Received: 22 November 2023 | Revised: 28 May 2024 | Accepted: 23 June 2024

DOI: 10.1002/mp.17287

RESEARCH ARTICLE

MEDICAL PHYSICS

Can input reconstruction be used to directly estimate uncertainty of a dose prediction U-Net model?

Margerie Huet-Dastarac[1] | **Dan Nguyen**[2] | **Eleonore Longton**[3] | **Steve Jiang**[2] | **John Lee**[1] | **Ana Barragán Montero**[1]

[1]Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, Institut de Recherche Expérimentale et Clinique (IREC), UCLouvain, Belgium

[2]Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, USA

[3]Department of Radiotherapy, Cliniques Universitaires Saint-Luc, Brussels, Belgium

Correspondence
Margerie Huet-Dastarac, Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, Institut de Recherche Expérimentale et Clinique (IREC), UCLouvain, Belgium.
Email: margerie.huet@uclouvain.be

Funding information
Fédération Wallonie-Bruxelles, Grant/Award Number: 7289

Abstract

Background: The reliable and efficient estimation of uncertainty in artificial intelligence (AI) models poses an ongoing challenge in many fields such as radiation therapy. AI models are intended to automate manual steps involved in the treatment planning workflow. We focus in this study on dose prediction models that predict an optimal dose trade-off for each new patient for a specific treatment modality. They can guide physicians in the optimization, be part of automatic treatment plan generation or support decision in treatment indication. Most common uncertainty estimation methods are based on Bayesian approximations, like Monte Carlo dropout (MCDO) or Deep ensembling (DE). These two techniques, however, have a high inference time (i.e., require multiple inference passes) and might not work for detecting out-of-distribution (OOD) data (i.e., overlapping uncertainty estimate for in-distribution (ID) and OOD).

Purpose: In this study, we present a direct uncertainty estimation method and apply it for a dose prediction U-Net architecture. It can be used to flag OOD data and give information on the quality of the dose prediction.

Methods: Our method consists in the addition of a branch decoding from the bottleneck which reconstructs the CT scan given as input. The input reconstruction error can be used as a surrogate of the model uncertainty. For the proof-of-concept, our method is applied to proton therapy dose prediction in head and neck cancer patients. A dataset of 60 oropharyngeal patients was used to train the network using a nested cross-validation approach with 11 folds (training: 50 patients, validation: 5 patients, test: 5 patients). For the OOD experiment, we used 10 extra patients with a different head and neck sub-location. Accuracy, time-gain, and OOD detection are analyzed for our method in this particular application and compared with the popular MCDO and DE.

Results: The additional branch did not reduce the accuracy of the dose prediction model. The median absolute error is close to zero for the target volumes and less than 1% of the dose prescription for organs at risk. Our input reconstruction method showed a higher Pearson correlation coefficient with the prediction error (0.620) than DE (0.447) and MCDO (between 0.599 and 0.612). Moreover, our method allows an easier identification of OOD (no overlap for ID and OOD data and a Z-score of 34.05). The uncertainty is estimated simultaneously to the regression task, therefore requires less time and computational resources.

Conclusions: This study shows that the error in the CT scan reconstruction can be used as a surrogate of the uncertainty of the model. The Pearson correlation coefficient with the dose prediction error is slightly higher than state-of-the-art techniques. OOD data can be more easily detected and the uncertainty metric is

computed simultaneously to the regression task, therefore faster than MCDO or DE. The code and pretrained model are available on the gitlab repository: https://gitlab.com/ai4miro/ct-reconstruction-for-uncertainty-quatification-of-hdunet

KEYWORDS
AI uncertainty estimation, deep learning, radiation therapy dose prediction

1 | INTRODUCTION

The reliable and efficient estimation of uncertainty in artificial intelligence (AI, see Table 1 for all abbreviations) models poses an ongoing challenge in many fields such as radiation therapy. AI models are intended to automate manual steps involved in the treatment planning workflow.[1] We focus in this study on dose prediction models that predict an optimal dose trade-off for each new patient for a specific treatment modality. They can guide physicians in the optimization,[2] be part of automatic treatment plan generation[3] or support decision in treatment indication.[4]

The clinical implementation of AI models is relatively slow compared with other fields. It is primarily due to the lack of reliable and easy to implement quality assurance tools.[5] Indeed, in critical environments like radiotherapy, AI must be used with caution, since large uncertainty in the predictions can jeopardize patients safety. Uncertainty estimation techniques must be reliable but also consider the constraints of the clinic such as limited time and resources. The two prominent methods are deep ensembling (DE)[6] and Monte Carlo dropout (MCDO).[7] DE uses the variance of the predictions of multiple networks while MCDO combines several forward passes with randomly dropped out weights. Both require high inference time and MCDO has been argued to be unreliable on real world datasets.[8]

Direct uncertainty estimation methods try to solve this issue by proposing one shot uncertainty estimation. Methods such as direct epistemic uncertainty prediction[9] or layer ensembling[10] have been proposed. The direct epistemic uncertainty prediction method subtracts an estimation of the aleatoric uncertainty (i.e., uncertainty inherent in the data, not the model) to the prediction of the error. The aleatoric uncertainty is approximated with the variance of labels for the same input by oracles. In applications like radiotherapy dose prediction, it would require the generation of multiple treatment plans by different experts for the same set of patients. Unfortunately, it is currently very difficult and unrealistic to obtain such data due to the time constraints of clinical environments. Layer ensembling estimates the uncertainty of contour segmentation by computing the variance of segmentation at different depths of the model and has been adapted to dose prediction.[11]

In this work, we present an alternative direct uncertainty estimation method using the input reconstruction error. A CT reconstruction branch stems from the bottleneck of the U-Net based dose prediction model (see Figure 1). The mean squared error (MSE) of the CT reconstruction, masked on the body, can be used as a surrogate of the uncertainty. We investigate its performance both in-distribution (ID) uncertainty estimation and out-of-distribution (OOD) detection. For the latter, we test the model on patients with a head and neck cancer type different from those in the training set. In the literature, we often see articles where the OOD is a completely different pathology (e.g., Head vs. non head CT,[12] liver vs. hippocampus[13]), which we find less clinically relevant as automation workflows are typically designed by location. A situation where a patient with different cancer location will enter the workflow is rare; and these situations could be easily spotted with bare eyes. In our study, we considered a case closer to a real clinical situation: selecting as OOD patients from the same location (i.e., head and neck) but with a different sub-location (i.e., nasopharynx) and treatment protocol. This represents a rather realistic situation, with a negative impact on the dose prediction task and more easily missed by clinicians in an automated routine.

TABLE 1 Abbreviations used in the article in alphabetical order.

AI: Artificial intelligence	MCDO: Monte Carlo dropout
CT: Computed tomography	MSE: Mean square error
CTV: Clinical target volume	OAR: Organ at risk
DE: Deep ensembles	OOD: Out-of-distribution
DL: Deep learning	PT: Proton therapy
DVH: Dose volume histogram	TPS: Treatment planning software
ID: In-distribution	

2 | METHODS

2.1 | Architecture

Our model is based on HDUNet,[14] a convolutional neural network, with a hybrid architecture between U-Net[15] and DenseNet.[16] HDUNet has become the state-of-the art for dose prediction.[17,18] We add a second decoder branch to the bottleneck to reconstruct the CT scan given in input (see Figure 1). This second decoder is

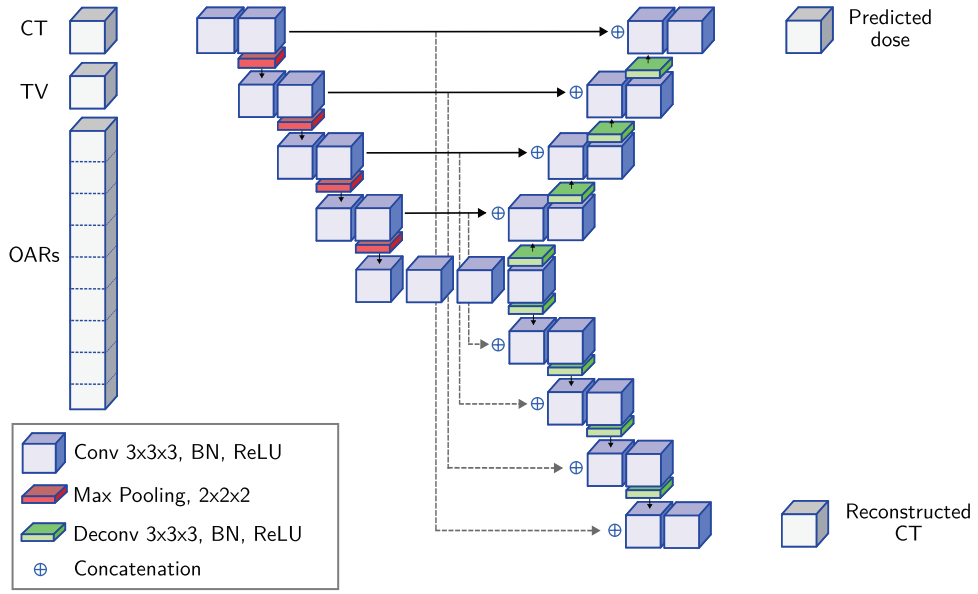

FIGURE 1 Architecture of the network with CT reconstruction branch.

composed of four upsampling operations with dense convolutions and skip connections stemming from the encoder. The training loss is the sum of two MSE for dose prediction and CT reconstruction.

2.2 | Data

2.2.1 | Dataset 1 - ID

The training set of the DL dose prediction network is an anonymized database of 60 patients with oropharynx cancer. The data was obtained from Cliniques Universitaires of Saint Luc (Belgium) from patients treated between 2015 and 2019 with bilateral Clinical target volume (CTV). For each patient, a proton therapy (PT) plan was manually generated by an experienced dosimetrist. Local ethics committee approval has been obtained for the collection and use of this retrospective database (reference: CEHF 2019/16SEP/402). The treatment planning system (TPS) used to generate the plans is RayStation 11B. For consistency regarding the planning strategy and the beam configuration, all selected patients are bilateral cases and they have two dose prescription levels: 70 Gy (TV_HIGH) and 54.25 Gy (TV_LOW) on tumor and elective nodal volumes, respectively. PT plans are planned for pencil beam scanning delivery, using four beams at (10°, 60°), (10°, 120°), (350°, 240°) and (350°, 300°) for couch and gantry angles respectively. The PT plans are generated using Monte Carlo dose calculation and robust worst-case optimization[19] on the CTV volume with 21 scenarios (2.6% proton range error and 4mm systematic error in the three space directions).

Each patient has a PT dose distribution, a CT image and a set of contours for the target volumes and the Organ at risks (OARs): brainstem, spinal cord, parotids, esophagus, oral cavity, pharyngeal constrictor muscles inferior, medium and superior, submandibular glands, and supraglottic larynx. Before being fed to the AI model, the CT and RTStruct are converted from DICOM to Nifti format, which are 3D matrices. We obtain a 3D binary matrix for each OAR, with the value of a voxel equaling 1 if inside the organ or 0 otherwise. For the target volume, we set the value of each voxel equal to the prescription level (i.e., 54 or 70 Gy for ID data, see Table S1 for OOD). All matrices are resampled to a 3 mm × 3 mm × 3 mm resolution to follow the dose grid of the TPS.

2.2.2 | Dataset 2 - OOD

The second database, used to investigate model performance with OOD, is a set of 10 extra patients, obtained from the same institution. The cancer sites are also head and neck but not oropharynx: larynx, hypopharynx, nasopharynx, supraglottis and pyriform sinus. Each patient in this dataset has a CT image and the same set of contours as the ID dataset. Details about the OOD patients are reported in Table S1.

2.3 | Training

Networks were trained for 150 epochs minimizing the MSE loss functions from the two branches. We applied random flip as data augmentation and fed the network random patches of size 128 × 128 × 128 voxels

sampled around the target volume. Because of the small size of the dataset, we want to use of as many predictions as possible to assess the performance of the uncertainty metrics. As we followed the common partitioning of 80/10/10 for the training, validation and test set, we would only have five test patients with one model. To have a more consequent population in order to draw conclusions, we train 11 models, each testing on 5 different patients to amount to the 55 patients. The pitfall of considering multiple models is that the different patients in the training set or training process could impact the predictions and therefore bring bias to the analysis. Therefore we take 3 patients out of all training, validation and testing to check the variability of these 11 models. This same training process is detailed in Huet-Dastarac et al.[4] For OOD data, we selected one of the trained models for each method. These operations were performed by a Tesla A100 GPU.

2.4 | Uncertainty estimation

The proposed surrogate of the uncertainty is the MSE between the clinical and reconstructed CT images. We only consider the voxels in the body contour to keep relevant information in our metric and avoid a positive bias caused by zero values of the background around the patient. Moreover, couch setup can vary and should not be considered by our metric. For MCDO experiments, a Bernoulli law was used to drop the weights at inference time with probabilities of 0.1–0.5. For DE, we trained 20 models with the weights following the He initialization.[20] For both techniques, the standard deviation is computed voxelwise on the inference of the models and we consider the mean over the body.

2.5 | Evaluation

We evaluated the performance of the input reconstruction method in three parts: its impact on the model performance, the accuracy of the uncertainty metric when fed ID data, and its ability to perform OOD detection.

1. Impact of the additional branch on the model: A dose prediction model performance is generally evaluated in terms of difference of dose-volume histogram (DVH) metrics between clinical and predicted dose distributions. We compute the difference in DVH metrics (Dmean for most OARs, D2 for spinal cord and brainstem, D95 and D99 for target volumes) between predictions of standard HDUNet and our model. A paired-samples Wilcoxon test was performed to assess the statistical significance.
2. ID analysis: The dose prediction model is trained and tested on the ID dataset following the training process described in the Training subsection. At inference time, we computed the Pearson correlation coefficient between the uncertainty estimation and the dose prediction error and compared it with MCDO and DE.
3. OOD analysis: The models are trained on the ID dataset and the inferences are performed on the ID and OOD datasets. The distributions are displayed as histograms for CT reconstruction, MCDO, and DE. We computed the Z-score, a metric typically used to measure the distance between distributions.[21] Finally, we computed the number of patients contained in the overlap between the ID and OOD histograms for the different methods.
4. Clinical impact analysis: To assess the clinical relevance of the uncertainty metrics, we considered a compound clinical goal deviation metric, Equation (1). It penalizes the goals not met in Table 5. We took Barragan-Montrero et al.[22] as reference to select the goals and added submandibular glands and upper esophagus following the TPS optimization function. The higher the value of this metric, the more we deviate from the clinical goals considered.

$$\text{compound clinical goal deviation} = \sum_{\text{i in max goals}} max(0, DVH_i - DVH_{achieved}) + \sum_{\text{i in min goals}} max(0, DVH_{achieved} - DVH_i) \quad (1)$$

3 | RESULTS

The performance of the dose prediction model is shown in Figure 2. The median absolute error is within 0.1% for the highest prescribed dose for the D95 and D99 of the target volumes (Figure 2, bottom) and less than one percent for Dmean and D2 of OARs (Figure 2, top). The addition of the CT reconstruction branch does not impact in a statistically significant way the dose prediction of the network, as seen in Figure 2 and Table 2 for the results of the paired-samples Wilcoxon test. The standard variation of all 11 models of DVH metrics on the three outer test set patients were below 2% of the prescription value, Figure S3. The small variation enables us to use the 55 predictions for the analysis of uncertainty methods. The uncertainty estimation of our method is obtained simultaneously to the main task which takes 10 seconds. MCDO and DE, on the other hand, require 20 inference passes.

The Pearson correlation coefficients between the different uncertainty metrics and the dose MSE for ID data are reported in Table 3. The correlation coefficient

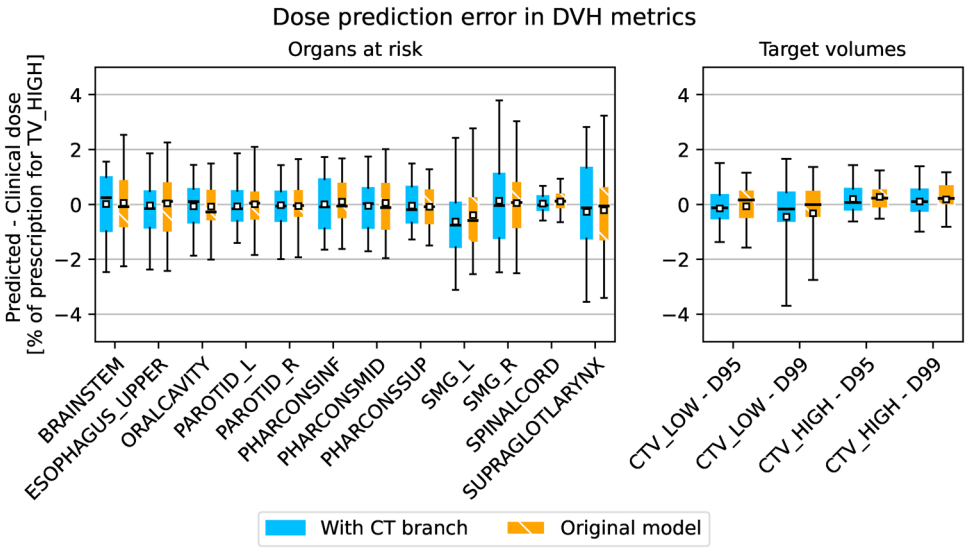

FIGURE 2 Dose prediction performance in terms of clinical DVH metrics. The mean is indicated by the white squares. DVH, dose volume histogram.

TABLE 2 Clinical goals included in the computation of the compound clinical deviation metric.

Structure	Clinical goal
Brainstem	D2 < 60 Gy
Spinal Cord	D2 < 48 Gy
Left parotid	Dmean < 28 Gy
Right parotid	Dmean < 28 Gy
Left submandibular gland	Dmean < 40 Gy
Right submandibular gland	Dmean < 40 Gy
Inferior pharyngeal constrictor muscle	Dmean < 50 Gy
Medium pharyngeal constrictor muscle	Dmean < 50 Gy
Superior pharyngeal constrictor muscle	Dmean < 45 Gy
Oral cavity	Dmean < 30 Gy
Upper esophagus	Dmean < 25 Gy
Larynx	D5 < 45 Gy
TV HIGH	D99 > 90% Dpre D95 > 95% Dpre
TV LOW	D99 > 90% Dpre D95 > 95% Dpre

obtained with all methods are significant as the p values are way below 0.05. The correlations are positive from 0.447 with Deep Ensemble to 0.62 with our method. For MCDO, the one with a probability 0.4 to drop yields the highest correlation of 0.613.

In Figure 3, we see the distribution of the different uncertainty methods for the ID and OOD databases. While the five MCDO methods fail to distinguish OOD from ID, DE is almost successful with only one patient on the right side of the histogram. Our CT reconstruction technique on the other hand allows us to easily set a threshold between ID and OOD patient data. Table 4 shows distance metrics between ID and OOD for each uncertainty estimation method. Figure 4 shows a scatter plot representing the compound clinical goal deviation versus uncertainty metric for MCDO, DE, and our CT reconstruction method. We display only MCDO with dropout rate of 0.4 as it presented the highest correlation in Table 3. Values for ID data are indicated by a dot and OOD data by a cross. The dose distribution and DVH curves of the patients circled in red can be found in the supplementary materials (Figures S1 and S2).

4 | DISCUSSION

In this study, we investigate a novel architecture to directly estimate the global uncertainty and detect OOD for a regression U-Net-like architecture by adding an input reconstruction branch. We applied it to an increasingly popular application in the medical field: the prediction of radiation therapy dose distribution. The surrogate of the uncertainty is the MSE of the CT reconstructed by the second decoding branch, grafted to the regular encoder-decoder architecture of a U-Net, as illustrated in Figure 1. The addition of the CT reconstruction branch and duplicate skip connections does not impact the prediction of the dose in a statistically significant way, as seen in Figure 2 and Table 2.

Table 3 shows that the CT reconstruction method presents the highest Pearson correlation coefficient compared with DE and MCDO with different dropout rates for ID data. Layer Ensembling adapted to dose

TABLE 3 Paired-samples Wilcoxon Test between standard HDUNet and the CT reconstruction branch.

Structure	*p*-Values	structure (Dmean)	*p*-Values
CTV_5425 - D95	0.440	Right parotid	0.712
CTV_5425 - D99	0.524	Inferior pharyngeal constrictor muscle	0.540
CTV_7000 - D95	0.407	Middle pharyngeal constrictor muscle	0.384
CTV_7000 - D99	0.404	Superior pharyngeal constrictor muscle	0.307
Brainstem - D2	0.597	Left submandibular gland	0.360
Spinal cord - D2	0.657	Right submandibular gland	0.377
Oral cavity - Dmean	0.497	Esophagus upper	0.535
Left parotid - Dmean	0.412	Supraglottic larynx	0.492

Note: Statistical significance is achieved if the *p*-value is inferior to 0.05.

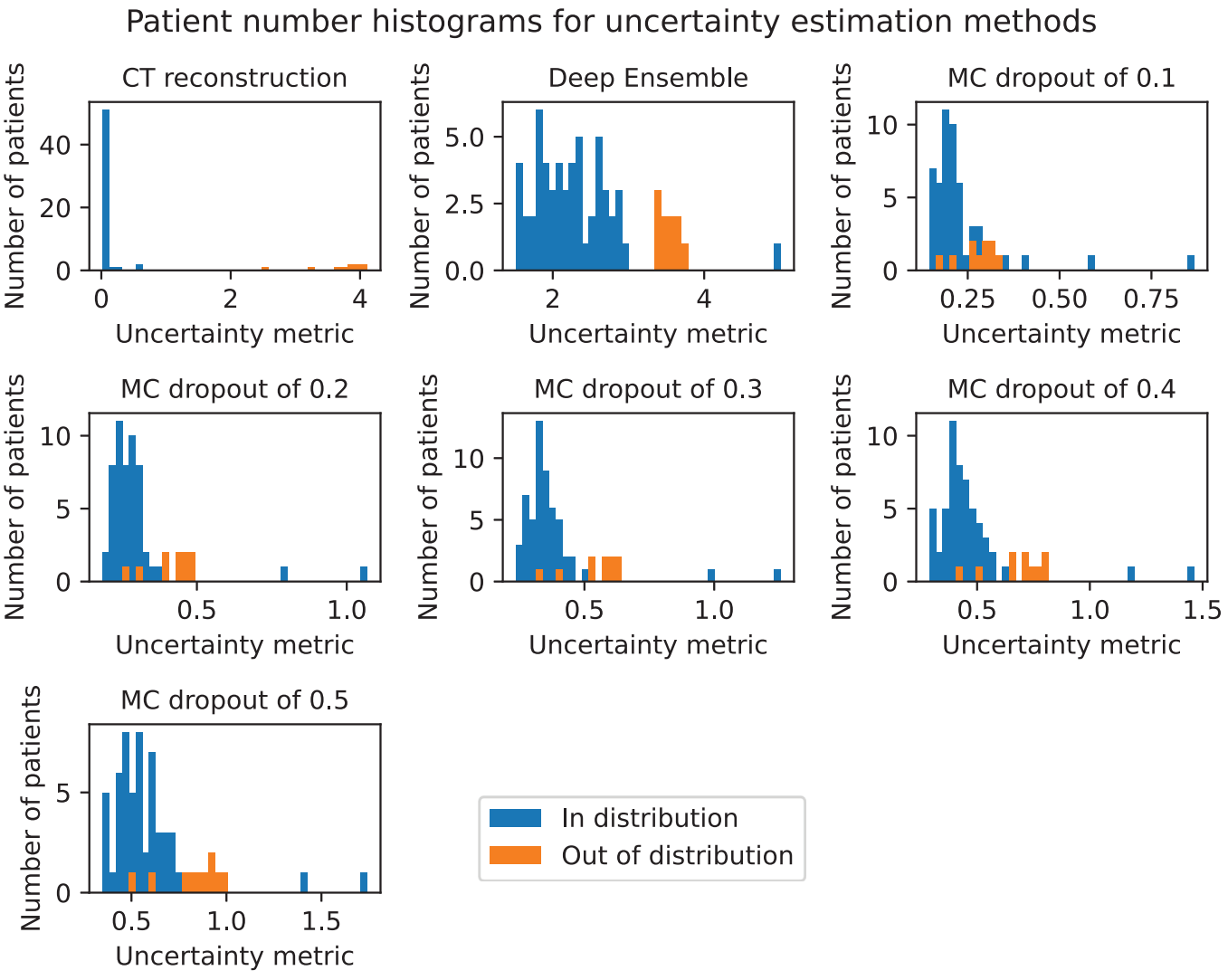

FIGURE 3 Distributions of the number of patient per uncertainty estimation metrics between in-distribution and out-of-distribution datasets.

prediction has shown to have slightly less good correlation between uncertainty estimation technique and MSE loss on the body contour than MCDO and DE with a difference of 3%.[11] Therefore, our method is a better indicator of patients susceptible to get a prediction with larger error.

In Figure 3, we notice that CT reconstruction is the only method allowing a clear separation with no over-

TABLE 4 Pearson correlation (PC) coefficients of MSE between the predicted and clinical dose in the body and mean squared error between predicted and clinical CT scan.

	DE	MCDO (0.1)	MCDO (0.2)	MCDO (0.3)	MCDO (0.4)	MCDO (0.5)	CT reconstruction
PC	0.447	0.599	0.606	0.609	0.613	0.612	0.620
p-value	6.2e-4	1.3e-6	9.1e-7	8.0e-7	6.4e-7	6.6e-7	4.5e-7

Note: The data was computed for in-distribution patients.
Abbreviations: DE, deep ensemble; MCDO, Monte Carlo dropout.

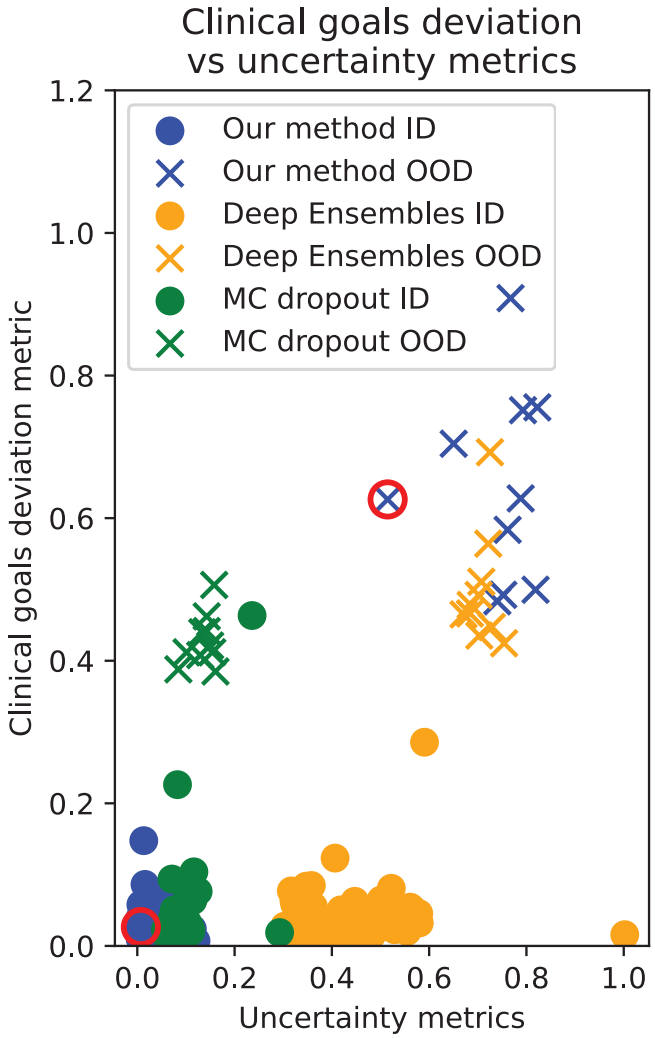

FIGURE 4 Compound clinical deviation metric in function of uncertainty. Both axes have been normalized between 0 and 1. For the patients indicated with the red circles, the dose details are displayed in the supplementary materials.

lap between the ID and OOD data. This observation is confirmed with the metrics in Table 4, where CT reconstruction has a Z-score of 34 whereas MCDO are less than 1.2 and about 2.2 for DE. Distribution shapes also differ. CT reconstruction distribution is positively skewed, DE is difficult to define and the five MCDO distributions are also positively skewed but with a wider spread. Two patients present much higher MCDO variances. They would be falsely considered as outliers if we used MCDO for OOD flagging. Despite having been used effectively for OOD detection in 2D image classification tasks in ref. [23], MCDO does not appear to perform well for our regression task with 3D clinical data. DE also has one patient that would have been classified as an outlier. Interestingly, the ID patient with highest DE or MCDO variances is the same for all metrics. After a visual check by a physician, no striking difference with other oropharynx patients was detected. The patient with the lowest CT reconstruction MSE value from the OOD dataset is a patient with a centered tumor site of supraglottis whereas the furthest apart has a tumor in the lowest part of the larynx. This observation suggests a connection between the high uncertainty and the distance between ID and OOD tumor location.

We compared the uncertainty metrics with MSE of the dose. One of the pitfalls of comparing correlation coefficients of MSE is that they might not be a good indicator of more clinically relevant metrics, such as DVH values. To have more insight about the clinical acceptability of the predicted doses, we have considered a compound clinical goal metric, which is a sum of the difference between predicted dose-volume metrics and target/tolerance doses for relevant clinical goals, Equation (1). The list of clinical goals involved in this compound metric is displayed in Table 5. In Figure 4, we show a scatter plot representing the compound clinical goal deviation versus uncertainty metric for MCDO with dropout probability of 0.4, which had the highest correlation in Table 3, DE, and CT reconstruction method. Similarly to the observation made on Figure 3, CT reconstruction allows an easier separation of the ID and OOD data for a given uncertainty threshold value. DVH curves and dose distribution slices of two patients, one from ID and one from OOD, which are circled in red on Figure 4 are shown in the supplementary materials (Figures S1 and S2). The dose distribution of the OOD patient is clearly not conformal, and abnormal areas of dose are located on the lateral and anterior sides of the patient. The dose distribution presents hot spots reaching 130 Gy, while the highest level of prescription is 70 Gy. On the other hand, the predicted dose distribution of the ID patient (Figure S1) is typical for this type of treatment.

One of the reasons MCDO and DE sparked interest in the research community lies in how easy it is to adapt them to any model. The architecture stays identical and either several models are trained with different

TABLE 5 Z-score and intersection (in number of patients) between in and out-of-distribution data for each method.

	DE	MCDO (0.1)	MCDO (0.2)	MCDO (0.3)	MCDO (0.4)	MCDO (0.5)	CT reconstruction
Z-score	2.347	0.368	0.877	1.098	1.177	1.124	34.050
Intersection	0	9	4	2	2	3	0

Abbreviations: DE, deep ensemble; MCDO, Monte Carlo dropout.

initialization weights like for DE or the model is inferred several times with randomly sampled weights dropped out for MCDO. However, both techniques require multiple inference passes, which means that either you need more computational resources to parallelize inferences or more time. Our solution allows us to compute the uncertainty for free, since it is computed simultaneously to the main task. This provides a huge time-gain with respect to MDCO and DE, which require on average 20–50 inference passes. In a real-time scenario such as adaptive radiotherapy workflow, reducing the time spent by a patient on the couch is crucial.

Contrary to MCDO and DE, our method outputs a single value as uncertainty measure whereas the former techniques can produce 3D uncertainty maps. Reliable uncertainty maps are definitely an interesting tool to help physicians apprehend the regions of uncertainty and would be a great addition to the interface of the clinical workflow. There are still applications where a single value can be used and where our model could be beneficial. For example, to flag ID patients with high uncertainty or OOD patients. Based on that flag, predictions from other networks trained on closer patients could be propose or to seek the attention of a clinician to proceed with the treatment. In an adaptive workflow, the key is to be fast, so a quickly visible flagging system could really improve the quality assurance of these treatments. Another application could be in the curation of networks with active learning. Active learning is the task of optimizing the selection of the training set among a pool of unlabeled data.[24,25] It is particularly useful when the labeling task is expensive in time and resources as it is the case in treatment planning to generate the dose distribution. Current active learning methods use MCDO or DE, rank unlabeled data according to their uncertainty and select highly uncertain samples for oracles to label.[24] We think that active learning would also benefit from our fast method.

One limitation of our study is that the addition of the CT reconstruction branch has only been investigated for a regression task. This implies that both outputs (predicted dose and CT reconstruction) are trained with the same MSE loss and therefore the correlation between the error and the uncertainty is more straightforward. It would be interesting to investigate the behavior of such an addition for a task where the main loss and the reconstruction loss are different. For example, also in the field of radiation therapy, a segmentation task where Dice or cross-entropy are used as the main loss function.

5 | CONCLUSION

In this study, we investigate a novel architecture to directly estimate the global uncertainty and detect OOD detection for a regression U-Net-like architecture by adding an input reconstruction branch. We applied it to the radiation therapy field where uncertainty estimation is crucial to ensure the patients safety. The task considered is dose prediction of PT treatments and the reconstruction of the CT scan is the surrogate of the model uncertainty. The input reconstruction method showed a higher Pearson correlation coefficient between the dose prediction error and uncertainty estimation metric than state-of-the-art techniques like MCDO and Deep ensemble. Moreover, our method allows an easier identification of OOD data. Contrary to MCDO and Deep ensemble, the uncertainty is computed simultaneously to the main task, which is ideal in time-sensitive applications. This technique could be investigated in the future for other U-Net like architecture like segmentation or detection.

ACKNOWLEDGMENTS

Margerie Huet Dastarac was funded by Federation Wallonie Bruxelles, grant number: 7289. Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. de Biase A, Sourlos N, van Ooijen PMA. Standardization of artificial intelligence development in radiotherapy. *Semin Radiat Oncol*. 2022;32:415-420.
2. Mashayekhi M, McBeth R, Nguyen D, et al. Artificial intelligence guided physician directive improves head and neck planning quality and practice Uniformity. *Clin Transl Radiat Oncol*. 2023;40:100616.
3. Maes D, Holmstrom M, Helander R, Saini J, Fang C, Bowen SR. Automated treatment planning for proton pencil beam scanning using deep learning dose prediction and dose-mimicking optimization. *J Appl Clin Med Phys*. 2023;24:e14065.
4. Huet-Dastarac M, Michiels S, Rivas ST, et al. Patient selection for proton therapy using normal tissue complication probability

with deep learning dose prediction for oropharyngeal cancer. *Med Phys*. 2023;50:6201-6214.
5. Barragán-Montero A, Bibal A, Dastarac MH, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. *Phys Med Biol*. 2022;67(11):11TR01. https://doi.org/10.1088/1361-6560/ac678a
6. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon I, Von Luxburg U, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
7. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proc Mach Learn Res*. 2016;48:1050-1059.
8. Liu Y, Pagliardini M, Chavdarova T. The peril of popular deep learning uncertainty estimation methods. *NeurIPS*. 2021.
9. Lahlou S, Jain M, Nekoei H, et al. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research* 2023. https://openreview.net/forum?id=eGLdVRvvfQ
10. Kushibar K, Campello V, Garrucho L, Linardos A, Radeva P, Lekadir K. Layer ensembles: a single-pass uncertainty estimation in deep learning for segmentation. In: *MICCAI*. Springer; 2022:514-524.
11. Tilman R, Huet-Dastarac M, Barragán-Montero AM, Lee JA. Single-pass uncertainty estimation with layer ensembling for regression: application to proton therapy dose prediction for head and neck cancer. In: 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, {ESANN} 2023, Bruges, Belgium, October 4–6, 2023. 2023. https://doi.org/10.14428/esann/2023.ES2023-115
12. Graham M, Tudosiu P-D, Wright P, et al. Transformer-based out-of-distribution detection for clinically safe segmentation. *Proc Mach Learn Res*. 2022;172:457-476.
13. Karimi D, Gholipour A. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Trans Artif Intell*. 2023;4:383-397.
14. Nguyen D, Jia X, Sher D, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-Net deep learning architecture. *Phys Med Biol*. 2019;64:065020.
15. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *MICCAI*. Springer; 2015:234-241.
16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *CVPR*. IEEE; 2017.
17. Liu R, Bai J, Zhou J, Zhang K, Ni C. A feasibility study for predicting 3D radiotherapy dose distribution of lung VMAT Patients. In: *ICTAI*. IEEE; 2020.
18. Barragán-Montero A, Nguyen D, Lu W. 3D dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations. *Med Phys*. 2019;49:3679-3691.
19. Souris K, Montero AB, Janssens G, Perri DD, Sterpin E, Lee JA. Monte Carlo methods to comprehensively evaluate the robustness of 4D treatments in proton therapy. *Med Phys*. 2019;46:4676-4684.
20. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *ICCV*. IEEE; 2015.
21. z-Score. In: Kirch W, ed. *Encyclopedia of Public Health*. Springer Netherlands; 2008;1484-1484. https://doi.org/10.1007/978-1-4020-5614-7_3826
22. Barragán-Montero A, Van Ooteghem G, Dumont D, Rivas ST, Sterpin E, Geets X. Dosimetrically triggered adaptive radiotherapy for head and neck cancer: considerations for the implementation of clinical protocols. *J Appl Clin Med Phys*. 2023;24:e14095.
23. Nguyen AT, Lu F, Munoz GL, Raff E, Nicholas C, Holt J. Out of distribution data detection using dropout bayesian neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 2022;36(7):7877-7885. https://doi.org/10.1609/aaai.v36i7.20757
24. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal*. 2021;71:102062.
25. Abdar M, Pourpanah F, Hussain S. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion*. 2021;76:243-297.

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: Huet-Dastarac M, Nguyen D, Longton E, Jiang S, Lee J, Montero AB. Can input reconstruction be used to directly estimate uncertainty of a dose prediction U-Net model? *Med Phys*. 2024;51:7369–7377. https://doi.org/10.1002/mp.17287