Comparison of Ethos template-based planning and AI-based dose prediction: general performance, patient optimality, and limitations

Benjamin Roberfroid[a,*], Ana M. Barragán-Montero[a], David Dechambre[b], Edmond Sterpin[a,c,d], John A. Lee[a] and Xavier Geets[a,b]

[a]Université catholique de Louvain – Center of Molecular Imaging, Radiotherapy and Oncology (MIRO), Brussels, Belgium

[b]Cliniques universitaires Saint-Luc, Department of Radiation Oncology, Brussels, Belgium

[c]Particle Therapy Interuniversity Center Leuven – PARTICLE, Leuven, Belgium

[d] KU Leuven – Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium

*Corresponding author at: Université catholique de Louvain (UCLouvain), Molecular Imaging, Radiotherapy and Oncology (MIRO), Avenue Hippocrate 54/B1.54.07, 1200 Brussels, Belgium.

E-mail address: benjamin.roberfroid@uclouvain.be

1. Introduction

In the field of radiation therapy (RT), adaptive RT (ART) has been a major research topic for many years [1-3]. Unlike conventional radiation therapy that relies on single computed tomography scans (CT) acquired prior to treatment, ART aims at accounting for anatomical changes that can occur in the tumor (shrinkage, progression) and organs at risk (OAR) (weight loss, air cavity filling, organs deformation, …), over the course of a treatment. In practice, ART requires repeated image acquisitions to capture the daily anatomy, to delineate the updated patient volumes, and to re-optimize the dose based on the current snapshot of the patient anatomy.

Varian®(Varian Medical system, Palo Alto, USA) has recently released a new generation of linear accelerator: Ethos, which proposes a commercial implementation of Online-ART (OART) guided by cone beam computed tomography (CBCT) [4]. With CBCT iterative reconstruction, daily automatic organ segmentation, and automatic dose optimization, Ethos allows the treatment plan to be adapted within a time frame of 20 minutes [5-9]. Ethos OART approach has already proven its efficiency and utility in several contexts and setups [6-11].

The automatic plan optimization enabled in Ethos uses an "*Intelligent Optimization Engine*" (IOE), an algorithm that optimizes objectives directly coming from a list of clinical goals, ordered by importance and created by the user prior to the treatment. Any list of clinical goals can be saved as a template of objectives and constraints to be used for other patients requiring the same treatment intent. Preliminary studies have shown that Ethos generated plans quality was globally satisfying and, when compared to manually generated plans, dose metrics differences were quite small [5-7, 12, 13]. However, this general perspective does not point out specific cases where optimization difficulties can arise from more challenging anatomies: although the template-based optimization of Ethos can produce clinically acceptable plans at the patient population level, it might encounter difficulties to generate the optimal dose distribution for any given patient.

Concurrently to Ethos, various other automatic planning methods have arisen [14, 15], some of which might intrinsically better integrate patient features to produce anatomy-specific optimal plans. Notably, the recently popular dose prediction (DP) approach, based on deep convolutional neural networks, has demonstrated the ability to predict optimal, patient-specific three-dimensional dose distribution [16-19]. Following this dose prediction, a subsequent step of inverse planning, also called "dose mimicking" (DM), allows machine parameters to be determined such that the predicted dose can actually be delivered [16, 20]. Such a sequence of steps is meant to generate anatomy-specific plans, which might turn out to be more tailored to individual patients than template-based plans.

Therefore, the goal of this study is twofold. On the one hand, we aim at assessing the general performance of the Ethos solution (i.e., the template-based IOE), by analyzing the global quality of the Ethos generated plans (EG) with respect to our clinical standards. On the other hand, we aim at investigating the capability of Ethos solution to produce patient-optimal plans and compare it to the aforementioned sequence "dose prediction & mimicking" (DP+DM) workflow. These two state-of-the-art automatic treatment planning approaches are compared by looking at their performance on selected challenging plans and their capability of generating specific dose trade-off when necessary.

Notice that this study looks only at the quality of the initial treatment plans and does not analyze the quality of the per-fraction plans.

2. Materials and methods

To achieve the goals of this study, 2 planning studies using 4 different planning methods were designed. These will be presented in the following sections. Figure 1 illustrates the schematic study design and data distribution.

2.1. Data

A database of 45 prostate cancer patients, without pelvic nodal irradiation, was used in this study. All patients were previously treated on our Halcyon (Varian Medical system, Palo Alto, USA) and initially planned on the Eclipse treatment planning system (TPS). CT images were acquired on a Toshiba Acquilion CT scanner with a slice thickness of 2mm. The dose prescription was 60Gy in 20 fractions, with a CTV-to-PTV isotropic expansion of 7mm.

Local ethics committee approval was obtained for the use of all patient data under the agreement "Learning from the past: the MIRO treatment planning database" which received the ethical approval by the CEHF (Comité d'Ethique hospitalo-facultaire).

2.2. Treatment planning

For this study, new treatment plans were retrospectively generated on the planning CT using the different dose optimizations methods described in this section. All planning methods used the 9-fields IMRT configuration and isocenters from Ethos TPS. Concerning automatic methods, they were assessed without any further manual optimization: no other operations than loading the CT scans, the contours, and optimization goals were performed.

2.2.1. Ethos planning with initial template (EG_init plans)

As previously mentioned, the Ethos IOE solution generates a plan by using a set of optimization objectives that are directly coming from a template of clinical goals. Within a given template, the clinical goals have priorities ranking from 1 (most important) to 4 (less important), and internal priorities within their own group between each other. The IOE converts this list into objective functions for the photon optimizer and then monitors the optimization process: it initializes weights for all objective functions and regularly adapts them with respect to the established priorities. It also resolves the possible conflicts between the target volumes and nearby or overlapping organs at risk, and it generates optimization structures to decrease dose to normal tissues [4].

All 45 patients were retrospectively planned using Ethos IOE based on our 60Gy prostate template (Table 1). The CT scans and the planning contours were exported from Eclipse TPS V16.01.10 to a remote Ethos emulator TPS (V02.01.00).

The Ethos TPS typically generates several plans with different IMRT and VMAT configurations. Nine-fields IMRT configuration was selected for this study, because VMAT has been reported to achieve lower plan quality in the early version of the Ethos TPS [6, 12, 13].

Table 1: IOE initial clinical goals template. Each line refers to an optimization goal for a defined volume with a defined optimization priority. Femur_L = left femoral head, Femur_R = right femoral head.

Priority	Volume	Goal
1: Most Important	PTV	D95% >= 95%
1: Most Important	PTV	D2% < 104%
1: Most Important	CTV	D98% >= 98%
2: Very Important	Rectum	V60Gy < 1%
2: Very Important	Bowel	D0.1cm^3 < 60Gy
2: Very Important	Anal canal	V60Gy < 1%

2: Very Important	Bladder	V60Gy < 3%
2: Very Important	Rectum	V52.8Gy < 15%
2: Very Important	Rectum	V30Gy < 35%
2: Very Important	Bladder	V50Gy < 20%
2: Very Important	Bladder	V40.8Gy < 20%
3: Important	PTV	D50% >= 99%
3: Important	PTV	D50% <= 101%
3: Important	Anal canal	V35Gy <25%
3: Important	Anal canal	V20Gy < 50%
3: Important	Bowel	V55Gy <3%
3: Important	Bowel	V20Gy <300cm^3
4: Less Important	Femur_L	V40Gy <20%
4: Less Important	Femur_L	V35Gy < 5%
4: Less Important	Femur_R	V40Gy <20%
4: Less Important	Femur_R	V35Gy < 5%
4: Less Important	Penile bulb	V42.5Gy < 50%
4: Less Important	Penile bulb	V54.1Gy < 10%

2.2.2. Manually-generated planning (MG plans)

A single planner manually generated plans for the 45 patients, with the goal of achieving the optimal clinical trade-offs. The beam configuration is the same as Ethos 9 fields IMRT. The plans were specifically re-optimized with this configuration for this study. For this reason, MG plans are considered as the gold standard, that is, benchmark plans for the dose comparisons that are reported in Section 2.3. The Eclipse Photon Optimizer algorithm was used for plan optimization and the Acuros External Beam 16.1.0 algorithm for the final volumetric dose calculation with a 2.5mm calculation resolution.

2.2.3. Ethos planning with updated template

Based on the acquired experience from manual planning and Ethos TPS, we updated the initial template from Table 1. The V20Gy constraints for the anal canal was decreased to 40% and optimization goals "V20Gy<40%" were added for rectum, bladder, and penile bulb with priority 4: "*less important*". Later in our experimentations (see section 2.3), 10 challenging patients were selected, and the updated template was applied in order to investigate whether the plan clinical quality could be improved for these 10 specific patients. Those are the 10 testing patients mentioned in Figure 1.

2.2.4. Dose prediction and dose mimicking planning (DP+DM plans)

The dose prediction model used in this work is a hierarchically densely connected U-Net (HD-UNET). The in-house implementation of the architecture is publicly available in https://gitlab.com/ai4miro/ntcp_predicted_dose [21]. Dose prediction with U-net architectures has already been studied over many treatment locations [22-26] and demonstrated good performances across several architecture variations including the HD-UNET version [27-29]. Our HD-UNET model contains 11 input channels: one channel for the CT-scan; one for the dose prescription mask, which consists of the prescription dose (60Gy) for the voxels inside the PTV and 0 for voxels outside; and 9 channels for the masks corresponding to the organs at risk, namely, the anal canal, bladder, rectum, left femoral head (Femur_L), right femoral head (Femur_R), colon, sigmoid, small bowel, and penile bulb.

To prevent exceeding the GPU (Nvidia A100 40GB) memory limit, the CT scan, masks, and dose were resampled at a resolution of [3x3x3 mm^3] and model training with image patches size of [176 x 176 x 96] was used. Data were augmented by flipping training data in the left-right direction to double the dataset size. The model was trained, validated, and tested using the 45 MG plans, with a 5 folds cross-validation. The mean squared error between the predicted and ground-truth doses (from MG plans) was used as a loss function. The training set included 30 patients, and the validation set 5 patients (those are the "training patients" on Figure 1). The test set contained the 10 selected patients. The model was trained for 300 epochs for each fold. After model training, the weights yielding the lowest mean squared error on the validation set were kept. This was done for all 5 folds, which resulted in 5 different models. For each test patient, the dose was predicted with these 5 different models, producing 5 different dose maps. A final dose map was then computed from the mean of these 5 dose maps. This way of aggregating the different dose maps over several folds is somehow analogous to bootstrap aggregation (bagging), which has already been reported as increasing the accuracy of dose prediction [30]. The test patients were implicated in neither the model training nor the model selection process.

Concerning dose mimicking, the goal is to optimize machine parameters that best achieve the predicted dose. For that purpose, a template was devised in the Ethos TPS to reproduce closely the dose prediction into a deliverable plan (see Table 2). This mimicking template sets up an isodose-based optimization (IBO) [31]. This template consists in requiring a classical target coverage (PTV:D95%>95% ; CTV:D98%>98% ; PTV:D2%<=104%), to then apply maximum dose constraints on the predicted isodose levels. Isodose volumes were computed from 57-60Gy (95% and 100% of the prescription dose) to 0Gy by steps of 10 Gy (60, 57, 47, 37, 27, 17, 7, 0) with an in-house Python script. The isodose template was refined by including maximum dose constraints on the intersection of some close OARs and isodose levels, to enable a more accurate mimicking on these specific anatomical regions. This mimicking template was set-up and tested on prostate plans that were not part of this study.

Table 2: Dose mimicking template for Ethos IOE. "AnoRectum" refers to the merging volume of anal canal and rectum contours. "Dose[D1-D2 Gy]" refers to the predicted dose volume ranging from dose D1 to dose D2 . AnoRectum, bladder, or bowel followed by values in square brackets [D1,D2 Gy] refers to the intersection volume between the anatomical volume and the predicted dose volume ranging from dose D1 to dose D2.

PRIORITY	VOLUME	GOAL
1: Most important	PTV	D95%>= 95%
1: Most important	CTV	D98%>=98%
1: Most important	AnoRectum[60-57Gy]	D0.1cm^3<60Gy
1: Most important	Bladder[60-57Gy]	D0.1cm^3<60Gy
1: Most important	PTV	D2%<=104%
2: Very Important	AnoRectum[47-57Gy]	D0.1cm^3<57Gy
2: Very Important	Bladder[47-57Gy]	D0.1cm^3<57Gy
2: Very Important	Bowel[+57Gy]	D0.1cm^3<60Gy
2: Very Important	AnoRectum[37-47Gy]	D0.1cm^3<47Gy
2: Very Important	Bladder[37-47Gy]	D0.1cm^3<47Gy
2: Very Important	AnoRectum[27-37Gy]	D0.1cm^3<37Gy
2: Very Important	Bladder[27-37Gy]	D0.1cm^3<37Gy
2: Very Important	AnoRectum[17-27Gy]	D0.1cm^3<27Gy
2: Very Important	Bladder[17-27Gy]	D0.1cm^3<27Gy
2: Very Important	AnoRectum[7-17Gy]	D0.1cm^3<17Gy
2: Very Important	Bladder[7-17Gy]	D0.1cm^3<17Gy
2: Very Important	AnoRectum[0-7Gy]	D0.1cm^3<7Gy

2: Very Important	Bladder[0-7Gy]	D0.1cm³<7Gy
3: Important	Dose[+57Gy]	D0.1cm³<60Gy
3: Important	Dose[47-57Gy]	D0.1cm³<57Gy
3: Important	Dose[37-47Gy]	D0.1cm³<47Gy
3: Important	Dose[27-37Gy]	D0.1cm³<37Gy
3: Important	Dose[17-27Gy]	D0.1cm³<27Gy
3: Important	Dose[7-17Gy]	D0.1cm³<17Gy
3: Important	Dose[0-7Gy]	D0.1cm³<7Gy

2.3. Planning studies

Two planning studies were performed, in order to investigate the general performance of the Ethos template-based and its patient-optimality (i.e., the capability to produce plans that achieve the best possible dose distribution for a patient, with the highest number of met clinical goals) against DP+DM plans, with respect to reference MG plans. All plans were scaled either to D95=57Gy on the PTV or D98=58.8Gy on the CTV so that both PTV and CTV coverage constraints could be met.

Study S1 -Performance of ETHOS using initial template

This first study evaluates the global performance of Ethos planning with our initial clinical template (EG_init) against conventional manual planning (MG) over the 45 patients.

The 45 EG_init plans and MG plans were compared through the OAR clinical metrics used in our Ethos template goals and on the mean doses to the OARs. Metrics were extracted from Ethos TPS for EG_init plans and from Eclipse TPS for MG plans. Homogeneity index (HI) and conformity number (CN) were also computed to provide supplementary information about quality of the dose distribution. HI and CN were computed with the following formulas: $HI = \frac{D_{2\%} - D_{98\%}}{D_{50\%}}$, $CN = \frac{V_{95\%,PTV}^2}{V_{PTV} \times V_{95\%,Body}}$ where $V_{95\%,PTV}$ and $V_{95\%,Body}$ are the volume receiving 95% or more of the prescribed dose for the PTV and the body, respectively.

Statistical significance between MG and EG_init plans is reported for the clinical goals and mean dose to OARs with a Wilcoxon signed rank test. All the statistical tests are operated with the "stats.Wilcoxon" function from the scipy Python package.

From the results of this comparison, we noticed 10 EG_init plans showing specific difficulties compared to MG plans, either they did not meet some clinical goals that were met with manual planning or the dose to OARs could be overall further improved.

Study S2 - Performance of ETHOS using an updated template for a selected set of patients and comparison with deep learning-based auto-planning

The second study highlights the 10 patients from previous study where Ethos IOE shows optimization difficulties with generic clinical constraints typically used in template-based optimization. Consequently, a second template which has been updated on basis of results of the first study is used to evaluate if clinical plan quality might be substantially increased. The 10 EG_init plans were kept (EG_init_selected) and corresponding patients were re-planned with the updated template (EG_upd_selected). Moreover, these 10 test patients were also planned with the alternative DP+DM approach. The 10 EG_init_selected, EG_upd_selected and DP+DM plans were compared using homologous MG plans as benchmark. Clinical metric and mean dose to OAR for EG_init_selected

plans, EG_upd_selected plans and DP+DM plans were compared. In addition, the numbers of clinical goals that are met were also provided for each approach.

HI and CN are also provided following the same formulas than for Study S1. Modulation complexity score (MCS) was computed for each Ethos multileaf collimator (MLC) with an in-house python script. Details about MCS computation can be found in supplementary materials S1 and in the original MCS article [32]. Deliverability was reported with Mobius gamma index. Recent studies have demonstrated and used the capability of Mobius to serve as a reliable indicator of plan failure [33, 34].

Statistical significance between EG_init_selected and DP+DM as well as between EG_upd_selected and DP+DM is reported for the clinical goals, mean doses to OARs, monitor unit (MU), MCS, and gamma index with a Wilcoxon signed rank test. Statistical tests are operated with the "stats.Wilcoxon" function from the scipy Python package.

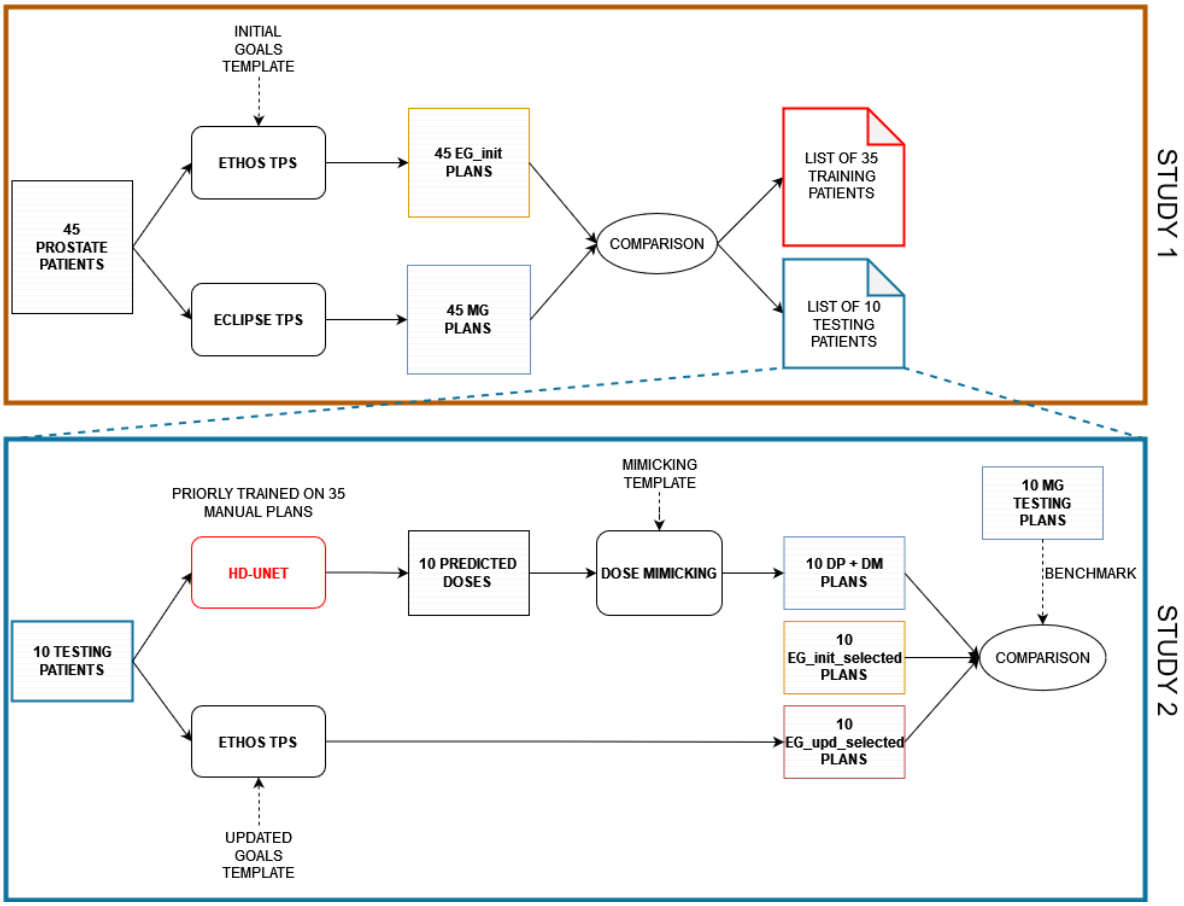

Figure 1: Workflow of the planning studies and data distribution. The patients used in the study 2 come from a manual selection of 10 patients from study 1. The 10 EG_init_selected plans did not undergo any modifications between study 1 and study 2 as for the 10 MG plans used as benchmark in the specific comparison. The 10 EG_upd_selected plans were generated the same way as the 10 EG_init_selected, excepted that the updated goals template was used.

3. Results

Study S1 - Performance of ETHOS using initial template

Comparison of the metrics between MG plans and EG_init plans are reported in Table 3. For metrics close to prescription dose (50Gy and higher), EG_init plans achieved similar mean values than MG

plans, absolute differences (|EG_init minus MG|) never exceeding 0.6%. Some of them are established as statistically significant, however, we consider these differences of less than 1% to be clinically insignificant.

For Lower dose levels, EG_init plans tend to show larger dose differences with MG plans: 13%, 6.6%, and 5.3% more, for V20Gy the anal canal, V30Gy to rectum and V42.5Gy to penile bulb, respectively. For the rest of the considered metrics, results achieved by EG_init plans were deemed satisfactory.

The previously observed trend about lower dose levels seems to be confirmed when looking at the mean dose distribution for each OAR in Figure 2. Again, OARs that show noticeable differences are the anal canal, bladder, rectum, and penile bulb with median mean dose difference of 5.0Gy, 3.7Gy, 4.9Gy, and 5.8Gy, respectively.

Table 4, reports the HI and CN for both planning approaches. We can notice that the mean HI for MG plans is slightly better than for EG_init plans. Mean CN looks comparable between the two approaches.

The V20Gy<40% constraints for the updated template were chosen because precise mean V20Gy for the anal canal, bladder, rectum, and penile bulb are, respectively, 21.3%, 28.1%, 35.6%, and 23.2% over the 45 MG plans. These values over the 10 MG plans for the selected patients rise to 30.0%, 32.9%, 42.5%, and 23.4% for the anal canal, bladder, rectum, and penile bulb, respectively. Such additional V20Gy constraints, aim at reasonably reducing the most extreme differences for lower dose levels of EG_init plans. Results of this template update can be found in the following section for the 10 selected patients.

Table 3: Mean metrics achieved for EG_init plans and MG plans over the 45 patients. Differences (EG_init minus MG) are also provided. The p-values are extracted from the comparison of EG_init plans and MG plans. Those presented in bold font are deemed statistically significant (p < 0.05). EG_init = Ethos generated plan (using initial template), MG = Manually generated plan.

Volume	Constraint	EG_init plans	MG plans	EG_init – MG	P-value
Anal canal	V60Gy (%)	0.3 ± 0.7	0.5 ± 0.6	-0.2	**0.006**
	V35Gy (%)	17.4 ± 9.3	14.3 ± 10.2	3.1	**$<10^{-3}$**
	V20Gy (%)	34.3 ± 14.5	21.3 ± 13.2	13.0	**$<10^{-3}$**
Bladder	V60Gy (%)	0.8 ± 1.7	1.0 ± 0.8	-0.2	**$<10^{-3}$**
	V50Gy (%)	10.8 ± 4.0	10.3 ± 4.6	0.5	**0.002**
	V40.8Gy (%)	15.8 ± 5.3	14.1 ± 6.0	1.7	**$<10^{-3}$**
Bowel	D0.1cm³ (Gy)	23.0 ± 24.1	21.4 ± 23.8	1.6	**0.020**
	V55Gy (%)	0.2 ± 0.7	0.2 ± 0.6	0.0	**0.020**
	V20Gy (cm³)	5.0 ± 9.7	3.7 ± 7.8	1.3	**0.002**
Rectum	V60Gy (%)	0.7 ± 1.6	0.7 ± 0.7	0.0	**0.040**
	V52.8Gy (%)	14.1 ± 2.8	14.3 ± 3.9	-0.2	0.780
	V30Gy (%)	34.1 ± 2.9	27.5 ± 6.5	6.6	**$<10^{-3}$**
Femur_L	V40Gy (%)	0.0 ± 0.1	0.1 ± 0.1	-0.1	**0.009**
	V35Gy (%)	0.0 ± 0.2	0.5 ± 1.0	-0.5	**$<10^{-3}$**
Femur_R	V40Gy (%)	0.0 ± 0.1	0.1 ± 0.2	-0.1	**0.004**
	V35Gy (%)	0.0 ± 0.3	0.4 ± 1.0	-0.4	**$<10^{-3}$**
Penile Bulb	V54.1Gy (%)	3.8 ± 5.5	4.4 ± 7.7	-0.6	**$<10^{-3}$**
	V42.5Gy (%)	15.5 ± 12.8	10.2 ± 11.4	5.3	0.120

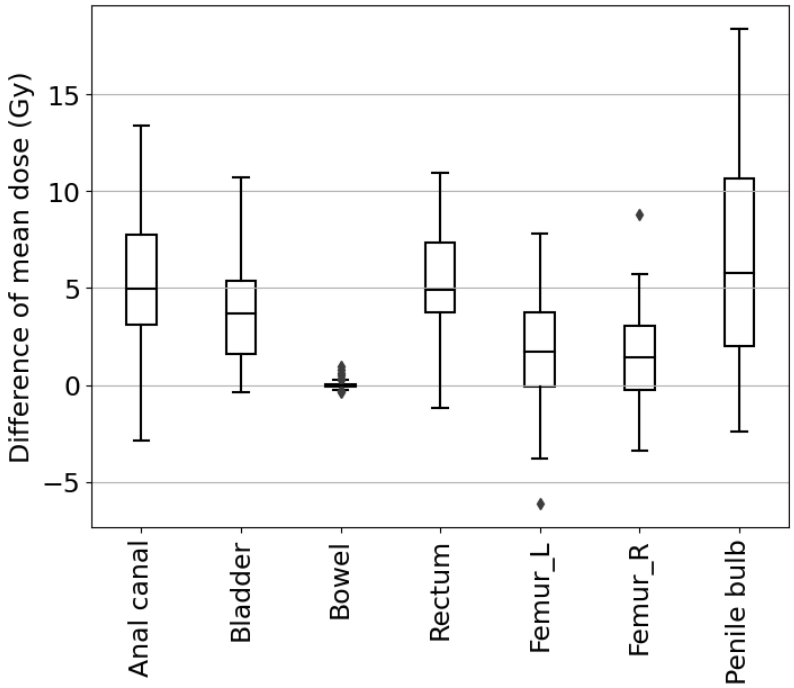

Figure 2: Difference of mean dose to OAR between EG_init and MG plans over the 45 patients. The inner line represents the median of the dataset, the box are the quartiles, the whiskers show the range of the distribution without outliers which are denoted by a diamond. Outliers are determined by the boxplot function of the seaborn Python package. All mean dose differences are statistically significant.

Table 4: Mean homogeneity index (HI) and conformity number (CN) for Ethos generated (EG_init) and manually generated (MG) plans

	EG_init	MG
HI	0.12 ± 0.02	0.10 ± 0.02
CN	0.89 ± 0.01	0.89 ± 0.03

Study S2 - Performance of ETHOS using an updated template for a selected set of patients and comparison with deep learning-based auto-planning

Metric differences for DP+DM, EG_init_selected and EG_upd_selected are reported in Table 5. Concerning dose levels above 50Gy, several of these metrics are now above 1% difference for EG_init_selected plans: V50Gy to bladder, V60Gy and V52.8Gy to rectum and V54.1Gy to penile bulb. As for S1, large mean differences can be observed for lower dose levels such as 16.9%, 8.5%, and 11.2%, for V20Gy to the anal canal, V30Gy to the rectum, and V42.5Gy to the penile bulb, respectively. This repeats the same trends as for S1, although the differences with MG plans get amplified due to patient selection.

Concerning EG_upd_selected plans, we noticed a slight increase of the V60Gy to rectum compared to EG_init_selected, while lower dose levels differences with MG dropped: from 16.9% to 11.5% for the anal canal V20Gy, from 8.5% to 5.7% for the rectum V30Gy, and from 11.2% to 3.4% for the penile bulb V42.5Gy.

The DP+DM approach overall achieved metrics closer to those from MG, in comparison with EG_init_selected and EG_upd_selected. Noticeable exceptions are the penile bulb V54.1Gy of 4.6% instead of 1.9% and 0.5% for EG_init_selected and EG_upd_selected, respectively. However, these

differences did not reach statistical significance between DP+DM and EG_init_selected and between DP+DM and EG_upd_selected.

Figure 3 illustrates the difference of mean dose to OARs with regards to MG plans for the three automatic approaches. As previously observed, the updated template reduced the dose difference for the low dose levels, and consequently for the mean dose to OAR. However, DP+DM outperformed the other automatic methods, and achieved the closest dose metrics to the MG method. Statistical difference between EG_upd_selected and DP+DM plans is moreover established for the anal canal and bladder.

Table 6 depicts when clinical goals were met or not for each approach. While the 10 MG plans could not meet all clinical goals, they were deemed clinically acceptable by our physicians. As to the automatic approaches, DP+DM succeeds in meeting the largest number of "most important" and "important" goals, followed by EG_init_selected and EG_upd_selected. The 10 DP+DM plans were also deemed clinically acceptable, but 1 EG_init_selected plan and 3 EG_upd_selected plans were not, due to critical unmet goals. Among these unmet goals, there is notably the "V60Gy<1%" to rectum that could not be met for 2 EG_init_selected and 4 EG_upd_selected plans. Figure 4 illustrates the distribution of V60Gy to rectum for each approach. We can notice that V60Gy of 2 EG_init_selected plans were significantly above 1% constraint and that template update resulted in a noticeable increase of V60Gy for 2 other patients, compromising the clinical quality of these EG_upd_selected plan.

Table 7 reports the HI and CN for the different planning methods, it can be noticed that MG and DP+DM achieved a better homogeneity while their conformity number is slightly lower than for EG_init_selected and EG_upd_selected plans.

Table 8 reports the mean total MU, MCS and Mobius gamma index for each approach. There are statistically significant differences between MU for MG plans and EG as well as for DP+DM and EG plans. MCS for both Ethos MLCs confirmed the higher complexity of DP+DM plans. This is expected since the dose metrics are probably more demanding for some constraints than the general Ethos template. However, no prostate plans were flagged as problematic while the mean gamma index is slightly lower for the MG plans than for other approaches.

Figure 5 illustrates the dose distribution for the different optimization methods for a specific patient where the V60Gy constraint to rectum was difficult to meet. Some facts already mentioned above can be observed: the DP+DM and MG methods reduce at best the V60Gy to rectum while ensuring PTV coverage. Such rectum sparing seems to have a cost in terms of dose conformity when compared to EG_init_selected and EG_upd_selected.

Table 5: Mean difference of dose metric between automatic plans (EG_init_selected, EG_upd_selected; and DP+DM) and MG plans for the 10 selected patients. P-values between DP+DM and EG_init_selected metrics distribution as well as

between DP+DM and EG_upd_selected are provided. Those presented in bold font are deemed statistically significant ($p < 0.05$).

OAR	Constraint	DP+DM - MG	EG_init_selected - MG	P-value DP+DM;EG_init_selected	EG_upd_selected - MG	P-value DP+DM;EG_upd_selected
Anal canal	V60Gy (%)	0.3 ± 1.0	0.1 ± 0.7	0.079	0.4 ± 1.3	0.500
	V35Gy (%)	1.2 ± 3.1	6.0 ± 5.9	**0.006**	4.8 ± 4.8	**0.004**
	V20Gy (%)	4.6 ± 4.7	16.9 ± 11.4	**0.004**	11.5 ± 7.9	**0.006**
Bladder	V60Gy (%)	-0.3 ± 0.5	0.1 ± 0.6	0.062	0.4 ± 1.0	**0.028**
	V50Gy (%)	0.6 ± 1.2	1.4 ± 0.9	0.130	1.3 ± 0.8	**0.014**
	V40.8Gy (%)	0.6 ± 1.8	2.9 ± 2.9	**0.010**	2.2 ± 1.3	**0.004**
Rectum	V60Gy (%)	-0.2± 0.3	1.4 ± 3.1	0.058	1.9 ± 3.1	**0.034**
	V52.8Gy (%)	0.0 ± 1.2	1.0 ± 2.6	0.193	0.7 ± 2.5	0.275
	V30Gy (%)	3.4 ± 2.8	8.5 ± 5.1	**0.027**	5.7 ± 5.0	0.375
Bowel	V55Gy (%)	-0.1± 0.8	-0.2 ± 0.8	0.625	-0.2 ± 0.8	0.769
	V20Gy (cm^3)	-1.3± 5.6	0.8 ± 5.8	0.180	-0.6 ± 5.9	0.102
	D0.1cm^3(Gy)	0.5 ± 2.8	1.0 ± 2.5	0.109	0.1 ± 1.8	0.108
Femur_L	V40Gy (%)	0.0 ± 0.2	-0.1 ± 0.2	0.180	-0.1 ± 0.2	0.180
	V35Gy (%)	0.0 ± 1.1	-0.6 ± 1.1	**0.028**	-0.4 ± 0.9	**0.043**
Femur_R	V40Gy (%)	0.0 ± 0.1	0.0 ± 0.1	0.157	0.0 ± 0.1	0.157
	V35Gy (%)	0.0 ± 0.6	-0.3 ± 0.5	**0.043**	-0.2 ± 0.5	**0.043**
Penile Bulb	V54.1Gy (%)	4.6 ± 7.7	1.9 ± 4.4	0.080	0.5 ± 5.5	0.080
	V42.5Gy (%)	6.6 ± 9.0	11.2 ± 11.8	**0.028**	3.4 ± 9.4	0.068

Table 6: The total number of met OAR goal for the 10 patients are reported for each method. For each of the 10 testing plans there are 8 "very important" goals, 4 "Important" goals and 6 "less important" OAR goals to satisfy. Highest number of met goals per importance for automatic planning methods are bolded. EG_init_selected = Ethos generated (using initial template) for patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

Priority	2: Very important	3: Important	4: Less Important
EG_Init_selected	66/80	37/40	**59/60**
EG_upd_selected	62/80	38/40	**59/60**
DP+DM	**70/80**	**39/40**	58/60
MG	74/80	40/40	60/60

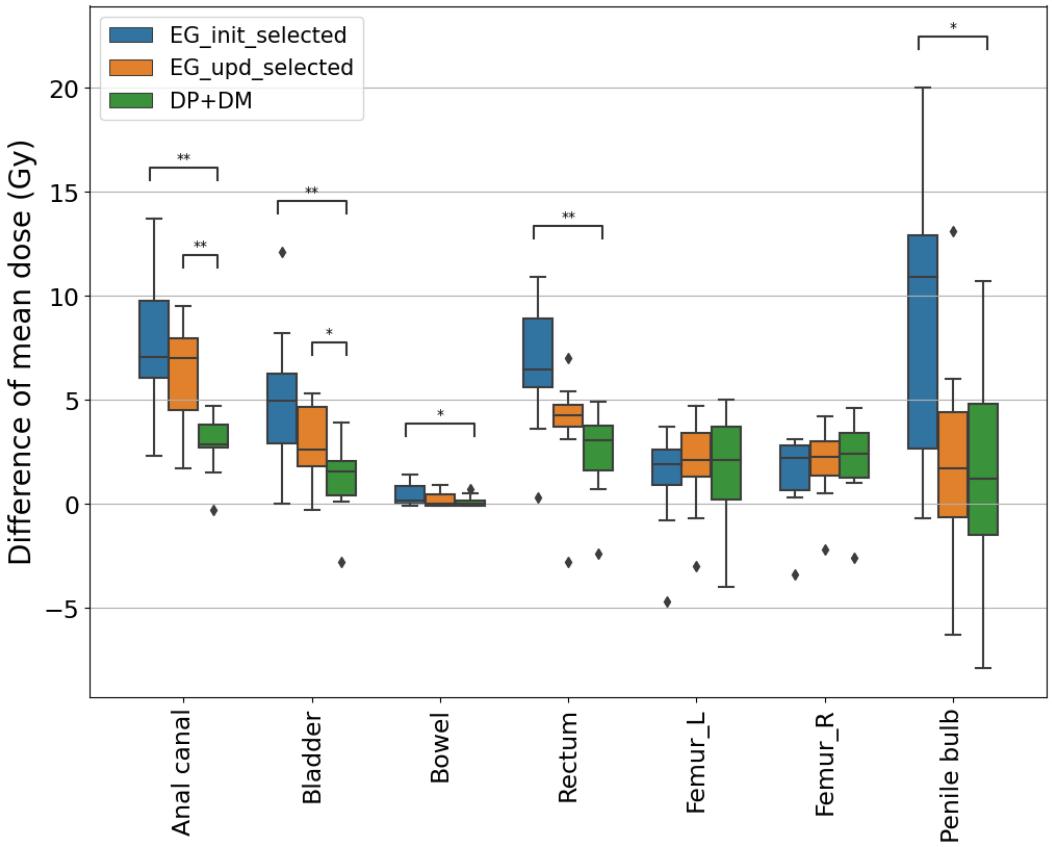

*Figure 3: Difference of mean dose to OAR between automatic planning methods and manual ground truth planning. The horizontal links between boxes denote a statistically significant mean dose difference between the 2 linked boxes (p-values lower than 5% are denoted by *; p-values lower than 1% are denoted by **). The inner line represents the median of the dataset, the box are the quartiles, the whiskers show the range of the distribution without outliers which are denoted by a diamond. Outliers are determined by the boxplot function of the seaborn Python package. EG_init_selected = Ethos generated (using initial template) for patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking.*

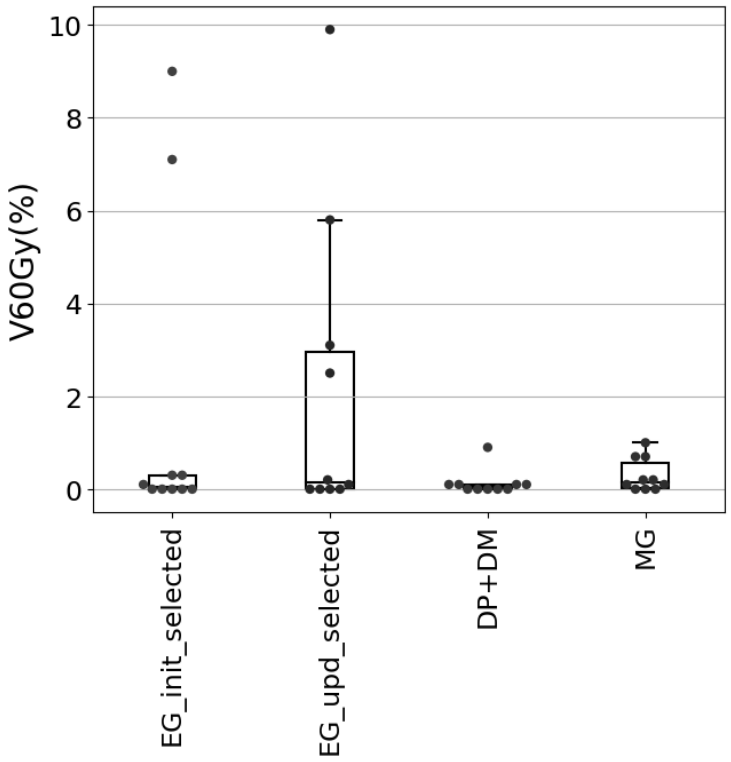

Figure 4: Rectum V60Gy for each optimization method over the 10 selected patients, individual values are denoted by a dot. The inner line represents the median of the dataset, the box are the quartiles, the whiskers show the range of the distribution without outliers. We can notice two patients having their V60Gy noticeably increasing above 1% when using the updated template instead of the initial one. EG_init_selected = Ethos generated (using initial template) for patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

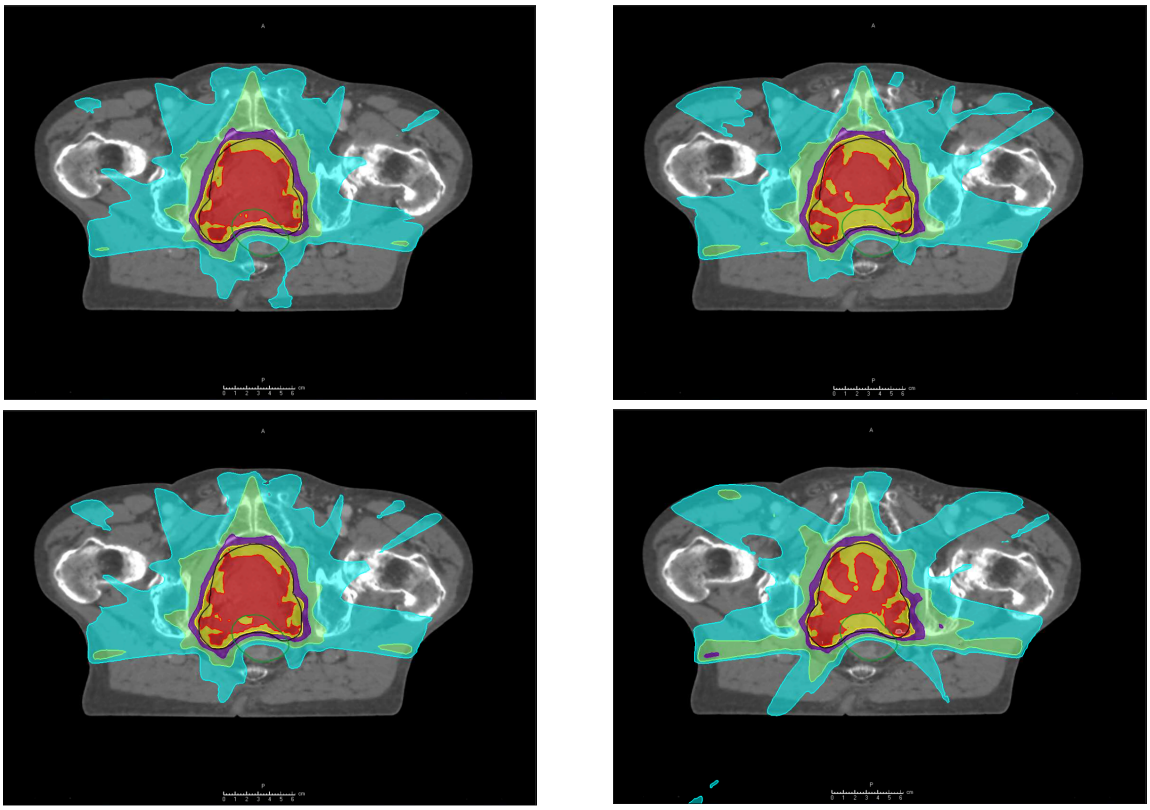

Figure 5: Dose distribution of each optimization method for a patient where the V60Gy constraint to the rectum was difficult to meet. Black contour: PTV; green contour: rectum. Red isodose: dose ⩾ 60Gy; yellow isodose: 57Gy ⩽dose <60Gy; purple isodose: 48Gy ⩽dose <57Gy; light green isodose: 36Gy ⩽ dose <48Gy; light blue isodose: 24Gy ⩽dose <36Gy.Top left: EG_init_selected ; bottom left: EG_upd_selected ; top right: DP+DM ; bottom right : MG. EG_init_selected = Ethos generated (using initial template) for a patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for a patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

Table 7: Mean homogeneity index (HI) and conformity number (CN) for the different planning methods over the 10 selected patients. EG_init_selected = Ethos generated (using initial template) for a patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

	EG_init_selected	EG_upd_selected	DP+DM	MG
HI	0.12 ± 0.02	0.12 ± 0.02	0.08 ± 0.01	0.09 ± 0.02
CN	0.89 ± 0.01	0.89 ± 0.01	0.88 ± 0.02	0.87 ± 0.02

Table 8: Mean total monitor unit (MU), modulation complexity score (MCS) for the two Ethos multileaf collimators and Gamma index (Mobius γ). EG_init_selected = Ethos generated (using initial template) for a patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for a patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

	EG_init_selected	P-value EG_init_selected-DP_DM	EG_upd_selected	P-value EG_upd_selected-DP_DM	DP+DM	P-value DP+DM-MG	MG

Total MU	1947.83 ± 434.07	0.010	1940.96 ± 301.85	0.002	2397.48 ± 332.56	0.002	2920.03 ± 564.37
MCS 1	0.26 ± 0.04	0.049	0.26 ± 0.04	0.002	0.23 ± 0.03	0.014	0.20 ± 0.03
MCS 2	0.27 ± 0.05	0.020	0.28 ± 0.04	0.002	0.24 ± 0.02	0.014	0.21 ± 0.03
Mobius γ (%)	99.9 ± 0.00	0.020	99.9 ± 0.00	0.020	99.8 ± 0.06	0.008	99.4 ± 0.43

4. Discussion

This work consisted in 2 planning studies aiming at evaluating the general performance of Ethos template-based automatic plan optimization and its patient-optimality against another popular approach: deep-learning dose prediction followed by dose mimicking. For study S1, we generated plans for 45 patients using our initial Ethos clinical goals template (EG_init), and compared them to manually generated plans (MG). Study S2 used 10 specific patients from S1 where IOE with initial goals template faced optimization difficulties when compared to MG plans. For these 10 selected patients, plans were generated with 3 different automatic approaches: Ethos IOE with initial template (EG_init_selected), Ethos IOE with updated template (EG_upd_selected) and DP+DM. The plans from these 3 automatic approaches were then compared to their homologous manual plan.

We first showed that the performance of the Ethos template-based approach over our database of 45 patients, using our initial clinical template (EG_init), was globally satisfactory for dose levels close to the prescription dose, although lower levels looked loosely optimized, especially for the anal canal, rectum, bladder, and penile bulb. We then compared the Ethos template-based approach to the DP+DM approach on 10 selected challenging patients and showed that while most EG_init_selected plans were still clinically acceptable, the DP+DM plans outperformed them. Number of clinical goals that were met, doses to OARs, homogeneity index and clinical acceptability were favorable to our DP+DM method while the deliverability did not seem to be impacted according to our results. MG plans performance suggested that quality of EG_init plans might be enhanced by updating our Ethos template. We investigated that hypothesis by adding low priorities constraints to the clinical template. While this enhanced some clinical metrics, this also deteriorated some other high-priority metrics, such as the V60Gy to the rectum. EG_upd_selected plans unfortunately resulted in fewer clinically acceptable plans and highlighted the limitation of using generic constraints from a template when optimizing plans with IOE.

The V60Gy to rectum that were noticeably above 1% for 2 EG_init_selected and 4 EG_upd_selected plans might stem from the limited number of iterations allotted for the dose optimization process for such challenging plans. During dose optimization, even if the IOE sequentially updates the weights of goals from higher to lower priority [4], low-priority goals still affect the global solution convergence, just due to their simple presence in the objective function from its initialization. Considering that the more complex the goals are to achieve, the more iterations are necessary, lower-priority goals might make the objective function very complex and prevent its convergence from occurring within the iteration budget allotted by the manufacturer [35]. This would explain why 2 plans that were initially clinically deliverable were no longer acceptable after the update: adding an extra dose objective likely required more iterations, which in turn exceeded the preset default limit for some plans. Such assumption about an insufficient number of iterations might also lead to additional conclusions related to our DP+DM approach. DM consists in optimizing constraints derived from predicted dose volumes through IOE. It might thus be supposed that with unrealistic predicted doses and

optimization objectives, the IOE would naturally have needed an extremely high number of iterations to try fulfilling conflicting goals. However, it seems that DP+DM plans were not much affected by this iteration limit, which tends to confirm the reliability of the predictions made by our HD-UNET models. Moreover, this would also highlight the capability of the IOE to produce qualitative plans for challenging cases within the iteration budget when patient-specific dose constraints are used instead of generic ones.

From these results, DP+DM seems to better approximate optimality for a wider range of patients than the Ethos template. The overall quality of Ethos plans is not questioned, but this study illustrates the difficulty of having an Ethos template that can both deliver clinically acceptable plans for a vast population of patients and spare OARs as much as possible for every individual case. Nonetheless, it is important to analyze from a broad perspective the potential and limitations of both automatic planning approaches (Ethos template and DP+DM):

First, DP+DM requires some time to set up. It is necessary to have access to a database of clinical plans large enough to be able to train the model, this can take several weeks of data collection or generation, as well as model testing. Dose mimicking can also consume time to best reproduce the predictions. From our experience, mimicking set-up and testing can take from days to weeks. Moreover, it is noteworthy that our model was exclusively trained for Ethos 9 fields IMRT configuration, while if 12 fields IMRT or VMAT plans were to be also considered, then new dose prediction models should be trained for each new beam configuration. In contrast, a single IOE template can be used for all beam configurations. Even using transfer learning from 9 fields IMRT model to a new configuration model, this would require additional plans with the new beam configuration [23]. These processes lead to an even longer implementation and contrast with the Ethos template solution designed to be fast and easy to implement.

Second, the DP+DM workflow is inherently longer than Ethos initial planning workflow. Both methods lead to a deliverable plan through the Ethos IOE, however, DP+DM necessitates firstly to load CT and contours in a trained HD-UNET to then predict the dose. This prediction step in our case takes approximately 50 to 60 seconds per plan (this accounts for the 5 predictions from the cross-validation models on a single GPU used to produce the final prediction of the dose map). Then, the generated isodose volumes are computed with our in-house script and uploaded to Ethos server in less than 10 seconds. Last, the dose mimicking step needs to be performed, which takes a time similar to the optimization with a template for EG plans. This results in a plan generation for our DP+DM method that is about 1 min longer than the EG approach in our case. However, such additional time might probably be optimized, for example by computing the 5 predictions in parallel.

Third, DP+DM was not evaluated over the whole patient database. In order to demonstrate the global superiority of DP+DM over Ethos template approach, it would require proving superiority over a larger dataset and not only on plans were Ethos IOE templates faced difficulties. Nevertheless, this does not detract from the conclusion that with an approach such as DP+DM, a more optimal automatic dose optimization seems possible at the cost of a longer and more demanding (in terms of training database generation) implementation of the workflow.

Concerning the adaptive workflow, clinical goals defined at initial planning are then further used for dose optimization at each adaptive session, consequently a question arises: "*Are challenging plans suited for the Ethos adaptive workflow?*". In the case where dose optimization would fail to converge

to an acceptable solution for initial planning, we can suppose that results would not be substantially better during the adaptive session and that there is a need for specific tuning of clinical objectives. However, excessive specific tuning of clinical goals during initial planning step might be risky later in adaptive workflow since patient anatomy changes over time. Passed a tipping point of tuning, IOE could no longer converge to a satisfying dose distribution for each plan of the day before reaching the supposed-critical number of iterations.

To prevent scenarios where dose would not converge for adaptive sessions, a recommendation based on our results would be to stick to general constraints when setting-up Ethos clinical goal templates. If even with a generic template, dose optimization during initial planning proves to be challenging, clinicians should consider excluding these patients from Ethos adaptive workflow. If the IOE cannot converge towards a dose optimum for most fractions, it is likely that the adaptive workflow will increase treatment time without clear benefits from plan adaptation. Keeping general constraints would, however, mean that some plans will inevitably be sub-optimal in comparison with what is manually achievable.

Note that the adaptive workflow in Ethos needs to be selected before the first fraction and is then maintained throughout the entire treatment. Considering the current limitations of the template-based approach, an interesting option could be to allow for a switch-up or switch-down of the adaptive workflow during the treatment course, based on the observed evolution of patient anatomy. This flexibility could address the cases where Ethos plan adaptation may not necessarily provide a clear added value for a specific patient. Alternatively, it could also be beneficial to directly visualize and select the initial plan recomputed on the daily image without having to wait for the re-optimized plan.

In the end, this is likely to be one of the keys to exploit the full potential of OART on Ethos: getting as patient-optimal as possible in a robust way. Although there seems to be room for improvement, and even if added value of treatment adaption cannot be ensured for every patient, recent studies have shown promising results in demonstrating the added value of Ethos plan adaptation for the global patients' cohort [6-11]. To achieve further value for adaptive treatments, future steps could involve utilizing DP as a reference of specific achievable metrics similarly to the DVH prediction of RapidPlan (Varian Medical system, Palo Alto, USA) but with extra-information such as the predicted 3D dose map. Also, having an indicator of how close to the iteration limit the optimizer gets during the initial planning could convey an interesting and complementary piece of information when trying to fine-tune dose constraints or when considering whether a patient should enter in the adaptive workflow or not.

Conclusion

This study has analyzed the general performances and capability to produce patient-optimal plans for Ethos template-based optimization in comparison with deep-learning dose prediction followed by dose mimicking (DP+DM). Our results suggest the existence of a limitation of the Ethos template-based optimization with increasing complexity of the template (i.e. higher number and more aggressive clinical goals). This limitation might be related to a limit in the number of iterations. Consequently, the clinical goals in the Ethos template must be defined very carefully: too restrictive or too stringent goals might lead to a significant proportion of plans that will not have time to converge to an optimal dose distribution; conversely, too lenient goals will lead to an important proportion of loosely optimized plans. Concerning DP+DM, this approach outperformed the classical template-based approach of Ethos for initial and updated template of clinical goals, showing that a

more patient-optimal automatic dose optimization seems possible at cost of a longer and more demanding implementation workflow.

To some extent, these investigations reflect the antagonism between explicit formulation of objectives using human expertise and brute-force connectionism relying on big data to blindly reach similar objectives. Although the first approach might be more interpretable and thus reassuring for users, the second approach raises more and more interest due to its flexibility and capability to cover very specific or rare cases, although it might struggle to gain the experts' trust [36]. Considering these advantages and shortcomings, as well as the clinical convenience of a template-based approach, the choice of such approach in the Ethos automatic optimization appears to be rational and relevant, until AI systems further progress with, for instance, the capability to interact with natural language.

Acknowledgements

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".

Funding

This study was supported by the Televie Grant from the Belgian Fonds National pour la Recherche Scientifique F.R.S-FNRS. Ana Barragán is funded by the Walloon region (PROTHERWAL/CHARP, grant number 7289)

References

1. de Jong, R., et al., *Online adaptive radiotherapy compared to plan selection for rectal cancer: quantifying the benefit.* Radiat Oncol, 2020. **15**(1): p. 162.
2. Ahunbay, E.E., et al., *Online adaptive replanning method for prostate radiotherapy.* Int J Radiat Oncol Biol Phys, 2010. **77**(5): p. 1561-72.
3. Vestergaard, A., et al., *Adaptive plan selection vs. re-optimisation in radiotherapy for bladder cancer: a dose accumulation comparison.* Radiother Oncol, 2013. **109**(3): p. 457-62.
4. Archambault, Y., et al., *Making on-line adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning.* Med Phys Intl J, 2020. **8**.
5. Byrne, M., et al., *Varian ethos online adaptive radiotherapy for prostate cancer: Early results of contouring accuracy, treatment plan quality, and treatment time.* J Appl Clin Med Phys, 2022. **23**(1): p. e13479.
6. Sibolt, P., et al., *Clinical implementation of artificial intelligence-driven cone-beam computed tomography-guided online adaptive radiotherapy in the pelvic region.* Phys Imaging Radiat Oncol, 2021. **17**: p. 1-7.
7. Yoon, S.W., et al., *Initial Evaluation of a Novel Cone-Beam CT-Based Semi-Automated Online Adaptive Radiotherapy System for Head and Neck Cancer Treatment - A Timing and Automation Quality Study.* Cureus, 2020. **12**(8): p. e9660.
8. Astrom, L.M., et al., *Online adaptive radiotherapy of urinary bladder cancer with full re-optimization to the anatomy of the day: Initial experience and dosimetric benefits.* Radiother Oncol, 2022. **171**: p. 37-42.
9. Zwart, L.G.M., et al., *Cone-beam computed tomography-guided online adaptive radiotherapy is feasible for prostate cancer patients.* Physics and Imaging in Radiation Oncology, 2022. **22**: p. 98-103.
10. Mao, W., et al., *Evaluation of Auto-Contouring and Dose Distributions for Online Adaptive Radiation Therapy of Patients With Locally Advanced Lung Cancers.* Pract Radiat Oncol, 2022.
11. Moazzezi, M., et al., *Prospects for daily online adaptive radiotherapy via ethos for prostate cancer patients without nodal involvement using unedited CBCT auto-segmentation.* J Appl Clin Med Phys, 2021. **22**(10): p. 82-93.
12. Pokharel, S., A. Pacheco, and S. Tanner, *Assessment of efficacy in automated plan generation for Varian Ethos intelligent optimization engine.* J Appl Clin Med Phys, 2022. **23**(4): p. e13539.
13. Calmels, L., et al., *Evaluation of an automated template-based treatment planning system for radiotherapy of anal, rectal and prostate cancer.* Tech Innov Patient Support Radiat Oncol, 2022. **22**: p. 30-36.
14. Ge, Y. and Q.J. Wu, *Knowledge-based planning for intensity-modulated radiation therapy: a review of data-driven approaches.* Medical physics, 2019. **46**(6): p. 2760-2775.
15. Wang, C., et al., *Artificial intelligence in radiotherapy treatment planning: present and future.* Technology in cancer research & treatment, 2019. **18**: p. 1533033819873922.
16. McIntosh, C., et al., *Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method.* Phys Med Biol, 2017. **62**(15): p. 5926-5944.
17. Babier, A., et al., *The importance of evaluating the complete automated knowledge-based planning pipeline.* Phys Med, 2020. **72**: p. 73-79.
18. Babier, A., et al., *OpenKBP-Opt: An international and reproducible evaluation of 76 knowledge-based planning pipelines.* arXiv preprint arXiv:2202.08303, 2022.
19. Fan, J., et al., *Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique.* Medical physics, 2019. **46**(1): p. 370-381.
20. Petersson, K., et al., *Evaluation of dual-arc VMAT radiotherapy treatment plans automatically generated via dose mimicking.* Acta Oncol, 2016. **55**(4): p. 523-5.

21. Huet-Dastarac, M., et al., *Patient selection for proton therapy using Normal Tissue Complication Probability with deep learning dose prediction for oropharyngeal cancer.* Med Phys, 2023. **50**(10): p. 6201-6214.
22. Nguyen, D., et al., *A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning.* Scientific reports, 2019. **9**(1): p. 1-10.
23. Kandalan, R.N., et al., *Dose prediction with deep learning for prostate cancer radiation therapy: model adaptation to different treatment planning practices.* Radiotherapy and Oncology, 2020. **153**: p. 228-235.
24. Gronberg, M.P., et al., *Dose prediction for head and neck radiotherapy using a three-dimensional dense dilated U-net architecture.* Medical physics, 2021. **48**(9): p. 5567-5573.
25. Zimmermann, L., et al., *Technical Note: Dose prediction for radiation therapy using feature-based losses and One Cycle Learning.* Med Phys, 2021. **48**(9): p. 5562-5566.
26. Liu, S., et al., *A cascade 3D U-Net for dose prediction in radiotherapy.* Medical Physics, 2021. **48**(9): p. 5574-5582.
27. Nguyen, D., et al., *Three-dimensional radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture.* arXiv preprint arXiv:1805.10397, 2018.
28. Barragan-Montero, A.M., et al., *Three-dimensional dose prediction for lung IMRT patients with deep neural networks: robust learning from heterogeneous beam configurations.* Med Phys, 2019. **46**(8): p. 3679-3691.
29. Draguet, C., et al., *Automated clinical decision support system with deep learning dose prediction and NTCP models to evaluate treatment complications in patients with esophageal cancer.* Radiotherapy and Oncology, 2022. **176**: p. 101-107.
30. Nguyen, D., et al., *A comparison of Monte Carlo dropout and bootstrap aggregation on the performance and uncertainty estimation in radiation therapy dose prediction with deep learning neural networks.* Physics in Medicine & Biology, 2021. **66**(5): p. 054002.
31. Villarroel, E.B., X. Geets, and E. Sterpin, *Online adaptive dose restoration in intensity modulated proton therapy of lung cancer to account for inter-fractional density changes.* Physics and imaging in radiation oncology, 2020. **15**: p. 30-37.
32. McNiven, A.L., M.B. Sharpe, and T.G. Purdie, *A new metric for assessing IMRT modulation complexity and plan deliverability.* Medical physics, 2010. **37**(2): p. 505-515.
33. Shen, C., et al., *Clinical experience on patient-specific quality assurance for CBCT-based online adaptive treatment plan.* Journal of applied clinical medical physics, 2023. **24**(4): p. e13918.
34. Visak, J., et al., *Evaluating machine learning enhanced intelligent-optimization-engine (IOE) performance for ethos head-and-neck (HN) plan generation.* Journal of applied clinical medical physics, 2023: p. e13950.
35. Varian, *Ethos Algorithms Reference Guide*. Publication ID: P1035867-003-C, 2019.
36. McIntosh, C., et al., *Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer.* Nat Med, 2021. **27**(6): p. 999-1005.

Supplementary materials

S1

The modulation complexity score was calculated following the same formula than the original article:

Pos is the leaf positions, N is the number of open leaves.

$$pos_{max} = \langle \max(\mathrm{pos}_{N\in n}) - \min(\mathrm{pos}_{N\in n}) \rangle_{\text{leaf bank}}$$

$$LSV_{segment} = \left\langle \frac{\sum_{n=1}^{N}\left(pos_{max} - (pos_n - pos_{n+1})\right)}{N \times pos_{max}} \right\rangle_{left\ bank} \times \left\langle \frac{\sum_{n=1}^{N}\left(pos_{max} - (pos_n - pos_{n+1})\right)}{N \times pos_{max}} \right\rangle_{right\ bank}$$

$$AAV_{segment} = \frac{\sum_{a=1}^{A}\langle pos_a \rangle_{left\ bank} - \langle pos_a \rangle_{right\ bank}}{\sum_{a=1}^{A}\langle max(pos_a) \rangle_{left\ bank \in beam} - \langle max(pos_a) \rangle_{right\ bank \in beam}}$$

$$MCS_{beam} = \sum_{i=1}^{I} AAV_{segment\ i} \times LSV_{segment\ i} \times \frac{MU_{segment\ i}}{MU_{beam}}$$

$$MCS_{plan} = \sum_{j=1}^{J} MCS_{beam\ j} \times \frac{MU_{beam\ j}}{MU_{plan}}$$

S2

Table A1: Study-S2, detailed met goals for each approach. EG_init_selected = Ethos generated (using initial template) for patient manually selected among 45 initial patients, EG_upd_selected = Ethos generated (using updated template) for patient manually selected among 45 initial patients, DP+DM = dose prediction followed by dose mimicking, MG = manually generated.

Volume	Constraint	EG_init plans /10	EG_upd_selected /10	DP+DM /10	MG /10
Anal canal	V60Gy (%)	9	8	9	10
	V35Gy (%)	8	8	9	10
	V20Gy (%)	9	10	10	10
Bladder	V60Gy (%)	10	9	10	10
	V50Gy (%)	10	10	10	10
	V40.8Gy (%)	6	6	7	9
Bowel	D0.1cm³ (Gy)	9	9	10	10
	V55Gy (%)	10	10	10	10
	V20Gy (cm³)	10	10	10	10
Rectum	V60Gy (%)	8	6	10	10
	V52.8Gy (%)	7	7	7	7
	V30Gy (%)	7	7	7	8
Femur_L	V40Gy (%)	10	10	10	10
	V35Gy (%)	10	10	10	10
Femur_R	V40Gy (%)	10	10	10	10
	V35Gy (%)	10	10	10	10
	V54.1Gy (%)	9	9	8	10

Penile Bulb	V42.5Gy (%)	10	10	10	10