International Journal of Radiation Oncology biology • physics

www.redjournal.org

Physics Contribution

Deep Convolution Neural Network (DCNN) Multiplane Approach to Synthetic CT Generation From MR images—Application in Brain Proton Therapy

Maria Francesca Spadea, PhD,* Giampaolo Pileggi, PhD,*,† Paolo Zaffino, PhD,* Patrick Salome, MSc,† Ciprian Catana, PhD,‡ David Izquierdo-Garcia, PhD,‡ Francesco Amato, PhD,§ and Joao Seco, PhD†,‖

**Department of Experimental and Clinical Medicine, Magna Graecia University of Catanzaro, Catanzaro, Italy; †Biomedical Physics in Radiation Oncology, DKFZ—Deutsches Krebsforschungszentrum, Heidelberg, Germany; ‡Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, Massachusetts; §Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli Federico II, Naples, Italy; and ‖Department of Physics and Astronomy, Heidelberg University, Germany*

Received Nov 28, 2018. Accepted for publication Jun 21, 2019.

Summary

The paper describes a multiplane deep convolution neural network approach to predict synthetic computed tomography from T1-weighted magnetic resonance imaging. The method was tested in the framework of brain proton therapy, where Hounsfield Unit inaccuracies and steep density gradients easily lead to range shift errors. Results proved that the predicted synthetic

Purpose: The first aim of this work is to present a novel deep convolution neural network (DCNN) multiplane approach and compare it to single-plane prediction of synthetic computed tomography (sCT) by using the real computed tomography (CT) as ground truth. The second aim is to demonstrate the feasibility of magnetic resonance imaging (MRI)-based proton therapy planning for the brain by assessing the range shift error within the clinical acceptance threshold.

Methods and Materials: The image database included 15 pairs of MRI/CT scans of the head. Three DCNNs were trained to estimate, for each voxel, the Hounsfield unit (HU) value from MRI intensities. Each DCNN gave an estimation in the axial, sagittal, and coronal plane, respectively. The median HU among the 3 values was selected to build the sCT. The sCT/CT agreement was evaluated by a mean absolute error (MAE) and mean error, computed within the head contour and on 6 different tissues. Dice similarity coefficients were calculated to assess the geometric overlap of bone and air cavities segmentations. A 3-beam proton therapy plan was simulated for each patient. Beam-by-beam range shift (RS) analysis was conducted to assess the proton-

Corresponding author: Joao Seco, PhD; E-mail: j.seco@dkfz.de

Maria Francesca Spadea and Giampaolo Pileggi made equal contributions to this study.

This study was supported by DKFZ. Authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU.

Disclosures: No conflict of interests to declare.

Supplementary material for this article can be found at https://doi.org/10.1016/j.ijrobp.2019.06.2535.

Int J Radiation Oncol Biol Phys, Vol. 105, No. 3, pp. 495–503, 2019
0360-3016/$ - see front matter
https://doi.org/10.1016/j.ijrobp.2019.06.2535

computed tomography can be suitable for proton beam planning in the vision of magnetic resonance imaging −only proton therapy.

stopping power estimation. RS analysis was performed using clinically accepted thresholds of (1) 3.5% + 1 mm and (2) 2.5% + 1.5 mm of the total range.

Results: DCNN multiplane statistically outperformed single-plane prediction of sCT ($P < .025$). MAE and mean error within the head were 54 ± 7 HU and −4 ± 17 HU (mean ± standard deviation), respectively. Soft tissues were very close to perfect agreement (11 ± 3 HU in terms of MAE). Segmentation of air and bone regions led to a Dice similarity coefficient of 0.92 ± 0.03 and 0.93 ± 0.02, respectively. Proton RS was always below clinical acceptance thresholds, with a relative RS error of 0.14% ± 1.11%.

Conclusions: The multiplane DCNN approach significantly improved the sCT prediction compared with other DCNN methods presented in the literature. The method was demonstrated to be highly accurate for MRI-only proton planning purposes.

Introduction

With the availability of new magnetic resonance imaging (MRI) and positron emission tomography (PET) hybrid scanners and MRI linear accelerator (linac) machines, MRI has recently increased in popularity among radiation therapy clinical practices.[1] The high soft-tissue contrast offered by this imaging modality makes it the preferred choice for accurate target and region of interest identification in specific anatomic sites.[2] However, the MRI signal is not directly correlated to tissue electron density, thus posing difficulty in using MRI for PET attenuation correction and radiation therapy planning. This issue has driven researchers to develop routines and algorithms for converting MRI into synthetic computed tomography (sCT) (often referred as "pseudo" or "substitute" computed tomography [CT]). Several methods to relate MRI intensities to CT Hounsfield units (HU) have been developed and tested, especially for head and pelvis. MRI-to-sCT techniques are usually classified as being atlas-based, voxel-based, or hybrid methods, and 2 systematic reviews of the available methodologies have been recently published.[3,4] A short and a more updated summary of the state of the art is also available.[5] In general, atlas-based methods[6,7] have the advantage of using standard MRI sequences (T1 and/or T2 weighted only); this means the imaging protocol for sCT generation can easily be incorporated into the clinical routine, limiting the acquisition time and cost. On the other hand, they rely heavily on the quality of image registration of the atlas onto the patient's MRI and cannot accurately handle atypical anatomy or variable patient populations (eg, pediatric patients). If multiple atlases are used, the registration errors can be reduced, but extra time is required to carry out all atlas-to-patient matching. In addition, the MRI field of view (FOV) has to be large enough to obtain a robust registration to the atlas. Voxel-based strategies[8,9] have the advantage of not requiring very accurate image registration, but they rely on multisequence MRI acquisition, including ultrashort echo-time to visualize bone, to improve tissue classification. The ultrashort echo-time sequences are not commonly part of standard clinical MRI protocol. Finally, hybrid methods[10] overcome most of the issues of the previous approaches.

More recently, deep learning (DL)-based techniques have been successfully developed for sCT generation from MRI. The advantages of using DL for generating sCT include the following: (1) just standard MRI sequences (T1, T2) can be used as input, and (2) atypical anatomy can be managed because neural networks are able to generalize and extrapolate results; Wolterink et al[11] showed that it is also possible to use unpaired data by training a generative adversarial neural network (GANN) to synthesize CT images. Several DL-based methods have been tested for generating sCTs of the head and pelvis site. For the head, which is the focus of this work, we focused only on sCT prediction for the brain, and methods tested in this anatomic site are presented here.

7Han[12] and Xiang et al[13] proposed the use of a deep convolution neural network (DCNN) to generate sCT scans from T1 MRI and obtained better results than atlas-based methods. A comparison between DCNN and GANN was presented by Emami et al,[14] who reported that GANN achieved slightly superior results compared with DCNN. Dinkla et al[15] proposed a dilated DCNN with multiple plane inputs and achieved the best results within the brain reported in the current literature. They also tested the sCT in a photon radiation therapy scenario by including dosimetric analysis on volumetric modulated arc therapy plans. The interest in testing MRI-to-sCT conversion for radiation therapy is highly motivated by the current clinical availability of MRI linac,[16] with several clinical studies having demonstrated the high potential for adaptive treatments.[17-20] In addition, some centers are also working toward the development of in-room MRI guided proton therapy.[21-24] This will combine the ability of MRI to perform real-time organ tracking with the ability of protons to significantly reduce total delivered dose to surrounding organs. However, the Bragg peak positioning within the patient is very sensitive to inaccuracies in the sCT prediction. To achieve MRI proton planning, strict constraints are needed on the uncertainty of the stopping power values obtained directly from MRI, because of the sharp gradient beyond the Bragg peak.[25] Pileggi et al[10] showed that performing a beam-by-

beam analysis of range shift is more accurate than computing γ-index and dose-volume histograms on the global dose to evaluate the accuracy of HU reassignment.

In this work, we describe a new DCNN multiplane approach for enhanced sCT prediction. The main novelty introduced in the DL algorithm is that 3 U-nets are independently trained to predict the HU of each pixel by using axial, sagittal, and coronal inputs, respectively. The final output is obtained by computing the median value of each HU triplet. We also propose a robust protocol to assess sCT accuracy in the air cavities, which are mostly prone to interscan error between the MRI and the ground truth (GT) CT. Finally, we tested our method in the framework of proton brain therapy by evaluating the dosimetric difference in terms of proton range shift between sCT and CT.

Methods and Materials

The data set consisted of 15 pairs of T1-weighted MRI ($T1_{MR}$) and CT image volumes that were used in previous studies.[10,26-28] Images were acquired with the patients in the supine position. All patients, treated with radiation therapy after surgical resection of glioblastoma, gave written informed consent, and the local institutional review board approved the study. MRI scans were obtained with a 3 Tesla MAGNETOM Trio (Siemens Healthcare GmbH, Erlangen, Germany), with a voxel size of $1 \times 1 \times 1$ mm^3 and reconstruction matrix of $256 \times 256 \times 176$. The scanned region included the whole head down to the C3 vertebra, except for 5 patients, in whom the FOV ended at the level of the base of the skull. Sequence parameters were 3-dimensional acquisition with gradient recalled inversion recovery sequence, flip angle $= 7°$, TE = 1.64 ms, TR = 2530 ms, TI = 1200 ms, pixel bandwidth = 650 Hz; Nf = 256; and readout bandwidth = 166.4kHz (256*650Hz). MRI scans were acquired using an in-house transmission and receiver (8-channel) coil specially built-in for a simultaneous brain PET/MR scanner called brainPET.[10,26-28] The position of the head was not restricted using the treatment facial mask, as in the case of the CT, for 2 reasons: first, patient comfort (patients were inside the MRI scanner for more than 1 hour at a time), and second and most importantly for MRI safety to avoid any potential overheating of the facial mask owing to the radiofrequency currents produced by the MRI scanner. Patient head position was, however, restricted within the magnetic resonance coil using comfortable pads. For CT scans a LightSpeed QX/i scanner (GE Healthcare) was used. Voxel size ranged from $0.49 \times 0.49 \times 2.5$ mm^3 to $0.67 \times 0.67 \times 2.5$ mm^3. CT scans were acquired using a facial mask to immobilize the patient head position for treatment purposes. CT images were used to train the network and served as GT for testing purposes. The interscan interval between $T1_{MR}$ and CT acquisition ranged from 11 to 20 days.

The patient head was masked out by using Plastimatch software,[29] and background image intensities were set to 0 and −1000 for $T1_{MR}$ and CT data, respectively. A rigid transformation was computed to register the CT onto the $T1_{MR}$ by using Plastimatch and setting Mattes Mutual Information as metric to optimize. Thus, the final image resolution for training was that of the MRI. Finally, $T1_{MR}$ was tissue normalized using Free Surfer, and the full range of intensities was kept. Inspired by Han,[12] an improved version of the 2D_VGG-16 network[30] was implemented for this work. Details about the multilayer structure of the network are given as supplementary material S1 (available online at https://doi.org/10.1016/j.ijrobp.2019.06.2535). $T1_{MR}$ and CT volumes were sliced along the 3 orthogonal planes, axial (Ax), sagittal (Sag), and coronal (Cor), and each group of 2-dimensional images was given as input to train a different network. As a result, from each plane, a sCT prediction was obtained independently. This means that for each voxel V_n (with n = 1,…, N number of voxels) of the $T1_{MR}$, a synthetic HU (sHU) triplet estimation was available $(sHU_{Ax_n}; sHU_{Sag_n}; sHU_{Cor_n})$. The final value of each V_n was obtained as

$$sHU_{V_n} = median\left(sHU_{Ax_n}; sHU_{Sag_n}; sHU_{Cor_n}\right) \; \forall \, n \in \{1, ..., N\} \qquad \text{Equation 1}$$

Training, validation and testing sets consisted of 12, 2, and 1 patient, respectively. To increase the testing size, leave-one-out cross validation was executed on each of the 15 patients: Cyclically, 1 patient was taken out and used as independent data for testing, and the other 14 served to train and validate the network.

To assess the quality of sCT in comparison to CT, estimation mean absolute error (MAE) and mean error (ME) were calculated inside the patient head (excluding the background) on a total of M < N voxels as

$$MAE = \frac{1}{M}\sum_{i=1}^{M} |sHU_n - HU_n| \qquad \text{Equation 2}$$

$$ME = \frac{1}{M}\sum_{n=1}^{M} (sHU_n - HU_n) \qquad \text{Equation 3}$$

where sHU and HU are the Hounsfield Unit values measured on sCT and CT, respectively.

A tissue-based analysis was also conducted to assess the accuracy of prediction by creating the following labels, according to the method described by Huang and Parra: FAT, gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), bone, and intracranial air. FAT, GM, WM and CSF segmentations were obtained by using Morphologically and Anatomically accuRate Segmentation toolbox,[31] an extension of SPM8,[32] on $T1_{MR}$ images. Segmented labels were checked by an MRI/PET clinician expert. Air and bone regions were segmented by thresholding the CT at <-800 HU and >200 HU, respectively.[33] As pointed out by Edmund and Nyholm,[3] bone HUs range from 200 (spongy bone) to about 1200 (cortical bone). Thus, the MAE spectrum for bone (MAE_{bone}) was calculated by binning the voxels with same HU value (bin size was 20 HUs) in the CT and averaging their absolute error. HU

values up to 1500 were included (these mainly represent metal clips).

Because CT and MRI interscan time was on average 14 days (mean value), we noticed relevant discrepancies between the 2 acquisitions in the nasal cavities (presence or absence of mucus) and naso/oro-pharynx. In this work, we defined reliable regions to measure errors for air by considering only frontal, maxillary, and sphenoidal sinus. These cavities are less prone to interscan variability and are large enough to minimize partial volume errors.

The Dice similarity coefficient (DSC) metric was used to assess the geometric accuracy of bone and air cavities on the sCT according to the following definition

$$DSC_{air,bone} = \frac{2\left(V_{pCT} \cap V_{CT}\right)}{V_{pCT} + V_{CT}} \qquad \text{Equation 4}$$

Air and bone volumes were segmented both on sCT and CT for comparison by thresholding at <-800 HU and >200 HU, respectively. As for MAE_{bone}, also DSC_{bone} was computed by varying the threshold value from >200 to >1000. We did not segment bone with threshold >1000 because the number of voxels would have been too small to compute the DSC.

For each patient, a tumor in the range of 2.8 to 3.5 mm in diameter was simulated behind the frontal sinus, where bone−air−soft tissue gradients are high and challenging for proton therapy treatments. Three beams were planned with the couch at 90°: 1 vertical, along the craniocaudal direction at 0° gantry angle (GA), 1 anteroposterior 270° GA, and 1 anterosuperior oblique at 315° GA. It is important to underline that the plan was not aiming at the optimal clinical set-up but was used to study proton range shift sensitivity to bone−air−soft tissue variations. The anteroposterior 270° traversed about 8 to 12 mm of bone, a few millimeters of air cavity, and additional bone before reaching the tumor. This path is particularly challenging for proton therapy, in which small variations in density along the range affect the final proton range.

Because the assessment method was a beam-by-beam range shift analysis, each beam delivered a single field uniform dose to allow homogeneous planning target volume (PTV) coverage of 2 Gy. A pencil beam dose algorithm, available in MatRad (www.matrad.org),[34] was used for dose calculation. As described previously,[10] range at 80% of maximum dose (R_{80}) was computed on the dose profile of 10 central slices of the dose cube dose both for sCT and CT. For all 15 patients, this resulted in 450 range shifts to analyze.

Range shift (RS) and relative range shift (RRS) were defined as

$$RS = (R_{80}sCT - R_{80}CT) \qquad \text{Equation 5}$$

$$RRS = RS/R_{80}CT \qquad \text{Equation 6}$$

where R_{80} is the beam range computed at 80% of the normalized Bragg peak.

Guidance on an acceptable level of RRS came from range uncertainty criteria currently used in clinical applications at the Massachusetts General Hospital (MGH) and University of Florida Proton Therapy Institute (HPTI)[35]:

$$RS < 3.5\%R_{80}CT + 1mm(MGH) \qquad \text{Equation 7}$$

$$RS < 2.5\%R_{80}CT + 1.5mm(HPTI) \qquad \text{Equation 8}$$

Finally, the voxel-by-voxel difference between CT and sCT dose cubes was computed within the PTV.

Statistical tests (1-way analysis of variance [ANOVA]) was performed in MATLAB 2017a (The MathWorks, Natick, MA).

Results

Training and testing time were on average 60 hours and 30 seconds, respectively, for each leave-one-out cross validation round on a workstation equipped with an NVIDIA p6000 GPU.

The mean ± standard deviation (stdev) values of MAE and ME of multiplane approach for the 15 patients were 54 ± 7 HU (range, 45-66 HU) and −4 ± 17 (range, −42-19), respectively, computed inside the skin contour. In the single-plane approach, MAE were 65 ± 8 HU for Ax, 59 ± 7 for Sag, and 60 ± 8 for Cor.

A 1-way ANOVA test found statistical difference in terms of MAE (Fig. 1) between the multiplane and the single-plane predictions ($P < .025$). ME was not significantly different: −2 ± 21 HU for Ax, −4 ± 17 for Sag, and 4 ± 17 Cor.

In Figure 2, the best (patient 11) and the worst case (patient 14) are shown, respectively, for visual comparison. In both cases, the algorithm is able to predict the sCT very well, with the residual errors being partially due to magnetic field distortion (magnetic ring artifacts are visible in coronal and sagittal views and discussed in more detail by Walker et al[36]) and misregistration inaccuracies between the $T1_{MR}$ and the GT CT (see also supplementary material; available online at https://doi.org/10.1016/j.ijrobp.2019.06.2535).

In Table 1, we report the tissue-by-tissue analysis of HU and sHU in terms of MAE and ME. For soft tissue, errors for FAT were statistically higher than those for CSF, WM, and GM (1-way ANOVA, $P < 10^{-6}$), with an average MAE of about 44 HU. Analysis of air cavities and bone structures showed very good agreement between sCT and CT, as measured by the DSC. Although for air, a comparison with the state of the art cannot be performed because of the lack of standard assessment, for bone >200 HU both MAE and DSC were better than those presented by other authors.[3,4,14,15]

Analysis of bone as a function of the HU is shown in Figure 3, where a comparison with other published results is also presented. The average DSC was above 0.85 for thresholds up to 900 HU, thus highlighting the accuracy of HU prediction for dense bone. In addition, the MAE was

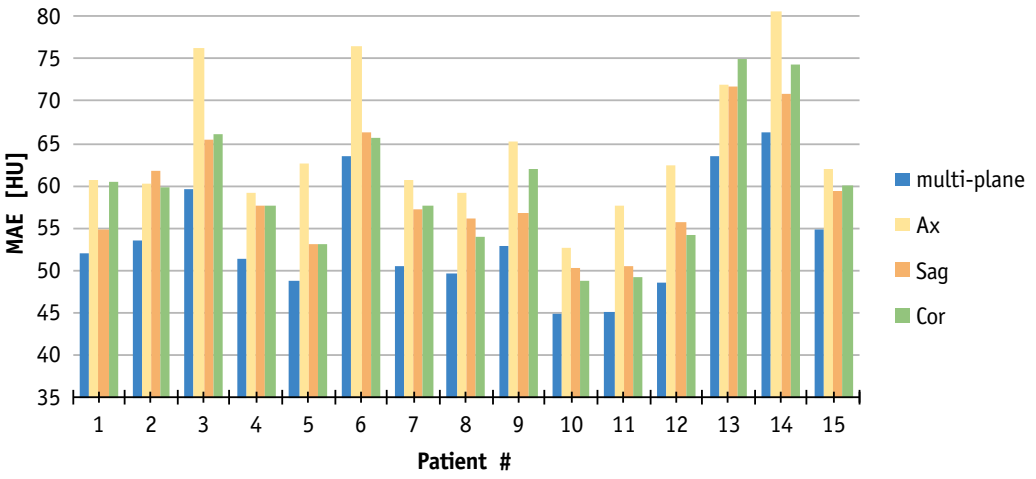


Fig. 1. Mean absolute error, computed within the skin contour, for each patient by using the multiplane approach compared with single plane.

calculated for the bone as a function of HU, up to a maximum of 1500 HU. The MAE value is approximately constant between 100 HU and 1400 HU.

Figure 4 shows the RRS boxplots (median ± quartiles, fifth to 95th percentile) for the 3 angles, 0°, 270°, and 315°. Slightly worse results were obtained for the 0° beam, as confirmed by 1-way ANOVA ($P = .001$), where the density step was higher. Mostly, RRS errors were within 2%. An example of RS analysis for a beam passing through bone and air cavity is also shown. All planned beams passed the tests of equations 7 and 8.

Figure 5 reports the relative distribution of dose cube difference within the PTV on the whole patient population. Mean ± stdev is 0.00 ± 0.01 Gy.

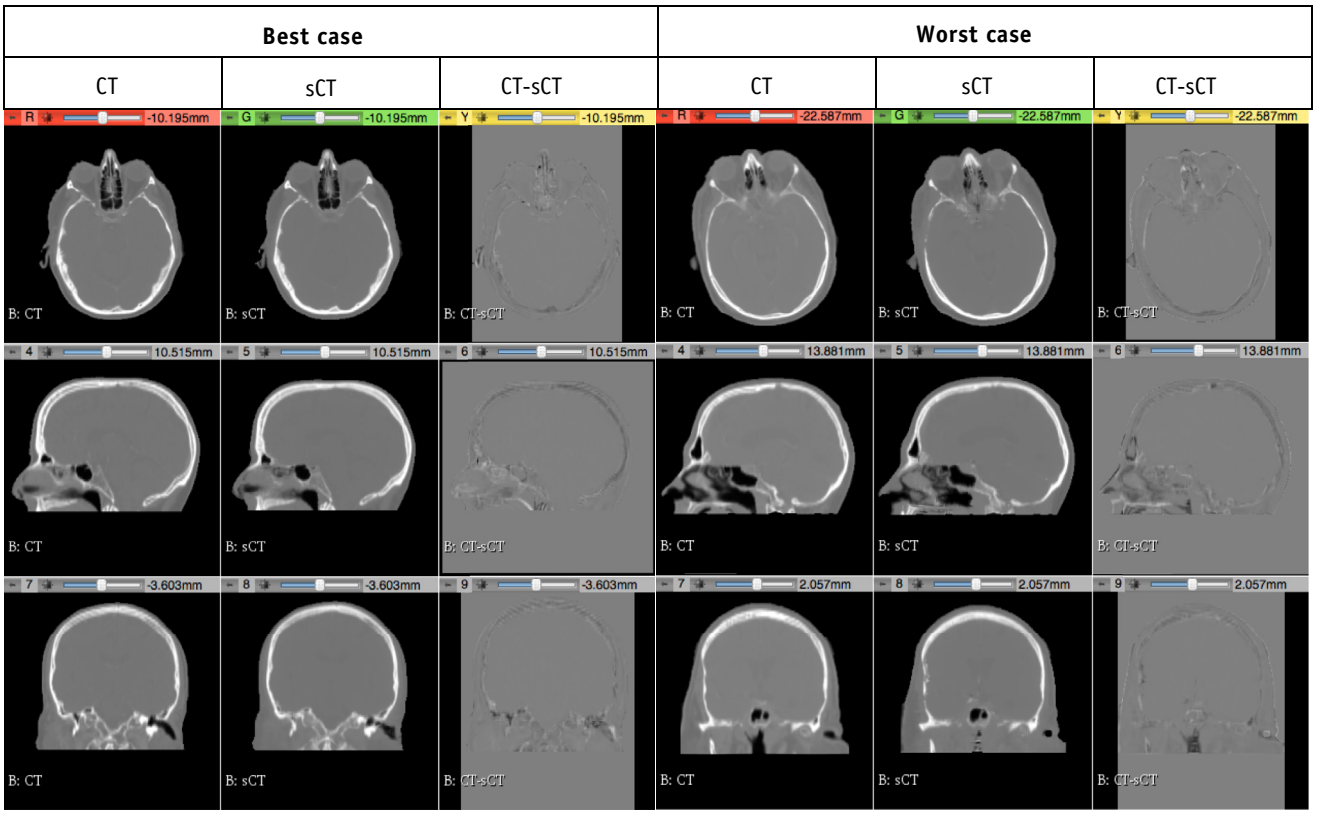


Fig. 2. Visual assessment for the best (patient 11: mean absolute error within head = 45 HU, $DSC_{AIR} = 0.96$, $DSC_{BONE} = 0.94$) and worst (patient 14: mean absolute error within head = 66 HU, $DSC_{AIR} = 0.87$, $DSC_{BONE} = 0.87$) synthetic computed tomography predictions. The image difference between computed tomography and synthetic computed tomography in the top part of the skull; most of the residual errors are due to magnetic field distortions. These artifacts are visible in the figures from the ring-artefact shape, as described previously.[36]

Table 1 Air, soft tissues, fat, and bone HU, MAE, ME, and DSC comparison between CT and sCT (mean ± standard deviation)

Tissue	Mean HU on CT	Mean HU on sCT	MAE (HU)	ME (HU)	DSC
Air (HU <−800)	−940 ± 14	−928 ± 18	53 ± 32	−37 ± 39	0.92 ± 0.03
FAT	−73 ± 3	−49 ± 13	44 ± 8	−4 ± 5	—
CSF	24 ± 2	28 ± 5	10 ± 3	0 ± 9	—
WM	37 ± 3	37 ± 3	6 ± 2	0 ± 4	—
GM	48 ± 4	48 ± 2	8 ± 2	0 ± 6	—
Bone (HU >200)	769 ± 71	767 ± 33	119 ± 17	20 ± 52	0.93 ± 0.02

Abbreviations: CT = computed tomography; DSC = Dice similarity coefficient; HU = Hounsfield units; MAE = mean absolute error; ME = mean error; sCT = synthetic computed tomography.

Discussion

MRI guided radiation therapy is now a reality for photon therapy but still a dream for proton therapy. At present, several groups are actively developing in-room MRI for proton therapy, with the added advantage of tumor dose escalation while keeping radiation side effects low.[21-24] In such a context magnetic resonance proton-based planning is desirable to close the loop of magnetic resonance−only proton therapy. This paper aimed at developing an MRI-to-sCT algorithm for accurate planning in brain proton therapy.

The MRI-to-sCT algorithm was based on training multiple DCNNs to predict the HU from $T1_{MR}$ intensities and comparing them with a standard CT HU estimation. The main novelty introduced in the procedure was that the input $T1_{MR}$ volume was sliced across the 3 orthogonal planes (axial, sagittal, and coronal) and that 3 CNNs were independently trained, 1 per plane, as has been successfully done for segmentation purposes.[37] A voting system, based on the median value, was then used to select the best of the 3 DCNN results on a voxel-by-voxel basis. Very recently, the multiplane approach was also proposed by Dinkla et al[15]; however, only 1 DCNN was trained with 3 equally weighted inputs representing the axial, sagittal, and coronal planes. Although from a mathematical point of view, in terms of network convergence, the 3 individual networks versus 1 network should perform similarly, from a computational point of view the 3-network training approach converges much faster. This leads to the possibility to parallelize the processes and to reduce the batch size even with limited computational resources or larger data sets. In addition, the 3-network training is preferable especially in case of nonisotropic voxel size and nonisotropic reconstruction matrix (where each view can have a different dimension). Moreover, the sagittal view is different compared with the other 2 views because the head is narrower and the number of informative slices in the sagittal plane is smaller than in the other 2 planes. Finally, sagittal slices are symmetrical to the central sagittal slice, so this view has less variability than the others.

In our experiments, the training of 3 DCNNs independently, 1 for each plane, reduced the average MAE from 67

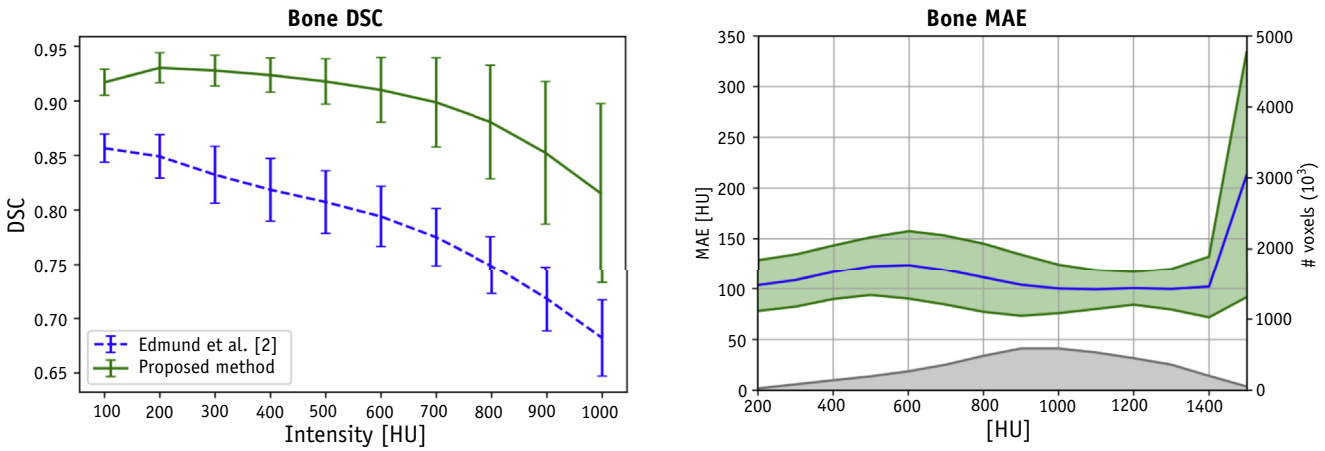


Fig. 3. Left panel: Dice similarity coefficients as a function of Hounsfield units for the proposed method (solid line) and as reported previously[3] (dotted line). According to our calibration curve, bone density should be considered above 200 HU. It is appreciable that bone Dice similarity coefficients depends on the threshold value set for the computed tomography number of the bone. Right panel: Mean absolute error (mean ± standard deviation, blue line and green area) for the full bone density spectrum, gray curve. (A color version of this figure is available at https://doi.org/10.1016/j.ijrobp.2019.06.2535).

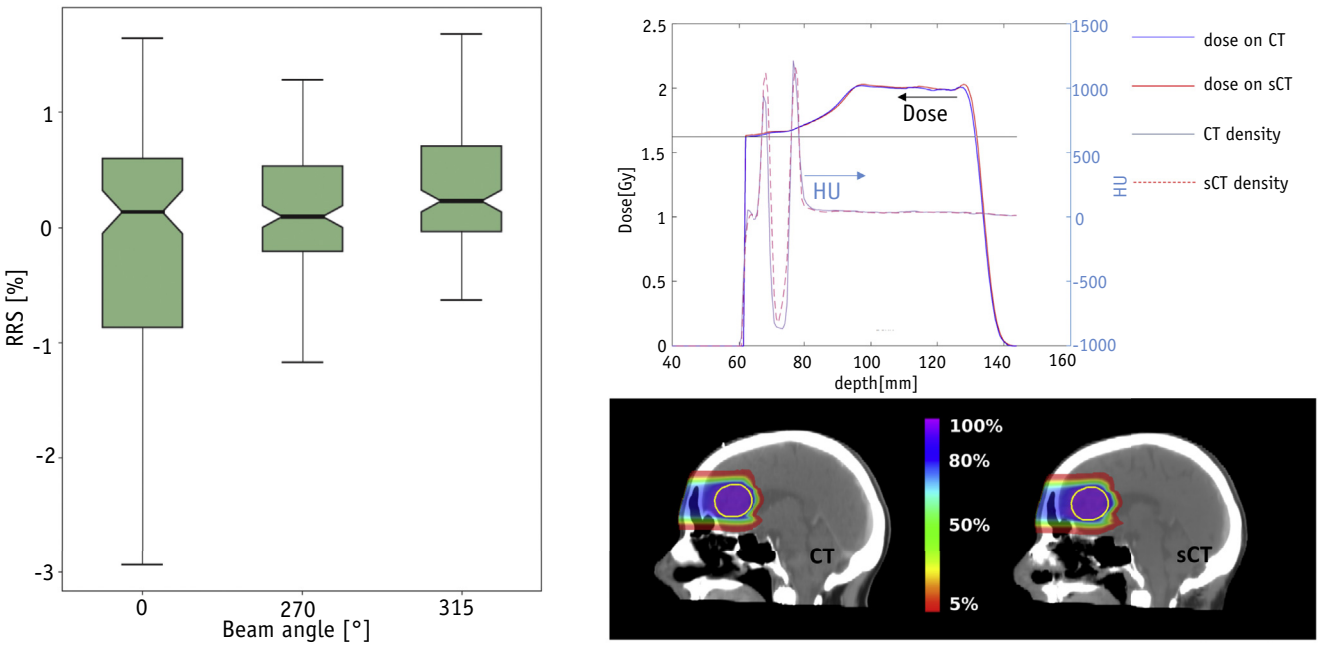


Fig. 4. Left panel: Boxplots (median±quartiles, fifth to 95th percentile) of relative range shift for each beam directions. Data are averaged over the 15 patients, with 10 slices per patient. Right panel: Example of range shift analysis on patient 12 (mean absolute error within head = 49 HU, $DSC_{AIR} = 0.93$, $DSC_{BONE} = 0.94$). The beam (couch at 90°, gantry at 0°) was planned and optimized on the synthetic computed tomography (sCT). Then it was copied on the computed tomography (CT) volume. The right panel shows the beam range profile on CT (blue) and sCT (red) relative to the white line shown on left panels. The Hounsfield unit profile is also shown (solid line and dotted line for CT and sCT, respectively). (A color version of this figure is available at https://doi.org/10.1016/j.ijrobp.2019.06.2535).

± 11 (1 DCNN with 3 inputs and batch size 18) to 54 ± 7 HU (3 DCNN with 1 input each and batch size 1) and the ME from −12 ± 30 to −4 ± 17 HU. The batch size reduction in the 3-network training led to a 20% improvement in the MAE and allowed us to train the algorithm on a not very high-performance GPU. Table 2 presents an overview of the MAEs obtained by recent works in which DL networks were implemented to generate sCT of the head. We also report numbers we previously obtained by 2 hybrid methods on the same patient cohort.

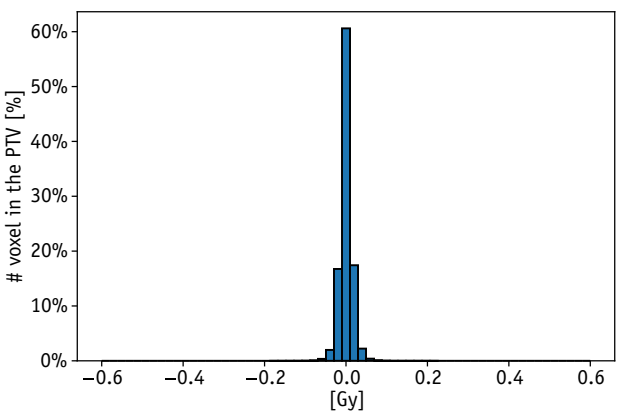


Fig. 5. Normalized distribution of dosimetric error computed within the planning target volume. The error was computed by performing a dose difference between computed tomography and synthetic computed tomography on a voxel-by-voxel basis and normalized by the total number of voxels in the planning target volume. Each histogram bin represents 0.02 Gy.

Aside from the small number of patients used to train the network, our results represent a significant improvement in the estimation of MAE compared with the current state of the art. It is possible that results can be improved by enlarging the training size. However, it is very difficult to compare methods because of the different metrics and procedures used to assess the sCT estimations algorithms, which are very dependent on the analyzed region (entire image FOV vs patient contour or smaller regions of interest) and the interscan time between MRI and CT (see supplementary material; available online at https://doi.org/10.1016/j.ijrobp.2019.06.2535). For instance, the presence of metal inserts (eg, dental fillings) and changes in anatomy between the 2 acquisitions can overestimate the error. The latter aspect is particularly relevant for air cavities (nasal cavities, mouth, pharynx), which are prone to intra- and interscan deformation, so that interscan interval plays a major role in image comparison. Ideally, the 2 scans should be acquired within a few hours for fair comparison. However, in many clinics this is not possible because the MRI scanner is located in a different department or because the clinical workflow does not allow close acquisitions. Finally, the DSC metric, which is usually used to assess geometric

Table 2 Comparison in terms of MAE computed within the skin contour in recent works implementing deep learning techniques for sCT generation

Method	MRI sequence	FOV (pixel) spatial resolution (mm)	No. of patients	Testing	MAE (HU) Mean ± standard deviation
Multiplane CNN	T1	256 × 256 × 176 1 × 1 × 1	15	LOOCV	54 ± 7
Dinkla et al[15]	T1	288 × 288 × 183 1.1 × 1.1 × 1	52	2-fold cross validation	67 ± 11
Wolterink[11]	T1	288 × 288 × 183 0.87 × 0.87 × 1	24	18 training and 6 testing	74 ± 2
Xiang et al[13]	T1	234 × 181 × 149 1.2 × 1.2 × 1	16	LOOCV	85 ± 9
Han[12]	T1	256 × 256 × 160 1 × 1 × 1	18	6-fold cross validation	85 ± 17
Emami et al[14]	T1	NA 0.9 × 0.9 × 1.25	15	4-fold cross validation	89 ± 10
Speier et al[26]	T1, T2 T1	256 × 256 × 176 1 × 1 × 1	15	LOOCV	73 ± 2 83 ± 6
Pileggi et al[10]	T1, T2 T1	256 × 256 × 176 1 × 1 × 1	15	LOOCV	118 ± 10 126 ± 9

Abbreviations: CNN = convolutional neural network; FOV = field of view; HU = Hounsfield units; LOOCV = leave-one-out cross validation; MAE = mean absolute error; MRI = magnetic resonance imaging; sCT = synthetic computed tomography.
In the last 2 rows of the table, results of 2 hybrid methods tested on the same patient cohort are also reported.

agreement (especially for bony anatomy) is dependent on the HU threshold chosen to segment, which sometimes is not even reported. In proton therapy, in this work, we partly addressed the issue of robust sCT versus GT CT comparison by proposing the following workflow for the brain district:

- compute MAE and other metrics inside the patient skin and from the top of the head to the base of skull to avoid major metal artifacts due to dental fillings (which are very common) and interscan position variation of the mouth and neck (see supplementary material; available online at https://doi.org/10.1016/j.ijrobp.2019.06.2535);
- perform a tissue-by-tissue analysis by segmenting soft tissue on the MRI and by thresholding air and bone on the CT. For air we also propose to select air cavities that are less prone to variations over days, such as frontal, maxillary, and sphenoidal sinus. It is very important to avoid the nares and pharynx because of different fillings and position of the tongue and to avoid the ear cavities, in which the small bone structures generate partial volume effects with air at standard CT resolution;
- analyze performances on bone tissue as a function of HU thresholding. Small metal clips that do not generate artifacts can be included in the analysis because they do not interfere too much. However, if there are large implants or metal screws through the skull, these should be removed from the analysis to avoid overestimating the error.
- perform range shift analysis when proton planning has to be tested. This method has been shown to be more sensitive than dose difference to HU inaccuracies and high step gradients.[10,25]

The dosimetric analysis conducted on our patient cohort showed very promising results in light of MRI-only–based proton therapy. The ability to accurately predict air and bone density from T1-weighted MRI led to accurate dose planning also in regions where the density step is very high (as behind the frontal sinus). A limitation of this study is that, because the tumor did not physically exist behind the frontal sinus, the tumor was assumed to have the properties of healthy brain, which has lower density than that of a normal brain tumor. The work was mainly aimed at testing the prediction of the air/bone interface with an MRI-to-CT conversion, where proton energy deposit is very sensitive to the high-density gradient.

Conclusions

The major advantage of our method is its ability to accurately predict shape and density of the tissues from air (0.001 g/cm^3) to bone (2 g/cm^3), which includes tumor densities that can vary between 1.0 g/cm^3 and 1.2 g/cm^3. However, a weakness of the present method is that the algorithm requires a paired set of CT and MRI scans for training; these scans should be taken under identical conditions (patient position and immobilization) and on the same day to minimize conversion errors.

References

1. Pollard JM, Wen Z, Sadagopan R, et al. The future of image-guided radiotherapy will be MR guided. *Br J Radiol* 2017;90:20160667.
2. Arivarasan I, Anuradha C, Subramanian S, et al. Magnetic resonance image guidance in external beam radiation therapy planning and delivery. *Jpn J Radiol* 2017;35:417-426.
3. Edmund JM, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol* 2017;12:28.
4. Johnstone E, Wyatt JJ, Henry AM, 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.
5. Owrangi AM, Greer PB, Glide-Hurst CK. MRI-only treatment planning: Benefits and challenges. *Phys Med Biol* 2018;63:05TR01.
6. Sjölund J, Forsberg D, Andersson M, et al. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Phys Med Biol* 2015;60:825-839.
7. Dowling JA, Lambert J, Parker J, 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* 2012;83:E5-E11.
8. Edmund JM, Kjer HM, Leemput K, et al. A voxel-based investigation for MRI-only radiotherapy of the brain using ultra short echo times. *Phys Med Biol* 2014;59:7501-7519.
9. Johansson A, Karlsson M, Nyholm T. CT substitute derived from MRI sequences with ultrashort echo time. *Med Phys* 2011;38:2708-2714.
10. Pileggi G, Speier C, Sharp GC, et al. Proton range shift analysis on brain pseudo-CT generated from T1 and T2 MR. *Acta Oncol* 2018;57:1521-1531.
11. Wolterink J, Dinkla A, Savenije M, et al. Deep MR to CT synthesis using unpaired data. *Simul Synth Med Imaging* 2017;14-23.
12. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys* 2017;44:1408-1419.
13. Xiang L, Wang Q, Nie D, et al. Deep embedding convolutional neural network for synthesizing CT image from T1-weighted MR image. *Med Image Anal* 2018;47:31-44.
14. Emami H, Dong M, Nejad-Davarani SP, et al. Generating synthetic CTs from magnetic resonance images using generative adversarial networks [e-pub ahead of print]. *Med Phys* https://doi.org/10.1002/mp.13047. Accessed August 6, 2019.
15. Dinkla AM, Wolterink JM, Maspero M, et al. MR-only brain radiotherapy: Dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network. *Int J Radiat Oncol Biol Phys* 2018;102:801-812.
16. Liney GP, Whelan B, Oborn B, et al. MRI-linear accelerator radiotherapy systems. *Clin Oncol (R Coll Radiol)* 2018;30:686-691.
17. Fischer-Valuck BW, Henke L, Green O, et al. Two-and-a-half-year clinical experience with the world's first magnetic resonance image guided radiation therapy system. *Adv Radiat Oncol* 2017;2:485-493.
18. Henke L, Kashani R, Robinson C, et al. Phase I trial of stereotactic MR-guided online adaptive radiation therapy (SMART) for the treatment of oligometastatic or unresectable primary malignancies of the abdomen. *Radiother Oncol* 2018;126:519-526.
19. Henke LE, Contreras JA, Green OL, et al. Magnetic resonance image-guided radiotherapy (MRIgRT): A 4.5-year clinical experience. *Clin Oncol (R Coll Radiol)* 2018;30:720-727.
20. Raaymakers BW, Jürgenliemk-Schulz IM, Bol GH, et al. First patients treated with a 1.5 T MRI-Linac: Clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment. *Phys Med Biol* 2017;62:L41-L50.
21. Fuchs H, Moser P, Gröschl M, et al. Magnetic field effects on particle beams and their implications for dose calculation in MR-guided particle therapy. *Med Phys* 2017;44:1149-1156.
22. Oborn BM, Dowdell S, Metcalfe PE, et al. Future of medical physics: Real-time MRI-guided proton therapy. *Med Phys* 2017;44:e77-e90.
23. Padilla-Cabal F, Georg D, Fuchs H. A pencil beam algorithm for magnetic resonance image-guided proton therapy. *Med Phys* 2018;45:2195-2204.
24. Schellhammer SM, Hoffmann AL. Prediction and compensation of magnetic beam deflection in MR-integrated proton therapy: A method optimized regarding accuracy, versatility and speed. *Phys Med Biol* 2017;62:1548-1564.
25. Seco J, Panahandeh HR, Westover K, et al. Treatment of non-small cell lung cancer patients with proton beam-based stereotactic body radiotherapy: Dosimetric comparison with photon plans highlights importance of range uncertainty. *Int J Radiat Oncol Biol Phys* 2012;83:354-361.
26. Speier C, Pileggi G, Izquierdo-Garcia D, et al. Advanced multimodal methods for cranial pseudo-CT generation validated by IMRT and VMAT radiation therapy plans. *Int J Radiat Oncol Biol Phys* 2018;102:792-800.
27. Izquierdo-Garcia D, Hansen AE, Forster S, et al. An SPM8-based approach for attenuation correction combining segmentation and nonrigid template formation: Application to simultaneous PET/MR brain imaging. *J Nucl Med* 2014;55:1825-1830.
28. Chen KT, Izquierdo-Garcia D, Poynton CB, et al. On the accuracy and reproducibility of a novel probabilistic atlas-based generation for calculation of head attenuation maps on integrated PET/MR scanners. *Eur J Nucl Med Mol Imaging* 2017;44:398-407.
29. Zaffino P, Raudaschl P, Fritscher K, et al. Technical Note: Plastimatch mabs, an open source tool for automatic image segmentation. *Med Phys* 2016;43:5155.
30. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2016. arXiv:1409.1556v6.
31. Huang Y, Parra LC. Fully automated whole-head segmentation with improved smoothness and continuity, with theory reviewed. *PLoS One* 2015;10:e0125477.
32. Ashburner J, Friston K. Segmentation. In: Penny W, Friston K, Ashburner J, Kiebel S, Nichols T eds. 1st Statistical Parametric Mapping: Great Britain Academic Press; 2007. p. 81-91.
33. White DR, Griffith RV, Wilson IJ. Photon, electron, proton and neutron interaction data for body tissues. *ICRU Report* 1992;46:24.
34. Wieser HP, Cisternas E, Wahl N, et al. Development of the open-source dose calculation and optimization toolkit matRad. *Med Phys* 2017;44:2556-2568.
35. Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. *Phys Med Biol* 2012;57:R99-R117.
36. Walker A, Liney G, Metcalfe P, et al. MRI distortion: Considerations for MRI based radiotherapy treatment planning. *Australas Phys Eng Sci Med* 2014;37:103-113.
37. Kang G, Liu K, Hou B, et al. 3D multi-view convolutional neural networks for lung nodule classification. *PLoS One* 2017;12:e0188290.

S1 Network architecture and training

The architecture of the CNN used in this work is depicted in figure S1 and it is similar to the U-net variant proposed by Han [12]. The network, structured on 5 levels, can be mainly split in two mirrored parts: the encoding and the decoding branch. In the encoding path the aim is to extract the most representative features from the $T1_{MR}$, meanwhile in the decoding part the identified features are mapped back to generate a sCT.
In the encoding part the number of filters, as well as the amount of convolutional layers, increase by going deeper; at the opposite, in the decoding side, they decrease as close the output layer approaches. As it is possible to see from figure S1, the top level is made of 64 filters and 2 convolutions, versus 512 and 3 of the bottom one.
When a level change happens, a max pool/upsample layer is used to reduce/expand the feature map for encoding and decoding respectively.
All convolutions are executed by using a 3x3 kernel and a leaky rectify activation function except for the last one where, in order to collapse the multidimensional feature map into a single channel (that is the 2D sCT slice), kernel size equal to 1 and linear activation function are used. Batch normalization is always adopted.
Finally, in order to keep and preserve the high resolution features from the encoding side of the network, they are directly reported to the decoding part and provided as additional input. This linkage is named “direct connection”.
Mean absolute error is used as cost function to train the network and L1 regularization is adopted to prevent overfitting.

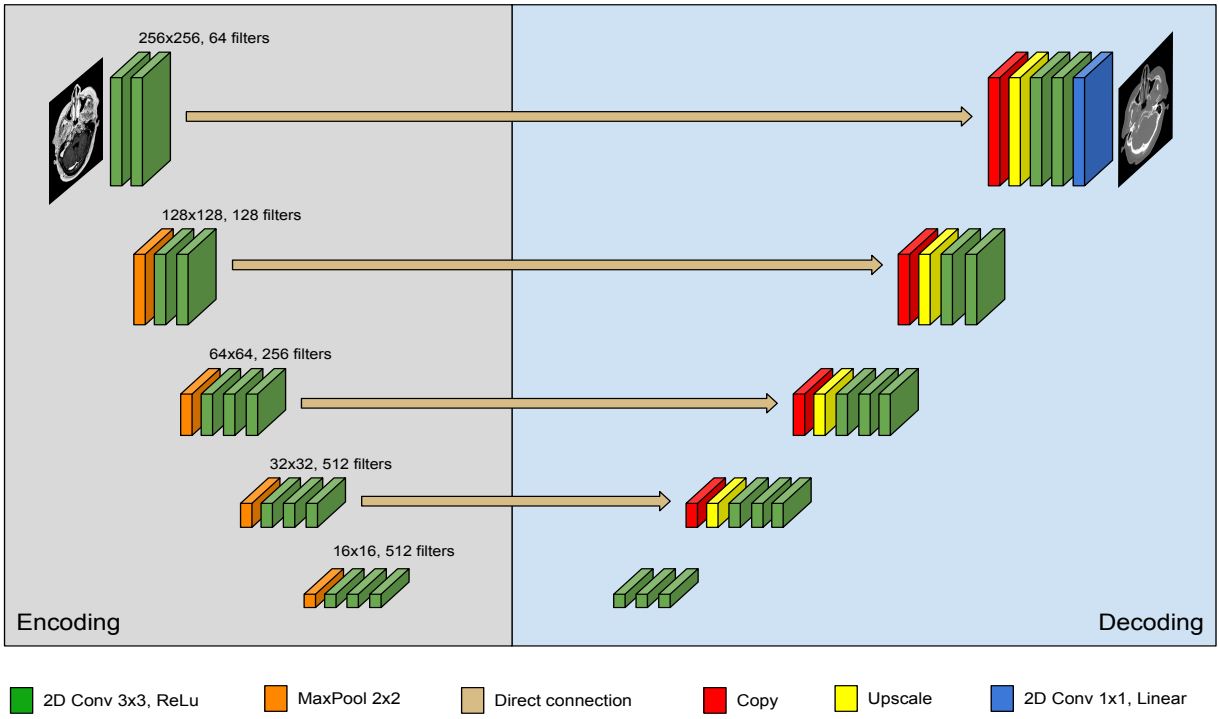


Figure S1. Architecture of the U-net used in this work

The upper panel of figure S2 shows our first attempt to introduce multiplane CNN for sCT prediction. The axial, sagittal and coronal images were introduced simultaneously and in shuffled order to train a single CNN. The batch size was set to 18, to allow sufficient computational performance on the hardware used in this study. By observing the results, we noticed that the sagittal view was statistically worse (p=0.05, one-way Anova). Our hypothesis was that the number of informative slices in the sagittal view is different compared to the other two views because the head is narrower in the latero-lateral direction and broader in the antero-posterior. Moreover, sagittal slices are symmetrical to the central sagittal, so for this view there is less variability than the others. One view may converge earlier if trained separately.

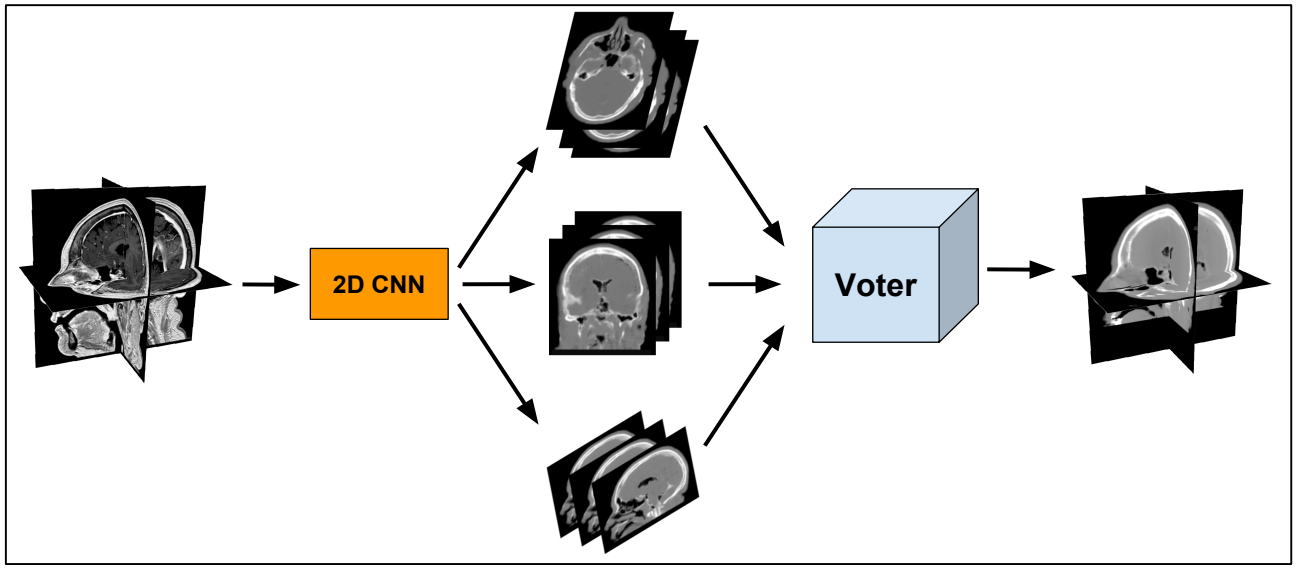


Figure S2. First approach for multiplane training and voting

So, we decided to carry out independent training, as depicted in figure S3. This allowed us to parallelize the process and, as a consequence, to decrease the batch size to 1. Results confirmed the improvement in image conversion by adopting this approach.

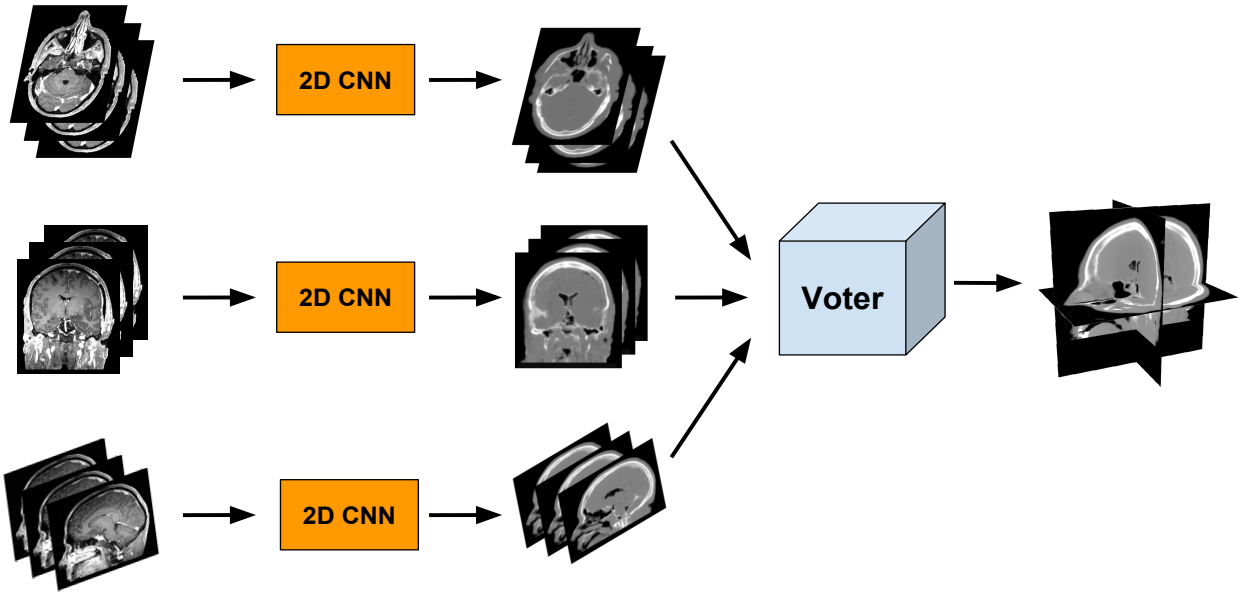
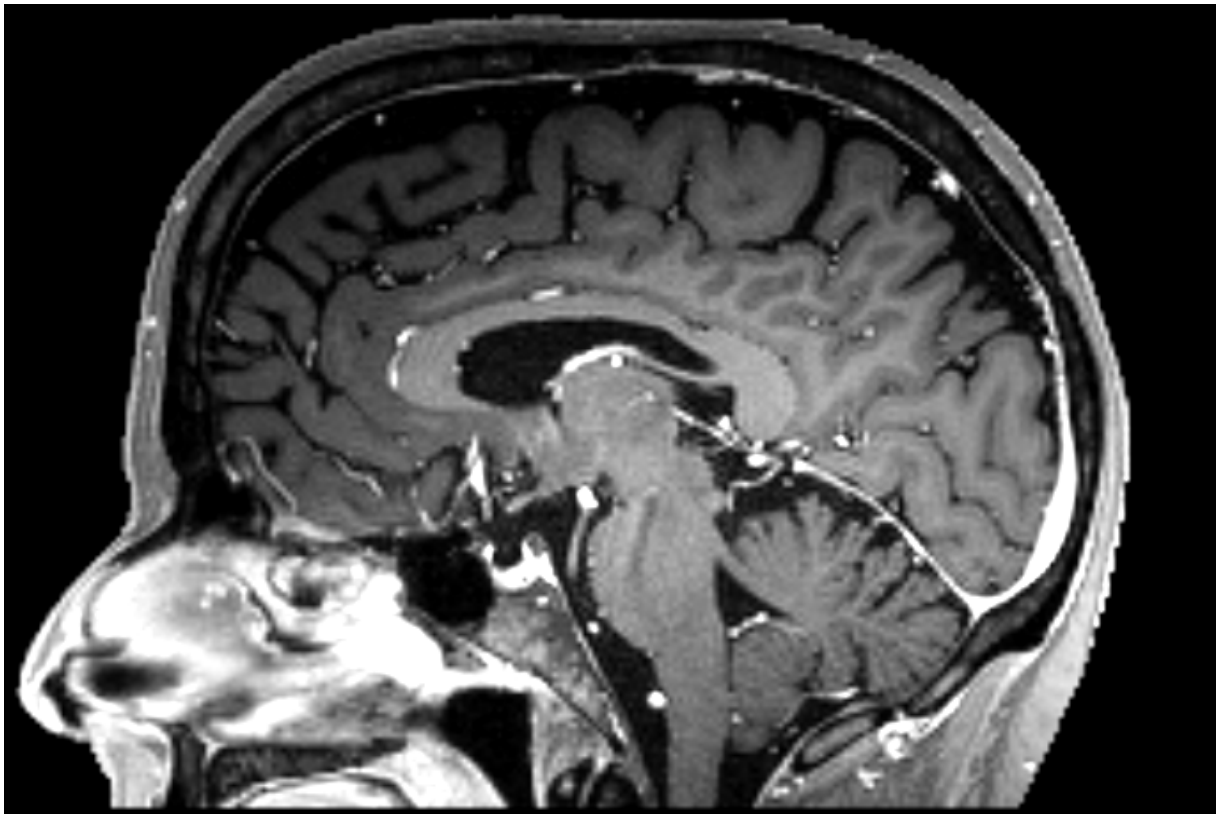


Figure S3. Adopted approach for multiplane training and voting

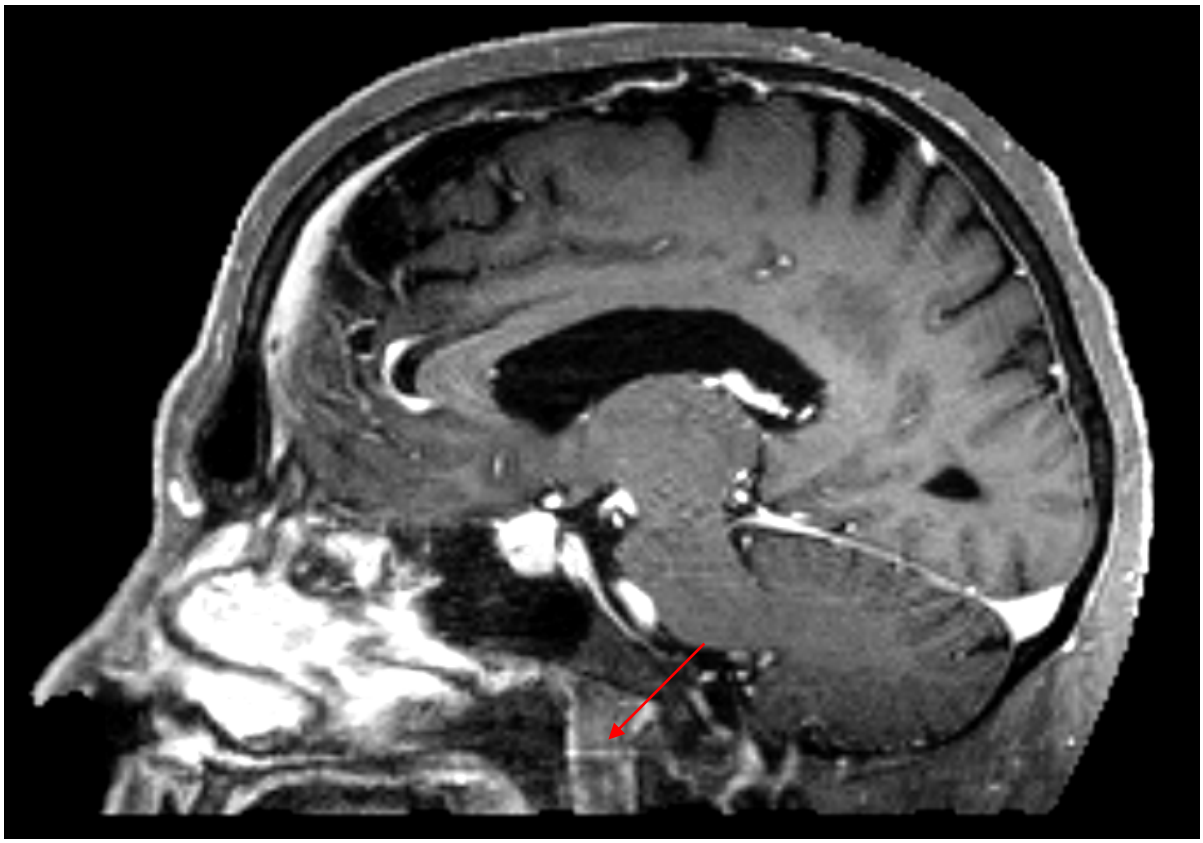
S2 Impact of image quality on sCT

In the case of patient 14 (worst case) we noticed motion artifacts in the facial regions of $T1_{MR}$. The difference in image quality between patient 11 and patient 14 can be appreciated in figure S5.

Patient #11



Patient #14



Patient #14

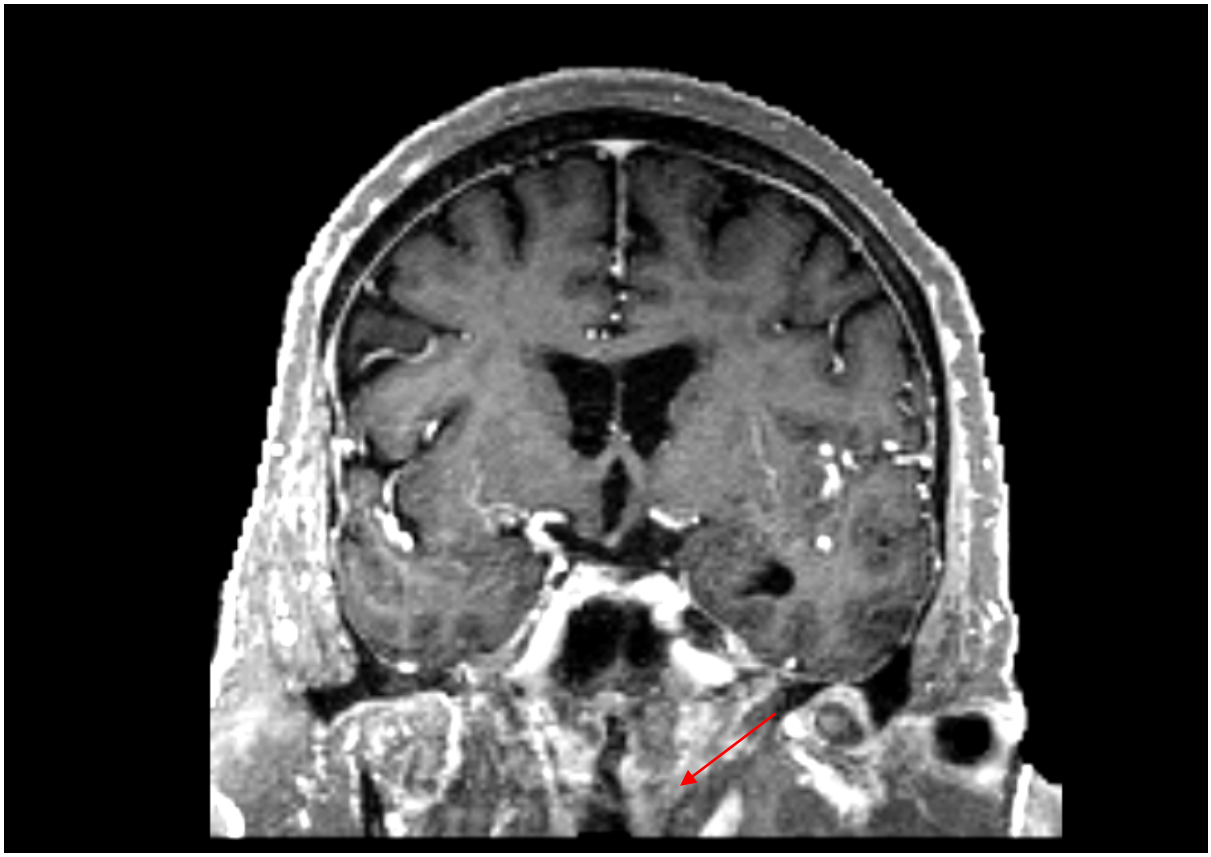


Figure S4. Visual comparison between patient 11 (upper panel) and 14 (middle and lower panel). Patient 14 presented some motion artifacts clearly visible in the mouth and pharynx. This is in agreement with the inaccuracies visible in figure 1 of the manuscript.

S3 Air and bone segmentation

In figure S6 an example of sinus and bone segmentation (patient #12) is showed. Table S1 reports DSC results for each patient.

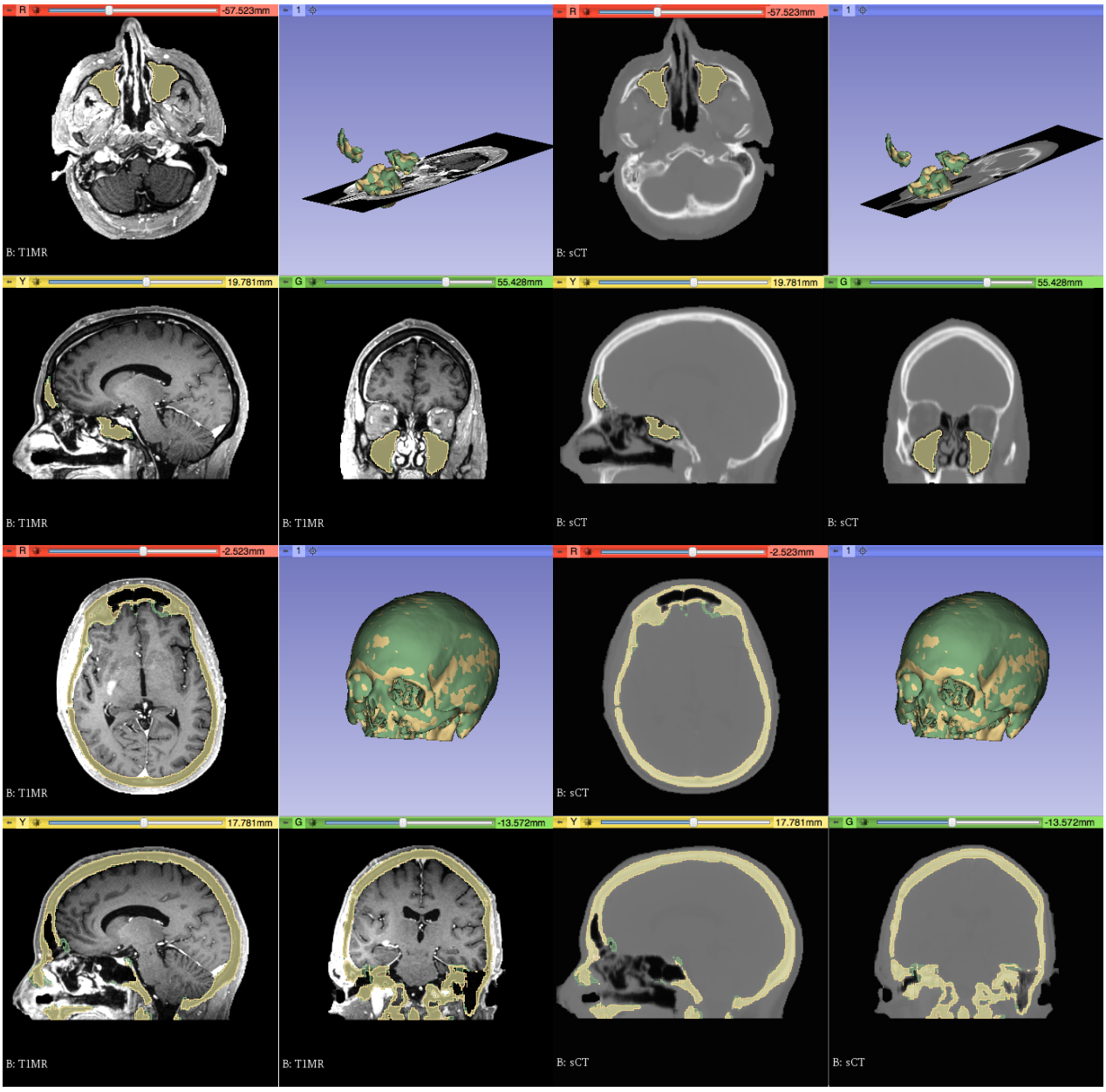


Figure S5. Sinus cavities and bone structures obtained by thresholding the CT (green labels) and the sCT (yellow labels). Results are overlaid both on $T1_{MR}$ in order to show the high accuracy in converting similar MR intensities to air and bone and on sCT to show the quality of estimation.

Table S1. DSC for bone and air cavities for each patient

Patient	DSC Bone	DSC Air
1	0.94	0.93
2	0.93	0.91
3	0.94	0.91
4	0.94	0.92
5	0.94	0.94
6	0.92	0.93
7	0.93	0.94
8	0.93	0.94
9	0.93	0.95
10	0.95	0.90
11	0.94	0.96
12	0.95	0.93
13	0.96	0.83
14	0.87	0.86
15	0.92	0.94

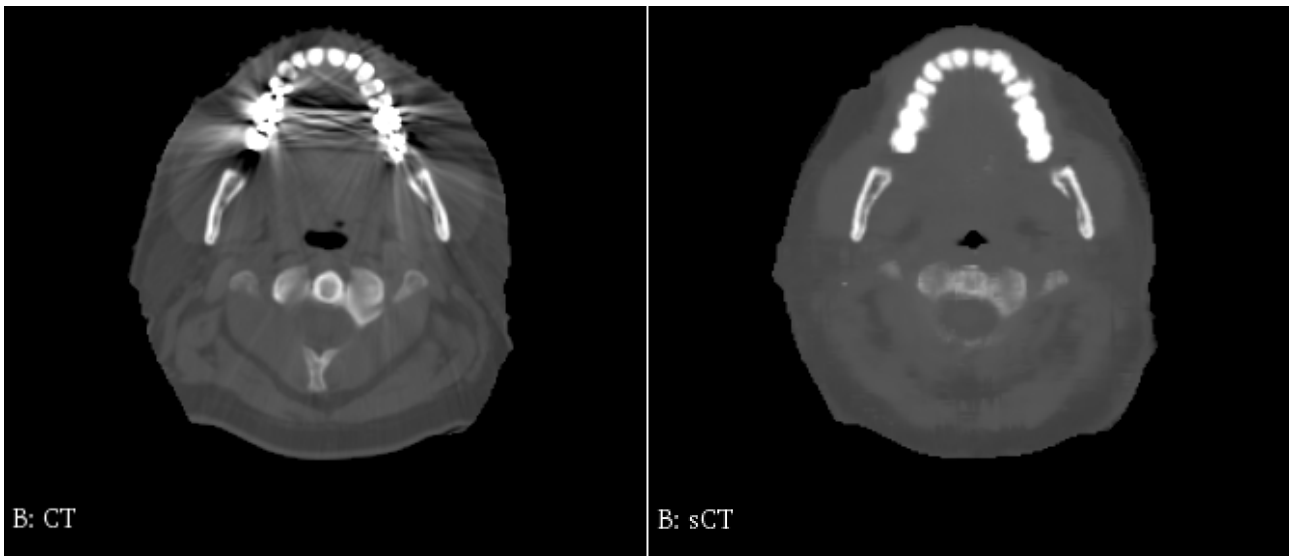
S4 Impact of metal inserts and mis-registration

The motivations for assessing the sCT conversion excluding mouth and neck are due to the presence of metal artifacts in the CT in the oral cavity (patients #1, 2, 3, 4, 5, 11, 15) and slightly different position of the neck between $T1_{MR}$ and CT acquisitions (patients # 1, 2, 3, 6, 12, 14). Table S2 reports the MAE error excluding and including mouth and neck and the interscan interval In figure S7, three examples of the axial view of teeth in the CT vs, sCT are showed. Interestingly, the DCNN method was able to predict the oral cavity anatomy without streak artifacts.
In figure S8, an example of neck misregistration is showed for patient #1.

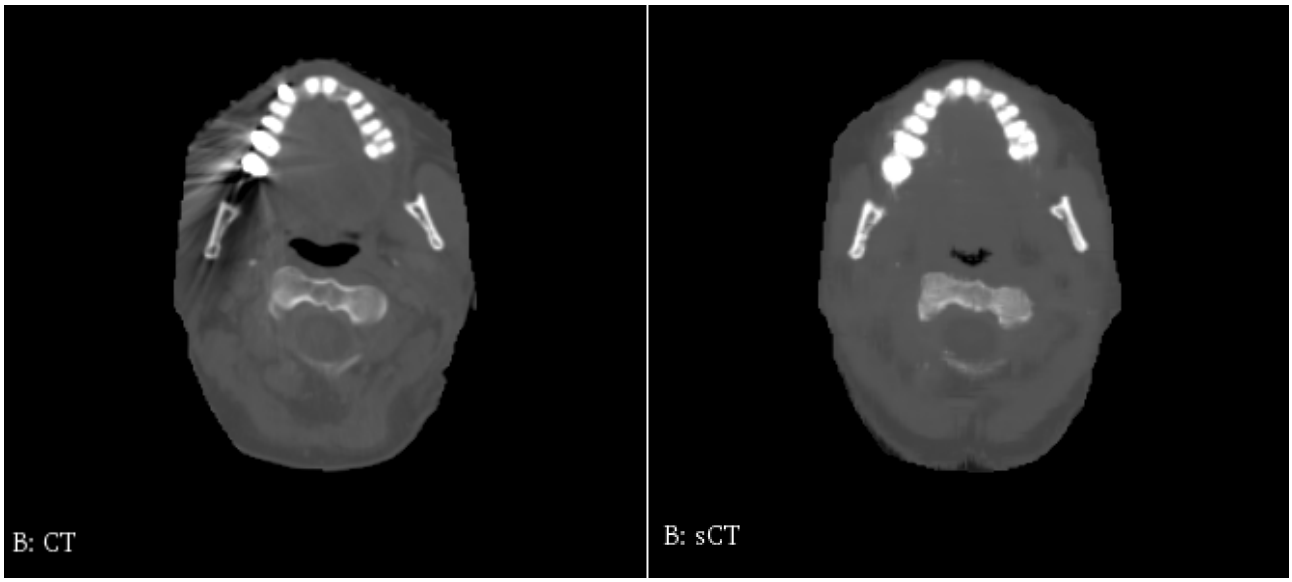
Table S2. Comparison in terms of MAE considering the whole field of view and excluding the neck and the mouth. Patients # 7, 9, 10, and 12 there is no difference because the FOV was already cropped at the base of skull.

Patient #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Crop at base of skull	52	54	60	51	49	64	50	50	53	45	45	49	64	66	55
Including mouth and teeth	61	58	66	55	51	81	50	52	53	45	47	49	64	80	60
Interscan interval (days)	11	15	10	11	11	12	16	20	15	11	19	14	17	16	16

Patient #1



Patient #6



Patient #15

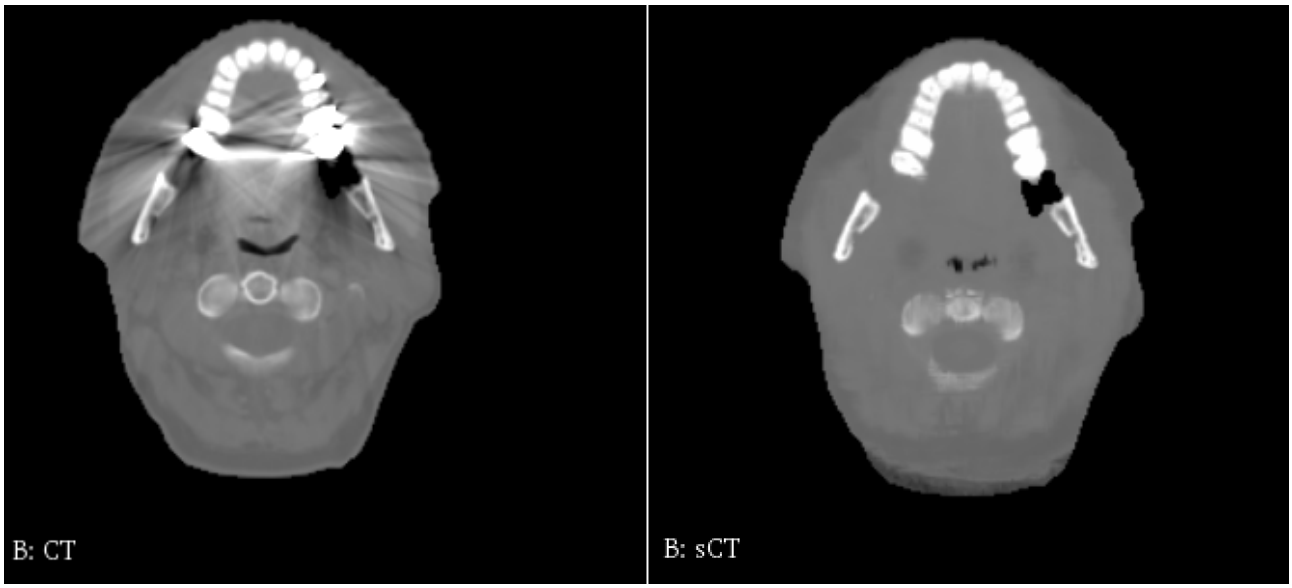


Figure S6. Effects of metal artifacts on reference CT

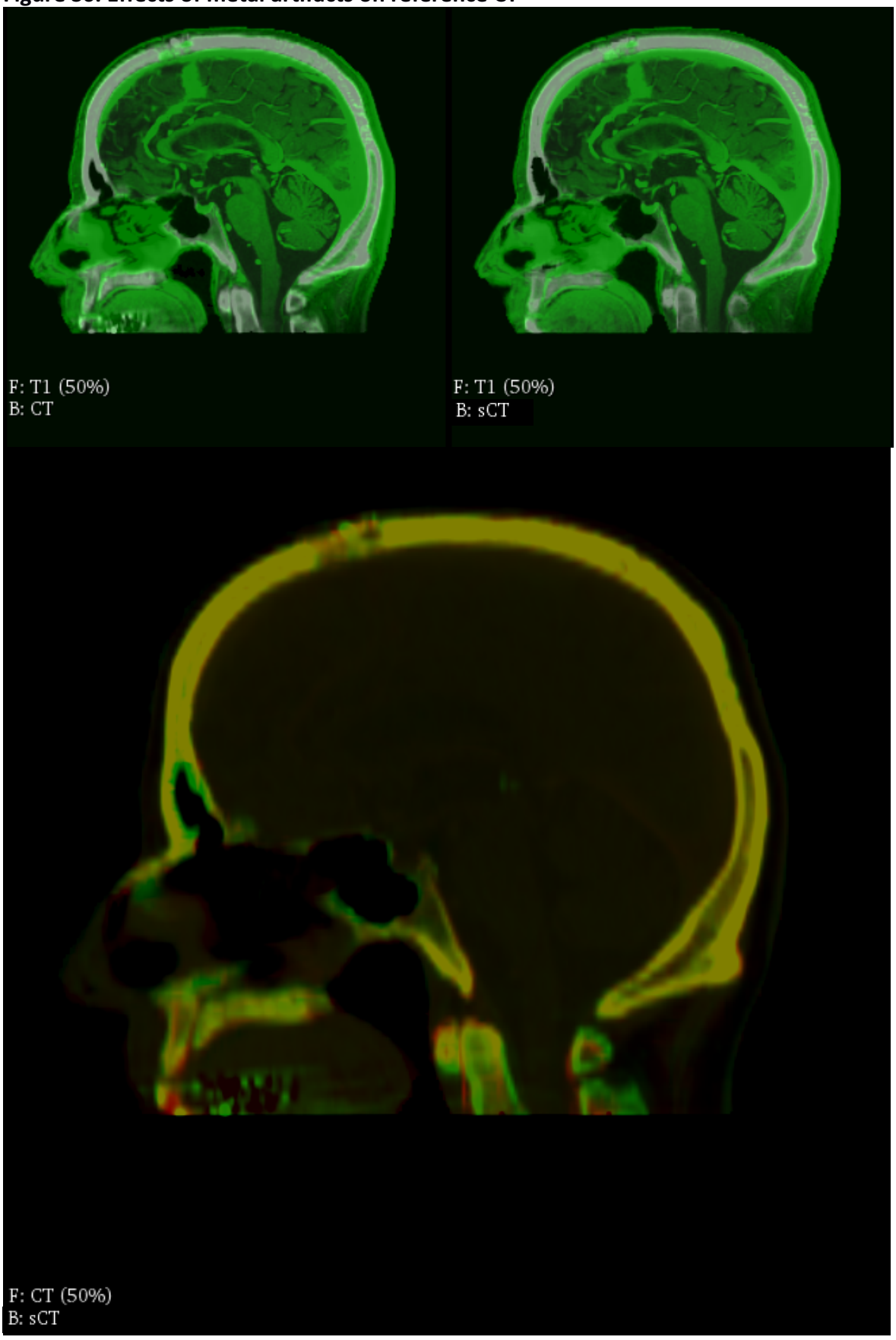


Figure S7. Inaccuracies between CT and $T1_{MR}$ for patient 1 in the neck region. While sCT geometry is consistent with $T1_{MR}$, CT and $T1_{MR}$ were not perfectly matching, because only rigid registration was performed between the two datasets. So errors differences between sCT and CT are visible.

S5 Relative range shift results

In table S3 the RRS for each analyzed slice are reported. All beams passed both MGH and HTPI acceptance tests.

Table S3. RRS slice by slice

	Relative RS											RRS test	
PAT	Beam	Slice 01	Slice 02	Slice 03	Slice 04	Slice 05	Slice 06	Slice 07	Slice 08	Slice 09	Slice 10	MGH	HTPI
1	0	2.22	1.64	1.07	1.01	1.01	0.83	0.80	0.74	0.26	0.53	✓	✓
	270	1.60	1.15	0.85	0.90	0.93	0.68	0.55	0.68	0.71	0.62	✓	✓
	315	0.21	0.20	0.15	0.20	0.29	0.27	0.22	0.17	0.13	0.09	✓	✓
2	0	1.36	0.95	1.06	1.12	1.02	0.86	0.46	0.04	0.57	0.73	✓	✓
	270	0.01	0.02	0.02	0.03	0.01	0.01	-0.01	-0.02	0.04	0.03	✓	✓
	315	0.13	0.14	0.12	0.12	0.03	-0.04	-0.02	0.06	0.01	-0.18	✓	✓
3	0	-1.05	-1.15	-0.57	-0.12	-0.35	-0.28	-0.08	0.16	0.29	0.78	✓	✓
	270	0.01	0.27	0.06	0.01	0.14	-0.04	0.00	0.03	-0.03	-0.02	✓	✓
	315	1.60	1.55	1.51	1.38	1.13	0.93	0.79	0.71	0.55	0.40	✓	✓
4	0	-1.51	-1.35	-0.19	1.29	-0.23	-2.22	-1.47	-0.62	-0.03	0.39	✓	✓
	270	0.49	0.62	0.65	0.63	0.58	0.51	0.52	0.44	0.45	0.45	✓	✓
	315	1.26	1.40	1.57	1.63	1.64	1.55	1.50	1.45	1.38	1.38	✓	✓
5	0	0.20	-1.16	-1.12	0.39	0.74	0.60	0.47	0.50	0.49	0.57	✓	✓
	270	0.38	0.31	0.63	0.71	0.62	0.50	0.48	0.50	0.54	0.51	✓	✓
	315	1.05	1.08	1.14	1.12	1.12	1.11	1.02	0.97	0.88	0.79	✓	✓
6	0	-0.58	-0.54	-2.08	-1.34	-2.79	-4.16	-3.76	-3.40	-2.93	-2.23	✓	✓
	270	-1.17	-1.32	-1.25	-1.06	-1.07	-1.29	-1.42	-1.41	-1.63	-1.57	✓	✓
	315	-0.63	-0.62	-0.64	-0.94	-1.08	-1.06	-1.05	-1.03	-1.06	-1.15	✓	✓
7	0	1.04	0.58	-0.09	-0.58	-0.88	-0.99	-0.83	-0.60	-0.37	0.02	✓	✓
	270	0.14	0.11	0.08	0.06	0.05	-0.06	-0.11	-0.02	0.01	0.08	✓	✓
	315	-0.43	-0.36	-0.19	-0.16	0.00	0.10	0.17	0.21	0.21	0.22	✓	✓
8	0	1.76	2.04	2.51	2.31	1.37	0.46	0.00	-0.08	0.04	0.58	✓	✓
	270	-0.51	-0.41	-0.36	-0.27	-0.12	-0.21	-0.32	-0.36	-0.35	-0.31	✓	✓

	315	0.02	0.03	0.03	0.03	0.00	0.03	0.12	0.23	0.36	0.30	✓	✓
9	0	-0.35	-0.35	-0.29	-0.13	-0.05	0.03	0.08	0.05	0.09	0.03	✓	✓
	270	0.48	0.59	0.46	0.25	0.19	0.29	0.23	0.20	0.23	0.32	✓	✓
	315	-0.19	0.05	0.20	0.35	0.44	0.45	0.51	0.57	0.65	0.70	✓	✓
10	0	0.75	0.57	0.68	0.60	0.39	0.12	0.02	0.15	0.16	-0.06	✓	✓
	270	-0.12	-0.19	-0.25	-0.17	-0.10	0.01	0.13	-0.06	-0.08	0.04	✓	✓
	315	-0.19	0.05	0.20	0.35	0.44	0.45	0.51	0.57	0.65	0.70	✓	✓
11	0	0.35	0.28	0.95	0.64	0.46	0.51	0.52	0.42	0.37	0.19	✓	✓
	270	-0.57	-0.58	-0.42	-0.51	-0.45	-0.44	-0.39	-0.27	-0.16	-0.02	✓	✓
	315	-0.05	-0.15	-0.22	-0.24	-0.25	-0.32	-0.35	-0.21	-0.18	-0.19	✓	✓
12	0	-3.22	-3.22	-3.30	-3.52	-3.47	-2.79	-1.97	-1.41	-1.21	-1.39	✓	✓
	270	1.47	1.44	1.41	1.34	1.31	1.31	1.28	1.26	1.29	1.12	✓	✓
	315	0.23	0.23	0.25	0.32	0.39	0.47	0.55	0.57	0.58	0.58	✓	✓
13	0	1.41	2.78	2.13	1.38	1.31	0.90	0.04	-0.38	-0.16	0.34	✓	✓
	270	0.27	0.40	0.39	0.34	0.36	0.23	0.14	0.19	0.37	0.35	✓	✓
	315	-0.41	-0.24	-0.22	0.02	0.04	0.11	0.21	0.30	0.31	0.40	✓	✓
14	0	-2.93	-2.93	-1.82	-1.49	-1.88	-1.79	-1.73	-1.85	-1.69	-1.59	✓	✓
	270	-1.19	-1.14	-1.04	-0.86	-0.63	-0.61	-0.66	-0.76	-0.66	-0.84	✓	✓
	315	-0.51	-0.51	-0.50	-0.50	-0.51	-0.51	-0.43	-0.40	-0.47	-0.55	✓	✓
15	0	2.44	1.40	0.77	0.48	0.38	0.32	0.25	0.38	0.30	0.19	✓	✓
	270	0.96	0.82	0.71	0.46	0.24	0.38	0.48	0.39	0.23	0.23	✓	✓
	315	0.18	0.19	0.22	0.34	0.39	0.47	0.49	0.52	0.49	0.50	✓	✓