Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy

Galapon, Arthur; Thummerer, Adrian; Langendijk, Johannes A.; Wagenaar, Dirk; Both, Stefan

Published in:
Medical Physics

DOI:
10.1002/mp.16838

Publication date:
2024

Received: 2 June 2023 | Accepted: 28 October 2023

DOI: 10.1002/mp.16838

RESEARCH ARTICLE

MEDICAL PHYSICS

Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy

Arthur Villanueva Galapon Jr[1] | **Adrian Thummerer**[1,2] | **Johannes Albertus Langendijk**[1] | **Dirk Wagenaar**[1] | **Stefan Both**[1]

[1]Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

[2]Department of Radiation Oncology, LMU University Hospital, LMU Munich, Germany

Correspondence
Arthur Villanueva Galapon Jr, Department of Radiation Oncology, University Medical Center Groningen, PO Box 30001, 9700 RB Groningen, The Netherlands.
Email: a.v.galapon@umcg.nl

Funding information
European Union's Horizon 2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 955956

Abstract

Background: Deep learning has shown promising results to generate MRI-based synthetic CTs and to enable accurate proton dose calculations on MRIs. For clinical implementation of synthetic CTs, quality assurance tools that verify their quality and reliability are required but still lacking.

Purpose: This study aims to evaluate the predictive value of uncertainty maps generated with Monte Carlo dropout (MCD) for verifying proton dose calculations on deep-learning-based synthetic CTs (sCTs) derived from MRIs in online adaptive proton therapy.

Methods: Two deep-learning models (DCNN and cycleGAN) were trained for CT image synthesis using 101 paired CT-MR images. sCT images were generated using MCD for each model by performing 10 inferences with activated dropout layers. The final sCT was obtained by averaging the inferred sCTs, while the uncertainty map was obtained from the HU variance corresponding to each voxel of 10 sCTs.

The resulting uncertainty maps were compared to the observed HU-, range-, WET-, and dose-error maps between the sCT and planning CT. For range and WET errors, the generated uncertainty maps were projected along the 90-degree angle. To evaluate the dose distribution, a mask based on the 5%-isodose curve was applied to only include voxels along the beam paths. Pearson's correlation coefficients were calculated to determine the correlation between the uncertainty maps and HUs, range, WET, and dose errors. To evaluate the dosimetric accuracy of synthetic CTs, clinical proton treatment plans were recalculated and compared to the pCTs

Results: Evaluation of the correlation showed an average of $r = 0.92 \pm 0.03$ and $r = 0.92 \pm 0.03$ for errors between uncertainty-HU, $r = 0.66 \pm 0.09$ and $r = 0.62 \pm 0.06$ between uncertainty-range, $r = 0.64 \pm 0.06$ and $r = 0.58 \pm 0.07$ between uncertainty-WET, and $r = 0.65 \pm 0.09$ and $r = 0.67 \pm 0.07$ between uncertainty and dose difference for DCNN and cycleGAN model, respectively. Dosimetric comparison for target volumes showed an average 3%/3 mm gamma pass rate of 99.76 ± 0.43 (DCNN) and 99.10 ± 1.27 (cycleGAN).

Conclusion: The observed correlations between uncertainty maps and the various metrics (HU, range, WET, and dose errors) demonstrated the potential of MCD-based uncertainty maps as a reliable QA tool to evaluate the accuracy of deep learning-based sCTs.

KEYWORDS
adaptive proton therapy, deep learning, MRI, quality assurance, synthetic CT

1 | INTRODUCTION

Daily adaptive radiotherapy (ART) utilizes daily imaging to deliver therapeutic doses to the target with precision and accuracy while minimizing exposure to surrounding healthy tissues.[1,2] This approach allows for continuous adaptation to changes in patient geometry throughout the treatment. Besides conventional computed tomography (CT) imaging, which is routinely used for treatment planning and verification, cone beam CT (CBCT) and magnetic resonance imaging (MRI) can provide valuable information within adaptive treatment workflows.[3,4]

The direct implementation of CBCT and MRI in adaptive radiotherapy is challenging. In-room CBCT systems offer daily imaging for monitoring patient position and anatomical changes. However, there is a need to enhance the image quality due to the presence of scattering artifacts. Moreover, daily CBCT imaging contributes to an increase in the overall dose delivered to the patient. The routine use of MRI-guided imaging is becoming increasingly popular in photon therapy due to the integration of MRI with a linear accelerator. While the technology has yet to be available in proton therapy, there is no doubt about the advantages of MRI in the adaptive workflow. MRI offers excellent soft tissue contrast, which provides a more straightforward delineation of target tumors and organs at risk (OARs). However, MRI images are not readily useable for treatment planning due to the lack of electron density information. Accurate synthetic CT (sCT) generation from MRI or CBCT is essential to overcome these limitations.

The availability of large, paired MRI and CT image datasets makes it possible to use deep learning (DL) methods to generate sCTs. In recent years, several DL approaches have been developed to create sCTs from CBCT[5,6] and MRI images,[3,7–9] enabling proton dose calculations in adaptive workflows.[10] A fast sCT conversion process is essential for integration into routine clinical practice. The use of DL methods to create sCT images has produced accurate results in a fraction of the time required by conventional sCT generation techniques.[4,6–8,8,10] Despite this, the 'black box' nature of these artificial intelligence (AI) models carries potential risks to patients and has hampered clinical implementation in adaptive proton therapy workflows. Especially for proton therapy, where the dose calculation depends on accurate stopping power conversion from CT Hounsfield units (HU), even minor variations in the HU number caused by sCT conversion can lead to significant deviations in the proton dose distributions.[11–13] This uncertainty raises the need to develop quality assurance (QA) tools to evaluate the quality of the generated sCT images towards clinical implementation.

Several techniques have been developed to evaluate the quality of sCTs generated via DL models. The application of CBCT as a clinically feasible QA method for MRI-based sCTs was demonstrated in photon radiotherapy by Palmer et al.[14] However, the method relies on the CBCT system's consistency in maintaining CT HU, which is susceptible to drift over time. An alternative method is the density override method to create patient-specific HU to mass density curves for each CBCT image.[15] However, these methods cannot be used in proton therapy as they do not guarantee the accuracy in HUs required for PT. Proton radiography (PR) was also proposed to evaluate sCT quality, as demonstrated by Seller Oria et al.[16] In their study, the HU accuracy of the sCT was assessed by evaluating the agreement between the measured integral depth dose (IDD) profile and the simulated IDDs. However, the feasibility of this approach depends on the availability of detectors and the limited beam angles for PR measurements.

Another approach to assess sCT quality is using uncertainty maps directly derived from DL models. Hemsley et al. used a Bayesian neural network (BNN) by replacing the deterministic weights with a probability distribution.[17] They incorporated the aleatoric or data-dependent uncertainty in the model's loss function, while the epistemic or model-dependent uncertainty was derived from multiple forward inferences. Results showed that voxels with large errors also showed large uncertainty in their uncertainty map. However, it was exhibited that incorporating uncertainty modifications to the network decreased the HU accuracy, while the complexity of Bayesian models makes them difficult to implement and requires more computational resources. Instead of directly using BNNs, Gal et al. proposed using dropout to approximate Bayesian inference.[18] Dropout is a regularization technique that reduces the overfitting of deep neural networks (NN).[19] It involves randomly deactivating each neuron and its connections with a certain probability during training. In essence, Monte Carlo dropout (MCD) offers a way to efficiently approximate different neural 'architectures' without having to train multiple models, enabling uncertainty estimation by performing multiple forward inferences while the model's dropout layer is active. A similar method involves training multiple models using multiple views of the input image,[7] allowing for the quantification of uncertainties by averaging the final prediction across the models.[20]

Methods to assess the quality of DL-based sCTs have been previously explored. However, most of these evaluations were restricted to assessing the accuracy of HUs and relied on external imaging and measurements as ground truth. A need still exists for a quick, integrated, reliable, and computational QA of DL-based

sCT generation to support its clinical adoption, alone or in combination with measurement-based methods. This study investigates the application of Monte Carlo-based uncertainty maps for QA of DL-based sCTs using more advanced metrics based on proton radiography and proton dose calculations.

2 | MATERIALS AND METHODS

2.1 | Dataset

A dataset containing 101 paired planning CT (pCT) and MRI of head and neck cancer patients treated with proton therapy at the University Medical Center Groningen (UMCG, Netherlands) was used to train and evaluate sCTs generated using two different deep-learning models, deep convolutional neural network (DCNN) and cycle generative adversarial network (cycleGAN). The MRI images were acquired from a Siemens MAGNETOM Skyra 3T or a Siemens MAGNETOM Prisma 3T with a T1-weighted VIBE Dixon imaging sequence using a gadolinium contrast agent. The pCT images were acquired using a Siemens SOMATOM Definition AS Open scanner with a 0.98 mm × 0.98 mm resolution and slice thickness of 2 mm. The CT-MRI image pairs were taken within the same day or a maximum of one day. Both images were taken in the treatment position using a 5-point mask as immobilization. The dataset was grouped into 71, 10, and 20 patients for the training, validation, and testing.

2.2 | Pre-processing

The pCT images were resampled to have a 1 mm × 1 mm × 1 mm spacing before performing an initial rigid registration with the MRI images. To account for changes due to anatomical variations, the MRI and pCT were deformably registered using the Elastix registration toolbox.[21,22] A mask was segmented on the MRI images before applying it to the pCT to remove the patient's couch and immobilization devices. The MRI image intensities were corrected using the N4 Bias Field Correction algorithm to minimize intensity variations between MRI slices.[23] The histogram matching method was applied to each MRI image using a randomly selected template from the dataset to standardize the intensity across different patients.[24] CT and MRI images were resized to 512 pixels × 512 pixels before extracting the individual slices. For the cycleGAN training, the MRI and CT intensities were linearly scaled to [0,1] using min-max normalization. No normalization was applied to the dataset for DCNN model training.

2.3 | Neural network

Two models were used to generate sCTs and the corresponding uncertainty maps – a deep convolutional Neural Network (DCNN) and a cycleGAN. The DCNN is based on a UNET architecture, similar to those used by Spadea et al. and Thummerer et al., implemented using the Pytorch framework.[4,5,9,25–27] The DCNN model was trained using the SGD optimizer and a dropout probability activation of 0.1. As loss function the mean absolute error (MAE) between the pCT and MRI was used. The cycleGAN model was modified to convert MRI images into sCTs.[28] The generator uses ResNet blocks with dropout layers based on the UNET architecture. In contrast, the discriminator is based on a PatchGAN classifier with a patch size of 61×61 pixels.[28,29] Unlike the DCNN, the cycleGAN utilizes adversarial loss, identity loss, and cycle-consistency loss as the loss function and a dropout probability activation of 0.5. Additional information on the network architectures and training parameters of DCNN and cycleGAN can be found in the Supplementary materials (A1).

Both models were trained on a workstation with an NVIDIA A40 GPU with 48GB of memory using 2D axial slices with a batch size of 1. The initial learning rate was set to 1e-4 before linearly decreasing after 100 and 50 epochs for the cycleGAN and DCNN, respectively. The training was stopped when the validation loss did not improve further.

2.4 | sCT and uncertainty map generation using Monte Carlo dropout (MCD)

sCTs and uncertainty maps were generated from each model using the MCD method by performing multiple inferences with active dropout layers.[18] The fourth and final layers of convolutional blocks were designated as the active layers for the DCNN model with a dropout rate of 0.1, while the bottom layer of Resnet blocks were selected as the active dropout layer for the cycleGAN with a dropout rate of 0.5. For each patient in the test set, a final sCT was obtained by averaging the generated inferences. The corresponding uncertainty map was obtained from the standard deviation of the voxels of the multiple inferences. Six inference frequencies (5, 10, 15, 20, 25, and 30) were used to determine whether the number of inferences affects the quality of the final sCT and uncertainty map.

To assess the impact of multiple inferences on the quality of the sCT, the mean absolute error (MAE) and structural similarity index (SSIM)[4,9,10] were used to evaluate the image similarity between the sCT and deformed pCT. Additionally, the dice coefficient and 95% Hausdorff distance were calculated using the segmented bone structures (HU > 250).

Of the 20 test patients, 14 were treated with intensity-modulated proton therapy (IMPT) and six with volumetric modulated arc therapy (VMAT). Only the proton plans were considered for the dosimetric evaluation in this study. The clinical proton treatment plans were

recalculated on both sCT and pCT. Dose calculations were performed using the clinical Monte Carlo dose engine with a dose grid of 3 mm × 3 mm × 3 mm in Raystation Research 11B. To evaluate the dosimetric performance of the sCTs, the D99, D95, D50, and mean dose of the clinical target volume (CTV) were measured, along with the mean dose to relevant OARs. In addition, the dose distribution corresponding to each beam angle was extracted, and the dose error was calculated using the voxel difference between the calculated dose of the pCT and both sCTs. Moreover, a gamma index analysis was conducted to evaluate the agreement of the dose distribution between pCT and both sCTs and for each beam angle using 2%/2 mm and 3%/3 mm criteria using the image processing software, Plastimatch (version 1.9.4, www.plastimatch.org).[30]

2.5 | Evaluation of uncertainty maps

The effectiveness of uncertainty maps in identifying areas of potential errors was assessed by comparing the generated uncertainty maps to the observed HU errors between sCT and pCT, proton range errors, water equivalent thickness (WET) errors, dose errors, and the corresponding gamma index maps. All evaluation metrics used the deformed pCT as the 'ground truth.' Pearson's correlation coefficients were calculated between the uncertainty map and all other metrics to quantify the correlation.

2.5.1 | Proton range error (RE) and water equivalent thickness error (WET)

The proton range and WET error maps were obtained using the openREGGUI proton radiography toolbox.[31,32] The algorithm simulates a proton radiography acquisition on the entire CT image based on a multilayer ionization chamber (MLIC), as described by Farace et al.[33] This study simulated proton radiographies (PR) on the pCT and sCT using a 210-MeV proton beam from a 270-degree gantry angle, a spot spacing of 5 mm, and a beam sigma of 3 mm. Similarly, the corresponding WET maps were generated using the same algorithm and pCT-sCT image sets. Corresponding range and WET error maps were obtained from the difference between the pCT and sCT PR and WET maps.

To ensure a consistent comparison between the uncertainty maps and evaluation metrics, the uncertainty and HU error maps (3D) were projected along the 90-degree angle, similar to a digitally reconstructed radiograph (DRR), and compared to the 2D PR and WET error maps. Similarly, to evaluate against the dose distribution, the dose, and uncertainty maps were projected along the clinical beam angles for comparison. A mask based on the 5%-isodose curve was applied to the uncertainty map to focus the analysis on the voxels along the beam's trajectories.

2.6 | Effects of dropout rates and dropout layers

To investigate the effects of the dropout rate on the uncertainty map and sCT, synthetic CTs were generated using the same model architecture and trained weights but with varying dropout rates. The dropout rates range from 0.1 to 0.5, indicating a 10% to 50% probability of deactivating the dropout nodes. Dropout rates greater than 0.5 were not used, as this can lead to more connections being severed, thus affecting the overall quality of the inference.[34]

Furthermore, three configurations were investigated to evaluate the effects of differing active dropout layers. The first configuration (A) sets the initial half of the dropout layers as active while the latter remains deactivated. The second configuration (B) sets the second half of the dropout layers as active, and the third configuration (C) sets alternating active and deactivated layers (Supplementary A2).

For each sCTs generated, the image quality and correlation coefficients were evaluated and compared to the sCTs generated using the same dropout rate during training, which utilized dropout rates of 0.1 for DCNN and 0.5 for cycleGAN.

3 | RESULTS

3.1 | sCT Evaluation

A total of 16 932 slices from the training set and 2029 slices from the validation set were employed to train the DCNN and cycleGAN models. The training process for the DCNN and cycleGAN was concluded after 135 and 180 epochs when the validation loss did not decrease further. This corresponds to 4 days of continuous runtime for the DCNN and 14 days for the cycleGAN. Regarding the sCT inference time, both models showed efficient processing with a synthesis time of 7 s for the DCNN and 3 s for the cycleGAN.

Table 1 summarizes the calculated metrics for each sCT generated from the different number of inferences and the corresponding synthesis completion time. The MAE for the sCT generated after 10 inferences was 56.34 ± 8.6 HU for the DCNN and 86.55 ± 7.90 HU for the cycleGAN. The difference in MAE between 10 and 30 inferences was negligible, with 0.1 HU and 0.31 HU for the DCNN and cycleGAN, respectively. Therefore, the sCT generated after 10 inferences were used for this study's subsequent analysis.

In terms of dosimetric performance, no significant difference was observed between the pCT and sCT doses

TABLE 1 Image quality metrics between DCNN and cycleGAN sCTs for each number of forward inferences, the dose difference of D99, D95, D50, and D_{mean} to the CTV, and the gamma pass rate using 2%/2 mm and 3%/3 mm criteria.

	DCNN					cycleGAN				
Number of inference	**MAE (in HU)**	**DICE**	**SSIM**	**HD95**	**Time to complete (s)**	**MAE (in HU)**	**DICE**	**SSIM**	**HD95**	**Time to complete (s)**
5	56.52	0.86	0.94	1.68	35	87.61	0.79	0.89	2.7	15
10	56.34	0.86	0.94	1.66	70	86.55	0.79	0.89	2.72	30
15	56.3	0.86	0.94	1.68	105	86.21	0.79	0.89	2.73	45
20	56.27	0.86	0.94	1.68	140	86.02	0.79	0.89	2.73	60
25	56.25	0.86	0.94	1.66	175	85.93	0.79	0.89	2.73	75
30	56.24	0.86	0.94	1.68	210	85.84	0.79	0.89	2.73	90

% Dose difference (CTV)

	D99	D95	D50	D_{mean}	D99	D95	D50	D_{mean}
Mean DD (%)	0.18 ± 0.24	0.14 ± 0.13	0.08 ± 0.06	0.07 ± 0.05	0.43 ± 0.52	0.28 ± 0.23	0.22 ± 0.34	0.22 ± 0.31
[min, max] (cGy)	[0, 0.9]	[0, 0.41]	[0, 0.23]	[0.02, 0.2]	[0.02, 1.73]	[0.04, 0.88]	[0.02, 1.18]	[0.02, 1.15]

Gamma pass rate

	2%/2 mm	3%/3 mm	2%/2 mm	3%/3 mm
Pass rate (%)	99.29 ± 0.81	99.76 ± 0.43	97.39 ± 2.51	99.10 ± 1.27
[min, max] (%)	[97.19, 99.99]	[98.49, 100.00]	[89.57, 99.51]	[95.22, 99.96]

($p = 0.44$). The gamma index analysis indicated better performance of the DCNN sCT, particularly with the 2%/2 mm criteria, achieving a 99.20 ± 0.81% gamma pass rate compared to the 97.39 ± 2.51% for cycleGAN, as illustrated in Table 1 ($p < 0.05$). Additionally, regarding the dose difference between the pCT and sCT on the target volume, the DCNN had a dose difference for the target D95 of 0.14 ± 0.13 %, while cycleGAN produced a D95 of 0.28 ± 0.23%. Moreover, it was observed that the DCNN shows a closer agreement with the pCT in terms of the mean dose in OARs. The full table on the mean DD of the target and OARs can be found in the Supplementary material (B).

3.2 | Uncertainty map correlation

Figure 1 shows an overview of the projections of the uncertainty map and HU error, mean absolute range error and WET difference maps for one test case. Visual inspection of Figure 1 shows that regions of high uncertainty correspond to regions of high errors in the image difference and range error maps, as highlighted in the figure. It can be observed that the cycleGAN uncertainty projection illustrates high uncertainty in the bone structures as compared to the DCNN, which suggests the uncertainty of the cycleGAN to convert bone tissues. Figure 2 shows the violin plot of the distribution of the R-values between the uncertainty map and HU error (UvHU), uncertainty and range error (UvRE), and uncertainty and WET difference (UvWET). Evaluation of the correlation coefficients shows an average of $r = 0.92 \pm 0.03$ and $r = 0.92 \pm 0.03$ for UvHU, $r = 0.66 \pm 0.09$ and $r = 0.62 \pm 0.06$ for the UvRE, and $r = 0.64 \pm 0.06$ and $r = 0.58 \pm 0.07$ for the UvWET for the DCNN and cycleGAN model, respectively. Overall results show a positive correlation between the uncertainties generated using MCD and all metrics investigated.

To calculate the correlation between the uncertainty and the dose difference and gamma index analysis, a mask was derived from the 5% isodose curve. This approach ensures that only uncertainties present within the beam's trajectory are incorporated into the projected map. Figure 3 shows an example of the projected uncertainty map within the beam's path and the projected dose difference and gamma analysis for the 200-degree beam angle for both models. Visual inspection of Figure 3 also reveals regions where there are dose discrepancies, which corresponds to regions of high uncertainties. Conversely, regions exhibiting high gamma index values correspond to regions of high uncertainties in the projected uncertainty map.

Our clinic's standard treatment plan for a head and neck patient consists of 4 beam angles. Typical beam angles range from 45–55, 110–165, 180–220, and 300–340 degrees. Figure 4 shows the distribution of the correlation coefficient per beam angle. The DCNN model resulted in an average correlation coefficient (R-value) of 0.57 ± 0.11 for the 45–55 degrees range (A1LV), 0.73 ± 0.06 for the 110–165 range (A1LA), 0.75 ± 0.06 for the 180–220 degrees (A1RA), and 0.59 ± 0.08 for the 300–340 degrees range (A1RV). The cycleGAN sCTs produced an average coefficient of 0.57 ± 0.1, 0.72 ± 0.07, 0.71 ± 0.06, and 0.62 ± 0.08 for the 45–55, 110–165, 180–220, and 300–340 degrees, respectively.

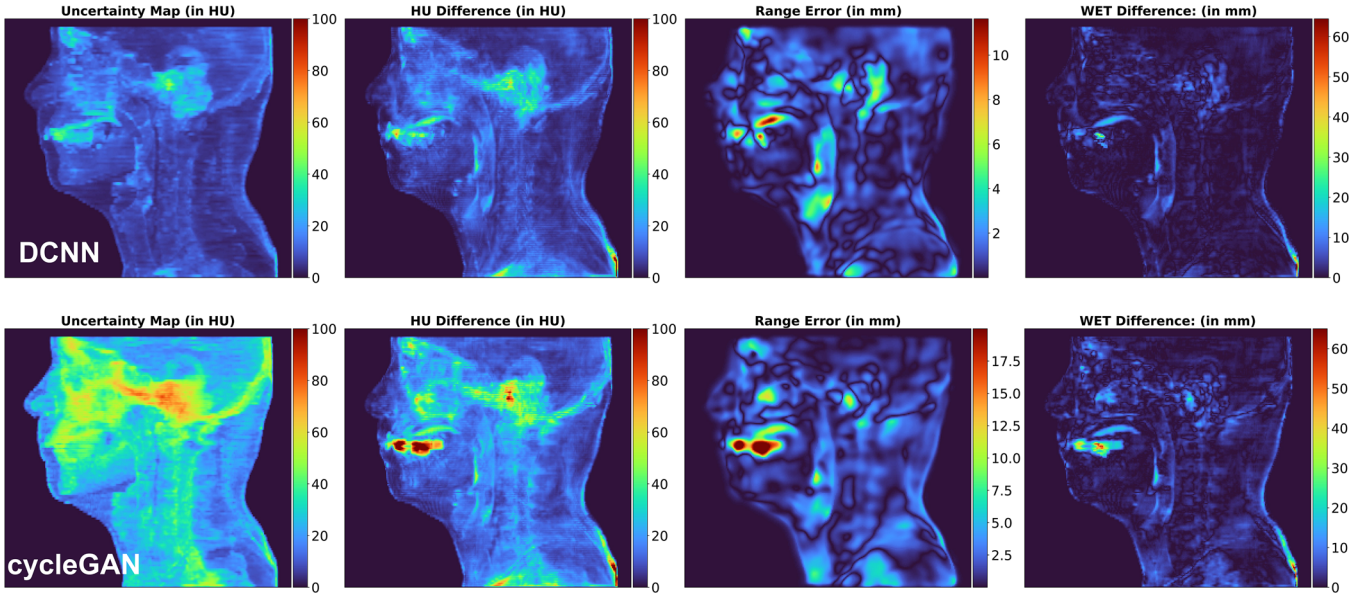

FIGURE 1 Overview of the uncertainty map and HU error between the pCT and sCT projected along the 90-degree angle, simulated range error, and water equivalent thickness for the DCNN (top) and cycleGAN model (bottom) for test patient 59.

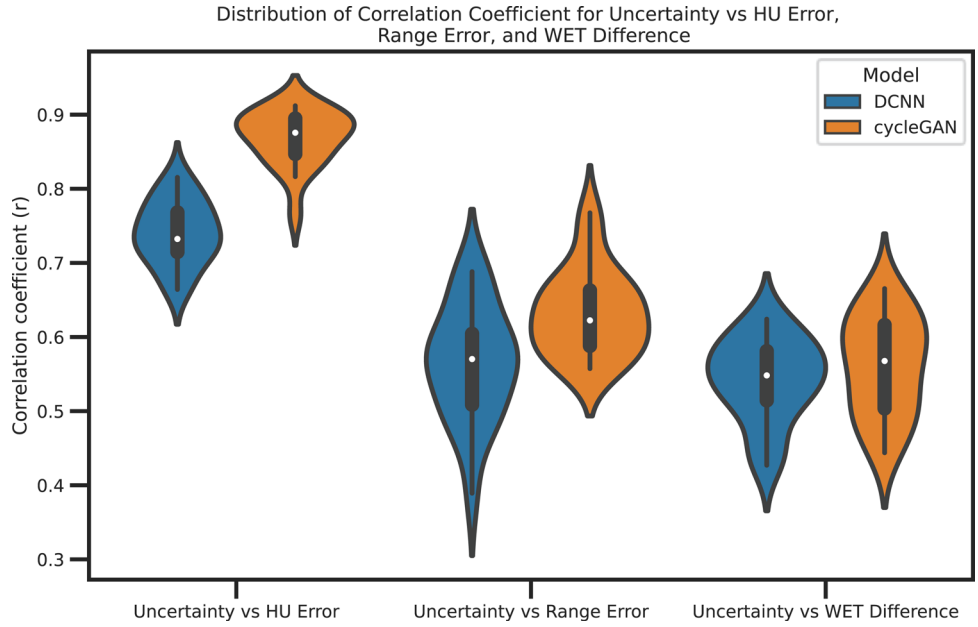

FIGURE 2 Violin plot of the correlation coefficient between the uncertainty and HU error, uncertainty and range error, and uncertainty and wet difference.

3.3 | Effects of dropout rates and active dropout layers

The impact of the dropout rates on the resulting sCT is more pronounced than the influence of the different configurations of the active dropout layers. Figure 5a presents the average mean absolute error and DICE similarity coefficients of sCTs generated using dropout rates of 0.1, 0.2, 0.3, 0.4, and 0.5 for both DCNN and cycleGAN models. In both instances, the sCTs generated with the dropout rate used during training exhibited better image quality than those sCTs using the other dropout rates investigated. The MAE ranges from 55.96 ± 8.33 HU at 10% dropout to 63.72 ± 8.11 HU at 50% dropout for the DCNN and 85.77 ± 7.44 HU at 50% dropout to 90.42 ± 8.31 HU at 10% dropout probability for the cycleGAN model. Statistical tests show no significant difference between the means of each dropout rate except for the DCNN 10% and 50% ($p = 0.005$).

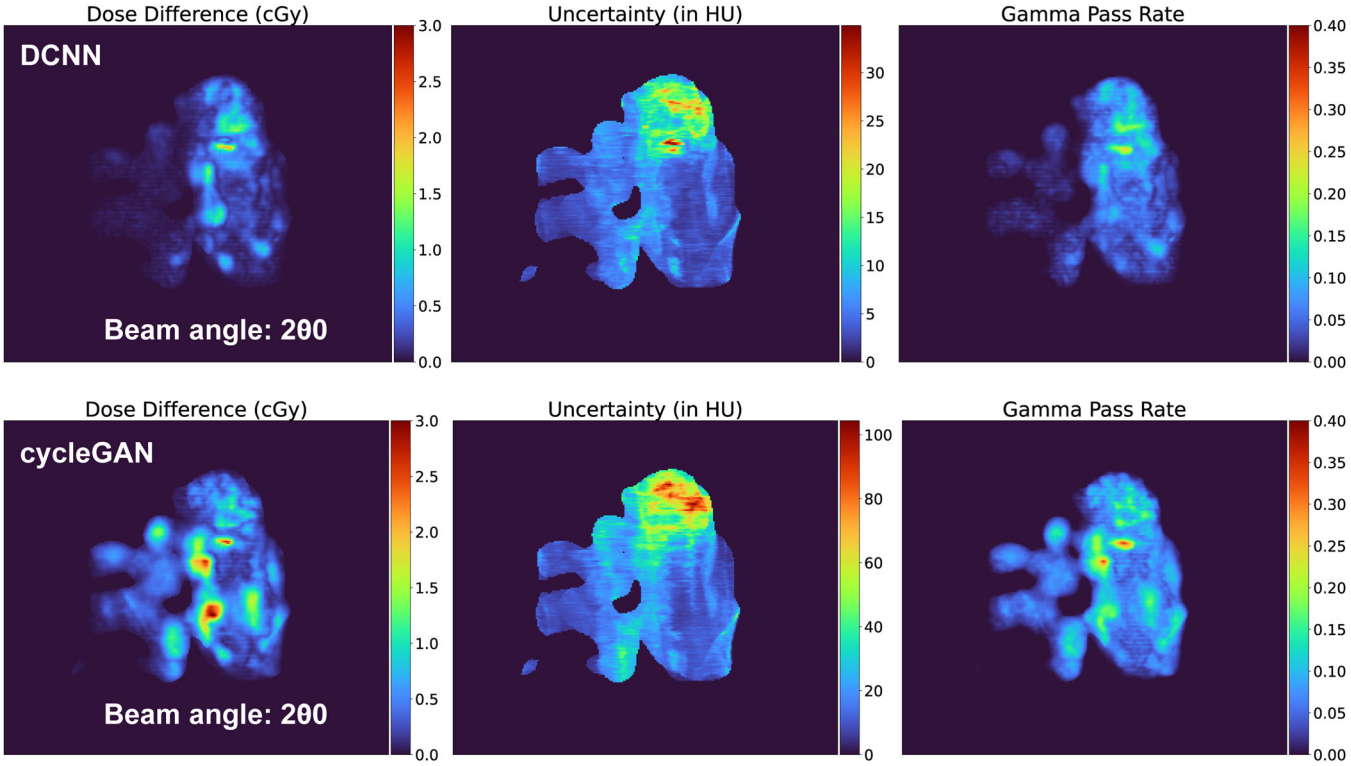

FIGURE 3 Projected absolute dose difference, uncertainty map within the beam path, and the projected gamma index along the 200-degree beam angle for test patient 59.

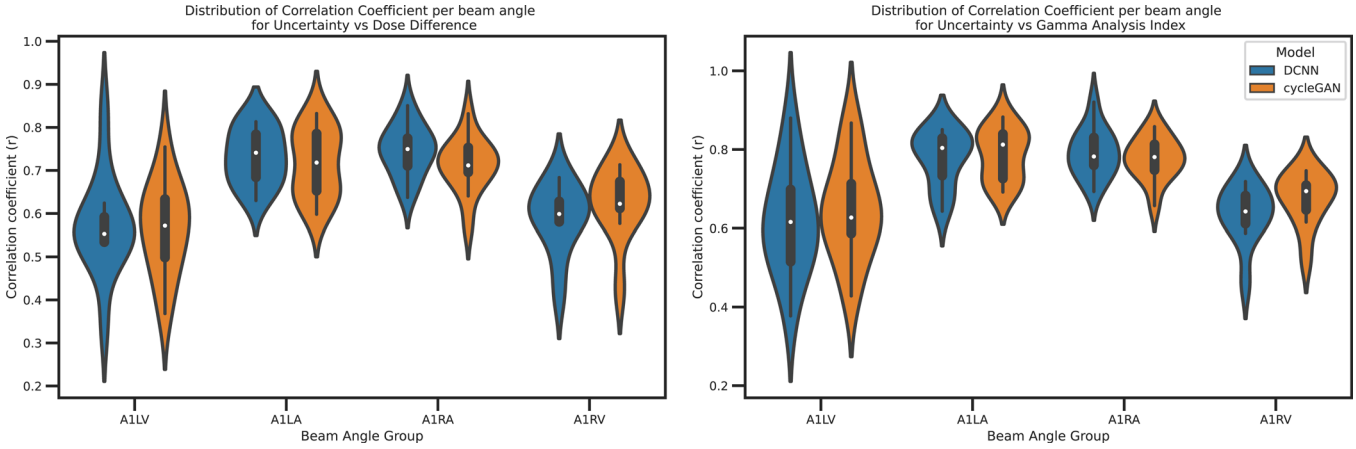

FIGURE 4 Distribution of the correlation coefficient between the uncertainty and dose difference (left) and uncertainty and gamma index (right) for each clinical beam angle group.

Figure 5b illustrates the correlation coefficients for UvHU, UvRE, and UvWET corresponding to each dropout rate. Similar to the case of the image quality metrics, the most favorable UvHU, UvRE, and UvWET values were obtained from sCTs generated using the 10% dropout probability for the DCNN model and the 50% dropout probability for the cycleGAN model. However, the difference between the 10% and 50% coefficients for DCNN and cycleGAN is insignificant ($p > 0.5$).

No significant difference in the image quality and correlation coefficients was observed for sCTs generated using different configurations of the active dropout layers for both DCNN and cycleGAN (Supplementary C).

4 | DISCUSSION

This work demonstrated the potential of using MCD-based uncertainty maps to assess the quality of

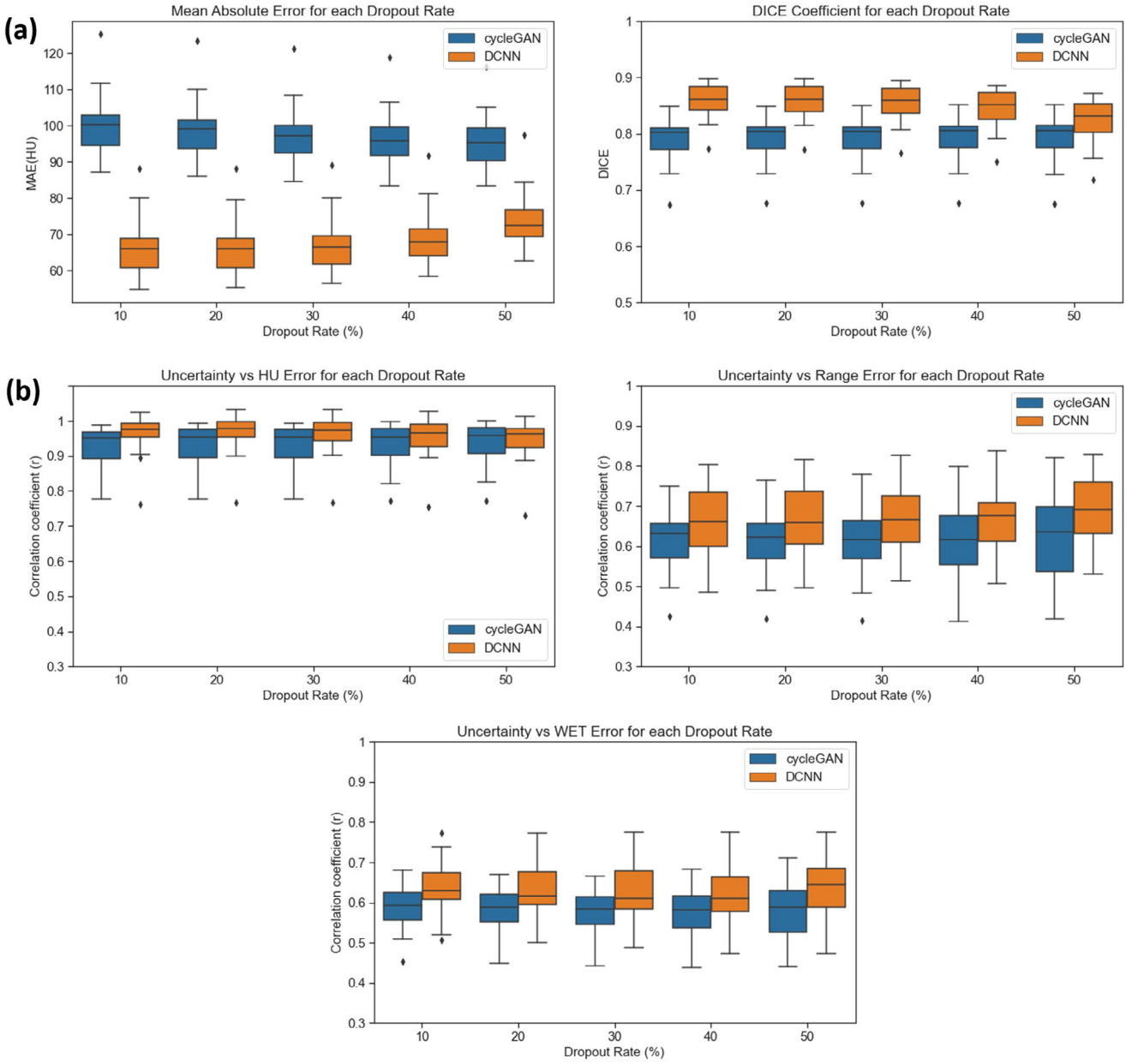

FIGURE 5 (a) Distribution of the mean absolute error (left) and DICE similarity coefficient (right)for each dropout rate. (b) Distribution of the correlation coefficient between the uncertainty and HU error (right), uncertainty and range error (left), and uncertainty and WET Error (bottom).

DL-based sCTs. The use of MCD made uncertainty estimation straightforward to implement. Existing sCT models trained using dropout layers can be easily modified to perform MCD-based inferences. The choice of dropout rates and active dropout layers does not significantly affect the image quality and correlation of the uncertainty to the errors evaluated. However, we recommend using the dropout rates used in fine-tuning the weights during training as these parameters were chosen to prioritize the quality of the sCTs. Verdoja et al. presented observations of the behavior of MCD using different parameters.[35] Verdoja et al. noted that the estimated epistemic uncertainties depend on the dropout rate and not on the amount of training data. In our application, the dropout rates during inference were set to the same values during training to reduce this dependency.

The conversion of MRI images to sCT images presents a greater challenge than the conversion from CBCT to sCT, as the underlying physical principles between MRI and CT imaging are fundamentally different.[9] Both models were trained on the same dataset and pre-processing steps, but the sCTs generated using the deep convolutional network (DCNN) showed superior image quality compared to the cycleGAN model. This superiority can be attributed to the cycleGAN model's inaccuracy in synthesizing bone

structures from MRI to CT intensities. This was demonstrated by the noticeable difference in HU, as shown in Figure 1, which results from the model's inability to convert some dental structures to HU values properly. Nonetheless, the uncertainty map produced by the cycleGAN model predicted this limitation. The results of the uncertainty map show the stability of the DCNN model where there are fewer regions of uncertainty, and these regions correspond with the actual difference in HU observed.

In this study, we utilized a method to visualize uncertainties efficiently and accurately by projecting the 3D uncertainty map along an axis. This approach allowed for the comparison of uncertainty maps, the simulated proton radiography, and the WET maps obtained from all possible angles. This implies that uncertainty maps can serve as a complementary validation tool for PR-based quality control methods such as that proposed by Seller Oria et al.[16] Moreover, the versatility of the method allows for the evaluation of the correlation between dose differences and uncertainties from actual beam angles. The application of the isodose mask ensures that only voxels along the beam paths are considered in the analysis. Areas such as the jaw are prone to misalignment even when images are taken closely together and with immobilization devices. However, treatment beams through these areas are avoided because of these issues. By limiting the analysis to the beam paths, only clinically relevant uncertainties are assessed. This is particularly beneficial in proton therapy as the dose distribution is more concentrated on the target region than in photon therapy.

Overall, a positive correlation between the uncertainty map and the various metrics was observed. Figure 2 illustrates that the correlation between the UvHU and UvRE were more favorable for the cycleGAN model than the DCNN. It is necessary to understand that the uncertainty map only highlights areas where the model has the highest level of uncertainty and does not necessarily imply errors. Therefore, even if the model is uncertain in some regions, but the sCT quality is superior, such as that of the DCNN, fewer errors lead to a slightly lower correlation.

Dosimetric quantities such as D99 and D95 of the target volume for both models were comparable, and the gamma analysis revealed an average pass rate of >95% for both the 2%/2 mm and 3%/3 mm criteria. There was no convincing disparity observed in the average correlation between the cycleGAN and DCNN for the dose difference and gamma index analysis. We have observed a lower correlation for the dose-based metrics, as shown in Figure 3, than for the HU difference metric. One plausible explanation is that the dose within a specific voxel also depends on other voxels along the beam's trajectory up to that voxel. We account for this dependence by projecting the dose and uncertainty along the beam delivery angles. However, this does not fully account for the physical process of dose deposition, which may contribute to the reduced overall correlation between the gamma index and HU uncertainty. On average, the correlation distribution was higher for beams originating from the posterior angles, which may be attributed to the more homogenous tissue distribution from the posterior direction.

The integration of MCD-based uncertainty maps into the adaptive proton therapy workflow provides a means of quality assurance for DL-based sCTs. Nevertheless, significant limitations must be addressed to ensure their effectiveness and reliability in clinical practice. Specifically, it is crucial to establish acceptable uncertainty thresholds that may indicate whether the sCT is suitable for clinical use. Additionally, the training and evaluation dataset utilized were carefully curated and did not contain significant outliers. Thus, it is essential to investigate further the behavior of the uncertainty maps for out-of-distribution datasets (e.g., differing imaging protocols or anatomical features). In addition, the method can also be applied to DL-based sCTs used in MR-guided adaptive photon therapy. However, since dose distribution for photons are different to that of protons, evaluation of the correlation between the uncertainty and dose metrics for photon dose distribution should be performed. Furthermore, this study confines its assessments to quantify model-dependent uncertainties. Further evaluation of the aleatoric or data-dependent uncertainties would require modifications in the loss function of the model. Incorporating the quantification of data-dependent uncertainties into models currently undergoing training would prove to be beneficial to obtain a comprehensive understanding of uncertainties involved in sCT generation using DL models.

5 | CONCLUSIONS

In this study, we demonstrated the feasibility of using MCD-based uncertainty maps in evaluating the quality of sCT images obtained from MRI. The results validated the correlation between the MCD-based voxel uncertainties and various metrics such as HU errors, range errors, WET difference, dose difference, and gamma index analysis. The positive correlation between the uncertainty map and the evaluated metrics suggests that uncertainty maps can be used to identify regions where potential range errors are likely to occur in sCT images. The integration of uncertainty maps with PR QA tools takes DL-based sCTs one step closer to clinical adoption within the adaptive proton therapy workflow.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Actions under Grant Agreement No. 955956.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Arthur Villanueva Galapon Jr
https://orcid.org/0000-0001-9264-3836

REFERENCES

1. Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A. Online daily adaptive proton therapy. *Br J Radiol*. 2020;93(1107):20190594. doi:10.1259/bjr.20190594
2. Paganetti H, Botas P, Sharp GC, Winey B. Adaptive proton therapy. *Phys Med Biol*. 2021;66(22):22TR01. doi:10.1088/1361-6560/ac344f
3. Wang C, Uh J, Merchant TE, Hua CH, Acharya S. Facilitating MR-guided adaptive proton therapy in children using deep learning-based synthetic CT. *Int J Part Ther*. 2022;8(3):11-20. doi:10.14338/IJPT-20-00099.1
4. Thummerer A, Seller Oria C, Zaffino P, et al. Clinical suitability of deep learning based synthetic CTs for adaptive proton therapy of lung cancer. *Med Phys*. 2021;48(12):7673-7684. doi:10.1002/mp.15333
5. Thummerer A, Zaffino P, Meijers A, et al. Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy. *Phys Med Biol*. 2020;65(9):095002. doi:10.1088/1361-6560/ab7d54
6. Yuan N, Rao S, Chen Q, Sensoy L, Qi J, Rong Y. Head and neck synthetic CT generated from ultra-low-dose cone-beam CT following Image Gently Protocol using deep neural network. *Med Phys*. 2022;49(5):3263-3277. doi:10.1002/mp.15585
7. Maspero M, Bentvelzen LG, Savenije MHF, et al. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. *Radiother Oncol*. 2020;153:197-204. doi:10.1016/j.radonc.2020.09.029
8. Gupta D, Kim M, Vineberg KA, Balter JM. Generation of synthetic CT images from MRI for treatment planning and patient positioning using a 3-channel U-Net trained on sagittal images. *Front Oncol*. 2019;9:964. doi:10.3389/fonc.2019.00964
9. Thummerer A, de Jong BA, Zaffino P, et al. Comparison of the suitability of CBCT- and MR-based synthetic CTs for daily adaptive proton therapy in head and neck patients. *Phys Med Biol*. 2020;65(23):235036. doi:10.1088/1361-6560/abb1d6
10. Boulanger M, Nunes JC, Chourak H, et al. Deep learning methods to generate synthetic CT from MRI in radiotherapy: a literature review. *Phys Med*. 2021;89:265-281. doi:10.1016/j.ejmp.2021.07.027
11. Taasti VT, Bäumer C, Dahlgren CV, et al. Inter-centre variability of CT-based stopping-power prediction in particle therapy: survey-based evaluation. *Phys Imaging Radiat Oncol*. 2018;6:25-30. doi:10.1016/j.phro.2018.04.006
12. Yang M, Zhu XR, Park PC, et al. Comprehensive analysis of proton range uncertainties related to patient stopping-power-ratio estimation using the stoichiometric calibration. *Phys Med Biol*. 2012;57(13):4095. doi:10.1088/0031-9155/57/13/4095
13. Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. *Phys Med Biol*. 2012;57(11):R99-R117. doi:10.1088/0031-9155/57/11/R99
14. Palmer E, Karlsson A, Nordstrom F, et al. Synthetic computed tomography data allows for accurate absorbed dose calculations in a magnetic resonance imaging only workflow for head and neck radiotherapy. *Phys Imaging Radiat Oncol*. 2021;17:36-42. doi:10.1016/j.phro.2020.12.007
15. Irmak S, Zimmermann L, Georg D, Kuess P, Lechner W. Cone beam CT based validation of neural network generated synthetic CTs for radiotherapy in the head region. *Med Phys*. 2021;48(8):4560-4571. doi:10.1002/mp.14987
16. Oria CS, Marmitt GG, Both S, Langendijk JA, Knopf AC, Meijers A. Classification of various sources of error in range assessment using proton radiography and neural networks in head and neck cancer patients. *Phys Med Biol*. 2020;65(23):235009. doi:10.1088/1361-6560/abc09c
17. Hemsley M, Chugh B, Ruschin M, et al. *Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions*. Springer; 2020:834-844.
18. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: Proceedings of The 33rd International Conference on Machine Learning. PMLR; 2016:1050-1059.
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15(56):1929-1958.
20. Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Adv Neural Inf Process Syst*. 2017:6405-6416.
21. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging*. 2010;29(1):196-205. doi:10.1109/TMI.2009.2035616
22. Shamonin D, Bron E, Lelieveldt B, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. *Front Neuroinformatics*. 2014;7:50. https://www.frontiersin.org/articles/10.3389/fninf.2013.00050
23. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging*. 1998;17(1):87-97.
24. Cox IJ, Roy S, Hingorani SL. Dynamic histogram warping of image pairs for constant image brightness. *IEEE*. 1995;2:366-369.
25. Spadea MF, Pileggi G, Zaffino P, et al. Deep Convolution Neural Network (DCNN) multiplane approach to synthetic CT generation from MR images-application in brain proton therapy. *Int J Radiat Oncol Biol Phys*. 2019;105(3):495-503. doi:10.1016/j.ijrobp.2019.06.2535
26. Spadea MF, Maspero M, Zaffino P, Seco J. Deep learning based synthetic-CT generation in radiotherapy and PET: a review. *Med Phys*. 2021;48(11):6537-6566. doi:10.1002/mp.15150
27. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. NIPS 2017 Workshop.
28. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. 2017:2223-2232. Accessed February 4, 2023. https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html
29. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. 2017:1125-1134.
30. Zaffino P, Raudaschl P, Fritscher K, Sharp GC, Spadea MF. Technical Note: plastimatch mabs, an open source tool for automatic image segmentation. *Med Phys*. 2016;43(9):5155. doi:10.1118/1.4961121
31. Deffet S, Cohilis M, Souris K, et al. openPR – A computational tool for CT conversion assessment with proton radiography. *Med Phys*. 2021;48(1):387-396. doi:10.1002/mp.14571
32. Farace P, Righetto R, Deffet S, Meijers A, Vander Stappen F. a direct ray-tracing method to compute integral depth dose in pencil beam proton radiography with a multilayer ionization chamber. *Med Phys*. 2016;43(12):6405-6412.
33. Farace P, Righetto R, Meijers A. Pencil beam proton radiography using a multilayer ionization chamber. *Phys Med Biol*. 2016;61(11):4078-4087. doi:10.1088/0031-9155/61/11/4078
34. Baldi P, Sadowski P. The dropout learning algorithm. *Artif Intell*. 2014;210:78-122. doi:10.1016/j.artint.2014.02.004

35. Verdoja F, Kyrki V. Notes on the behavior of MC dropout. Published online July 13, 2021. Accessed February 2, 2023. http://arxiv.org/abs/2008.02627

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Galapon AV, Thummerer A, Langendijk JA, Wagenaar D, Both S. Feasibility of Monte Carlo dropout-based uncertainty maps to evaluate deep learning-based synthetic CTs for adaptive proton therapy. *Med Phys*. 2023;1-11. https://doi.org/10.1002/mp.16838