NeuroImage 222 (2020) 117221

Contents lists available at ScienceDirect

NeuroImage

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

AI-driven attenuation correction for brain PET/MRI: Clinical evaluation of a dementia cohort and importance of the training group size

Claes Nøhr Ladefoged [1,*], Adam Espe Hansen [1], Otto Mølby Henriksen [1], Frederik Jager Bruun [1], Live Eikenes [2], Silje Kjærnes Øen [2], Anna Karlberg [2,3], Liselotte Højgaard [1], Ian Law [1], Flemming Littrup Andersen [1]

[1] Department of Clinical Physiology, Nuclear Medicine & PET, Rigshospitalet, University of Copenhagen, Denmark
[2] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway
[3] c Department of Radiology and Nuclear Medicine, St. Olavs hospital, Trondheim University Hospital, Trondheim, Norway

ARTICLE INFO

Keywords:
Attenuation correction
Deep learning
Convolutional neural network
Artificial intelligence
Brain
PET/MRI

ABSTRACT

Introduction: Robust and reliable attenuation correction (AC) is a prerequisite for accurate quantification of activity concentration. In combined PET/MRI, AC is challenged by the lack of bone signal in the MRI from which the AC maps has to be derived. Deep learning-based image-to-image translation networks present itself as an optimal solution for MRI-derived AC (MR-AC). High robustness and generalizability of these networks are expected to be achieved through large training cohorts. In this study, we implemented an MR-AC method based on deep learning, and investigated how training cohort size, transfer learning, and MR input affected robustness, and subsequently evaluated the method in a clinical setup, with the overall aim to explore if this method could be implemented in clinical routine for PET/MRI examinations.

Methods: A total cohort of 1037 adult subjects from the Siemens Biograph mMR with two different software versions (VB20P and VE11P) was used. The software upgrade included updates to all MRI sequences. The impact of training group size was investigated by training a convolutional neural network (CNN) on an increasing training group size from 10 to 403. The ability to adapt to changes in the input images between software versions were evaluated using transfer learning from a large cohort to a smaller cohort, by varying training group size from 5 to 91 subjects. The impact of MRI sequence was evaluated by training three networks based on the Dixon VIBE sequence (DeepDixon), T1-weighted MPRAGE (DeepT1), and ultra-short echo time (UTE) sequence (DeepUTE). Blinded clinical evaluation relative to the reference low-dose CT (CT-AC) was performed for DeepDixon in 104 independent 2-[^{18}F]fluoro-2-deoxy-D-glucose ([^{18}F]FDG) PET patient studies performed for suspected neurodegenerative disorder using statistical surface projections.

Results: Robustness increased with group size in the training data set: 100 subjects were required to reduce the number of outliers compared to a state-of-the-art segmentation-based method, and a cohort >400 subjects further increased robustness in terms of reduced variation and number of outliers. When using transfer learning to adapt to changes in the MRI input, as few as five subjects were sufficient to minimize outliers. Full robustness was achieved at 20 subjects. Comparable robust and accurate results were obtained using all three types of MRI input with a bias below 1% relative to CT-AC in any brain region. The clinical PET evaluation using DeepDixon showed no clinically relevant differences compared to CT-AC.

Conclusion: Deep learning based AC requires a large training cohort to achieve accurate and robust performance. Using transfer learning, only five subjects were needed to fine-tune the method to large changes to the input images. No clinically relevant differences were found compared to CT-AC, indicating that clinical implementation of our deep learning-based MR-AC method will be feasible across MRI system types using transfer learning and a limited number of subjects.

1. INTRODUCTION

Positron emission tomography (PET) images need to be corrected for photon attenuation to accurately quantify the measured radioactive tissue concentration (Andersen et al., 2014; Dickson et al., 2014). In a dual modality PET and Magnetic Resonance Imaging (MRI) scan-

* Corresponding author.
E-mail address: claes.noehr.ladefoged@regionh.dk (C.N. Ladefoged).

https://doi.org/10.1016/j.neuroimage.2020.117221
Received 8 November 2019; Received in revised form 15 July 2020; Accepted 28 July 2020
Available online 1 August 2020

ner, a density map for attenuation correction (AC) has to be derived from the MRI. This was initially not possible, which hampered the use of PET/MRI scanners, especially for brain studies in both clinical and research applications (Vandenberghe and Marsden, 2015). Several MRI-guided attenuation correction techniques were proposed as potential solutions (Chen and An, 2017; Izquierdo-garcia and Catana, 2016; Ladefoged et al., 2016; Mehranian et al., 2016). Eleven state-of-the-art AC-methods were studied in a large cohort of adult subjects with normal anatomy (Ladefoged et al., 2016), which concluded that AC was a solved topic in the brain when using one of the best performing methods. However, some of these methods, including our own segmentation-based RESOLUTE method (Ladefoged et al., 2015), were later found to be sensitive to specified MRI sequences, and, thus, vulnerable to system software updates.

Recently, artificial intelligence (AI) with deep learning convolutional neural networks (CNN) is being considered as an alternative, as they offer a number of advantages over the existing methods. Deep learning methods can confer robustness towards changes to the input caused by MRI hardware or system software updates, as well as cross platform compatibility between vendors through the process of transfer learning. Furthermore, methods based on CNNs are usually very processing intensive at the training step, but the generation of an attenuation map for a given subject occurs within seconds, making them attractive tools as a part of a clinical workflow where speed, accuracy, and robustness are key elements.

Since the first use of deep learning to convert MR images to CT (Han, 2017), numerous methods have been proposed, see e.g. (Teuho et al., 2020; Torrado-Carvajal, 2020)). Several state-of-the-art networks were employed, from traditional encoder-decoder architectures (Gong et al., 2018; Han, 2017; Torrado-Carvajal et al., 2019), to generative adversarial networks (GANs) (Arabi et al., 2019; Kazemifar et al., 2019), including variants accepting unpaired data (Ge et al., 2019; Lei et al., 2020; Wolterink et al., 2017; Yang et al., 2018). Most methods from the literature use small training group sizes (<30) even though larger sizes could increase generalizability and robustness. The methods are based on single or multiple MRI sequences, spanning the common T1-weighted MPRAGE as well as specialized sequences capable of visualizing bone such as zero echo time (ZTE) or ultra-short echo time (UTE). The possible advantages in the context of attenuation correction, especially in terms of robustness, from using large training cohorts as well as specialized sequences over traditional sequences remain to be thoroughly investigated systematically.

Recently, methods converting non-attenuation corrected (NAC) PET images directly to attenuation and scatter corrected PET images have emerged, mainly targeted for whole-body applications as paired data are readily available in large numbers (Arabi et al., 2020; Shiri et al., 2019; Van Hemmen et al., 2019; Yang et al., 2019). The drawbacks of these methods are their dependence towards choice of tracer and limited ability to extract structural information (Arabi et al., 2020). In the brain, the performance of these new methods remains to be evaluated thoroughly on a cohort with neurologic abnormalities.

The aim of this study was to implement a deep learning CNN for clinical MR-AC use, and investigate the potential impact on the quantitative accuracy and clinical reading of PET scans depending on training group size and choice of MRI input. This was achieved by utilizing a large cohort of subjects all examined on the same PET/MRI from two independent sites including common and specialized MRI, as well as low-dose CT images used as reference.[1]

[1] Code and models for inference available at https://github.com/CAAI/DeepMRAC

2. MATERIALS AND METHODS

The data included comprised studies acquired on two Siemens Biograph mMR systems (Siemens Healthineers, Erlangen, Germany) spanning two different software versions. A larger cohort, imaged with software version VB20P, was used to investigate the impact of cohort size. A smaller cohort, with the most recent software update (VE11P), was used to investigate the effect of transfer learning (based on VB20P data), impact of choice of MRI input, and to perform a clinical evaluation.

2.1. Patients

Data sets from 1037 adult subjects were obtained retrospectively from two different centers; n = 1007 from Rigshospitalet, University Hospital Copenhagen, Denmark, and n = 30 from St. Olavs hospital, Trondheim University Hospital, Norway. Rigshospitalet provided data sets from the complete cohort of subjects referred for a PET/MRI brain examination with matching same-day head CT between November 2013 and April 2019, examined with software version VB20P (n = 811) or VE11P (n = 196). Data comprised PET/MRI studies imaged with various tracers, but only the MRI sequences were used to develop the method. The subjects included from St. Olavs hospital were referred to a clinical 2-[^{18}F]fluoro-2-deoxy-D-glucose ([^{18}F]FDG) PET/MRI brain examination for dementia, all examined with VE11P, and had matching same-day head CT. Retrospective use of subjects from Rigshospitalet was approved by the Danish Patient Safety Authority (ref. 3–3013–1513/1). The study from St. Olavs hospital was approved by the Regional Committee for ethics in Medical Research (REC Central) (ref. 2013/1371) and all subjects gave written informed consent. Data were extracted only in fully anonymized form in compliance to The European General Data Protection Regulation (GDPR).

In each of the two groups (VB20P and VE11P), we divided the subjects into training, validation, and test cohorts. The train and validation cohorts were used to develop the method. The subjects in the independent test cohort were all imaged with [^{18}F]FDG; none of these subjects had e.g. bone modifying cranio-facial surgical interventions, cranial defects, hyperostoses, dysplasias, disfigurement or metal implants besides dental implants. For the VB20P group, the test cohort was identical to the patients recently used in our multi-center evaluation (Ladefoged et al., 2016), and the train/validation split was done 70/30. We initially developed the models for the VE11P group using 4-fold cross validation. Once the models were finalized, we fixed the training/validation cohorts to be the first cross validation. The independent test cohort was prospectively acquired after the models were trained. An illustration of the splits for each group is shown in Fig. 1.

2.2. Imaging protocols

2.2.1. MRI

The scan protocols always included a T1-weighted (T1w) MPRAGE, a UTE AC sequence, and a Dixon-VIBE sequence (the vendor default for MR-AC). The upgrade to VE11P included upgraded versions to all three sequences. The UTE AC sequence was re-implemented, changing the relationship between the two echo images, with consequences especially to the signal in bone (Suppl. Fig. 1A). Visually, the most noticeable change was to the Dixon-VIBE sequence, which is now available in high-resolution, targeted for brain purposes (Suppl. Fig. 1B). No apparent differences could be observed for the T1-weighted MPRAGE sequence. Nevertheless, inspection of the area representing bone showed a slight decrease in mean value following the upgrade (Suppl. Fig. 1C). Sequence details are available in Table 1.

2.2.2. CT

A reference low-dose CT scan (120 kVp, 36–40 mAs, 0.6 × 0.6 × 3 mm^3 voxels) of the head using a PET/CT system

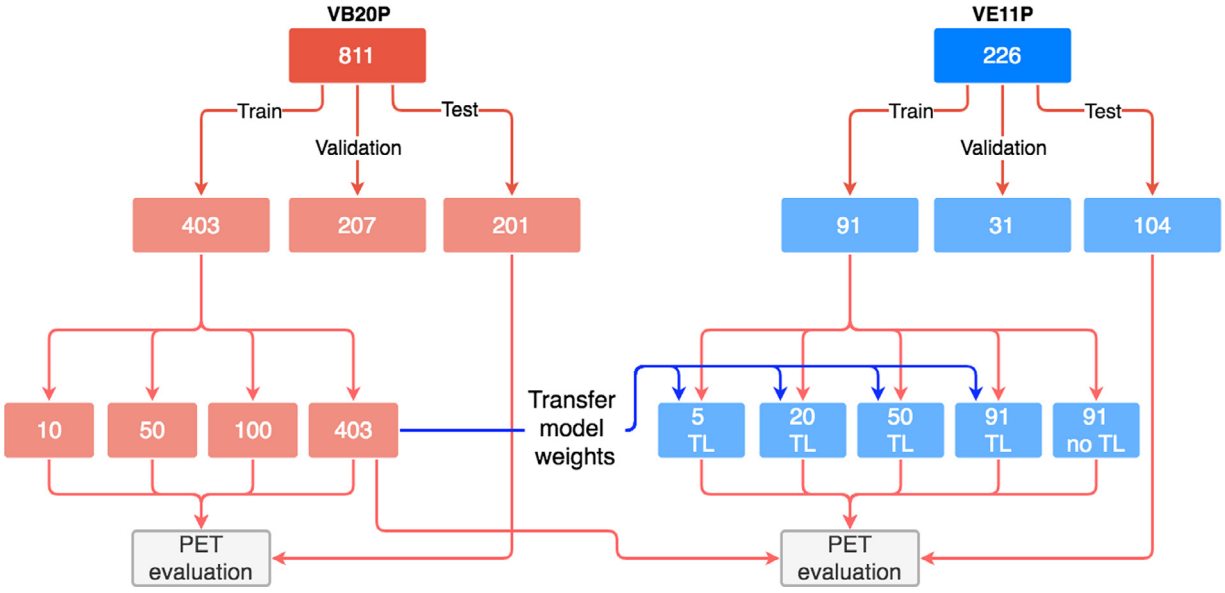

Fig. 1. Separation of subjects into train, validation and test cohorts within each group, and further split to investigate the impact of training group size. Note, all 30 patients from St. Olavs hospital were part of the 91 VE11P training cohort. For each MRI input type, four models are trained from the $n = 403$ patients of the VB20P cohort with increasing number of subjects. The performance of these models is evaluated using the independent VB20P test cohort ($n = 201$). For VE11P, using the $n = 91$ training cohort patients, a total of four models are trained using transfer learning (TL) from the $n = 403$ VB20P model. An additional model is trained using all $n = 91$ training patients, but without any transfer learning (no TL). All VE11P models, and the $n = 403$ VB20P model applied directly without re-training, are evaluated using the independent VE11P test cohort ($n = 104$). This setup is identical for DeepUTE, DeepDixon, and DeepT1.

Table 1
MRI sequence parameters.

MRI Sequence	Repetition time (TR) [ms]	Echo time (TE) [ms]	Flip angle [degrees]	Acquisition time [s]	Voxel size [mm^3]	Matrix size
VB20P						
Dixon	3.6	1.23/2.46	10	19	2.6 × 2.6 × 3.1	126 × 192 × 128
T1w	1900	2.44	9	300	0.5 × 0.5 × 1	512 × 512 × 192
UTE	11.94	0.07/2.46	10	100	1.6 × 1.6 × 1.6	192 × 192 × 192
VE11P						
Dixon	4.14	1.28/2.51	10	39	1.3 × 1.3 × 2	204 × 384 × 128
T1w	1900	2.44	9	300	0.5 × 0.5 × 1	512 × 512 × 192
UTE	4.64	0.07/2.46	10	118	1.6 × 1.6 × 1.6	192 × 192 × 192

Table 2
Patient characteristics in the [^{18}F]FDG PET test sets.

Software version	N	Male/Female	Age Mean (Range)	Injected dose Mean (SD)	Scan start p.i. Median (Range)
VB20P	201	108/93	68 (23–96) y	203 (+/- 20) MBq	51 (24–134) min
VE11P	104	52/52	73 (41–93) y	200 (+/- 11) MBq	47 (39–69) min

p.i.: post injection.

(Biograph TruePoint 40, 64, or Biograph mCT, Siemens Healthineers) was acquired for all patients on the same day as the PET/MRI examination.

2.2.3. [^{18}F]FDG PET

The test cohort included 201 (VB20P) and 104 (VE11P) subjects for quantitative and clinical evaluation of [^{18}F]FDG PET data. The patients were referred for suspected neurodegenerative disease as part of the clinical work-up. Patient characteristics are given in Table 2. The subjects were positioned head-first with arms down in the fully-integrated PET/MRI system. Data were acquired over a single bed position of 25.8 cm covering the head and neck for 10 min. For the purpose of this study, the PET data from the PET/MRI acquisition were reconstructed using 3D Ordinary Poisson-Ordered Subset Expectation Maximization (OP-OSEM) with 4 iterations, 21 subsets, and 3 mm Gaussian post-filtering on 344 × 344 matrices (2.1 × 2.1 × 2.0 mm^3 voxels) in line with the clinical protocols. Each MR-AC map was resampled to PET resolution as a part of the reconstruction. No additional filtering was applied.

2.3. Deep convolutional neural network

2.3.1. Network structure

The proposed network used in this study is shown in Supplementary Figure 2. The 3D convolutional network is based on an encoder-decoder structure with symmetry concatenations between corresponding states, inspired by the U-Net architecture (Çiçek et al., 2016; Ronneberger et al., 2015) but modified for an end-to-end image synthesis task. Specifically, each stage in the 3D-network consists of 3 × 3 × 3 kernels, batch normalization (BN), rectified linear unit (ReLU) activation, and a dropout layer with increasing fraction from 0.1–0.3 in the encoding part, and vice versa in the decoding part. The down sampling between stages was replaced by convolutions with stride 2. We used L_2 penalties for kernel regularization on the convolution layers.

2.3.2. Network training

The proposed networks were implemented in TensorFlow (version 2.1.0) (Abadi et al., 2016). Our experiments used mean squared error as loss function and the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1×10^{-4} trained for 100 epochs with a batch size of 16. All computations were performed on an IBM POWER9 server with four NVIDIA TESLA V100 GPUs. The networks uses 3D volumes as input consisting of 16 neighboring transaxial slices for each MRI scan (16 slices x 192 voxels x 192 voxels x C channels), where C denotes the number of images in the MRI sequence (in- and opposed-phase for Dixon (two channels), echo images for UTE (two channels), and MPRAGE for T1w (one channel)), and outputs the corresponding CT slices (16 slices x 192 voxels x 192 voxels x 1 channel). All MRI sequences were first resampled to the resolution of the UTE image, to ensure isotropic voxels and matrix size, and normalized to zero mean and unit variance. Subsequently, we extracted 3D volumes from the $192 \times 192 \times 192$ MRI scans with a stride of 4. The scanner bed and structures other than the patient was removed from the CT images, before they were converted to linear attenuation coefficients and moved into PET/MRI space using a 6-parameter rigid alignment procedure (minctracc, McConnell Imaging Center, Montreal, Canada) with normalized mutual information as objective function. A mask of the CT-coverage was applied to the three MRI sequences during the training phase.

2.3.3. Network prediction and post-processing

To generate the deep learning attenuation maps, we extracted the 3D stack-of-slices around each slice in the volume, and computed the average voxel values for each of the overlapping predicted slices.

2.4. Reference methods

The rigidly co-registered CT images were used as our gold standard AC reference during both training and evaluation following conversion of Hounsfield Units as implemented on the Siemens PET/CT system (Carney et al., 2006). Due to the limited coverage in the neck region by the acquired CT, we replaced the missing areas by the values from the vendor-provided UTE AC map. To ensure a fair comparison, this replacement was also performed in all the other attenuation maps. In addition, we also computed the RESOLUTE attenuation map (Ladefoged et al., 2015) for VB20P patients from Rigshospitalet. RESOLUTE is calibrated to VB20P UTE data, and was therefore not computed for the VE11P patients. As part of the VE11P software upgrade, a vendor-provided atlas-based MR-AC method was made available (Koesters et al., 2016; Paulus et al., 2015), and was used as the MR-based reference for the VE11P test cohort. This method is prone to bone artifacts related to misregistration in more than 20% of the cases (Øen et al., 2019). Therefore, patients with this type of artifacts were excluded from the analysis of the atlas-based method.

2.5. PET evaluation metrics

Due to the use of data from different software versions (VB20P and VE11P), causing differences in all MR images with varying degree, we evaluated the cohorts separately.

We first moved all data to common MNI space using ANTs (Avants et al., 2011) by diffeomorphic non-rigid registration of the patient's T1w MPRAGE image to the ICBM 152 2009a template (Fonov et al., 2009). Voxels inside the MNI brain mask was considered part of the brain mask if the PET activity was >20% of the maximum intensity value of the brain. The voxel-wise percent difference relative to PET with CT-AC, defined as:

$$Rel_{\%} = \frac{PET_x - PET_{CT}}{PET_{CT}} \times 100,$$

as well as the absolute relative percent difference, defined as:

$$Abs_{\%} = \frac{|PET_x - PET_{CT}|}{PET_{CT}} \times 100,$$

were calculated for the PET images corrected with each of the MRI-based AC's.

As a measure of robustness towards outliers, we used the metric introduced in Ladefoged et al. (Ladefoged et al., 2016) to estimate the number of outliers measured in the PET images. The metric calculates the percentage of patients within a 3% accuracy in the $Rel_{\%}$ images for varying voxel-wise fractions of the brain, varied from 0% to 100%. A perfect score for a method is therefore to have 100% of the voxels in the brain in 100% of the patients within ±3% of PET with CT-AC.

2.6. Effect of cohort size and changes to the input on [^{18}F]FDG PET

To evaluate the effect of training group size, we trained in total four networks with sizes of $n = \{10, 50, 100, 403\}$. The subjects were sampled with replacement.

The robustness towards changes to the input images was evaluated using images from the VE11P cohort. In recognition of the changes to the MR images following the software upgrade, it was expected that further fine-tuning of the network was needed to adapt to these changes. The purpose of the analysis was to test the number of subjects needed for this adaptation. We compared a network trained using a group of all available training subjects ($n = 91$) against $n = \{5, 20, 50\}$, all trained using transfer learning from the full VB20P training cohort ($n = 403$). In addition, we also trained a network without transfer learning on the full cohort ($n = 91$). The overview of the setup is shown in Fig. 1. We repeated the training of the two networks with lowest number of subjects ($n = 5$ and $n = 20$) a total of four times using different combinations of training subjects each time, to determine the robustness towards the selection of subjects. The comparisons were repeated for each MRI sequence type, using identical hyper parameters as presented in Section 2.3. We compared the networks based on the number of outliers measured in the PET images, representing the robustness.

2.7. Effects of MRI sequence on [^{18}F]FDG PET

We evaluated the effects of MRI sequence on accuracy by training three independent networks, one for each sequence: Dixon, T1w and UTE, respectively. Each network was trained on the full VB20P cohort, and subsequently fine-tuned using the full VE11P cohort, and designated: DeepDixon, DeepT1, and DeepUTE. We assessed the robustness dependent on MRI sequence by comparing the number of outliers in the VE11P cohort.

Full brain and regional performances of the networks were evaluated using anatomical predefined template regions from MNI space (Collins et al., 1999; Fonov et al., 2009), with extraction of mean $Rel_{\%}$ and $Abs_{\%}$ values. We furthermore generated parametric average and standard deviation $Rel_{\%}$-distribution images across all patients for each method for visual inspection.

2.8. Clinical evaluation

The [^{18}F]FDG PET images from the independent test cohort (VE11P, $n = 104$) reconstructed using CT-AC and DeepDixon were analyzed by MI Neurology (Siemens Healthineers, Erlangen, Germany). Statistical surface projections (z-score maps) were generated showing deviations from a vendor-provided database of healthy controls (46–79 years) using cerebellar gray matter as reference region. Statistical surface projections are widely used and accepted as the most sensitive method for the identification of metabolic reductions in [^{18}F]FDG PET. The projections are routinely used in the reading of clinical [^{18}F]FDG PET scans providing information on regional patterns and severity of hypometabolism. Statistical surface projections were produced for PET images created with CT-AC and DeepDixon, and for each patient presented (blinded and randomized) side by side to two expert nuclear medicine physicians (IL, OH). The readers first independently and then by consensus visually scored each pair of projections as "no difference", "minor, but not

significant", or "clinically significant" where the latter would indicate a change of diagnosis or difference indicative of disease progression in only one of the PET images. This strategy was selected as the differences in the images were expected to be small and barely discernible on direct visual inspection, and statistical surface projections is the most sensitive method to discrete changes in a clinical setting (Burdette et al., 1996). The reading, thus, simulates the clinical evaluation of a patient with follow-up imaging using standard clinical methodology, and includes also the indirect effects of perturbations in cortical uptake caused by AC induced effects on anatomical warp and reference region.

3. RESULTS

Fig. 2 shows the axial and sagittal views for each proposed attenuation method (DeepDixon, DeepT1 and DeepUTE) for a single sample patient from the VE11P test cohort. Notice especially the excellent performance in the skull-base and nasal cavities in the proposed methods replicating the morphology of even small anatomical details from CT. The network training time using the full VB20P cohort was 40 hrs, where the fine-tuning to the full VE11P cohort was 12 hrs. The inference time to predict an attenuation map for a new subject was 4 sec. A total of 13 patients (13%) had artifacts in their atlas-based attenuation map related to misplaced bone. These subjects were removed from the average performance evaluations of the atlas-based method only.

3.1. Effect of cohort size and changes to the input on [^{18}F]FDG PET

The effect of VB20P cohort size in DeepUTE training is shown in Fig. 3a, which shows a clear correlation between group size and model performance in terms of outliers at the 3% [^{18}F]FDG PET error-level. Training using $n = 10$ subjects results in inadequate bone representation, incorrect attenuation values in brain tissue, and an overall smoother AC map with an 8–10% negative bias relative to PET with CT-AC (Fig. 3b). Increasing the group size decreased the blurring and increased the image contrast and overall detail level in the AC images. Furthermore, the robustness clearly increased with group size. Thus, $n = 100$ was required to outperform RESOLUTE in the number of outliers. When training using the full cohort, $n = 403$, DeepUTE markedly reduced the number of outliers compared to RESOLUTE. The large amount of training data empowers our method to handle common artifacts such as signal voids from dental artifacts. An example of this is illustrated in Fig. 4. A similar relationship between training group size and number of outliers were found when using DeepDixon and DeepT1 (Suppl. Fig. 3). DeepT1 appeared more robust towards training group size, as 10–50 subjects were sufficient to achieve performance near RESOLUTE and increasing group size above 100 subjects did not improve robustness.

Fig. 5 shows the effect of fine-tuning the DeepUTE network to a significant change in the UTE MRI input sequence following the VB20P to VE11P software upgrade. The VB20P model without transfer learning is shown, where it is apparent that transfer learning is necessary. Transfer learning from VB20P cohort was performed on 5, 20, 50 and the full $n = 91$ VE11P cohort with UTE MRI as input. Here, too, robustness was correlated to the group size, but size needed for convergence was markedly reduced to $n = 5$ subjects. Incremental robustness improvements were achieved with increasing group size. For comparison, training the VE11P network without transfer learning using all $n = 91$ subjects resulted in similar model accuracy as when using between 5 and 20 subjects with transfer learning. Overall similar results were observed for DeepDixon, with the exception that all models *with* transfer learning outperformed the model *without* transfer learning (Suppl. Fig. 4A). As expected, DeepT1 trained only with VB20P patients generalized well to the VE11P cohort without re-training, with performance surpassing the atlas-based method (Suppl. Fig. 4B). The number of outliers was similar to training with all $n = 91$ VE11P training subjects without transfer learning, but fine-tuning with VE11P data further improved the robustness. Repeating model training using different training subjects for $n = 5$ and $n = 20$ appeared robust across all three MRI sequence types (Fig. 5 and Suppl. Fig. 4).

3.2. Effects of MRI input sequence on [^{18}F]FDG PET

The number of outliers at the ±3% level, representing the robustness of the method, was similar across all three proposed methods when evaluated on the VB20P test patients (Fig. 6A) and on the VE11P test patients (Fig. 6B) after applying transfer learning. The methods showed a substantial improvement over both RESOLUTE and the atlas-based method.

The relative and absolute relative percent difference regional analysis for the VE11P cohort with transfer learning from the VB20P cohort is shown in Fig. 7 and Supplementary Figure 5, respectively. None of the proposed methods exceeded ±1% average relative error ($Rel_{\%}$) in any region of the brain. The atlas-based method achieved a low full brain $Rel_{\%}$ of 0.8 ± 2.4%, with higher regional errors subcortically of up to

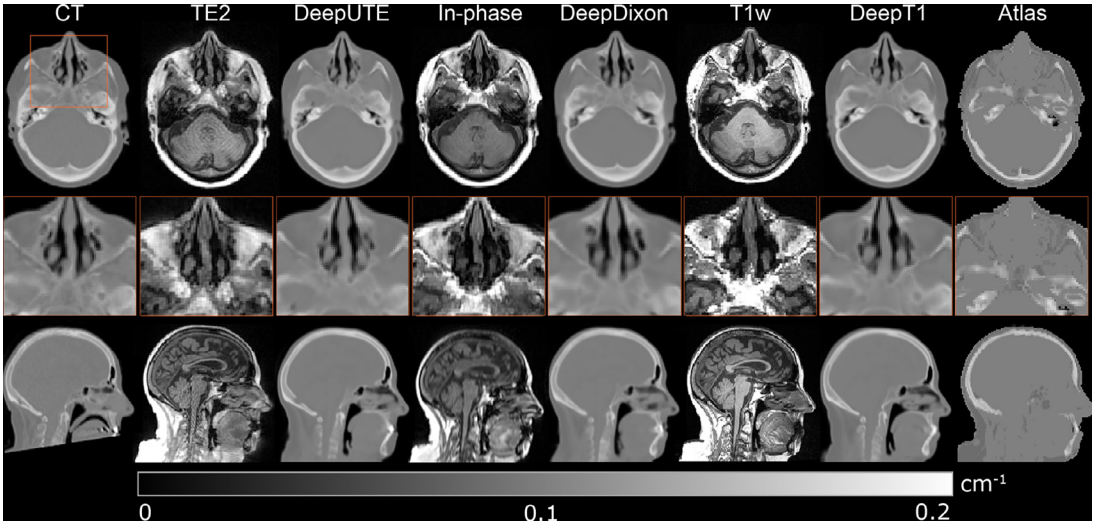

Fig. 2. Attenuation map comparison for a representative patient from the VE11P cohort. The attenuation images are shown prior to superimposing UTE values in the area outside the CT field-of-view. Each proposed MR-based attenuation map is preceded by the underlying MR image used for inference for reference. Note, for simplicity, only second echo (TE2) and in-phase is shown for DepeUTE and DeepDixon, respectively. All models were trained using the full cohort ($n = 91$) with transfer learning from the corresponding VB20P full cohort models ($n = 403$).

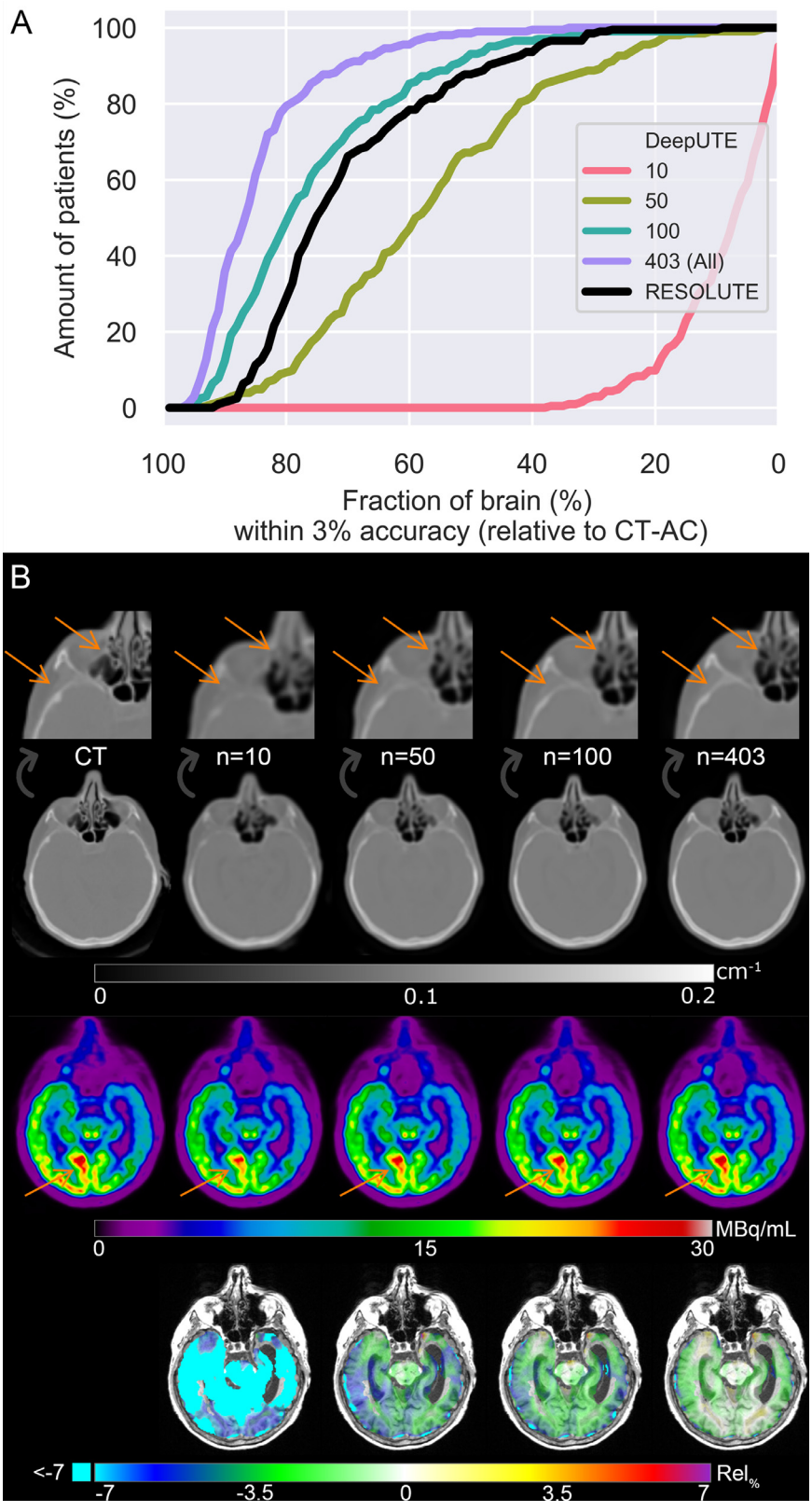

Fig. 3. The effect of training group size on model accuracy of DeepUTE. A) An outlier analysis for VB20P test subjects ($n = 201$) of model accuracy with increasing training group size. B) Axial images of a representative patient with [^{18}F]FDG PET and corresponding DeepUTE AC-maps, and %-difference maps relative to PET CT-AC. The arrows in the AC-maps point to the nasal cavity and bone with a more distinct resemblance to the reference CT with increasing group size. The arrows in the PET images point to an occipital lobe [^{18}F]FDG PET hyper-intense area with convergent resemblance to the reference standard PET CT-AC.

7%. The maximal outlier for a single patient in any region of the brain was below 6% for all proposed methods (DeepUTE range: −4% to 5%, DeepDixon range: −4% to 5%, DeepT1 range: −5% to 6%). For the atlas-based method, the errors ranged from −15% to 14%. Similarly, average absolute relative error ($Abs_{\%}$) was below 2.5% in any region of the brain for the proposed methods, and between 4% and 8% regionally for the atlas-based method. The results for the regional analysis for the VB20P cohort are shown in Supplementary Figure 6.

The averaged relative difference mean and standard deviation images are shown in Fig. 8 for the VE11P cohort and Supplementary

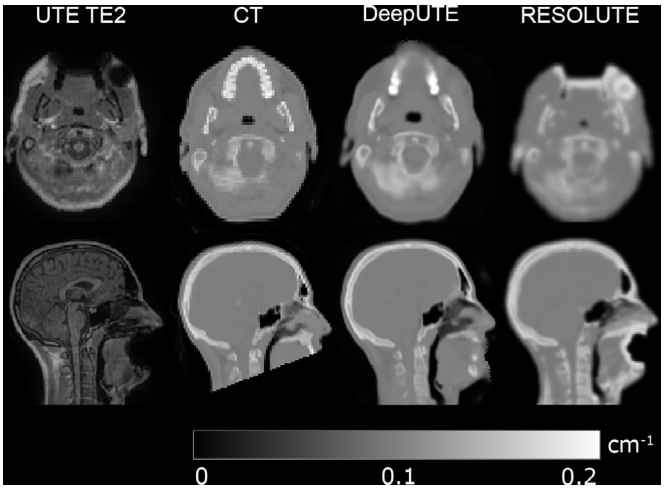

Fig. 4. Example case showing robustness to metallic dental implants for DeepUTE trained with the full VB20P training group ($n = 403$). Metal implants did not cause any noticeable artifacts in CT, but caused large signal voids in the UTE echo image. The artifacts resulted in large errors in the RESOLUTE attenuation map, whereas DeepUTE were able to largely correct for the artifact, as shown both in the axial and sagittal orientation. The attenuation images are shown prior to superimposing UTE values in the area outside the CT field-of-view.

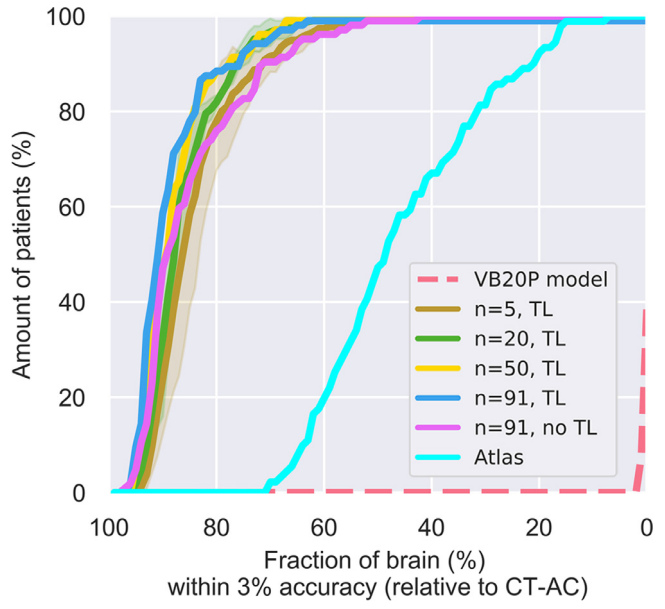

Fig. 5. Outlier analysis for the VE11P test patients ($n = 104$) showing the effects of increasing group size on transfer learning model accuracy after fine-tuning the DeepUTE model. The dashed lines represent the performance of the DeepUTE model from the VB20P cohort applied to the VE11P test patients without transfer learning (TL). The pink line represents the performance of training the network (DeepUTE) from scratch without TL, but with the full train cohort ($n = 91$), where the remaining lines represents the performance of fine-tuning of DeepUTE with increasing training group size after transfer learning from the VB20P cohort. The shaded areas around $n = 5$ and $n = 20$ represents the 95% confidence interval after repeating the training four times with different subjects in each repetition. The atlas-based MR-AC method, shown for comparison, was only based on subjects without registration-related artifacts ($n = 91$).

Figure 7 for the VB20P cohort. Again, near equal performance is achieved by applying either input MRI sequence to the deep learning method. Compared to RESOLUTE, especially cortical regions close to bone was more accurate with a lower standard deviation (Suppl. Fig. 7).

Table 3
Consensus scores from clinical evaluation of [^{18}F]FDG PET comparing attenuation correction using CT and DeepDixon (VE11P; $n = 104$).

Consensus score	Number
No difference	78 (75%)
Minor, not significant	25 (24%)
Clinically significant	1 (1%)*

* Difference caused by warp error in spatial normalization.

3.3. Clinical evaluation

The 104 pairs of [^{18}F]FDG PET reconstructions (CT and DeepDixon) were evaluated, and 1 pair (1%) was scored as "clinically significant different" based on the statistical surface projection where 103 pairs (99%) were scored as not clinically significantly different (Table 3). On direct clinical reading of the [^{18}F]FDG PET image of the single case rated as "clinically significant different" there was no visually discernable change in voxel activity. The differences could be traced to a defect spatial normalization warp that would be found on routine quality control. Presumably it was brought on by scanning in extreme neck flexion combined with small differences in extra-cerebral activity.

4. DISCUSSION

This study confirmed the usability of deep learning-based networks for MRI-based attenuation correction in a clinical setting, and demonstrated performances exceeding previous state-of-the-art non-deep learning-based methods. By training a common auto-encoder architecture using increasing group sizes, we showed a direct correlation between accuracy and size when the network was trained from scratch. Using transfer-learning from the large cohort of subjects, however, we showed the amount of training data needed to adapt to changes to the MRI sequence input could be reduced significantly to as low as 5 subjects. Furthermore, we demonstrated robustness towards the choice of MRI sequence input, with identical performance when using a common Dixon-based MR-AC sequence as with the specialized UTE sequence. Finally, we demonstrated a retained clinical value and accuracy of our methodology compared to our reference CT-AC.

The methodology employed in this study is not novel, as the auto-encoder architecture has been widely applied for MR-AC purposes already (Gong et al., 2018; Han, 2017; Liu et al., 2017). The novelty of our study lies with the unprecedented amount of training data utilized and the analysis of robustness with respect to the size of the training data set and type of MRI input. Deep learning is usually associated with large amounts of training data, something that is difficult to obtain in most health-care applications. Previous publications employing CNNs for MR-to-CT conversion are therefore often based on small cohorts, with a group size ranging between 10 and 30 (Gong et al., 2018; Han, 2017; Liu et al., 2017). To investigate the effect of size, we trained the network end-to-end from scratch using 10, 50, 100, and 403 subjects, respectively. While there was an impact on the average performance with an increasingly larger training group (Fig. 3B), a larger effect was determined to be in the number of associated outliers (Fig. 3A), with the best overall performance achieved for the largest cohort ($n = 403$). Interestingly, to achieve the performance of RESOLUTE, measured in number of outliers, a training group size between 50 and 100 subjects was needed (Fig. 3A and Suppl. Fig. 3). This suggests that the deep learning methods based on fewer than 50 subjects for training might be unstable, albeit having decent average errors. The model accuracy further improves with increasing training group size from 100 to 403 in DeepUTE and DeepDixon, confirming findings in other domains where deep learning were applied (Sun et al., 2017). Using T1w MPRAGE generally appears to be more stable (Suppl. Fig. 3B), which could be due to the sequence being more standardized compared to Dixon-VIBE and UTE.

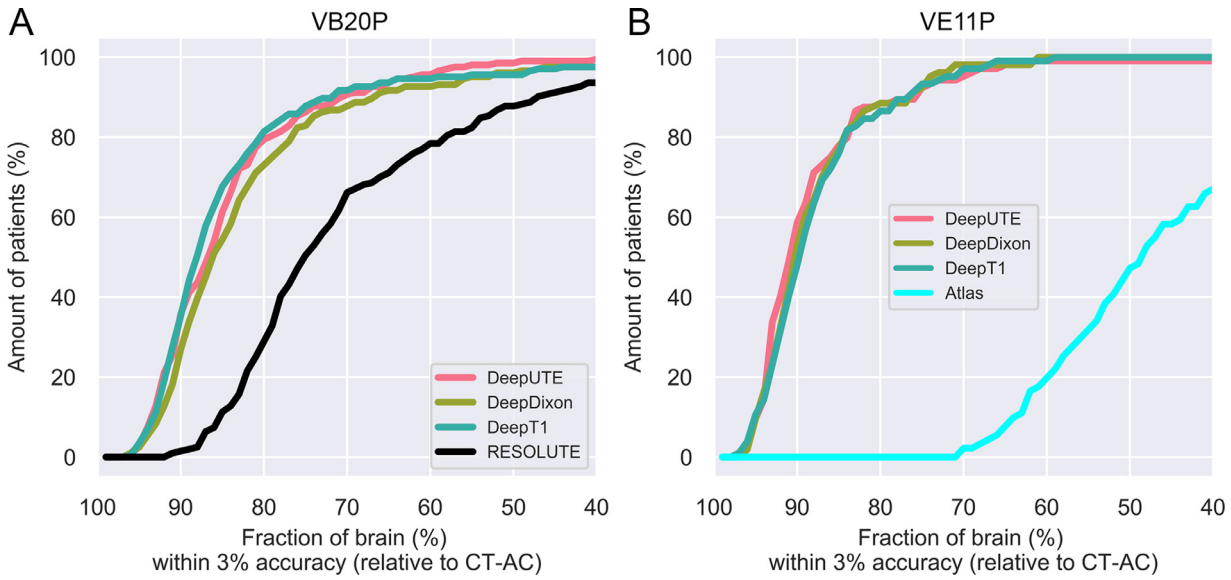

Fig. 6. Outlier analysis for the VB20P (left, $n = 201$) and VE11P (right, $n = 104$) test patients to show the effects on model robustness by varying the MRI sequence input type and across software upgrades. All models are trained using the full train cohorts, $n = 403$ for VB20P and $n = 91$ with transfer learning for VE11P. RESOLUTE and atlas-based methods are shown for comparison. Only subjects without registration-related artifacts were used to compute the outliers for the atlas-based method ($n = 91$).

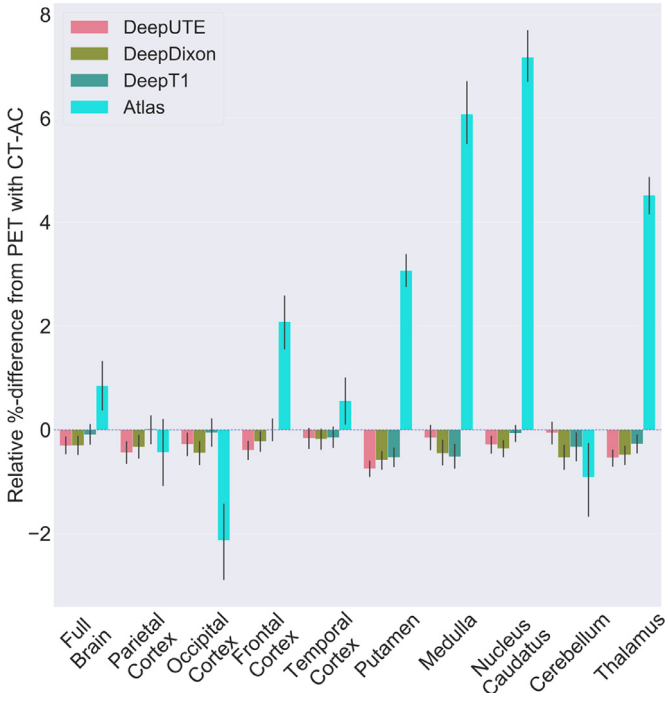

Fig. 7. Full brain and regional mean relative differences across all VE11P test patients ($n = 104$) for each of the three networks with MRI sequences UTE, Dixon, and T1-weighted MPRAGE, all trained using the full train cohort ($n = 91$) with transfer learning from the VB20P cohort, as well as the atlas-based MR-AC method for comparison. Only subjects without registration-related artifacts were used to compute the results for the atlas-based method ($n = 91$). The bars represent the average relative difference to PET with CT-AC across patients. The black line in each represents the 95% confidence interval.

A popular and useful strategy to overcome small training group sizes is to apply transfer learning (Bengio et al., 2013). This strategy was also used by Han to initiate part of their network from a pretrained VGG-16 layer model (Han, 2017), by Jang et al. to train a model using 6 patients transfer learned from a model with 30 patients (Jang et al., 2018), and by Torrado-Carvajal et al. to train a model pretrained on 19 T1w brain images to synthesize Dixon-VIBE pelvis images from 19 patients (Torrado-Carvajal et al., 2019). In this study, we employed transfer learning to re-calibrate the network to a new image appearance following a major software upgrade. The results showed little effect of increasing the number of subjects above 5, as 5–91 subjects for training yielded similar model accuracy (Fig. 5 and Suppl. Fig. 4). Training with transfer learning on only five subjects matched (DeepUTE) or exceeded (DeepDixon and DeepT1) the performance of training on all subjects ($n = 91$) without transfer learning, demonstrating that information from the original model trained on a large cohort is preserved and utilized. These findings have relevance not only for recalibrating methods after major software upgrades, but also for distribution of models between scanners and centers when the models do not generalize well. Using only a limited number of subjects with paired CT and MRI, the model can be adapted to match scanners at different locations, potentially even from different vendors. We hypothesize that such transfer learning will also apply to cohorts with different demographics (ethnicity etc.).

There were differences, to a various degree, in all three MRI sequences pre- and post-upgrade, see Table 1 and Supplementary Figure 1, impacting the ability of the methods to generalize across the system upgrade. The largest difference was observed with the Dixon sequence, mainly expressed in change of resolution, but nonetheless, DeepDixon achieved similar performance after transfer learning as DeepT1. This suggests that similar domain adaptation to MRI sequences from other vendors are feasible, as differences in T1 weighted implementations across systems are no greater than between VB20P and VE11P for the Dixon-VIBE or UTE sequence. DeepT1 trained with VB20P data generalized well to VE11P data, producing images that were objectively identical to the images produced after fine-tuning. The quantitative PET evaluation resulted in a 1–2% overestimation on average (results not shown). Further inspection revealed a general reduction in MRI intensity in the area representing bone in patients examined after the upgrade (Suppl. Fig. 1C), causing DeepT1 to predict denser bone, ultimately causing the overestimation. Despite this error being acceptable for most clinical purposes, we found that fine-tuning reduced the PET bias, and indicates that fine-tuning is needed after all major upgrades of the MRI system. Training the model with a more heterogeneous dataset with T1 weighted

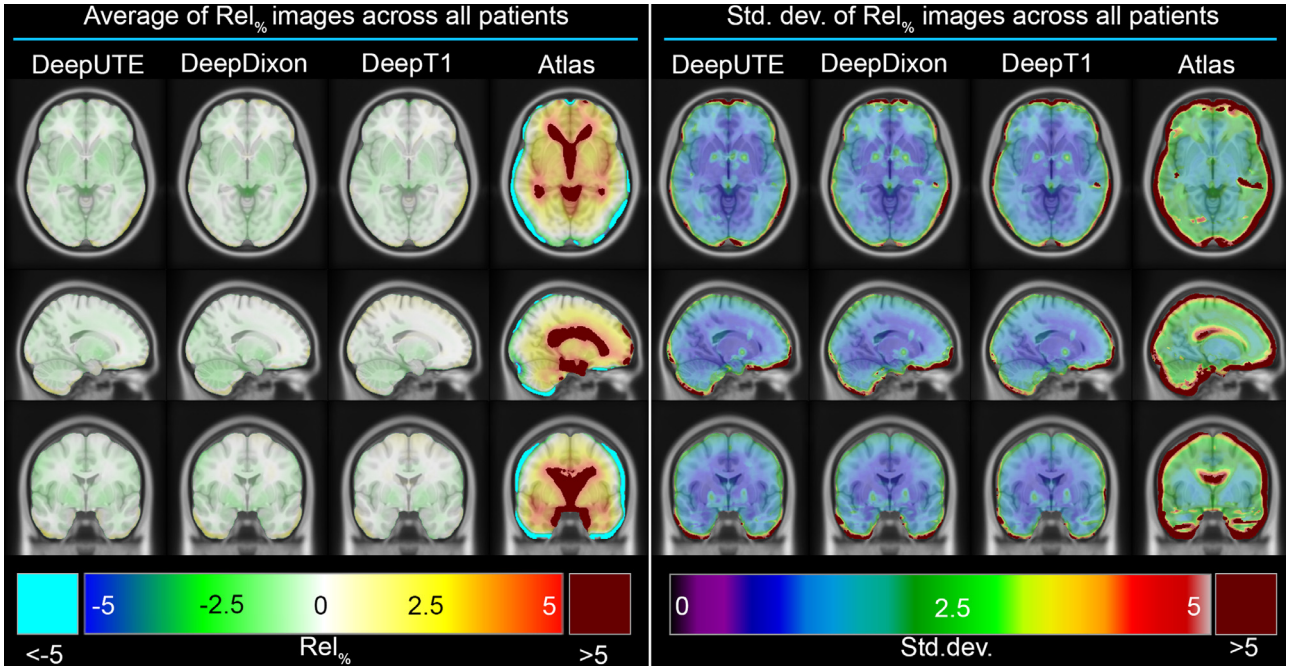

Fig. 8. Averaged relative difference (left four columns) and standard deviation (right four columns) images across all VE11P test patients $Rel_{\%}$ images (n = 104) for each of the three networks with MRI sequences UTE, Dixon, and T1-weighted MPRAGE, all trained with transfer learning from the VB20P cohort, as well as the atlas-based MR-AC method. Images computed for the atlas-based method were only based on subjects without registration-related artifacts (n = 91).

MPRAGE images from multiple sites and systems could potentially eliminate the need for fine-tuning completely.

Using our method, the average relative bias is within 1% from PET with CT-AC in any region of the brain with any of the MR images as input (Fig. 7). This is essential for clinical evaluation as in e.g. tumor delineation and treatment response assessment (Law et al., 2019), and for neurological applications using the cerebellum as reference region (Borghammer et al., 2010; Ishii et al., 2001; Yakushev et al., 2008). Utilizing the same patient cohort and metrics as was employed in a previous multi-center comparison (Ladefoged et al., 2016), allows us to compare not only to RESOLUTE, but also indirectly to the other best performing state-of-the-art methods for the Siemens PET/MRI (Burgos et al., 2014; Izquierdo-Garcia et al., 2014; Mérida et al., 2017). Across all metrics, our method was found to have similar or better performance than that of the most promising methods. The methods based on deep learning that have been proposed in the literature report comparable PET bias as was found in our work. Jang et al. (Jang et al., 2018) and Liu et al. (Liu et al., 2017) reported average regional $Rel_{\%}$ [^{18}F]FDG PET bias within ±2% across eight subjects and ±4% in 10 subjects, respectively, compared to a tissue-segmented three class (air, soft tissue, and bone) CT reference, where Gong et al. (Gong et al., 2018) reported ±3% in 12 subjects compared to a reference CT-AC. However note that no outlier analysis or clinical evaluations were performed in these publications, and a robust regional performance is critical for clinical use.

The atlas-based method had registration-related artifacts in 13 patients. Of these, four were positioned outside the patient volume, as previously reported (Øen et al., 2019), and could have been manually removed prior to reconstruction. The remaining errors corrupted the image, rendering a rescan the only option. Despite removing these patients from the PET evaluation, the atlas-based method still had a global absolute relative error of 5% (Suppl. Fig. 5), which is likely related to the absence of accurate air segmentation, see e.g. Fig. 2. The PET bias was higher than the previously reported 2.5% (Øen et al., 2019), but is most likely due to a difference in patient cohort.

Specialized sequences able to generate contrast in bone have little diagnostic value, and the added contrast comes at the cost of increased acquisition time, and thus less patient comfort and compliance. While the specialized sequences have proven pivotal for segmentation-based methods (Dickson et al., 2014), no evidence exists that such sequences are needed in order for deep learning-based methods to succeed. Our results demonstrate that traditional MRI sequences are sufficient for deep learning-based MR-AC, confirming the findings of several previous works (Teuho et al., 2020). Of the three networks we chose to clinically evaluate the more simplified and patient compliant DeepDixon. In terms of cross-vendor use, DeepT1 is the obvious choice, but on a Siemens mMR, the fast Dixon-VIBE sequence is always part of the PET acquisition, and therefore inherently has reduced motion and optimal alignment of PET and MR images. In the 104 patient examinations evaluated, two experienced expert readers found no cases with clinically significant differences between CT and DeepDixon. The spatial normalization was performed individually for each PET image, which could partly explain the minor non-significant differences in 24% of the cases (Table 3). However, limitations to DeepDixon in particular related to abnormal bone structures, surgical deformation and metallic implants should be kept in mind. It is recommended that evaluation of DeepDixon for the use in brain tumor evaluation is performed separately using tracer specific clinical metrics as done previously (Ladefoged et al., 2019, 2017). Nonetheless, the frequency of potential errors/differences related to using DeepDixon is very low, and probably smaller and less frequent than that introduced by dental artifacts and motion on the PET/CT system. In our center, we have now implemented DeepDixon MR-AC in routine clinical imaging and performed more than 200 [^{18}F]FDG PET scans in adult patients referred for suspected neurodegeneration without routine low-dose CT. To further minimize potential errors, attenuation-maps are carefully inspected for unusual structures and artifacts before the patient leaves the department and a low-dose CT is performed if errors are suspected following image inspection.

Our study had a number of limitations. We chose to focus on evaluating the effects of group size and MRI sequence input. The conclusions drawn here could potentially be different if other network types were applied. It was not the scope of this study to evaluate the effect of deep learning architecture, but we recognize the potential improved accuracy

associated with more sophisticated networks, such as the generative adversarial network (Goodfellow et al., 2014). The high accuracy and low number of outliers presented here suggests, however, that only minor improvements are to be found. Moreover, a limitation of the comparison is the use of identical training setups for each training group size. Tailoring the hyperparameters to each model, or investigating the use of 2D or 3D patches as input to boost amount of training samples could potentially improve the results of the networks with a low number of subjects.

5. CONCLUSION

We have described and evaluated a deep learning attenuation correction approach for PET/MRI neuroimaging using more than 1000 subjects. We showed that a requirement for accurate and robust MR-AC is a large group size of at least 50 subjects for training, but further increasing the size to 400 directly impacted the number of outliers significantly. However, using transfer learning from a large cohort, a group size of 5 subjects was sufficient to recalibrate to changes in the MRI sequences. Full robustness was achieved with only 20 subjects, with performance at the same level or even surpassing that of a larger training cohort (n = 91) without transfer learning. Furthermore, we demonstrated robustness towards the choice of MRI sequence input. The clinical evaluation showed no clinically relevant differences compared to CT-AC, although knowledge about MR-AC limitations is important when used in clinical routine. The combination of accuracy, outlier performance, clinical performance, robustness towards the choice of MRI sequence input, and low group size needed for re-training following a major software upgrade, indicates that the clinical implementation of our deep learning-based MR-AC method will be feasible across MRI system types.

CRediT author statement

Claes Nøhr Ladefoged: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing – Original Draft, **Adam Espe Hansen**: Conceptualization, Methodology, Formal analysis, **Otto Mølby Henriksen**: Conceptualization, Data Curation, Resources, **Frederik Jager Bruun**: Data Curation, **Live Eikenes**: Data Curation, **Silje Kjærnes Øen**: Data Curation, **Anna Karlberg**: Resources, **Liselotte Højgaard:** Resources, Funding Acquisition, **Ian Law**: Conceptualization, Data Curation, Resources, **Flemming Littrup Andersen**: Conceptualization, Methodology, Formal analysis, Supervision. All author participated in drafting and revising the manuscript.

Supplementary Figures

Supplementary Figure 1: Differences between VB20P and VE11P for the three sequences UTE (A), Dixon-VIBE (B), and T1-weighted MPRAGE (C). Using the CT bone area (linear attenuation coefficient > 0.103 cm^{-1}) as a mask, the mean bone surrogate signal, measured with $R2^*$ from UTE sequences, are higher after the upgrade (A). For T1w MPRAGE, there is a decrease in the signal in the area representing bone (C). The effects of the resolution improvement (Table 1) for the Dixon-VIBE sequence are clearly seen visually (B).

Supplementary Figure 2: CNN U-net-like architecture used in this study. The network takes a stack-of-slices from 16 neighboring MR slices, and outputs the corresponding pseudo-CT image. C represents the number of MR channels: 2 for UTE (TE1 and TE2), 2 for Dixon (in- and opposed-phase), and 1 for T1w.

Supplementary Figure 3: The effects of group size on model accuracy. Outlier analysis shown for VB20P test patients (n = 201) for increasing training group size for DeepDixon (A) and DeepT1 (B). RESOLUTE is added for comparison.

Supplementary Figure 4: Outlier analysis for the VE11P test patients (n = 104) showing the effects of increasing group size on transfer learning model accuracy after fine-tuning the DeepDixon (A) and DeepT1 (B) models. The dashed lines represent the performance of the model from the VB20P cohort applied to the VE11P test patients without transfer learning (TL). The pink line represents the performance of training the network from scratch without TL, but with the full train cohort (n = 91), where the remaining lines represents the performance of fine-tuning with increasing training group size after transfer learning from the VB20P cohort. The shaded area around n = 5 and n = 20 represents the 95% confidence interval after repeating the training four times with different subjects in each repetition. The atlas-based MR-AC method, shown for comparison, was only based on subjects without registration-related artifacts (n = 91).

Supplementary Figure 5: Global and regional mean absolute relative differences across all VE11P test patients (n = 104) for each of the three networks with MRI sequences UTE, Dixon, and T1-weighted MPRAGE, all trained with transfer learning from the VB20P cohort. The atlas-based MR-AC method, shown for comparison, was only based on subjects without registration-related artifacts (n = 91). The bars represent the average absolute relative difference to PET with CT-AC across patients. The black line in each represents the 95% confidence interval.

Supplementary Figure 6: Full brain and regional mean relative (upper) and absolute mean (lower) differences across all VB20P test patients (n = 201) for each deep learning model. The bars represent the difference to PET with CT-AC across patients. The black line in each represents the 95% confidence interval. RESOLUTE shown for comparison.

Supplementary Figure 7: Averaged relative difference (top three rows) and standard deviation (bottom three rows) images across all VB20P testing patients (n = 201). Please note the change of scale compared to Figure 8.

ACKNOWLEDGMENTS

The PET/MRI system at Rigshospitalet was kindly provided by the John and Birthe Meyer Foundation, Denmark. Special thanks to the bioengineers and radiographers at Rigshospitalet and St. Olavs Hospital for patient preparations and image acquisitions. We thank IBM Denmark for providing two POWER9 servers with 4 Tesla V100 GPUs in each system.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117221.

References

Abadi, M., Barham, P., Chen, J., et al., 2016. TensorFlow: a System for Large-Scale Machine Learning, in: 12th USENIX Conference on Operating Systems Design and Implementation (OSDI 16). pp. 265–283.

Andersen, F.L., Ladefoged, C.N., Beyer, T., et al., 2014. Combined PET/MR imaging in neurology: mR-based attenuation correction implies a strong spatial bias when ignoring bone. Neuroimage 84, 206–216. https://doi.org/10.1016/j.neuroimage.2013.08.042.

Arabi, H., Bortolin, K., Ginovart, N., Garibotto, V., Zaidi, H., 2020. Deep learning-guided joint attenuation and scatter correction in multitracer neuroimaging studies. Hum. Brain Mapp 1–13. https://doi.org/10.1002/hbm.25039.

Arabi, H., Zeng, G., Zheng, G., Zaidi, H., 2019. Novel adversarial semantic structure deep learning for MRI-guided attenuation correction in brain PET / MRI Novel adversarial semantic structure deep learning for MRI-guided attenuation correction in brain PET / MRI. Eur. J. Nucl. Med. Mol. Imaging 46, 2746–2759.

Avants, B.B., Tustison, N.J., Song, G., et al., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

Borghammer, P., Chakravarty, M., Jonsdottir, K.Y., et al., 2010. Cortical hypometabolism and hypoperfusion in Parkinson's disease is extensive: probably even at early disease stages. Brain Struct Funct 214, 303–317. https://doi.org/10.1007/s00429-010-0246-0.

Burdette, J.H., Minoshima, S., Vander Borght, T., Tran, D.D., Kuhl, D.E., 1996. Alzheimer disease: improved visual interpretation of PET images by using three-dimensional stereotaxic surface projections. Radiology 198, 837–843. https://doi.org/10.1148/radiology.198.3.8628880.

Burgos, N., Cardoso, M.J., Thielemans, K., et al., 2014. Attenuation correction synthesis for hybrid PET-MR scanners: application to brain studies. IEEE Trans Med Imaging 33, 2332–2341. https://doi.org/10.1109/TMI.2014.2340135.

Carney, J.P.J., Townsend, D.W., Rappoport, V., Bendriem, B., 2006. Method for transforming CT images for attenuation correction in PET/CT imaging. Med. Phys. 33, 976–983. https://doi.org/10.1118/1.2174132.

Chen, Y., An, H., 2017. Attenuation correction of PET/MR imaging. Magn Reson Imaging Clin N Am 25, 245–255. https://doi.org/10.1016/j.mric.2016.12.001.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (Eds.). In: (Eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science, Vol 9901. Springer, Cham, pp. 424–432.

Collins, D.L., Zijdenbos, A., Baaré, W.C., Evans, A., 1999. ANIMAL + INSECT: improved Cortical Structure Segmentation. In: Kuba, A., Šáamal, M., Todd-Pokropek, A. (Eds.), (Eds.), Information Processing in Medical Imaging. Springer Berlin Heidelberg, pp. 210–223.

Dickson, J.C., O'Meara, C., Barnes, A., 2014. A comparison of CT- and MR-based attenuation correction in neurological PET. Eur. J. Nucl. Med. Mol. Imaging 41, 1176–1189. https://doi.org/10.1007/s00259-013-2652-z.

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47, S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Ge, Y., Wei, D., Xue, Z., et al., 2019. Unpaired Mr to CT Synthesis with Explicit Structural Constrained Adversarial Learning, in: 2019. IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE 1096–1099.

Gong, K., Yang, J., Kim, K., et al., 2018. Attenuation correction for brain PET imaging using deep neural network based on dixon and ZTE MR images. Phys. Med. Biol. 63, 1–15. https://doi.org/10.1088/1361-6560/aac763.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., 2014. Generative adversarial nets, in: advances in Neural Information Processing Systems 27 (NIPS 2014). pp. 2672–2680.

Han, X., 2017. MR-based synthetic CT generation using a deep convolutional neural network method. Med. Phys. 44, 1408–1419. https://doi.org/10.1002/mp.12155.

Ishii, K., Willoch, F., Minoshima, S., et al., 2001. Statistical brain mapping of 18F-FDG PET in Alzheimer's disease: validation of anatomic standardization for atrophied brains. J Nucl Med 42, 548–557.

Izquierdo-garcia, D., Catana, C., 2016. MR Imaging-Guided Attenuation Correction of PET Data in PET/MR Imaging 11, 129–149. https://doi.org/10.1016/j.cpet.2015.10.002

Izquierdo-Garcia, D., Hansen, A.E., Forster, S., et al., 2014. An SPM8-based approach for attenuation correction combining segmentation and nonrigid template formation: application to simultaneous PET/MR brain imaging. J Nucl Med 55, 1825–1830. https://doi.org/10.2967/jnumed.113.136341.

Jang, H., Liu, F., Zhao, G., Bradshaw, T., Mcmillan, A.B., 2018. Deep learning based MRAC using rapid ultrashort echo time imaging. Med. Phys 45, 3697–3704. https://doi.org/10.1002/mp.12964.

Kazemifar, S., McGuire, S., Timmerman, R., et al., 2019. MRI-only brain radiotherapy: assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach. Radiother. Oncol. 136, 56–63. https://doi.org/10.1016/j.radonc.2019.03.026.

Kingma, D.P., Ba, J., 2015. Adam: a Method for Stochastic Optimization. 3rd Int. Conf. Learn. Represent. (ICLR).

Koesters, T., Friedman, K.P., Fenchel, M., et al., 2016. Dixon sequence with superimposed model-based bone compartment provides highly accurate PET/MR attenuation correction of the brain. J. Nucl. Med. 57, 918–924. https://doi.org/10.2967/jnumed.115.166967.

Ladefoged, C.N., Andersen, F.L., Kjær, A., Højgaard, L., Law, I., 2017. RESOLUTE PET/MRI Attenuation Correction for O-(2-18F-fluoroethyl)-L-tyrosine (FET) in Brain Tumor Patients with Metal Implants. Front. Neurosci. 11, 453. https://doi.org/10.3389/fnins.2017.00453.

Ladefoged, C.N., Benoit, D., Law, I., et al., 2015. Region specific optimization of continuous linear attenuation coefficients based on UTE (RESOLUTE): application to PET/MR brain imaging. Phys. Med. Biol. 60, 8047–8065. https://doi.org/10.1088/0031-9155/60/20/8047.

Ladefoged, C.N., Law, I., Anazodo, U., et al., 2016. A multi-centre evaluation of eleven clinically feasible brain PET/MRI attenuation correction techniques using a large cohort of patients. Neuroimage 147, 346–359. https://doi.org/10.1016/j.neuroimage.2016.12.010.

Ladefoged, C.N., Marner, L., Hindsholm, A., et al., 2019. Deep learning based attenuation correction of PET/MRI in pediatric brain tumor patients: evaluation in a clinical setting. Front. Neurosci. 12, 1005. https://doi.org/10.3389/fnins.2018.01005.

Law, I., Albert, N.L., Arbizu, J., et al., 2019. Joint EANM/EANO/RANO practice guidelines/SNMMI procedure standards for imaging of gliomas using PET with radiolabelled amino acids and [18 F] FDG: version 1.0. Eur. J. Nucl. Med. Mol. Imaging 46, 540–557. https://doi.org/10.1007/s00259-018-4207-9.

Lei, Y., Dong, X., Wang, T., et al., 2020. MRI-aided attenuation correction for PET imaging with deep learning, in: medical Imaging 2020: biomedical Applications in Molecular. Structural, and Functional Imaging, 1131723.

Liu, F., Jang, H., Kijowski, R., Bradshaw, T., McMillan, A.B., 2017. Deep learning MR imaging-based attenuation correction for PET/MR imaging. Radiology 286, 676–684. https://doi.org/10.1148/radiol.2017170700.

Mehranian, A., Arabi, H., Zaidi, H., 2016. Vision 20/20: magnetic resonance imaging-guided attenuation correction in PET/MRI: challenges, solutions, and opportunities. Med. Phys. 43, 1130–1155. https://doi.org/10.1118/1.4941014.

Mérida, I., Reilhac, A., Redouté, J., et al., 2017. Multi-atlas attenuation correction supports full quantification of static and dynamic brain PET data in PET-MR. Phys. Med. Biol. 62, 2834–2858. https://doi.org/10.1088/1361-6560/aa5f6c.

Øen, S.K., Keil, T.M., Berntsen, E.M., et al., 2019. Quantitative and clinical impact of MRI-based attenuation correction methods in [18F] FDG evaluation of dementia. EJNMMI Res 9, 83. https://doi.org/10.1186/s13550-019-0553-2.

Paulus, D.H., Quick, H.H., Geppert, C., et al., 2015. Whole-body PET/MR imaging: quantitative evaluation of a novel model-based MR attenuation correction method including bone. J. Nucl. Med. 56, 1061–1066. https://doi.org/10.2967/jnumed.115.156000.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional Networks for Biomedical Image Segmentation. Med. image Comput. Comput. Interv. 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Shiri, I., Ghafarian, P., Geramifar, P., et al., 2019. Direct attenuation correction of brain PET images using only emission data via a deep convolutional encoder-decoder (Deep-DAC). Eur. Radiol. 29, 6867–6879. https://doi.org/10.1007/s00330-019-06229-1.

Sun, C., Shrivastava, A., Singh, S., Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era, in: proceedings of the IEEE International Conference on Computer Vision. pp. 843–852. https://doi.org/10.1109/ICCV.2017.97

Teuho, J., Torrado-Carvajal, A., Herzog, H., et al., 2020. Magnetic Resonance-Based Attenuation Correction and Scatter Correction in Neurological Positron Emission Tomography/Magnetic Resonance Imaging—Current Status With Emerging Applications. Front. Phys 7. https://doi.org/10.3389/fphy.2019.00243.

Torrado-Carvajal, A., 2020. Importance of attenuation correction in PET/MR image quantification: methods and applications. Rev. Española Med. Nucl. e Imagen Mol. (English Ed. 39, 163–168. https://doi.org/10.1016/j.remnie.2020.03.002.

Torrado-Carvajal, A., Vera-Olmos, J., Izquierdo-Garcia, D., et al., 2019. Dixon-VIBE deep learning (DIVIDE) pseudo-CT synthesis for pelvis PET/MR attenuation correction. J. Nucl. Med. 60, 429–435.

Van Hemmen, H., Massa, H., Hurley, S., et al., 2019. A deep learning-based approach for direct whole-body PET attenuation correction., in: journal of Nuclear Medicine. Soc Nuclear Med 569.

Vandenberghe, S., Marsden, P.K., 2015. PET-MRI: a review of challenges and solutions in the development of integrated multimodality imaging. Phys. Med. Biol. 60, R115. https://doi.org/10.1088/0031-9155/60/4/R115.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., et al., 2017. Deep MR to CT synthesis using unpaired data. In: Tsaftaris, S., Gooya, A., Frangi, A, Prince, J (Eds.). In: (Eds) Simulation and Synthesis in Medical Imaging. SASHIMI 2017. Lecture Notes in Computer Science, Vol 10557. Springer, Cham, pp. 14–23.

Yakushev, I., Landvogt, C., Buchholz, H.G., et al., 2008. Choice of reference area in studies of Alzheimer's disease using positron emission tomography with fluorodeoxyglucose-F18. Psychiatry Res 164, 143–153. https://doi.org/10.1016/j.pscychresns.2007.11.004.

Yang, H., Sun, J., Carass, A., et al., 2018. Unpaired brain MR-to-CT synthesis using a structure-constrained cyclegan. In: Stoyanov, D., et al. (Eds.). In: (Eds) Deep Learning in Medical Image Analysis and Multimodal Learning For Clinical Decision Support. DLMIA 2018, ML-CDS 2018. Lecture Notes in Computer Science, Vol 11045. Springer, Cham. Springer, pp. 174–182 https://doi.org/doi.org/10.1007/978-3-030-00889-5_20.

Yang, J., Park, D., Gullberg, G.T., Seo, Y., 2019. Joint correction of attenuation and scatter in image space using deep convolutional neural networks for dedicated brain 18F-FDG PET. Phys. Med. Biol. 64, 075019. https://doi.org/10.1088/1361-6560/ab0606.