Can we trust synthetic CT algorithms? Uncertainty-aware evaluation of CBCT to CT synthesis for adaptive proton therapy

Blanca Rodriguez-Gonzalez[1][0009−0007−0982−1293], Ivan Escobar Corominas[2][0009−0009−5280−6229], Angel Torrado-Carvajal[1][0000−0002−1540−2809], Borja Rodriguez-Vila[1][0000−0003−4779−0225], Alejandro Mazal[2][0000−0003−1391−8446], and Norberto Malpica[1][0000−0003−4618−7459]

[1] Medical Image Analysis and Biometry Laboratory at Universidad Rey Juan Carlos, Madrid, 28933, Spain
[2] Medical Physics Department at Centro de Protonterapia Quironsalud, Madrid, Spain
blanca.rodriguez@urjc.es

Abstract. Synthetic CT (sCT) generation from cone beam CT (CBCT) has emerged as a key component of adaptive proton therapy (APT), enabling daily dose recalculation. However, image quality metrics alone provide limited information regarding prediction reliability. This work studies whether uncertainty estimates can identify regions where synthesis errors may have dosimetric consequences. A 2.5D Attention U-Net was trained using three uncertainty quantification strategies (Monte Carlo Dropout, Deep Ensembles and Evidential Deep Learning (EDL)). Beyond conventional image-quality evaluation, uncertainty was assessed through its spatial association with Hounsfield Unit (HU) errors and dose discrepancies. EDL achieved the best image-domain performance and the strongest uncertainty-error agreement. While global dosimetric differences between methods were modest, uncertainty maps localized regions of increased HU and dose error. These findings suggest that uncertainty provides complementary reliability information and should be considered as part of quality assurance for sCT-based APT.

Keywords: Uncertainty Quantification · Proton Therapy

1 Introduction

Adaptive radiotherapy (ART) is transforming radiation treatment from a static process based on a single planning image into a dynamic workflow that accounts for anatomical changes during the treatment [1]. This change relies on the availability of repeated imaging, routinely acquired for patient positioning, which provides updated information about patient's anatomy across fractions [2, 3]. In proton therapy, such anatomical variations are particularly relevant because small changes in tissue density or geometry can alter proton range, compromising dose accuracy [4]. Therefore, adaptive proton therapy (APT) requires not

only anatomical monitoring, but also reliable image data from which dose can be recalculated when relevant changes are observed [5].

CBCT is routinely acquired during treatment since it provides anatomical information for patient positioning and treatment monitoring; however, its image quality and HU inaccuracies limit its direct use for proton dose calculation [5]. To overcome these limitations, synthetic CT (sCT) generation has emerged as a key element of adaptive workflows, enabling dose recalculation from daily images [6, 7]. Deep learning (DL) approaches have become the main strategy due to their ability to preserve CT electron density information [8–12]. Despite these advances, image quality alone is insufficient for decision-making in ART. A sCT may appear visually accurate while still containing local errors with dosimetric consequences. Uncertainty quantification (UQ) can complement predictions by providing spatial information on prediction reliability, supporting quality assurance and helping identify cases that may require further review [13, 14].

Several approaches have been proposed to estimate uncertainty in DL models. Monte Carlo (MC) Dropout and Deep Ensembles (DE) quantify uncertainty through stochastic sampling, whereas Evidential Deep Learning (EDL) directly models predictive uncertainty in a single forward pass [15]. Although these methods have been increasingly explored in ART [14], their role in APT remains largely unexplored, particularly regarding whether image uncertainty can inform dose-calculation reliability. Existing studies on uncertainty propagation in sCT-based workflows are still scarce and have mainly focused on ART settings or on a single uncertainty strategy [16, 17]. Unlike prior sCT studies primarily focused on image similarity, this work investigates whether voxel-wise uncertainty estimates are spatially aligned with synthesis errors and whether they can support dosimetric reliability assessment in Head and Neck (H&N) APT.

2 Methods

2.1 The Dataset

Anonymized data was retrospectively retrieved from a real clinical H&N proton therapy cohort treated at Centro de Protonterapia Quironsalud (Madrid, Spain). The dataset included 84 patients, each containing 2-7 image pairs constructed by matching the reference CT, weekly acquired during treatment, with the closest CBCT. For each pair, the reference CT was non-rigidly registered to the corresponding CBCT using Elastix. Registered CTs were cropped to a standardized H&N FOV to match the smaller CBCT coverage. Intensities were clipped to $[-1024, 3071]$ HU and min-max normalized to $[0, 1]$. Data was split using five-fold patient-level strategy, with an independent test set of 10 patients.

2.2 Uncertainty Quantification Strategies

Monte Carlo Dropout estimates predictive uncertainty by enabling dropout at inference and performing T ($T = 15$) stochastic forward passes with a given dropout rate (0.25). The final sCT was computed as the mean prediction, while uncertainty was estimated as the voxel-wise variance across stochastic samples.

Deep Ensembles allow to estimate uncertainty from the variability among K ($K = 5$) independently trained models. The ensemble sCT was obtained as the mean prediction, while uncertainty was estimated as the voxel-wise variance across ensemble members.

Evidential Deep Learning To estimate predictive uncertainty in a single forward pass, we adopted an evidential regression formulation based on the Normal-Inverse-Gamma (NIG) distribution [18]. Instead of predicting only a deterministic intensity, the network outputs four voxel-wise parameters, $\mu, \lambda, \alpha, \beta$, which define a prior over the mean and variance of a Gaussian likelihood. The parameter μ represents the predictive mean (i.e., sCT intensity), whereas λ, α and β control the amount of associated evidence. For each voxel, the model defines:

$$p(y \mid \mu, \lambda, \alpha, \beta) = \mathrm{NIG}(y \mid \mu, \lambda, \alpha, \beta), \quad \text{with } \lambda, \beta > 0 \text{ and } \alpha > 1 \tag{1}$$

The evidential formulation allows the predictive uncertainty to be decomposed into aleatoric (data-dependent) and epistemic (model-dependent) terms.

$$\sigma^2_{\text{alea}} = \frac{\beta}{\alpha - 1} \qquad \sigma^2_{\text{epis}} = \frac{\beta}{\lambda(\alpha - 1)} \tag{2}$$

The total predictive uncertainty was obtained as:

$$\sigma^2_{\text{EDL}} = \sigma^2_{\text{alea}} + \sigma^2_{\text{epis}}. \tag{3}$$

The model was trained by minimizing the negative log-likelihood (NLL) of the marginal Student's t-distribution induced by the NIG prior combined with an evidential regularization term penalizing high evidence for large prediction errors. Additional details are provided at https://github.com/brodgon/MIART2026-UQsCT-APT.

2.3 The Network

All experiments build upon a 2.5D Attention U-Net [19] to isolate the effect of the UQ strategy. The network receives seven consecutive axial CBCT slices of size 512×512 and predicts the corresponding central CT slice, allowing the incorporation of anatomical context while avoiding the memory burden of 3D models. The architecture follows a U-Net design with four resolution levels, starting with 64 filters, doubling after downsampling. Attention gates are applied to the skip connections before concatenation with decoder features.

All models were trained in TensorFlow 2.12 on an NVIDIA A100 for up to 250 epochs, with batch size 8 and an initial learning rate 10^{-4}. A deterministic baseline was trained using the same backbone to assess if UQ strategies preserved image quality. Baseline, MC Dropout and DE were optimized using the same reconstruction loss, combining MAE, SSIM and gradient consistency.

2.4 Evaluating UQ tasks

Besides standard image-quality metrics, UQ performance was evaluated by comparing predicted uncertainty with the absolute HU error. Voxel-wise agreement was quantified using Spearman correlation, while spatial agreement was measured with the uncertainty-error overlap (UEO) Dice coefficient, obtained by thresholding both error and uncertainty maps at their respective 90th percentiles. Uncertainty was also evaluated as a score for high-error voxel detection using AUROC. Model calibration was assessed by comparing expected and empirical predictive interval coverage. For a confidence level c, predictions were considered covered when $e_i \leq z_c \sigma_i$, where z_c is the Gaussian quantile and σ_i is the predicted standard deviation. Empirical coverage was computed as

$$Cov(c) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left(e_i \leq z_c \sigma_i \right), \tag{4}$$

and summarized using the Expected Calibration Error (ECE):

$$ECE = \frac{1}{M} \sum_{m=1}^{M} |Cov(c_m) - c_m|. \tag{5}$$

NLL was additionally used to assess the probabilistic quality of the predicted uncertainty. UQ maps were not post-hoc calibrated; they were converted to HU standard deviation and evaluated as produced. Calibration and NLL were therefore interpreted as empirical comparisons of native uncertainty quality under a shared Gaussian assumption [15].

2.5 Dosimetric Evaluation

To assess the dosimetric impact of synthesis errors, treatment plans optimized on the reference CT were recalculated on each generated sCT using RayStation (v2025 SP1). Global dose agreement was quantified using the mean dose bias, MAE, the 95th percentile of the absolute dose difference ($P95|\Delta D|$), CTV dose coverage difference ($|\Delta D_{95}|$) and gamma passing rate using a 3%/3 mm criterion. To evaluate whether image-domain uncertainty was informative of dosimetric reliability, the same uncertainty-error localization analysis was repeated using voxel-wise absolute dose error instead of HU error. Sparsification analysis was additionally performed by progressively excluding voxels with the highest predicted uncertainty and measuring the remaining dose MAE, allowing to assess whether high-uncertainty regions concentrated larger dosimetric discrepancies.

Statistical comparisons were performed at patient level using Friedman tests followed by paired Wilcoxon signed-rank tests with Holm-Bonferroni correction.

3 Results

Quantitative sCT results are reported in Table 1. Overall, EDL achieved the best performance for most body metrics. Within the CTV, EDL also achieved

Table 1. Quantitative evaluation of sCT image quality over body and CTV, reported as mean ± standard deviation over the test set. Superscripts denote significant EDL differences against Baseline (^{B}), MC Dropout (^{M}) and DE (^{D}), corrected $p < 0.05$.

		Baseline	MC Dropout	Deep Ensemble	EDL
sCT	MAE [HU]	33.754 ± 7.317	36.611 ± 5.322	36.625 ± 7.132	**31.361 ± 5.849**^{B,M,D}
	PSNR [dB]	30.301 ± 1.570	29.622 ± 1.101	29.892 ± 1.367	**31.995 ± 1.538**^{M}
	MS-SSIM	**0.959 ± 0.011**	0.952 ± 0.010	0.955 ± 0.011	0.958 ± 0.012
	$\text{Dice}_{\text{HU}>300}$	0.908 ± 0.018	0.880 ± 0.014	0.891 ± 0.015	**0.911 ± 0.024**^{M,D}
CTV	MAE [HU]	43.735 ± 34.567	61.796 ± 43.494	52.272 ± 42.537	**37.425 ± 30.853**^{B,M,D}
	PSNR [dB]	37.624 ± 8.096	33.688 ± 8.048	36.558 ± 8.702	**38.654 ± 8.768**^{B,M,D}
	SSIM	0.949 ± 0.040	0.910 ± 0.066	0.924 ± 0.061	**0.965 ± 0.026**^{B,M,D}

the best metrics; however, the large standard deviations indicate substantial inter-patient variability in CTV sizes. Fig. 1 further support these findings. All studied methods reduced CBCT intensity deviations, but errors remained around bone-air interfaces and external artifacts.

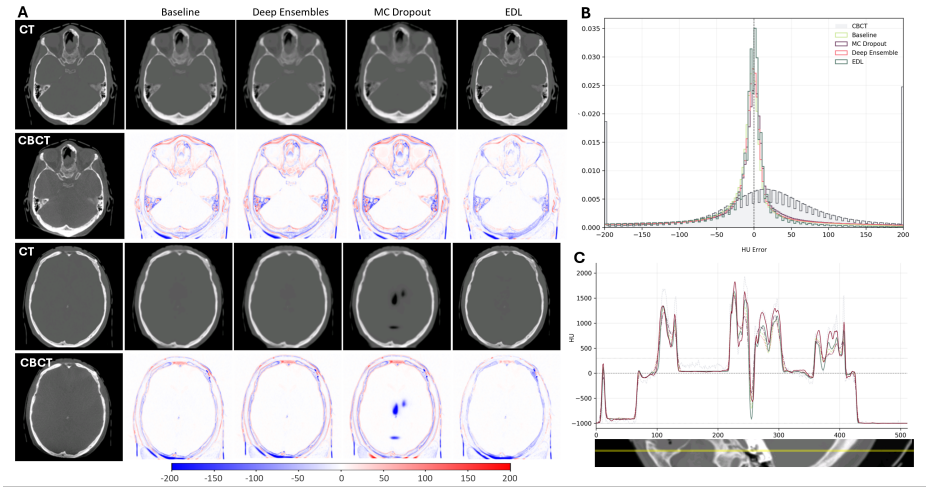

Fig. 1. Qualitative sCT comparison. (A) Representative axial slices from reference CT, CBCT and generated sCTs along with HU error maps. (B) Voxel-wise HU error distributions inside the body mask. (C) HU intensity profiles along the selected lin.e

EDL achieved the best performance across all UQ metrics (Table 2), with significantly higher Spearman correlation and UEO Dice, indicating stronger agreement between predicted uncertainty and voxel-wise HU errors. EDL also obtained the highest AUROC, although all methods showed good high-error detection capability. Naive calibration metrics further favored EDL, which achieved significantly lower ECE and NLL than MC Dropout and DE, suggesting improved probabilistic quality. As shown in Fig. 2, MC Dropout and DE tended

Table 2. UQ performance across the test cohort. Results are reported as mean ± standard deviation across patients. Superscripts indicate EDL's significant differences against MC Dropout (^{M}) and DE (^{D}), with corrected $p < 0.05$

	MC Dropout	Deep Ensemble	EDL
Spearman ($\lvert e \rvert, \sigma$)	0.713 ± 0.031	0.704 ± 0.057	**0.774 ± 0.016**^{M,D}
UEO Dice	0.718 ± 0.025	0.705 ± 0.030	**0.887 ± 0.034**^{M,D}
High-error AUROC	0.846 ± 0.032	0.826 ± 0.020	**0.874 ± 0.017**^{D}
ECE	0.331 ± 0.016	0.322 ± 0.035	**0.136 ± 0.022**^{M,D}
NLL	11.160 ± 5.683	9.995 ± 4.940	**7.857 ± 2.761**^{M,D}

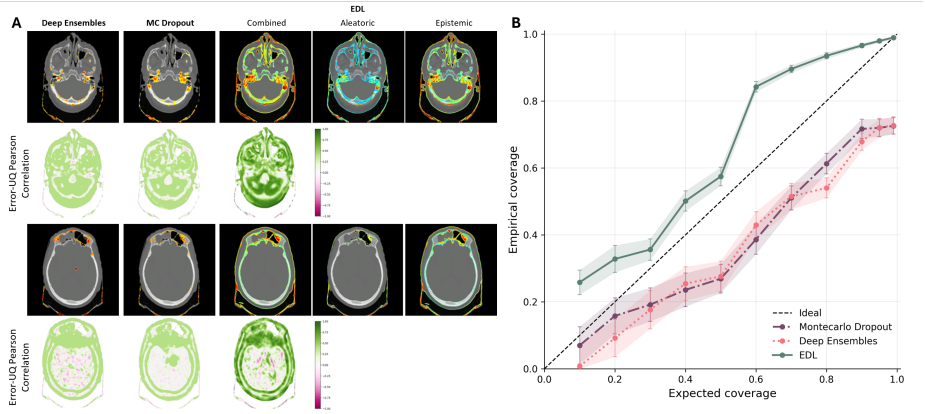

Fig. 2. Uncertainty evaluation. (A) Uncertainty maps and local error-uncertainty Pearson correlation maps. For EDL, combined, aleatoric and epistemic uncertainties are shown separately. (B) Calibration curves.

to underestimate uncertainty, with calibration curves below the ideal diagonal, whereas EDL was closer to ideal calibration but showed a conservative behavior. The EDL decomposition also revealed complementary uncertainty patterns, with aleatoric uncertainty more broadly distributed in ambiguous anatomical regions and epistemic uncertainty more localized in areas with lower model evidence.

Dosimetric agreement is reported in Table 3, with a representative example shown in Fig. 3. Overall, all methods produced similar global dosimetric agreement in the clinically relevant dose region. Differences between methods

Table 3. Global dosimetric agreement between sCT-based recalculated doses and reference CT dose, reported as mean ± standard deviation.

	Baseline	MC Dropout	Deep Ensemble	EDL
Bias [Gy(RBE)]	2.232 ± 3.763	2.306 ± 3.822	2.236 ± 3.735	**2.211 ± 3.707**
MAE [Gy(RBE)]	2.673 ± 3.454	2.707 ± 3.544	2.657 ± 3.419	**2.461 ± 3.405**
P95 $\lvert \Delta D \rvert$ [Gy(RBE)]	4.270 ± 6.885	4.337 ± 6.991	4.292 ± 6.826	**4.256 ± 6.706**
$\gamma_{3\%/3mm}$ pass rate [%]	99.47 ± 0.76	99.33 ± 0.59	99.47 ± 0.68	**99.60 ± 0.44**
CTV $\lvert \Delta D_{95} \rvert$ [Gy(RBE)]	3.897 ± 2.853	4.016 ± 2.875	3.921 ± 2.831	**3.668 ± 2.848**

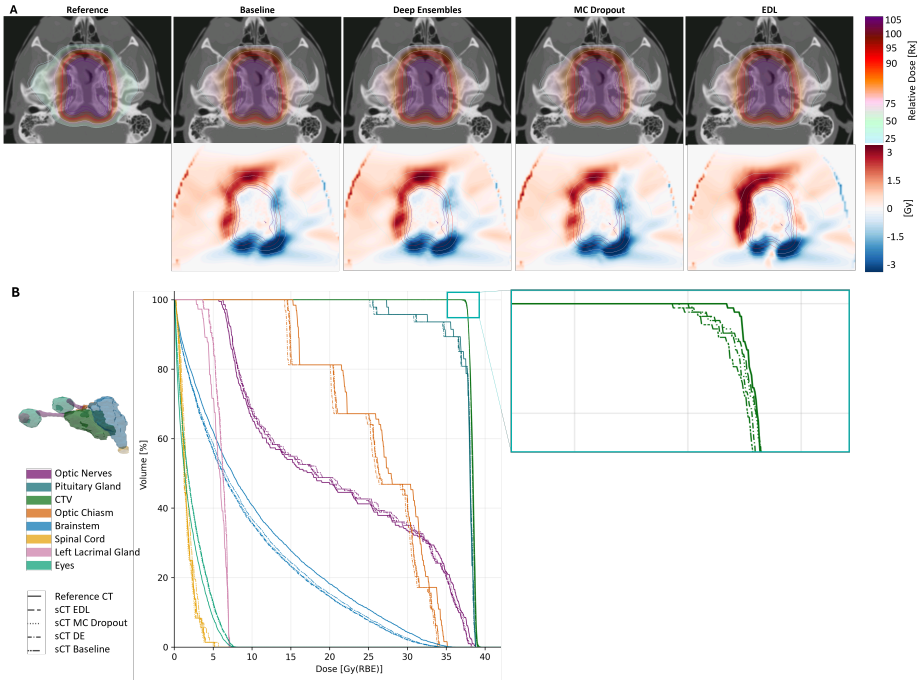

Fig. 3. Dosimetric comparison. (A) Relative dose distributions and dose difference maps. (B) DVH comparison for the CTV and selected organs at risk.

were smaller than those observed for image-domain metrics, suggesting that HU improvements do not necessarily translate into proportional reductions in dose discrepancies. Gamma analysis further confirmed the high dosimetric agreement between sCT and CT dose calculations. Similarly, CTV $|\Delta D_{95}|$ values were comparable across methods, indicating that target dose coverage was preserved despite residual local dose differences. Visual inspection of the dose maps and DVHs in Fig. 3 confirmed this trend.

Dose-error localization result are reported in Table 4 and illustrated in Fig. 4. EDL achieved the strongest metrics, indicating that its uncertainty maps were more strongly associated with regions of increased dose discrepancy. The sparsification analysis in Fig. 4 B further supports this observation. Removing voxels with the highest predicted uncertainty reduces dose error.

Table 4. UQ performance for dose-error localization (mean ± standard deviation).

	MC Dropout	**Deep Ensemble**	**EDL**
Spearman ($\|e\|, \sigma$)	0.698 ± 0.021	0.717 ± 0.047	**0.769 ± 0.018**^{M,D}
UEO Dice	0.616 ± 0.156	0.605 ± 0.269	**0.669 ± 0.120**^{M,D}
High-error (10%) AUROC	0.649 ± 0.019	0.693 ± 0.022	**0.724 ± 0.011**

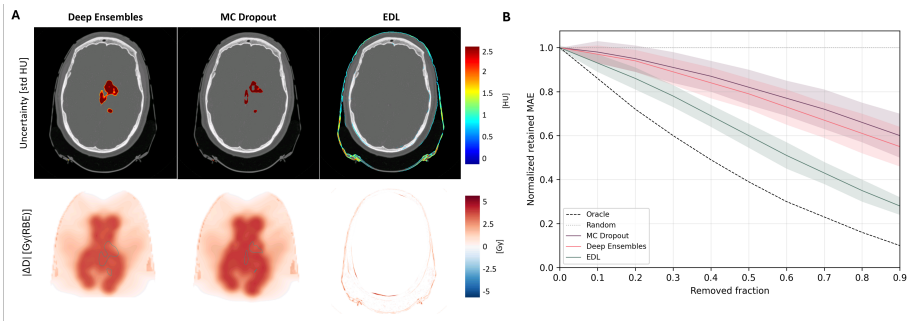

Fig. 4. Uncertainty-dose Analysis. (A) Dose-error maps and corresponding sCT uncertainty. (B) Sparsification curves. The oracle curve represents ideal removal pattern.

4 Discussion and Conclusions

As CBCT-based sCT generation moves closer to routine APT, the question is no longer limited to how accurately an image can be synthesized, but also how the prediction can be trusted, making UQ essential [14]. The obtained results support this perspective. While EDL achieved the best overall image performance, the most notable differences between methods were observed in the UQ analyses rather than in dosimetry. UQ maps showed a consistent association with synthesis and dose errors, with EDL achieving the strongest agreement. More importantly, the sparsification analysis demonstrated that regions assigned high uncertainty contained a substantial portion of the dose error. Uncertainty should therefore be considered as a quality-assurance signal that highlights regions deserving further inspection [22, 20]. This observation is relevant as proton dose discrepancies are not solely determined by the magnitude of HU errors, but also by their spatial relationship to beam paths [4]. Consequently, improvements in image-quality metrics do not necessarily translate into proportional gains in dose agreement. The ability to identify where errors are likely to occur may therefore be as important as reducing their average magnitude.

The superior performance of EDL should nevertheless be interpreted cautiously [23]. Unlike MC Dropout and DE, which estimate uncertainty from prediction variability across multiple stochastic realizations, EDL relies on an evidential formulation that infers uncertainty from a learned prior distribution. This provides substantial computational advantages through single-pass inference, an attractive property for time-constrained adaptive workflows. However, the resulting uncertainty estimates are strongly dependent on the evidential assumptions and regularization strategy adopted during training. Previous studies have reported that evidential models may accumulate excessive evidence and become overconfident when confronted with prediction errors or distribution shifts [24, 25]. Consequently, improved uncertainty-error correlation does not necessarily imply perfectly calibrated uncertainty, nor should the estimated aleatoric and

epistemic components be interpreted as a complete representation of all uncertainty sources present in APT.

Several limitations should be acknowledged. First, the study was performed on a single-center cohort and external validation remains necessary. Second, registered CTs were used as reference and registration errors may propagate to posterior analyses. Third, the proposed framework evaluates the association between image uncertainty and dose discrepancies rather than full uncertainty propagation through the dose-calculation process. Future work should study how image uncertainty interacts with other sources of variability, including stopping-power conversion, anatomical deformation and treatment delivery uncertainties.

Acknowledgments. This study was funded by MAGERIT-CM project (grant number TEC-2024/COM-44).

Disclosure of Interests. Authors declare no conflict of interests relevant to this research.

References

1. Green, O. L., Henke, L. E., Hugo, G. D.: Practical clinical workflows for online and offline adaptive radiation therapy. In Seminars in radiation oncology (2019), vol. 29, No. 3, pp. 219-227. WB Saunders.
2. Lim-Reinders, S., Keller, B. M., Al-Ward, S., Sahgal, A., Kim, A.: Online adaptive radiation therapy. International Journal of Radiation Oncology* Biology* Physics, 99(4), 994-1003 (2017).
3. Dawson, L. A., Jaffray, D. A.: Advances in image-guided radiation therapy. Journal of clinical oncology, 25(8), 938-946 (2007).
4. Paganetti, H., Botas, P., Sharp, G. C., Winey, B.: Adaptive proton therapy. Physics in Medicine & Biology, 66(22) (2021).
5. Albertini, F., Matter, M., Nenoff, L., Zhang, Y., Lomax, A.: Online daily adaptive proton therapy. The British journal of radiology, 93(1107) (2020).
6. Wang, Y., Qi, X. S.: Artificial intelligence in adaptive radiation therapy. IOP Publishing, Bristol, UK (2025).
7. Dona Lemus, O. M., Cao, M., Cai, B., Cummings, M., Zheng, D.: Adaptive radiotherapy: next-generation radiotherapy. Cancers, 16(6) (2024).
8. O'Hara, C. J., Bird, D., Al-Qaisieh, B., Speight, R.: Assessment of CBCT–based synthetic CT generation accuracy for adaptive radiotherapy planning. Journal of applied clinical medical physics, 23(11), (2022).
9. Çağlar, M., Ertaş, K. S., Cebe, M. S., Kara, I., Kheradmand, N., Metcalfe, E. Synthetic CT generation from CBCT using deep learning for adaptive radiotherapy in prostate cancer. Frontiers in Radiology, 5 (2025).
10. Raggio, C. B., Zabaleta, M. K., Skupien, N., Blanck, O., et al F.: FedSynthCT-Brain: a federated learning framework for multi-institutional brain MRI-to-CT synthesis. Computers in Biology and Medicine, 192(2025).
11. Spadea, M. F., Maspero, M., Zaffino, P., Seco, J.: Deep learning based synthetic-CT generation in radiotherapy and PET: a review. Medical physics, 48(11), 6537-6566 (2021)

12. Maspero, M., Bentvelzen, L. G., Savenije, M. H., Guerreiro, F., et al.: Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. Radiotherapy and Oncology, 153, 197-204 (2020)
13. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. Artificial Intelligence in Medicine, 150 (2024).
14. Hémon, C., Texier, B., Lafond, C., Nunes, J. C., Barateau, A.: Towards trustworthy AI in radiotherapy: a comprehensive review of uncertainty-aware techniques. Physics in Medicine & Biology, 71(1) (2026).
15. Huang, L., Ruan, S., Xing, Y., Feng, M.: A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. Medical Image Analysis, 97, (2024).
16. Quintero, P., Cerviño, L., Zhang, H., Harris, W. Evaluation of model uncertainty in AI-based synthetic CT generation from CBCT for abdominal adaptive radiotherapy. Medical physics, 52(6), 4054-4067 (2025)
17. Hémon, C., Cubero, L., Boussot, V., Martin, R. A.,et al.: Modeling dose uncertainty in cone-beam computed tomography: Predictive approach for deep learning-based synthetic computed tomography generation. Physics and Imaging in Radiation Oncology, 33, (2025).
18. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. Advances in neural information processing systems, 33, 14927-14937 (2020)
19. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K. et al.: Attention u-net: Learning where to look for the pancreas. arXiv (2018).
20. Cabitza, F., Jurman, G., Molinari, F., Bellazzi, R.: Why almost all ML models for medicine are wrong-and what we need for evidence-based medical AI. International Journal of Medical Informatics (2026)
21. Rusanov, B., Hassan, G. M., Reynolds, M., Sabet, M., Kendrick, J., Rowshanfarzad, P., Ebert, M.: Deep learning methods for enhancing cone-beam CT image quality toward adaptive radiation therapy: a systematic review. Medical Physics, 49(9), 6019-6054 (2022).
22. van Harten, L. D., Wolterink, J. M., Verhoeff, J. J., Išgum, I.: Automatic online quality control of synthetic CTs. In Medical Imaging 2020: Image Processing (Vol. 11313, pp. 399-405). SPIE (2020)
23. Shen, M., Ryu, J. J., Ghosh, S., Bu, Y., Sattigeri, P., Das, S., Wornell, G. W.: Are uncertainty quantification capabilities of evidential deep learning a mirage?. Advances in Neural Information Processing Systems, 37, 107830-107864 (2024).
24. Chen, M., Gao, J., Xu, C.: Revisiting essential and nonessential settings of evidential deep learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2025).
25. Pandey, D. S., Choi, H., Yu, Q. Generalized Regularized Evidential Deep Learning Models: Theory and Comprehensive Evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)