and 99.9 percentiles and [-1000HU, 1800HU] respectively. The maximum HU was empirically determined based on CT intensity histograms. Finally, the HU were rescaled between [-1, 1].

Lastly, 60%, 20% and 20% of the patients were randomly parsed into training, validation and

[1] https://github.com/biomedia-mira/drop2

testing sets, provided that the T1 and T1-Gd were equal in proportion. Patients with all CT, T1 and T1-Gd images were automatically assigned to the testing set, to be used for the dosimetry-based evaluation.

Standardization strategies

Three different approaches were adopted to standardize the MRI.

The first approach was a histogram-based standardization (HB) based on the method described in (20). HB consists in matching the percentiles (10, 20, 30, 40, 50, 60, 80, 90) of an image to per-defined template values that are computed using the MR images of the cohort. The intensity match is obtained via a piece-wise linear transformation applied to image intensities.

The second approach consisted in a normalization of the intensity distribution inside the head of each patient to zero mean and unit variance (ZMUV) (15).

The last method, namely White Stripe (WS) (21), was similar to the ZMUV approach, but based on the normal appearing white matter mean and standard deviation, as it is known to be homogeneous. Brain masks were first extracted with the HD-BET tool (22). The MR images were then normalized with the intensity-normalization package (15).

Network architectures

Following popular choices of network architectures in the literature, we decided to use the HighResNet 3D CNN presented by Li et al. (23) and the 3D UNet (24).

The HighResNet was originally designed for a segmentation task. In contrast to other networks, it preserves the image resolution (no pooling layers) and is compact (0.8 million parameters).

The main components of the network were the dilated 3D convolutions with kernels of size 3x3x3, the residual connections, the normalization layers and the Rectified Linear Unit (ReLU) activations.

These operations were organized into nine residual blocks based on convolution filter sizes dilated by one, two or four. Each block contained a series of normalization, ReLU and convolution, which was repeated twice before adding the block input to its output. The two last layers were not residual and were composed of 3x3x3 and 1x1x1 convolutions to obtain the final pCT volume.

The 3D UNet is a popular encoder-decoder neural network architecture in medical image computing. It is characterised by its long shortcut connections between layers output at different stages of the network architecture that give it a U-shape. These connections allow to combine features at different scales and different spatial resolutions. Contrary to the HighResNet, 3D UNet uses max-pooling layers and no dilated convolutions. This difference enables the 3D UNet to have more features and to use larger input patches than the HighResNet at the price of a lower spatial resolution in some layers of the 3D UNet. ReLU activation, 3x3x3 convolutions, instance normalization, and linear upsampling were used for the 3D UNet resulting to approximately 15 million parameters.

The final aim of this work was not to develop an original network but to provide guidelines for the future pCT studies by evaluating the impact of different parameters on the pCT quality in terms of image intensity and dosimetry. As a result, we adapted the HighResNet for pCT generation. We replaced the normalization layers by instance normalization (25), we removed the softmax layer after the last convolutional layer and we changed the output channel number to one. The modified network architecture is displayed in Figure A2.

To optimize the network parameters, we used the MAE loss function:

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} \left| I_{CT}(i) - I_{pCT}(i) \right| \quad (1)$$

Where $I_{CT}(i)$ and $I_{pCT}(i)$ are the intensities of the CT and the pCT at voxel i, and N is the considered number of voxels.

Due to memory constraints, patches of size 96x96x96 voxels and 136x136x136voxels were

used as input of the HighResNet and 3D UNet respectively. At inference, the 3D MRI were divided into patches to reconstruct the whole pCT. A patch margin of length 5 and 1 voxels for the HighResNet and 3D UNet respectively, was applied leading to predictions inside sub-patches of size 86 x 86 x 86 and 134 x 134 x 134. The motivation of the margins is to guarantee a smooth transition between patches prediction. Note that patches overlapped, contrary to sub-patches. The overlap process is described in Figure A3.

For both networks, the learning rate was set to 0.001. Early stopping on the validation set was used as stopping criterion to assess the convergence of the CNN. Dropout was used after the penultimate layer during training with a probability of 0.5.

Note that no data augmentation was used in this study.

Impact of key parameters

The first experiment consisted in quantifying the impact of the training set size. Five different HighResNet networks were trained using 242 (121 T1-121 T1-Gd), 121 (61 T1–60 T1-Gd), 60 (30 T1–30 T1-Gd, 30 (15 T1–15 T1-Gd) and 15 (8T1–7T1-Gd) patients respectively in the training set. The validation and testing cohorts were the same for all the training set sizes and included 81 (41 T1–40 T1-Gd) and 79 (39 T1–40 T1-Gd) cases respectively. All the MR images were standardized using the HB method.

A second experiment was conducted to determine the best suited T1 input sequence to generate pCT. We constituted two HB-standardized cohorts: 1) a T1-only cohort with 134, 44 and 40 T1 MRI cases for the training, validation and testing sets respectively, 2) a T1-Gd-only cohort with 133, 44 and 40 patients cases respectively. The cases included in the two testing cohorts were the same, for a fair comparison. For this experiment, different T1 and T1-Gd histograms templates were computed for the HB standardization, based on the 134 and 133 patients included in the training cohorts. Experiment two was based on the HighResNet.

The third experiment assessed the role of the MRI standardization using 242 (121 T1–121 T1-

Gd), 81 (41 T1–40 T1-Gd) and 79 (39T1 –40 T1-Gd) cases in the training, validation and testing sets respectively. The HighResNet architecture was used for this experiment. Four different approaches were investigated: HB, ZMUV, WS and no standardization (NS).

The fourth experiment was performed to evaluate the role of the bias field correction, using HighResNet. As a result, the N4 filter (26) was optionally applied on MR images. The best standardization technique defined by experiment 3 was used here. The training, validation and testing sets were those used in experiment 3.

The last experiment was conducted to analyse the influence of the network architecture on the quality of the generated pCT. To this aim, the HighResNet used in the previous experiments and the 3D UNet, were trained, validated and tested. Best preprocessing strategies highlighted by the third and fourth experiments were applied. The split of the dataset was the same as experiment 3.

A summary of the experiments is presented in Figure A4.

Evaluation criteria

First, the initial CT and the pCT were compared using the MAE (Equ. 1). It was computed in four different areas: whole head, air, bone and water. The head was segmented using the Otsu approach, described in (27). The other regions were obtained thresholding the CT: $x \leq -200HU$, $-200HU < x < 250HU$ and $250HU \leq x$ for the air, water and bone regions respectively. The MAE was calculated from the 3D intensities volumes or the 3D ED volumes obtained applying the HU-ED calibration curve.

Furthermore, we evaluated the pCT quality in terms of dose prediction for all the experiments, except the first one, by computing metrics used in clinics. 1%/1mm, 2%/2mm and 3%/3mm 3D global gamma indices were considered, and no dose threshold was applied. In addition, relative differences between CT and pCT DVH ($D_{02\%}$, $D_{50\%}$, $D_{95\%}$ and $D_{98\%}$) of the PTV were calculated. The dosimetry plans from the original CT were recalculated on the pCT, with the Pencil Beam (PB) dose calculation algorithm implemented in iPlan RT 4.5 Dose (Brainlab, Munich, Germany) (28). The default grid size was set to 2mm. It is worth noting the grid was adaptive, meaning that it became

finer for small object. This approach was combined with a ray-tracing technique which was applied during the radiological path length calculation. These two approaches resulted in a speed up of the dose calculation. For this dosimetry analysis, a subset cohort of the testing set, corresponding to cases whose dosimetry had been realized with iPlan, was used. It was composed of 39 grades III and IV glioma patients cases (19 T1 - 20 T1-Gd) treated with a sliding window Intensity Modulated Radiation Therapy (IMRT) approach, delivered with a 6 MV beam. 18, 11, 7, 2 and 1 patients cases were treated with 5, 6, 7, 8 and 10 beams respectively. An illustration of the overall workflow is presented in Figure A5.

Two-sided paired Wilcoxon tests, with a significance level set to 0.05, were performed as statistical analysis.

Only results computed on the testing set are reported.

Results

Figures 1A and 1B present examples of MRI, CT and pCT with soft tissues and bone windows and levels respectively. They were extracted from the third experiment, using the HighResNet and the HB intensities standardization. The first line corresponds to a low MAE case (head MAE=64HU) and the second line to a high MAE case (head MAE=110HU). Some air and bone areas appear to be less accurately reconstructed, as highlighted by the red squares.

The intensity-based MAE obtained from different training set sizes, is displayed in Figure 2A. For the head area, increasing the training dataset resulted in a decrease of the MAE mean+/-standard deviation (std) from 189HU+/-28HU for the 15 patients-training set model to 92HU+/-23HU corresponding to the 242 patients-training set model. Bone and air regions reported the highest MAE. Differences between all the training size models were significant for the head region ($p<0.0001$) except between 30 and 60 patients (Table A6).

The ED-based MAE is presented in Figure 2B, to more accurately assess the pCT quality with respect to its clinical use. A similar behaviour is observed, with a head MAE decrease from 0.10+/-0.01 to 0.05+/-0.01 when increasing the training set size from 15 to 242.

Table 1 presents the means+/-std of the MAE, gamma indices, DVHs differences and Wilcoxon tests values derived from the T1-only and T1-Gd only models. The maximum differences between the T1 and T1-Gd models obtained for the head MAE means and gamma indices means were equal to 3HU and 0.07 percent point (pp) respectively.

Means+/-std of the MAE, gamma indices and DVH differences obtained for the standardization experiment are provided in Table 2. The statistical analysis is presented in Table A7. WS led to a head MAE of 78HU+/-22HU, which was significantly lower than the three other methods (p-values<0.0001). Regarding the dosimetry, 3%/3mm gamma indices of 99.86%+/-0.16%, 99.83%+/-0.19%, 99.85%+/-0.17%, 99.86%+/-0.18% were achieved for the HB, ZMUV, WS and NS approaches.

Regarding the fourth experiment based on the combination of the HighResNet with the WS standardization, means+/-std of the MAE and dosimetry metrics are presented in Table 3. Applying the bias field correction led to a head MAE of 81HU+/-22HU. Concerning the DVH D02%, differences equal to 0.15%+/-0.12% and 0.20%+/-0.13% were achieved with and without the application of the N4 filter respectively (p-value=0.026).

Table 4 provides the MAE and dosimetry values for the last experiment, which was conducted to compare the HighResNet with the 3D UNet. For both networks, the WS MRI standardization and the N4 filter were applied. Means+/-std obtained for the head MAE were equal to 81HU+/-22HU and 90HU+/-21HU for the HighResNet and 3D UNet respectively (p-value<0.0001). Significantly higher gamma indices were obtained with the HighResNet (p-value<0.0001), with a pass rate of 97.92%+/-1.06% for the most restrictive 1%/1mm criterion.

Discussion

This study aimed at evaluating the impact of key parameters of brain pCT generation from T1 or T1-Gd images, namely the training set size, the MR input sequence, the standardization strategy, the application of a bias field correction and the network architecture. Best results were achieved when combining the WS MRI standardization with an inhomogeneity correction, the HighResNet, and all our 242 training patients cases. This suggests that more training cases could lead to further improvements.

Regarding the MR sequences experiment, a difference of 3HU was observed between the head MAE means of the T1 only and T1-Gd only models, suggesting that the contrast agent resulted in a negligible pCT improvement. The DVH differences led to a similar conclusion, as only 0.07pp maximum difference between the two models means was obtained. We conducted an extra experiment to evaluate the potential benefit of the T2 Fluid Attenuated Inversion Recovery (FLAIR) MR sequence. 134, 44 and 40 patients were included in the training, validation and testing sets respectively. The preprocessing described for the T1 only and T1-Gd only cohorts was similarly applied. A mean MAE+/-std of 115HU+/-22HU was obtained within the head area. Differences with the T1 only and T1-Gd only cohorts were found significant ($p<0.001$). Thus, T2 FLAIR appeared to generate largest pCT intensities-linked errors. It could be attributed to the lower contrast contained in T2 FLAIR images compared to T1/T1-Gd images. A second interpretation could be the slice thickness which was larger for most of the T2 FLAIR images compared to T1/T1-Gd images, resulting in a less informative spatial sampling. Future work includes the comparison of T1 and unusual sequences, such as zero echo time in which bone areas are more visible, to assess which combination of MRI sequences is optimal for an accurate pCT reconstruction in radiotherapy.

The third experiment concerned the MRI standardization, and used the HighResNet as network architecture. A mean+/-std of 78HU+/-22HU was obtained for the head MAE when applying the WS standardization, which slightly outperformed HB, ZMUV and NS ($p<0.0001$). Largest errors were located in the air and bone areas, with MAE of 253HU+/-65HU and 199HU+/-54HU respectively and seemed to correspond to misaligned regions or areas with high dose gradients.

Dinkla et al. (11) reported competitive head MAE of 67HU+/-11HU. All the CT and MR images used in their study were acquired on the same device. In this work, MR images were acquired from five different devices. Table A8 presents the composition of the training, validation and test sets in terms of MR devices. As one can notice, most of the MRI of the training set, namely 133, were acquired with the DISCOVERY MR750w - 3Teslas (T) device. To analyse the impact of this unbalance, the test set was split into two subsets: MRI from the DISCOVERY MR750w - 3T (57 patients) and MRI from the SIGNA EXCITE – 1.5T (21 patients). The default HB standardization and HighResNet were used for this experiment. Mean head MAE+/-std led to 86HU+/-22HU for the DISCOVERY MR750w - 3T and 106HU+/-16HU for the SIGNA EXCITE – 1.5T ($p<0.0001$). It showed pCT computed from the DISCOVERY MR750w - 3T device were of higher quality since more MRI acquired with such device composed the training set and since 3T devices offer a better images resolution. Thus, we think that the composition of the training set had a non-negligible impact on the generated pCT. Comparing the literature MAE is however not trivial due to the use of heterogeneous datasets, suggesting the need for publicly available datasets.

Concerning the dosimetry analysis, negligible differences were observed between the different standardization approaches. Regarding WS, a mean+/-std of 99.85%+/-0.17% was obtained for the 3%/3mm gamma index, which was not significantly different from the ZMUV, HB and NS gamma indices (p-values≥0.14). These non-significant dosimetry results can be attributed to the non-linearity of both the HU-ED curve and the radiation matter interactions effects. Very few studies reported dosimetry evaluations for brain pCT generated with a DL-based approach. Dinkla et al. (11) achieved 91.1%+/-3.0%, 95.8%+/-2.1% and 99.3%+/-0.4% for 1%/1mm, 2%/2mm and 3%/3mm head gamma indices with no threshold. A similar performance was obtained by Liu et al. (29) who reported 99.2% for the 3%/3mm gamma index. Recently, Kazemifar et al. (12) achieved state of-the-art 1%/1mm and 2%/2mm gamma indices of 94.6%+/-2.9% and 99.2%+/-0.8%. Eventually, dosimetry analyses are crucial as they are the only relevant metric for a use in clinics.

Fourth experiment evaluated the role of an inhomogeneity correction combined with the

HighResNet and the WS standardization. Although a slight increase of 3HU of the mean head MAE was obtained when applying the N4 filter, the DVH metrics analysis showed a negligible decrease in the means up to 0.08pp (p-values≤0.026). It could be justified by an acceptable MRI quality or the network ability to handle this issue.

The last experiment was the evaluation of two different network architectures, namely the HighResNet and the 3D UNet. For each model, the WS standardization and the N4 filter were applied. Mean head MAE+/-std were equal to 81HU+/-22HU and 90HU+/-21HU for the HighResNet and 3D UNet respectively. The lower HighResNet MAE may be attributed to two major advantages: the dilated convolution filters which enable a large spatial context while retaining the full image resolution and the residual connections which regularize the optimization of the model. Regarding the dosimetry, 3%/3mm gamma indices equal to 99.83%+/-0.19% and 99.74%+/-0.24% were obtained for the HighResNet and the 3D UNet respectively. As a result, no significant clinical impact was observed between the two architectures. In the literature, a lower MAE of 47HU+/-11HU was reported by Kazemifar et al. (12) using a 2D GAN. In the context of pCT generation, a GAN corresponds to the training of a second auxiliary neural network which learns a loss function to estimate the distance between a pCT and the distribution of all the true possible CT. This data-driven loss function is used to train the main neural network that learns the mapping from MRI to pCT. Therefore, pCT produced by a GAN are not guaranteed to respect the anatomy of the patient. To mitigate this issue, CycleGAN using an additional cycle-consistency penalization have been proposed (19,30). However, the cycle-consistency implies a one-to-one mapping between the MRI and CT, which is not realistic and can lead to artefacts in the pCT (31). As a result, further investigation of the errors specific to GAN and CycleGAN is needed for their clinical use in radiotherapy and is beyond the scope of this paper. The loss function used to train the network has a knock-on effect on the pCT quality. Here, the MAE was chosen since it was found to generate less blurry images than the mean squared error during preliminary experiments. Kazemifar et al. (12) trained two 2D GAN based on the MAE and the mutual information loss functions and obtained head MAE means+/-std of 60HU+/-

22HU and 47HU+/-11HU respectively. Therefore, exploring different loss functions is of interest as it can heavily impact intensities-linked errors.

Based on all the dosimetry results, very small discrepancies were obtained between all the preprocessing applied. For instance, 3%/3mm gamma indices equal to 99.83%+/-0.19% and 99.85%+/-0.17% were achieved for the experiments based on the combination of the HighResNet with the WS standardization and optionally applying the N4 filter (Table 3). Although a significant p-value of 0.012 was obtained, no major clinical impact is expected. As a result, it suggests that the proposed pCT generation method may be suitable for an introduction into clinics, regardless of the preprocessing applied.

The dose calculation algorithm used in this study was PB. An extra experiment was conducted to evaluate its relevance against Monte Carlo (MC), considered as more accurate in taking into account heterogeneities (32,33). Since the latter is not commissioned in our institution for IMRT plans, we constituted an additional cohort of 8 brain tumor patients treated with arctherapy. 4 out 8 patients had a CT and a T1 MRI, the rest had a CT and a T1-Gd MRI. The preprocessing previously described in the Materials and Methods section was similarly applied and pCT were generated. A dosimetry was performed on the pCT with the two different dose algorithms. No significant differences were observed for the DVH differences analysis ($p \geq 0.27$). A similar conclusion was obtained for the 3%/3mm and 2%/2mm gamma indices ($p \geq 0.40$). Concerning the 1%/1mm criterion, 98.94%+/-0.68% and 98.40%+/-0.84% gamma pass rates were achieved for the PB and MC algorithms respectively ($p=0.0078$). As a result, PB approach is a reliable technique for the head localization, due to the absence of large inhomogeneities.

Regarding the dataset, it was composed of 402 cases. To our knowledge, it is the largest cohort ever used in the head pCT generation field. Previous studies involved up to 77 patients (12). Our data were split into independent sets, namely training, validation and testing. Note that most of the published studies lack a validation set (11,13,14,19,29,30), potentially leading to biased results.

MRI-only radiotherapy could remove isotropic 2mm of margins due to registration errors (1).

However, distortions can also lead to errors up to 2mm even after applying a correction algorithm (34). Therefore, establishing a reliable quality assurance (35,36) is the key to unlock the full potential of radiotherapy.

Several limitations are present in this study. First, our DL pipeline necessitated paired images, and thus an intermodality registration which can introduce errors in the training set. To evaluate this error, an experienced radiologist placed three landmarks both on the CT and the MRI of ten patients. Registrating the CT onto the MR led to a mean distance error+/-std of 3.0mm+/-1.1mm. Further investigation may focus on rigid registration errors and evaluate different algorithms, such as the FLIRT (37,38) tool for comparison. Second, no analysis of the interplay effect of preprocessing steps and networks architecture was performed. Indeed, the use of a bias field correction and the selection of WS as the best standardization was based on experiments performed using HighResNet. This may have introduced a bias in the comparison between HighResNet and 3D Unet.

Conclusions

In this study, we aimed at optimizing relevant parameters to achieve high quality pCT for MR-only radiotherapy. The large variety of imaging devices and the considerable patients number constituting the training set appear to have a great impact on the pCT quality. All the parameters previously described, such as the MR sequence, intensities standardization, bias field correction, network architecture, have minor dosimetry influence as the gamma indices and DVH differences remained clinically convincing for every technique in our cohort. It suggests the efficiency of the model and its possible introduction into clinics. Future work include the extension of the current 3D network to integrate segmentation masks of target and organs at risk volumes and the development of a pCT generation model for a different anatomical site, such as pelvis.

References

1. Ulin K, Urie MM, Cherlow JM. Results of a multi-institutional benchmark test for cranial CT/MR image registration. Int J Radiat Oncol Biol Phys. 1 août 2010;77(5):1584-9.

2. Schmidt MA, Payne GS. Radiotherapy Planning using MRI. Phys Med Biol. 21 nov 2015;60(22):R323-61.

3. 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. 1 janv 2018;100(1):199-217.

4. Kang KM, Choi HS, Jeong BK, Song JH, Ha I-B, Lee YH, et al. MRI-based radiotherapy planning method using rigid image registration technique combined with outer body correction scheme: a feasibility study. Oncotarget. 15 août 2017;8(33):54497-505.

5. Wang C, Chao M, Lee L, Xing L. MRI-based treatment planning with electron density information mapped from CT images: a preliminary study. Technol Cancer Res Treat. oct 2008;7(5):341-8.

6. Sjölund J, Forsberg D, Andersson M, Knutsson H. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. Phys Med Biol. 21 janv 2015;60(2):825-39.

7. Demol B, Boydev C, Korhonen J, Reynaert N. Dosimetric characterization of MRI-only treatment planning for brain tumors in atlas-based pseudo-CT images generated from standard T1-weighted MR images. Med Phys. déc 2016;43(12):6557.

8. 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 CT from MRI. Med Phys. 20 juin 2019;

9. 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. Br J Radiol. 20 juin 2019;20190067.

10. 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. 30 mai 2019;

11. 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. 15 2018;102(4):801-12.

12. 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 J Eur Soc Ther Radiol Oncol. 2019;136:56-63.

13. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. Med Phys. avr 2017;44(4):1408-19.

14. Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. Med Phys. 14 juin 2018;

15. Reinhold JC, Dewey BE, Carass A, Prince JL. Evaluating the impact of intensity normalization on MR image synthesis. Med Imaging 2019 Image Process. 2019;10949:109493H.

16. Kläser K, Markiewicz P, Ranzini M, Li W, Modat M, Hutton BF, et al. Deep Boosted Regression for MR to CT Synthesis. In: Gooya A, Goksel O, Oguz I, Burgos N, éditeurs. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2018. p. 61-70. (Lecture Notes in Computer Science).

17. Kläser K, Varsavsky T, Markiewicz P, Vercauteren T, Atkinson D, Thielemans K, et al. Improved MR to CT Synthesis for PET/MR Attenuation Correction Using Imitation Learning. In: Burgos N, Gooya A, Svoboda D, éditeurs. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2019. p. 13-21. (Lecture Notes in Computer Science).

18. Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, et al. Medical Image Synthesis with Context-Aware Generative Adversarial Networks. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv. sept 2017;10435:417-25.

19. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I. Deep MR to CT Synthesis Using Unpaired Data. In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL, éditeurs. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2017. p. 14-23. (Lecture Notes in Computer Science).

20. Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med. déc 1999;42(6):1072-81.

21. Shinohara R, Sweeney E, Goldsmith J, Shiee N, Mateen F, Calabresi P, et al. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin. 15 août 2014;6:9-19.

22. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated Brain Extraction of Multisequence MRI Using Artificial Neural Networks. Hum Brain Mapp. 2019;40(17):4952–4964.

23. 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, et al., éditeurs. Information Processing in Medical Imaging. Springer International Publishing; 2017. p. 348-60. (Lecture Notes

in Computer Science).

24. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, éditeurs. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 424-32. (Lecture Notes in Computer Science).

25. Ulyanov D, Vedaldi A, Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. 2016;

26. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 Bias Correction - IEEE Journals & Magazine. IEEE Trans Med Imaging. juin 2010;29(6):1310-20.

27. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans Syst Man Cybern. janv 1979;9(1):62-6.

28. Mohan R, Chui C, Lidofsky L. Differential pencil beam dose computation model for photons. Med Phys. févr 1986;13(1):64-73.

29. 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. mars 2019;20(3):105-14.

30. 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. août 2019;46(8):3565-81.

31. Chu C, Zhmoginov A, Sandler M. CycleGAN, a Master of Steganography. ArXiv. 2017;abs/1712.02950.

32. Fragoso M, Wen N, Kumar S, Liu D, Ryu S, Movsas B, et al. Dosimetric verification and clinical

evaluation of a new commercially available Monte Carlo-based dose algorithm for application in stereotactic body radiation therapy (SBRT) treatment planning. Phys Med Biol. 21 août 2010;55(16):4445-64.

33. Petoukhova AL, van Wingerden K, Wiggenraad RGJ, van de Vaart PJM, van Egmond J, Franken EM, et al. Verification measurements and clinical evaluation of the iPlan RT Monte Carlo dose algorithm for 6 MV photon energy. Phys Med Biol. 21 août 2010;55(16):4601-14.

34. Weygand J, Fuller CD, Ibbott GS, Mohamed ASR, Ding Y, Yang J, et al. Spatial Precision in Magnetic Resonance Imaging-Guided Radiation Therapy: The Role of Geometric Distortion. Int J Radiat Oncol Biol Phys. 15 2016;95(4):1304-16.

35. Xing A, Holloway L, Arumugam S, Walker AR, Rai R, Juresic E, et al. Commissioning and quality control of a dedicated wide bore 3T MRI simulator for radiotherapy planning. In 2016.

36. Sun J, Barnes M, Dowling J, Menk F, Stanwell P, Greer PB. An open source automatic quality assurance (OSAQA) tool for the ACR MRI phantom. Australas Phys Eng Sci Med. mars 2015;38(1):39-46.

37. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal. juin 2001;5(2):143-56.

38. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage. oct 2002;17(2):825-41.

39. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. févr 1998;17(1):87-97.

40. Cox IJ, Roy S, Hingorani SL. Dynamic histogram warping of image pairs for constant image brightness. In: Proceedings of the 1995 International Conference on Image Processing (Vol2)-

Volume 2 - Volume 2. USA: IEEE Computer Society; 1995. p. 2366. (ICIP '95).

Figure captions

Fig. 1. (From left to right) Magnetic Resonance Imaging (MRI), original Computed Tomography (CT) and pseudo Computed Tomography (pCT) with soft tissues (1A) and bone (1B) windows and levels respectively for two patients. Red squares highlight some of the incorrect reconstructed areas.

Fig. 2. Evolution of the Mean Absolute Error (MAE) based on Hounsfield Units (HU, 2A) and Electron Densities (ED, 2B) when modifying the number of training subjects. The boxplot corresponds to the first and third MAE quartiles with the MAE median in the middle, while the whiskers correspond to the range of the MAE distribution after excluding the outliers.

Table 1. Means +/- standard deviations (std) of the Mean Absolute Error (MAE), gamma indices, Dose Volume Histograms (DVH) differences computed for the Planning Target Volume (PTV) and statistical analysis derived from the T1-weighted MRI (T1) and contrast enhanced T1-weighted MRI (T1-Gd) cohorts comparison.

	T1 only	**T1-Gd only**	**p-value**	**95% Confidence interval**
MAE head	84HU+/-25HU	87HU+/-28HU	**0.0047**	[-3.93, -0.76]
MAE air	274HU+/-63HU	306HU+/-74HU	**<0.0001**	[-36.51, -22.37]
MAE bone	228HU+/-63HU	236HU+/-71HU	0.066	[-11.38, 0.48]
MAE water	38HU+/-11HU	38HU+/-12HU	0.82	[-0.83, 0.73]
1%/1mm gamma index	97.87%+/-1.16%	97.94%+/-1.07%	0.59	[-0.12, 0.05]
2%/2mm gamma index	99.60%+/-0.33%	99.63%+/-0.30%	0.50	[-0.05, 0.02]
3%/3mm gamma index	99.84%+/-0.18%	99.85%+/-0.18%	0.44	[-0.03, 0.01]
Difference PTV $D_{02\%}$	0.20%+/-0.15%	0.15%+/-0.09%	**0.0041**	[0.02, 0.08]

Difference PTV $D_{50\%}$	0.20%+/-0.15%	0.13%+/-0.08%	**0.015**	[0.02, 0.12]
Difference PTV $D_{95\%}$	0.20%+/-0.17%	0.14%+/-0.10	**0.012**	[0.02, 0.12]
Difference PTV $D_{98\%}$	0.27%+/-0.37%	0.22%+/-0.41%	**0.026**	[0.01, 0.12]

Table 2. Means +/- standard deviations (std) of the Mean Absolute Error (MAE), gamma indices and Dose Volume Histograms (DVH) differences computed for the Planning Target Volume (PTV) derived from the histogram-based (HB), zero mean/unit variance (ZMUV), White Stripe (WS) and no standardization (NS) cohorts.

	HB	**ZMUV**	**WS**	**NS**
MAE head	92HU+/-23HU	83HU+/-22HU	78HU +/- 22HU	96HU+/-23HU
MAE air	297HU+/-73HU	284HU+/-62HU	253HU +/- 65HU	313HU+/-68HU
MAE bone	251HU+/-61HU	214HU+/-55HU	199HU +/- 54HU	252HU+/-60HU
MAE water	39HU+/-11HU	38HU+/-12HU	36HU +/- 11HU	43HU+/-11HU
1%/1mm gamma index	97.94%+/-1.06%	97.90%+/-1.10%	98.08% +/- 1.01%	97.80%+/-1.17%

2%/2mm gamma index	99.63%+/-0.28%	99.61%+/-0.30%	99.64% +/- 0.29%	99.61%+/-0.31%
3%/3mm gamma index	99.86%+/-0.16%	99.83%+/-0.19%	99.85% +/- 0.17%	99.86%+/-0.18%
Difference PTV $D_{02\%}$	0.22%+/-0.17%	0.22%+/-0.16%	0.20% +/- 0.13%	0.24%+/-0.20%
Difference PTV $D_{50\%}$	0.24%+/-0.16%	0.23%+/-0.16%	0.21% +/- 0.13%	0.27%+/-0.17%
Difference PTV $D_{95\%}$	0.27%+/-0.31%	0.21%+/-0.17%	0.19% +/- 0.15%	0.32%+/-0.32%
Difference PTV DVH $D_{98\%}$	0.38%+/-0.58%	0.27%+/-0.35%	0.20% +/- 0.17%	0.38%+/-0.46%

Table 3. Means +/- standard deviations (std) of the Mean Absolute Error (MAE), gamma indices, Dose Volume Histograms (DVH) differences of the Planning Target Volume (PTV) and statistical analysis derived from the White Stripe (WS) and WS combined with a bias field correction (N4) cohorts comparison.

	WS	**WS & N4**	**p-value**	**95% Confidence interval**
MAE head	78HU +/- 22HU	81HU +/- 22HU	**<0.0001**	[-4.79, -2.57]
MAE air	253HU +/- 65HU	244HU +/- 62HU	**<0.0001**	[5.23, 11.84]

MAE bone	199HU +/- 54HU	230HU +/- 56HU	**<0.0001**	[-35.81, -27.07]
MAE water	36HU +/- 11HU	34HU +/- 10HU	**<0.0001**	[2.02, 2.91]
1%/1mm gamma index	98.08% +/- 1.01%	97.92% +/- 1.06%	**0.0035**	[0.04, 0.19]
2%/2mm gamma index	99.64% +/- 0.29%	99.60% +/- 0.32%	**0.0026**	[0.01, 0.06]
3%/3mm gamma index	99.85% +/- 0.17%	99.83% +/- 0.19%	**0.012**	[0.00, 0.03]
Difference PTV $D_{02\%}$	0.20% +/- 0.13%	0.15% +/- 0.12%	**0.026**	[0.00, 0.13]
Difference PTV $D_{50\%}$	0.21% +/- 0.13%	0.13% +/- 0.10%	**0.0017**	[0.03, 0.15]
Difference PTV $D_{95\%}$	0.19% +/- 0.15%	0.11% +/- 0.12%	**0.0034**	[0.03, 0.14]
Difference PTV $D_{98\%}$	0.20% +/- 0.17%	0.13% +/- 0.13%	**0.0088**	[0.02, 0.14]

Table 4. Means +/- standard deviations (std) of the Mean Absolute Error (MAE), gamma indices, Dose Volume Histograms (DVH) differences computed for the Planning Target Volume (PTV) and statistical analysis derived from the White Stripe (WS) combined with a bias field correction (N4) and the initial HighResNet against WS associated with N4 and the 3D UNet cohorts comparison.

	WS & N4 & HighResNet	**WS & N4 & 3D UNet**	**p-value**	**95% Confidence interval**

MAE head	81HU +/- 22HU	90HU +/- 21HU	**<0.0001**	[-9.39, -6.99]
MAE air	244HU +/- 62HU	266HU +/- 66HU	**<0.0001**	[-27.18, -15.56]
MAE bone	230HU +/- 56HU	209HU +/- 54HU	**<0.0001**	[16.91, 25.79]
MAE water	34HU +/- 10HU	49HU +/- 11HU	**<0.0001**	[-15.81, -14.09]
1%/1mm gamma index	97.92% +/- 1.06%	97.28% +/- 1.46%	**<0.0001**	[0.42, 0.79]
2%/2mm gamma index	99.60% +/- 0.32%	99.39% +/- 0.47%	**<0.0001**	[0.10, 0.24]
3%/3mm gamma index	99.83% +/- 0.19%	99.74% +/- 0.24%	**<0.0001**	[0.03, 0.11]
Difference PTV $D_{02\%}$	0.15% +/- 0.12%	0.33% +/- 0.21%	**<0.0001**	[-0.28, -0.11]
Difference PTV $D_{50\%}$	0.13% +/- 0.10%	0.29% +/- 0.19%	**<0.0001**	[-0.22, -0.10]
Difference PTV $D_{95\%}$	0.11% +/- 0.12%	0.28% +/- 0.18%	**<0.0001**	[-0.24, -0.13]
Difference PTV $D_{98\%}$	0.13% +/- 0.13%	0.31% +/- 0.18%	**<0.0001**	[-0.26, -0.15]

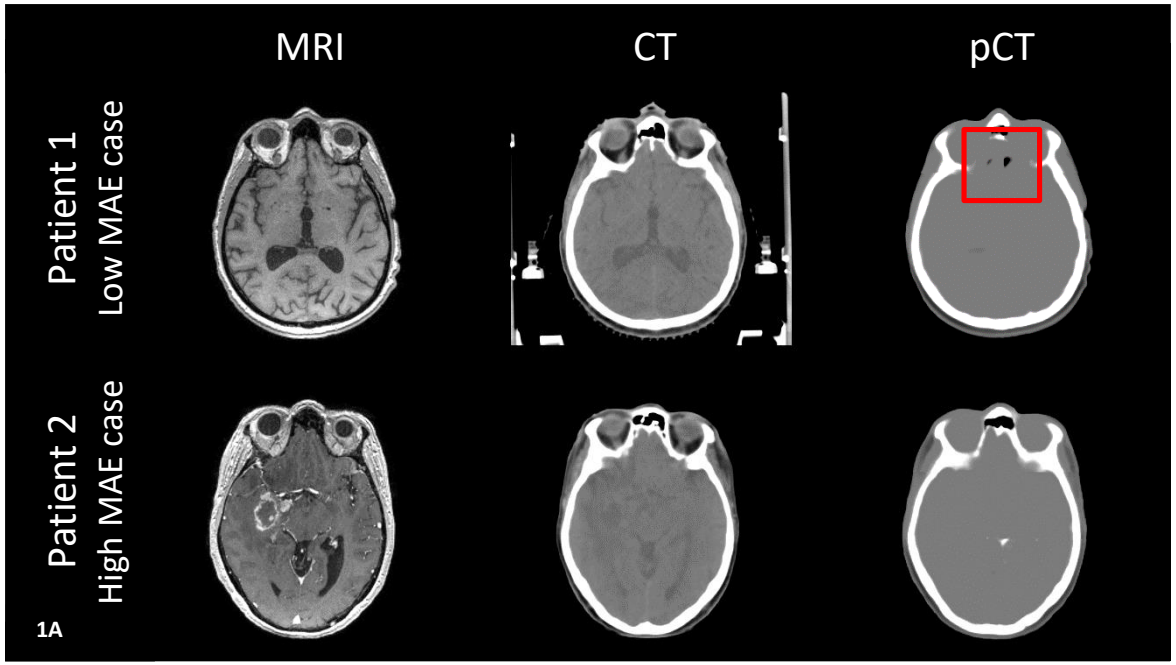

Figure 1A

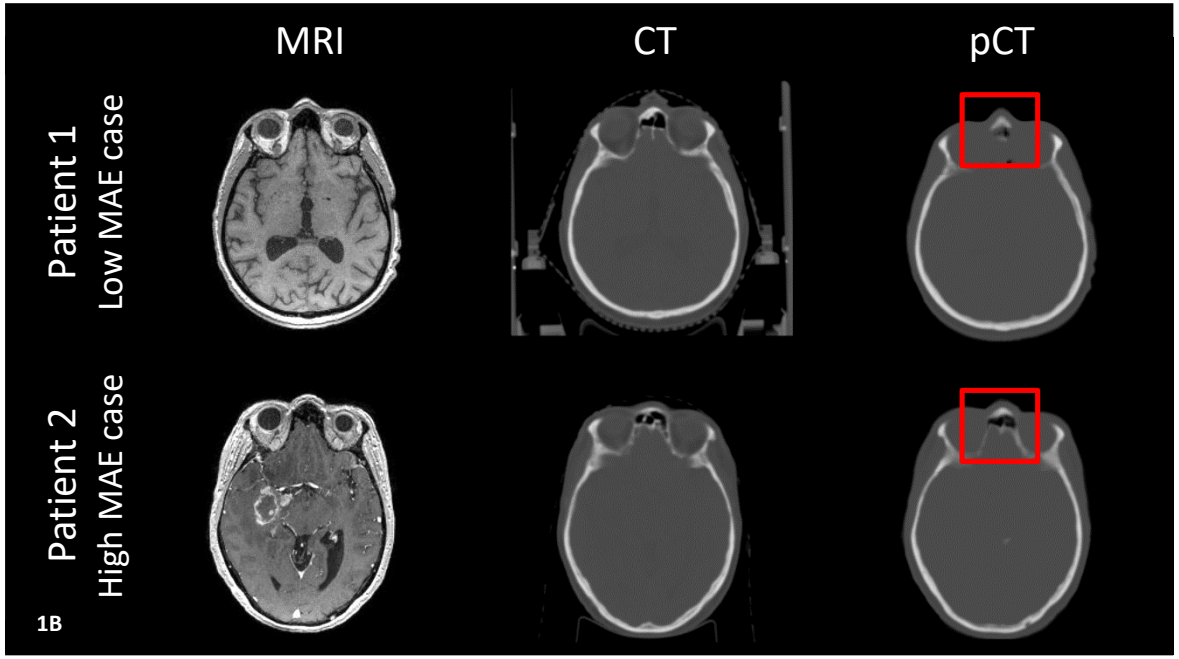

Figure 1B

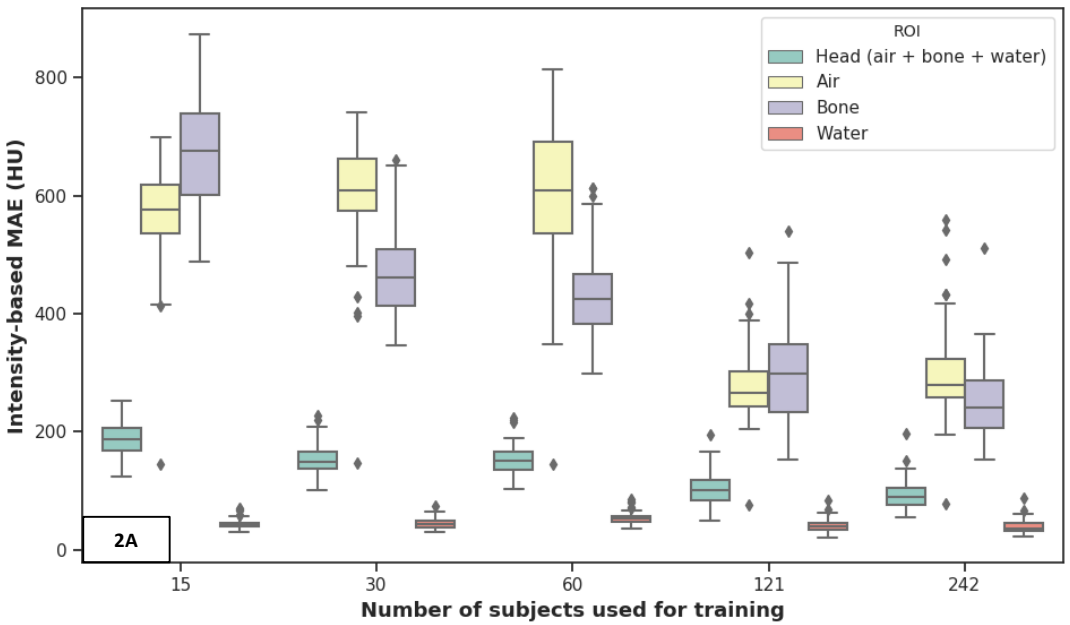
ROI
Head (air + bone + water)
Air
Bone
Water
Intensity-based MAE (HU)
0
200
400
600
800
15
30
60
121
242
Number of subjects used for training
2A

Figure 2A

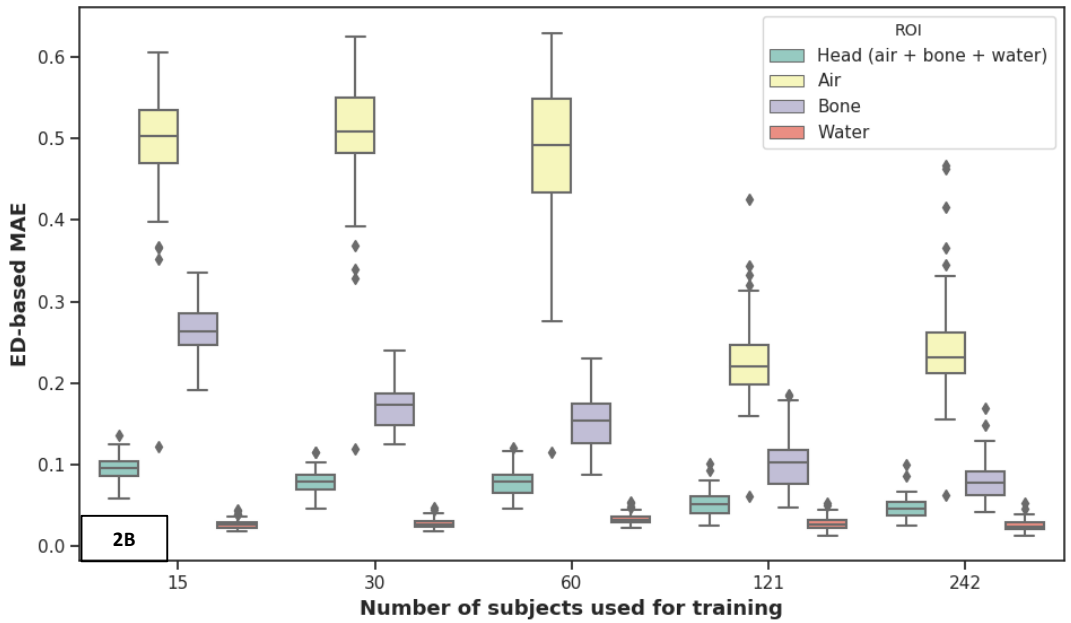
ROI
Head (air + bone + water)
Air
Bone
Water
ED-based MAE
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2B
15
30
60
121
242
Number of subjects used for training

Figure 2B