**Max 10000 mots.**

# Automated online adaptive proton therapy treatment planning for esophageal cancer: which combination of deformable registration and deep learning planning tools performs best?


**Abstract (max 300) - reste 0**

**Objective:** To demonstrate the feasibility of integrating various fully-automated online adaptive proton therapy strategies (OAPT) within a commercially available treatment planning system. These strategies utilize existing deformable image registration (DIR) algorithms and state-of-the-art deep learning (DL) networks for organ segmentation and proton dose prediction.

**Approach:** Four OAPT strategies featuring automatic segmentation and robust optimization were evaluated on a cohort of 17 patients, each undergoing a repeat CT scan. (1) DEF-INIT combines deformably registered contours with template-based optimization. (2) DL-INIT, (3) DL-DEF, and (4) DL-DL employ a nnU-Net DL network for organ segmentation and a controlling regions-of-interest (ROIs)-guided DIR algorithm for iCTV segmentation. DL-INIT uses this segmentation alongside template-based optimization, DL-DEF integrates it with a dose-mimicking (DM) step using a reference deformed dose, and DL-DL merges it with DM on a reference DL-predicted dose. All strategies were evaluated on manual contours and on the contours used for optimization, and compared with manually adapted plans. Key dose volume metrics like iCTV D98% are reported.

**Main results:** iCTV D98% was comparable in manually adapted plans and for all strategies in nominal cases but dropped to 20Gy in the worst-case scenarios for a few patients per strategy, highlighting the need to correct segmentation errors in the target volume. Evaluations on optimization contours showed minimal relative error, with some remaining outliers, particularly in template-based strategies (DEF-INIT and DL-INIT). DL-DEF achieves a good trade-off between speed and dosimetric quality and showed a passing rate (iCTV D98% > 94%) of 90% when evaluated against 2, 4 and 5mm setup error and of 88% when evaluated against 7mm setup error. While template-based methods are more rigid, DL-DEF and DL-DL have potential for further enhancements with proper DM algorithm tuning.

**Significance:** Among investigated strategies, DL-DEF and DL-DL demonstrated promising within-10-minutes OAPT implementation results and significant potential for future improvements.


## Introduction

Intensity-modulated proton therapy (PT) potentially yields improved dose conformity compared to conventional photon-based radiotherapy techniques (1,2). For esophageal cancer (EC), this may translate into lower risks of developing toxicities in important organs-at-risk (OARs), such as the lungs and heart, and a reduced 1-year mortality (3–5). The energy deposition pattern of PT is a double-edged sword that makes dose distributions highly conformal but very sensitive to treatment uncertainties and anatomical variations (6). Robust optimization and adaptive therapy are used to address these issues, respectively.

Currently, adaptive proton therapy is often performed offline with a few days between the acquisition of the new planning image and the delivery of the approved adapted plan. However, daily online adaptive planning better accounts for rapid anatomical changes and may allow to reduce the magnitude of the uncertainties taken into account in the robust optimization (i.e., reducing the robustness parameters) (7). However, clinically compatible online adaptive proton therapy (OAPT) requires a largely automated treatment planning workflow. Substantial research efforts have been made towards that objective in recent years. Investigations focused on three key aspects: the availability of daily imaging of near-CT quality, the ability to reliably propagate/delineate contours on repeated images with minimal manual intervention; and automatic optimization of a new treatment plan. This study focuses specifically on the last two aspects.

While deformable image registration (DIR) was a very popular tool for contour propagation and segmentation by deformation ten to fifteen years ago (8,9), deep learning (DL) models such as U-Net are now prevailing for OARs segmentation (10). For the target volume, although the number of studies is increasing (11–14), their reported accuracy has not reached a level suitable for clinical implementation. For EC, Jin et al. (15) introduce a DL-based method for clinical target volume (CTV) delineation that integrates contextual visual characteristics and encoded spatial distances to both the gross tumor volume and associated lymph nodes, thereby simulating the margins and appearance criteria utilized by radiation oncologists during the segmentation process. They achieved an absolute Dice Similarity Coefficient (DSC) of 0.826. Notably, none of the surveyed studies adhere to the current European consensus guidelines for esophageal CTV delineation as outlined by Thomas et al. (16). Besides, only a few studies provide segmentation metrics associated with dosimetric performances and most of them reveal that, even with a close conformity to the groundtruth volume, a reduction in target coverage is observed for auto-segmented contours (11,17). A recent commercial implementation of online adaptive intensity modulated radiotherapy (IMRT), ETHOS (Varian Medical Systems, Palo Alto, CA), relies on the combination of DIR and DL for CTV delineation, selecting a subset of "influencers" organs delineated by a DL network. These influencers, situated near the target volume, guide the DIR-based propagation of the target volume from the planning CT to the CT of the day (18).

DL has also gained popularity in automatic planning, using a DL model for automatic dose distribution generation, followed by a dose mimicking (DM) optimization step (19). Borderias et al. (20) compared three strategies for OAPT in head and neck patients, using DM for the optimization method. The best performing strategy, in terms of target coverage and complication probability values, involved a DL dose predicted on synthetic CTs. While this workflow has been proven effective for head and neck or prostate cases, especially for PT (21–23), it has yet to be studied in EC cases, making use of the existing PT dose prediction models for EC (24).

In order to minimize anatomical uncertainties, it is important that the time between image acquisition and treatment delivery is also minimized. Therefore, it might be of interest to fully automate treatment planning without any manual intervention, provided that the errors in the segmentation occurring along the way are minimal. However, such an approach is clinically viable only if the resulting treatment plan is still within clinical requirements in terms of target coverage and OAR sparing, considering a given robustness requirement. In this study, we

investigate whether the error introduced by the full automation could be compensated by increasing the robustness parameters compared to the ones that could be used in an ideal online adaptive strategy where only intra-fraction and range errors remain. We assume that such a strategy might still be more effective than investing too much time on manual editions, as these adjustments would also eventually require larger robustness parameters to accommodate anatomical modifications occurring between the new planning image and the effective anatomy of the patient during irradiation. Therefore, in this work, we investigate the proportion of our patient population for which an increase in the residual setup error might compensate for full automation-induced errors.

Given the respective development of DIR and DL and their frequent use in studies aimed at demonstrating the feasibility of an online adaptive workflow, this work aims to test well established techniques in an end-to-end OAPT process. We implement four fully automated OAPT techniques on a commercially available treatment planning system and compare them to reference manually adapted plans for EC patients. We investigate two auto-contouring techniques and combine them with optimization strategies involving template-based optimization or dose generation algorithms followed by a DM step. Additionally, we evaluate the impact of increasing the robustness setup parameters on the automation-induced error mitigation.

## Materials and methods

### *Patient cohort*

We used a retrospective patient database containing 40 patients diagnosed with locally advanced EC, treated with neoadjuvant chemoradiotherapy followed by surgery at UZLeuven (Leuven, 3000 Belgium). For this study, we used the average planning CT (pCT) scans and corresponding manually generated contours. To ensure database consistency and quality, all patients were planned for intensity modulated proton therapy (PT) with pencil beam scanning, using robust optimization on the internal CTV (iCTV) (7mm setup error and 2.6% range error). The plans were made in RayStation 11B Software (RaySearch Laboratories) by two experienced medical physicists and reviewed by the same radiation oncologist. Planning guidelines from the PROTECT clinical trial (25) were applied. We used two posterior beams: 150° left posterior oblique (LPO) and 180° posterior (PO). The prescribed treatment regimen involved a total dose prescription of 50.4Gy, administered in 28 fractions. Detailed information about the planning protocol can be found in Section A of the Supplementary Materials.

After about three weeks of treatment, 18 out of 40 patients in our patient cohort had a repeated CT (rCT) scan, out of which one patient was left out because of missing slices. On this remaining cohort of 17 patients, one radiation oncologist manually delineated OARs and target contours on the rCT, and the patients' intensity modulated radiotherapy (IMRT) and PT plans were consistently manually adapted by two medical physicists, again following the planning guidelines of the PROTECT clinical trial (25). The same radiation oncologist subsequently reviewed all these adapted plans, which will serve as baseline in this study. All subsequent evaluations are done on these average rCT scans.

### *Contour delineation*

Two different types of automatically segmented contours were made available on the rCT scans: 1) a DIR based technique, and 2) a combined DL and DIR technique.

The first strategy involved the hybrid ANACONDA deformable registration algorithm, built upon an intensity and anatomical information-based algorithm (26) implemented in RayStation 2023B Software (RaySearch Laboratories, Sweden). This algorithm was used to deformably map the contours of 9 OARs (fundus, heart, both kidneys separately, left ventricle, liver, lungs as a single organ, and spinal canal) and iCTV to the rCT scans.

The second technique involved the use of a nnU-Net model, a network considered as state-of-the-art in the field (10), which was trained to predict the contours of the 9 OARs from the corresponding CT scan only. A total of 8 models were trained, with a training set of 35 patients and the remaining 5 patients being used for testing, in a kind of ‘cross-testing’ scheme. The models were trained for 1000 epochs, which is the default number of epochs recommended for nnU-Net training. Inference of the 9 OARs contours for one patient can be made in approximately 50 seconds on a GPU Tesla A100 with 80 Gb of RAM. Careful selection of the model used for inference was done so that a model can only be used to predict on a rCT if the patient's pCT is part of the testing set, and not of the training set. The predicted OARs contours were compared to manual contours, or groundtruth, by computing DSC. However, we noticed that the nnU-Net model segmented the OARs in every slice where it considers that these structures are present, irrespective of the volume that will actually receive dose. For the spinal canal in particular, the nnU-Net provides a contour on every slice where vertebrae, surrounding the spinal canal, are present. This is inconsistent with the groundtruth contours provided by a medical doctor which, in the interest of time, will only segment structures on slices surrounding the target volume, such as noticeable in Figure 1. For a fair DSC computation, we cropped the spinal canal nnU-Net prediction to the slices where a groundtruth contour was present. After OAR prediction with our in-house trained nnU-Net, the iCTV was generated by the ANACONDA deformable registration algorithm, implemented in RayStation and, here, guided by a selected subset of this organs (also known as “controlling ROIs"), similar to what is done with the selection of “influencers” in the ETHOS system (Varian Medical Systems, Palo Alto, CA) (18). The selected controlling ROIs are soft tissues, directly surrounding the target volume. An additional criterion for selecting a controlling ROI is to have an average DSC greater than 0.8. The resulting deformed iCTV is compared to the manually defined iCTV.

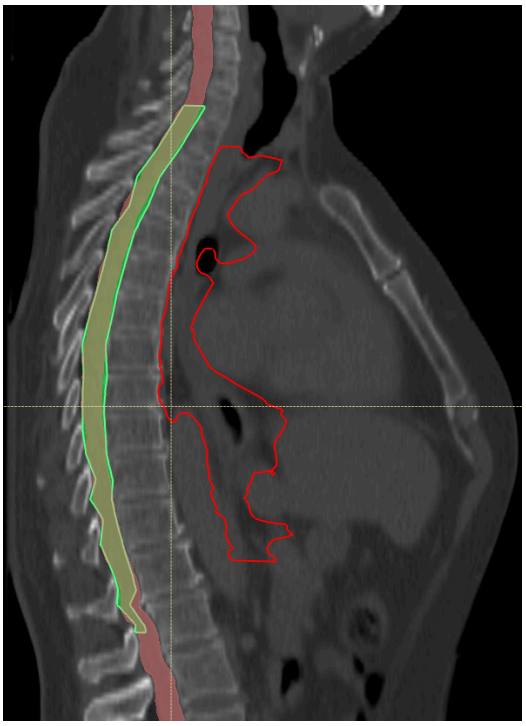

*Figure 1: Patient CT scan showing the iCTV in red, the manually delineated spinal canal in green (shaded area) and the DL predicted spinal canal contour in pink (shaded area).*

*Treatment plan optimization strategies*

All the presented strategies are fully automated in the scripting environment of RayStation 2023B Software. 1) a template-based optimization strategy, 2) a DIR + DM-based strategy, 3) a DL + DM-based strategy. All strategies are optimized on a 3mm dose grid, with the same robust optimization settings as for the initial plan.

The first strategy, here referred to as the template-based optimization strategy, the objective functions defined for the initial plan are copied in an adapted plan before the optimization process is run for 200 iterations, with a cold start, following the PROTECT protocol (25). The objectives are defined on the automatically segmented contours. In the second strategy, the dose of the initial plan is deformed with a hybrid ANACONDA DIR (26), subsequently followed by a DM optimization step giving us a deliverable plan. For the third strategy, we used the previously published DL model for PT dose prediction (ref). Multiple HDU-Net models were trained using the same database as for the nnU-Net segmentation model, following a cross-validation and cross-testing scheme. The patients' set is divided into 25 patients for training, 5 for validation and 10 for testing. The models take as input the contours of 9 OARs (fundus, heart, both kidneys separately, left ventricle, liver, lungs as a single organ and spinal canal as well as 3mm ring extension of the spinal canal) and the corresponding CT scan. More information about the model architecture and training can be found in (24) and in Section B of the Supplementary Materials. Inference for the PT dose distribution of one patient is obtained in approximately 50 seconds on a GPU Tesla A100 with 80 Gb of RAM, while predicting over 5-folds and averaging the prediction. The predicted dose distribution is subsequently translated into a deliverable plan using a DM optimization.

The DM optimization algorithm used in the second and third strategies, here referred to as the DM-based strategies, consists of a voxel-wise least-squared error minimization problem. Typically, voxels in the target volume or in a close proximity (high isodoses) are highly penalized for underdosage or overdosage as compared to a reference dose and this penalty decreases in lower isodoses. The reference dose corresponds to the dose distribution to be mimicked. In addition, we include two conventional radiotherapy robust objective functions on the CTV. On the OARs, no additional conventional radiotherapy objective functions were implemented. This is a simpler version of the DM algorithm used by Borderias et al. (20). Given the iteration counts required to achieve an optimization plateau, we have set the number of iterations for the DM algorithm to 60, instead of the 200 iterations fixed in the first optimization strategy. In the results section, we illustrate the accuracy of our DM algorithm in reproducing the reference dose by comparing DVH metrics.

*OAPT strategies*

Figure 2 illustrates the four proposed OAPT strategies, which are a combination of the two automatic delineation techniques and the three automatic optimization strategies. DEF-INIT combines the deformably registered contours with the template-based optimization strategy, optimizing on the objectives of the initial plan. The three other strategies, DL-INIT, DL-DEF, and DL-DL, involve predicting the contours of the OARs and then utilizing a selected subset of controlling ROIs to guide the deformation of the iCTV on the rCT. In terms of optimization strategies, the optimization function of DL-INIT includes all the objectives from the initial plan. In DL-DEF, the initial dose is deformably registered on the rCT, and the deformed dose is used as input for the DM algorithm. Finally, in DL-DL, the dose is predicted by a DL HD U-Net model before being used as the reference dose for the DM algorithm.

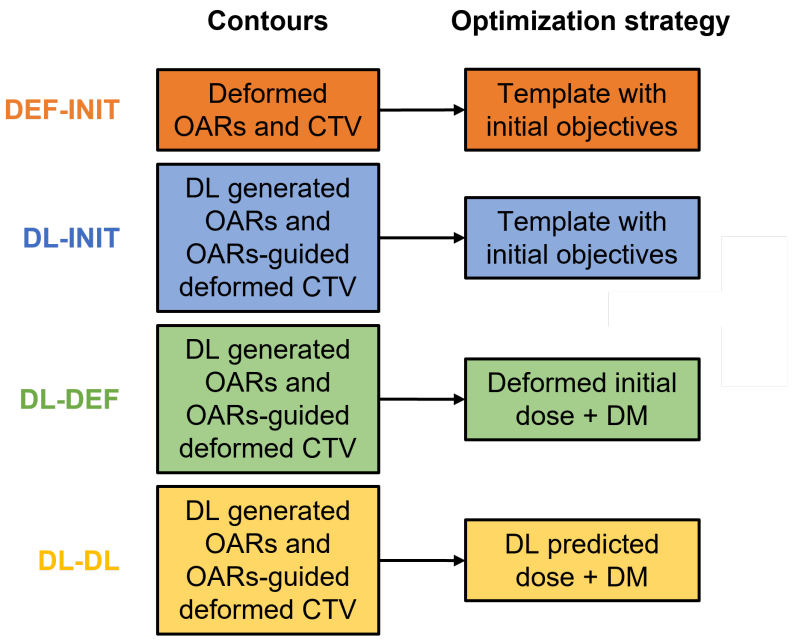


*Figure 2: Workflow illustrating the four proposed OAPT strategies. DL = deep learning, DM = dose mimicking*

*Plan evaluation*

The dosimetric effect of setup (7mm) and range (2.6%) uncertainties were evaluated for each strategy with respect to the manual contours (27). However, to discriminate between errors coming from the delineation strategy and errors coming from the optimization strategy, we additionally evaluate on the automatically segmented contours. Results of these evaluations are illustrated in comparison with the manually adapted plan. We present three metrics evaluating the target coverage, namely D98%, V95% and the mean dose (Dmean) to iCTV. For OARs, we show the mean lungs dose (MLD), mean heart dose (MHD) and D0.05 $cm^3$ to the spinal canal.

*Automation-induced errors mitigation*

In this work, we evaluated the proportion of our patient cohort achieving a worst-case iCTV D98% over 94%, under 2, 4, 5 and 7mm residual setup error. This value, here referred to as the “passing rate”, provides insight into the potential mitigation capabilities of increased robustness parameters against errors introduced by full automation. The selection of the threshold dose aligns with the findings of Visser et al. (28), indicating that acceptable coverage is typically achieved with iCTV D98% above 94%. Strategies exhibiting an acceptable target coverage for more than 90% of patients show robustness comparable to IMRT plans with PTV margins (29). The evaluation was performed on both the manual iCTV and on the deformed iCTV used for optimization. The latter offers insight into the failure of the optimization process to yield an acceptable plan, even with accurate contours, which may be mitigated by accounting for increased setup errors.

**Results**

*Contour delineation*

Both contour delineation methods, namely, the DIR and nnU-Net, are considered as state-of-the-art strategies (10,30,31). For every structure that is a part of the optimization objective function, we intended to determine which of these two benchmark delineation techniques works best. The DSCs between the groundtruth manual structures and the automatically segmented structures are reported in Figure 3 with boxplots.

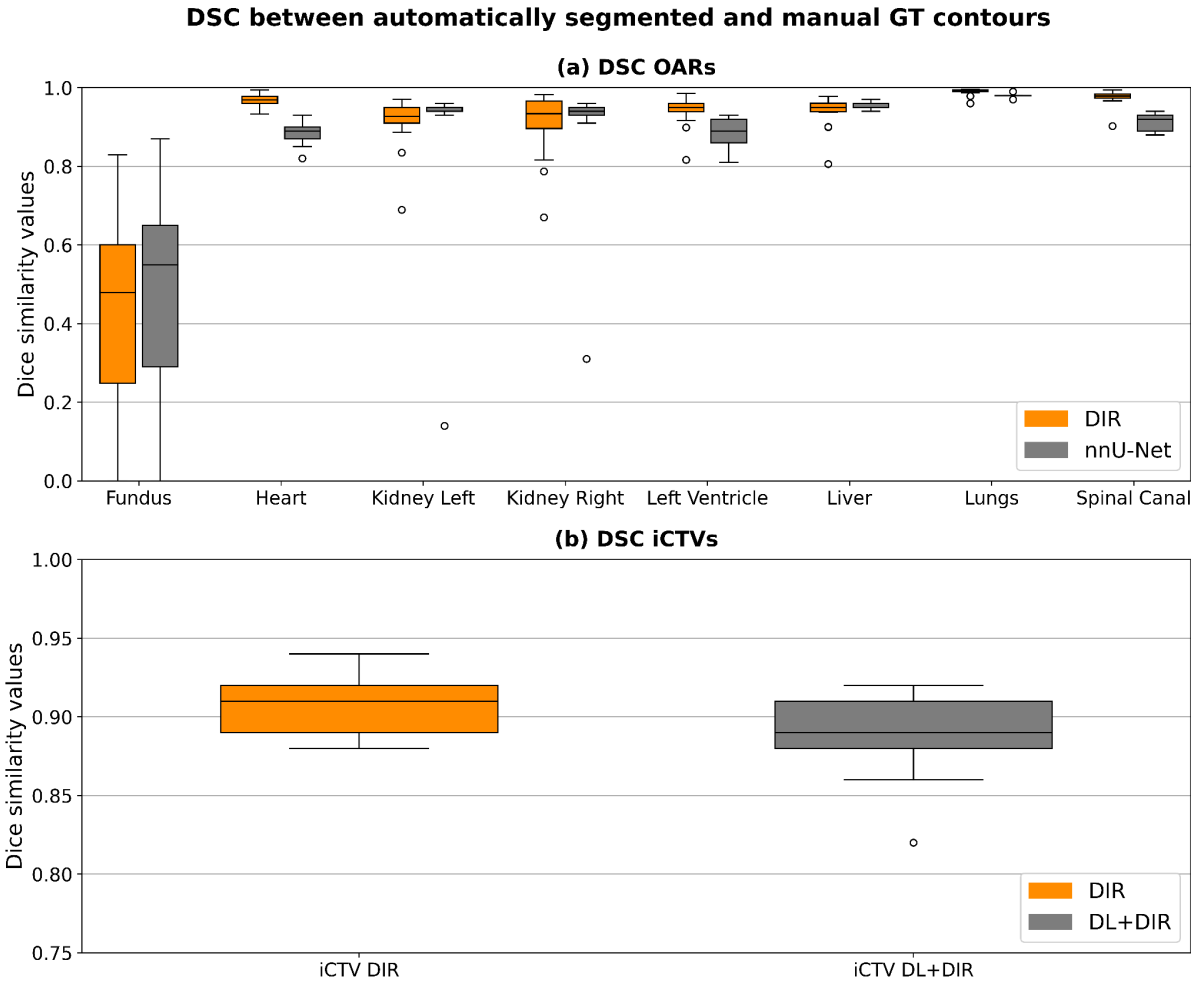


*Figure 3: Patient distribution of Dice similarity coefficients (DSC) between manual groundtruth contours and automatically segmented contours. In (a), orange box plots illustrate DSC obtained in comparison with DIR mapped OARs contours and gray box plots illustrate the comparison with nnU-Net segmented OARs contours. In (b), orange box plots represent DSC obtained in comparison with DIR mapped iCTV contours and gray box plots represent DSC computed on iCTV contours generated with a DL-OARs-guided DIR.*

The kidneys are generally better segmented by nnU-Net with median DSCs of 0.95, 0.94, and 0.96, respectively, displaying very little variations among the whole patient population. Still, we observe that our nnU-Net model entirely missed both kidneys in the case of one outlier patient. However, since the kidneys were not used in patient 17's initial optimization objective function or as controlling ROIs to guide the iCTV deformation, there were no consequences for the implementation of any strategies. Unlike other patients in the rCT cohort, this patient showed a less contrasted rCT scan, although none of these patients received a contrast agent injection prior to the rCT. Regarding the lungs and liver, results from the DIR algorithm and nnU-Net are comparable ($p = 0.003$ and $p = 0.1$ respectively), though the DIR algorithm did produce a few outliers for the liver contour. The DIR algorithm produces heart and spinal canal structures that are closer to the groundtruth contour than the contours predicted by nnU-Net. Regarding the fundus, none of the two approaches yields an accurate delineation.

Figure 3b shows the DSC for the two automatically obtained iCTV structures. Again, both strategies behave similarly with median DSC of 0.9 for the iCTV produced by the DIR algorithm and 0.89 for the iCTV produced by DL segmented OARs, guiding a DIR algorithm.

*Dose mimicking*

Two of our automated strategies, DL-DEF and DL-DL, rely on the ability of our mimicking algorithm to accurately reproduce a reference dose. We evaluate this accuracy by comparing DVH metrics before and after the mimicking step. These results can be observed in Figure 4.

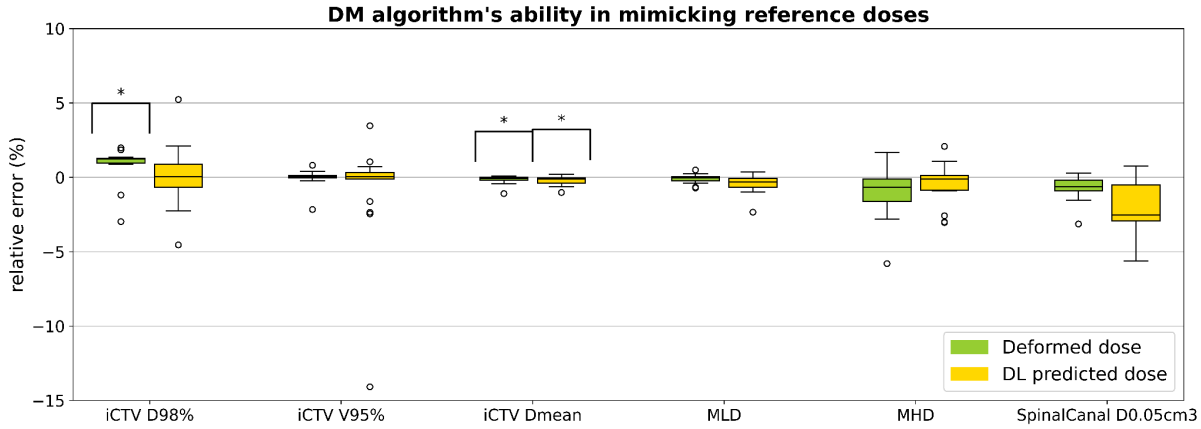


*Figure 4: Relative error on the DVH metrics (reference dose - mimicked dose), showing target coverage metrics, mean lungs dose (MLD), mean heart dose (MHD) and dose to the hottest 0.05 $cm^3$ of the spinal canal. Dose metrics are given in percentage of the prescribed dose (50.4 Gy). Green box plots show the comparison between deformed reference dose and mimicked deformed dose. Yellow box plots show the comparison between DL predicted reference dose and mimicked predicted dose. * indicates a statistically significant difference between the two distributions (p-value < 0.05).*

We observe that the difference between the reference dose and the dose after DM is less than 5% for each metric under study. For dose metrics, this corresponds to a dose difference of less than 2.5 Gy approximately. We can also notice that for the iCTV V95% and Dmean in particular, the DM algorithm seems to precisely recover high dose regions and average dose, except for a few outliers. Regarding the D98% however, we observe that the DM algorithm is less precise, in particular with the DL predicted reference dose, showing a larger standard deviation than for other iCTV DVH metrics (median +/- standard deviation = 0.06 +/- 2.0), which suggests that errors around the high dose region might be introduced by the optimization process. With the deformed dose used as reference, the median iCTV D98% presents a slight decrease after the DM process (1.25 +/- 1.16). Moreover, the DM algorithm performs less accurately for some OARs, especially in the spinal canal (-0.63 +/- 0.99 for the difference between the deformed dose before and after DM and -2.54 +/- 1.72 for the difference between the DL predicted dose before and after DM). It is however important to note that in the DM optimization algorithm, we integrated conventional radiotherapy objective functions solely on the target volume, alongside voxel-wise least-squared error minimization. This prioritizes target coverage over other OAR criteria during optimization. With our beam

configuration (150° LPO and 180° PO) and target aim, the spinal canal is typically found in lower dose regions than other OARs, which means that the DM algorithm considers it as less important. Although some dose distributions might be significantly different from each other from a statistical point of view, these differences are not clinically relevant.

*OAPT strategies*

In the interest of evaluating whether our strategies could be implemented in an online setting, we timed each strategy in every patient case, including contour delineation and optimization. Table 1 reports the results of these time records. Both DL prediction times are computed while running the inference on a GPU Tesla A100 with 80Gb of RAM. It is worth mentioning that the median time reported for an inference with HD U-Net accounts for the inference over 5-folds.

| **Strategies** | **Median time (seconds)** |
|---|---|
| **Contour delineation** | |
| DIR | 64 (+/- 10) |
| nnU-Net inference + DIR | 115 (+/- 30) |
| *nnU-Net inference* | 48 (+/- 17) |
| **Optimization strategy** | |
| Template-based optimization | 1281 (+/- 408) |
| Dose deformation + DM | 449 (+/- 161) |
| HD U-Net inference + DM | 517 (+/- 164) |
| *HD U-Net inference* | 53 (+/-) 18 |
| **OAPT strategy** | |
| DEF-INIT | 1324 (+/- 415) |
| DL-INIT | 1520 (+/- 421) |
| DL-DEF | 583 (+/- 181) |
| DL-DL | 622 (+/- 186) |

*Table 1: Median time for each strategy and sub-strategy. nnU-Net inference time and HD U-Net inference time are accounted for in nnU-Net inference + deformable image registration (DIR) time and in HD U-Net inference + dose mimicking (DM) time respectively. Rather than being a sum of the median delineation and median optimization times, the timings for the online adaptive proton therapy (OAPT) strategies - which include both a contour delineation strategy and an optimization strategy - are also median times computed over the timings required to complete each OAPT strategy for each patient. DEF-INIT combines structures deformation with the template-based optimization strategy, DL-INIT combines OARs contours prediction followed by controlling ROIs-guided iCTV deformation and a template-based optimization strategy, DL-DEF combines OARs contours prediction followed by controlling ROIs-guided iCTV deformation and dose deformation followed by a DM optimization algorithm while finally, DL-DL combines OARs contours prediction followed by controlling ROIs-guided iCTV deformation and dose prediction followed by a DM optimization algorithm.*

*Plan evaluation*

For the two DM-based strategies, namely DL-DEF and DL-DL, results after the DM optimization process are depicted in the following Figures. In Figure 5, results of the plan evaluation over the manual contours can be observed. Figure 5a shows the relative error (OAPT dose - manually optimized dose) in terms of DVH metrics of the iCTV. For each metric both the nominal and worst-case scenarios with a 7mm setup error and a 2.6% range are depicted. We noticed that the worst-case iCTV D98% drops down to approximately 20 Gy for one or two patients with every strategy, which is not clinically acceptable. For the other metrics and other scenarios, the relative error remains under 5Gy or under 10% of volume except for DL-DL iCTV V95% worst-case and of some outliers, which remain consistent across the different strategies. Figure 5b shows the relative error for some OARs, usually implicated in the initial optimization objective function, and known as high-risk OARs for EC. For the mean lungs dose (MLD) and the mean heart dose (MHD), the relative error stays below 5%, which corresponds to a threshold of approximately 2.5Gy.

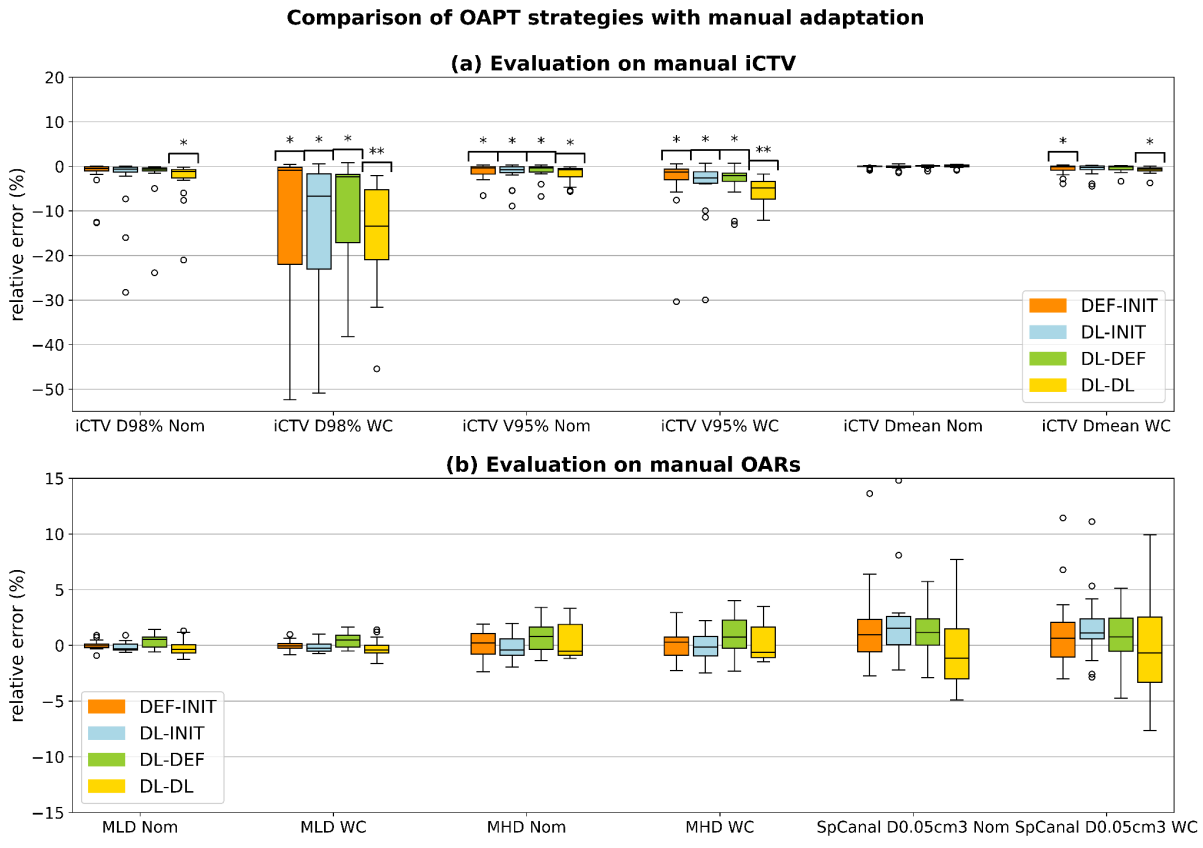


*Figure 5: Boxplots showing DVH metrics for the manually contoured iCTV and OARs. The relative error is computed between the OAPT dose and the manually optimized dose. Dose metrics are given in percentage of the prescription (50.4Gy). * indicates a statistically significant difference (p-value < 0.05) between the metric calculated on the corresponding OAPT strategy dose distribution and the metric calculated on the manually adapted dose distribution. ** indicates a high statistically significant difference (p-value < 0.01).*

We further evaluated on the automatically generated contours to distinguish between the dosimetric impact of the optimization error and the delineation error. The results are displayed in Figure 6. For the iCTV, we observe a consistent decrease in all high dose metrics errors. However, the template-based optimization strategies (DEF-INIT and DL-INIT) show strong outliers. Although the DL-DL strategy has a higher median relative error, it appears to handle these outliers better. For the OARs, we do not observe any important difference, when comparing the boxplot in Figure 6b to the boxplot in Figure 5b. Additionally, we did not observe any significant ($p > 0.05$) difference between the error made by the template-based strategies and the error made by the DM-based strategies (DL-DEF or DL-DL) on the iCTV DVH metrics. Regarding OARs, we did observe significantly higher differences for the MLD and worst-case MHD. These statistical results are not indicated on Figure 5. Results showing the impact of the DM step can be found in Supplementary Materials, Section C.

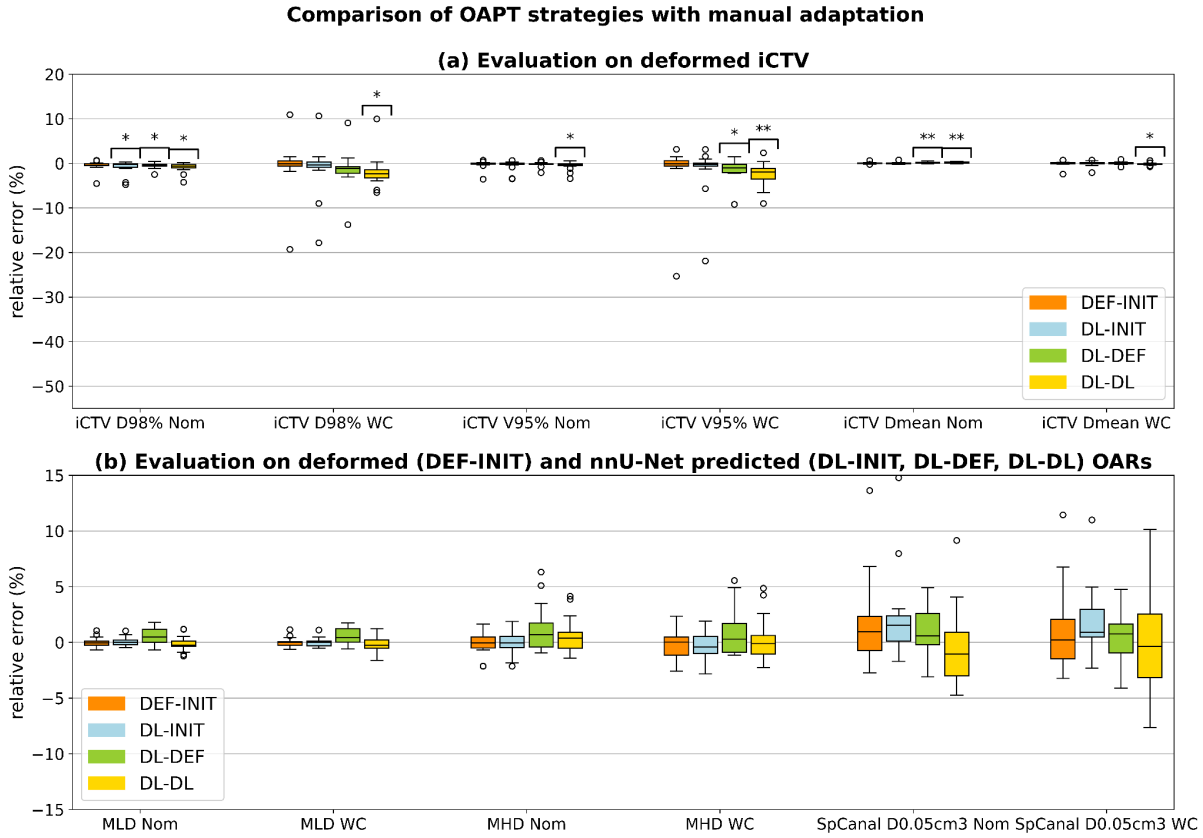


*Figure 6: Boxplots showing DVH metrics for the automatically contoured iCTV and OARs. The relative error is computed between the OAPT dose and the manually optimized dose. Dose metrics are given in percentage of the prescription (50.4Gy). * indicates a statistically significant difference (p-value < 0.05) between the metric calculated on the corresponding OAPT strategy dose distribution and the metric calculated on the manually adapted dose distribution. ** indicates a high statistically significant difference (p-value < 0.01).*

Remarkably, compared to DL-DL, we noticed a lower median relative error for some iCTV metrics when using DL-DEF. The only difference between these two strategies is the reference dose, which is computed by a DL network in DL-DL and is simply deformed on the anatomy of the day in DL-DEF. Both strategies use the same contour segmentation and DM optimization techniques.

*Automation-induced errors mitigation*

Finally, we investigated the ability of each strategy to compensate against various setup errors by evaluating the iCTV D98% on different setup margins. Figure 9 illustrates the proportion of patients for whom each strategy achieves an iCTV D98% above 94% of the prescription dose, denoted here as “passing rate”. Notably, our analysis reveals that none of the strategies, when evaluated on manual contours, achieve a passing rate over 90%, irrespective of the residual setup error considered. Specifically, the passing rate decreases to approximately 10% for the DL-DL strategy evaluated against a 7mm margin. However, when evaluated on the automatically deformed iCTV, the passing rate of DEF-INIT, DL-INIT and DL-DEF remain consistently around 90% across all residual setup errors. DL-DL, on the contrary, fails to adequately compensate for 5 and 7mm errors, showing a passing rate below 80% and just above 50% respectively.

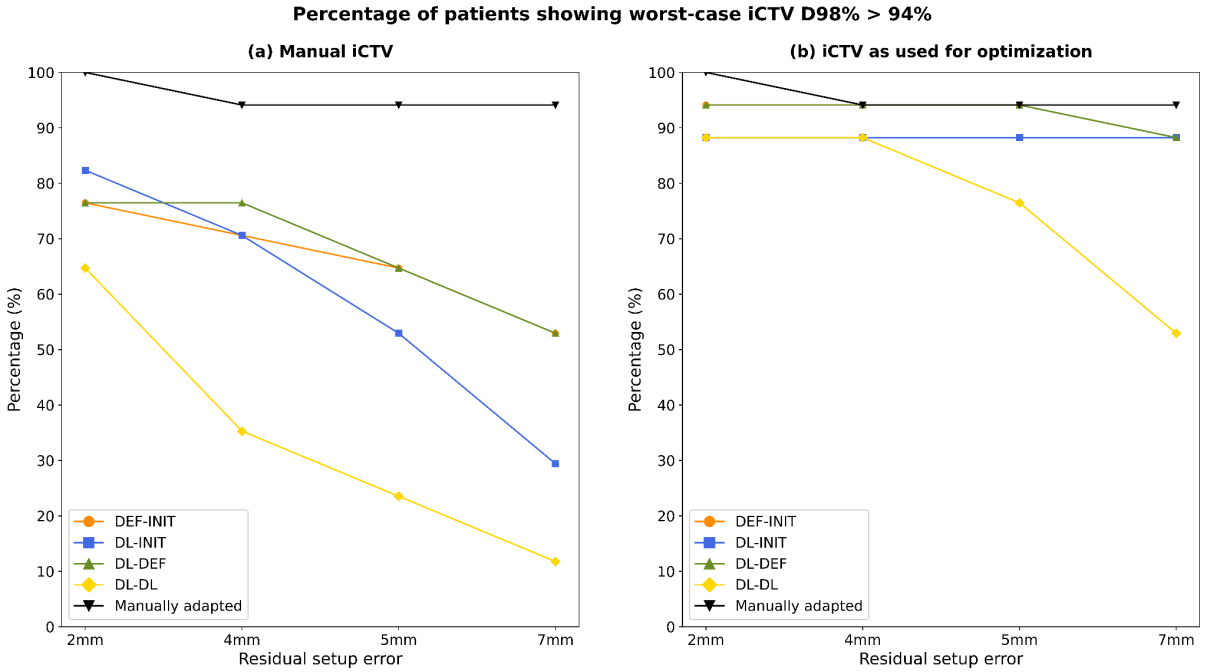


*Figure 9: Percentage of patients having iCTV D98% above the threshold dose (47.38Gy = 94% of the prescription) for each OAPT strategy and manual adaptation. The selected threshold dose follows the adequate target coverage criterion of Visser et al.* (28).

## Discussion

This study illustrates the feasibility of integrating a comprehensive online adaptive proton therapy (OAPT) workflow into a commercially available treatment planning system. We conducted a comparative analysis of four OAPT strategies, integrating contour delineation and optimization techniques, with a focus on time efficiency and clinical objectives requirements. Notably, two approaches, DL-DEF and DL-DL, emerge as promising alternatives, each offering distinct advantages. These methods not only enable

implementation within a median time of 10 minutes, including automatic contour delineation and optimization, but also offer great potential for improvements, which will be discussed later.

Evaluation of delineation methods indicates comparable precision between DIR and nnU-Net approaches, as measured by DSCs. While these strategies do not consistently achieve acceptable results across all anatomical structures, the DSCs for the critical structures, such as those surrounding the tumor volume or used in optimization objectives for EC treatment planning, are larger than 0.8. We observed an exception for the delineation of the spinal canal by nnU-Net, highlighting an inconsistency in our database. We noticed that long structures, such as the spinal canal, are manually not delineated on all slices but rather selectively on slices adjacent to the tumor volume, where a dose could be deposited. This practice is understandable, considering the time constraints faced by physicians, necessitating delineation efforts where clinically relevant. For a fair DSC comparison, we cropped the spinal canal predictions to slices containing the manual structures. Therefore, we believe that we could improve the performance of our nnU-Net model by enhancing the consistency across the database, namely by establishing precise guidelines for constructing databases aimed at training and evaluating AI models effectively, similarly to the conclusion of Barragan et al. (32). Additionally, the nnU-Net that is presented in this work was trained on a small database with only 40 patients, suggesting that a more extensive training dataset could potentially improve its performance. Besides, while the training database consisted of CT scans obtained immediately following a contrast-enhanced fan beam CT, our contours were predicted on non-enhanced rCT scans, which may have affected the results even further.

Moreover, despite achieving high median DSCs of 0.9 with DIR and 0.89 with nnU-Net+DIR for the comparison of manual and deformed iCTV contours, we observed significant dosimetric impact due to these minor delineation errors. Indeed, strong outliers were found in high dose metrics like D98% and V95% in the nominal scenarios of every strategy. Furthermore, these delineation inaccuracies compromise the robustness of the plan against the evaluated setup error (7mm), as evidenced by the iCTV D98% worst-case scenarios depicted in Figure 5a. This emphasizes a recurrent issue highlighted by many studies, wherein identifying a contour evaluation metric capable of accurately capturing all discrepancies between automatic and groundtruth contours remains difficult (33–35). Our study demonstrates again the importance of integrating geometric evaluation, such as DSC or Hausdorff distance, with a dosimetric evaluation. However, in line with the results of Smolders et al. (17), we do not observe substantial dosimetric impact on the OARs, with moderate differences between median DVH metrics evaluated on the manual contours and on the auto-segmented contours. These results suggest that editions of our critical OAR structures are not necessary, typically while editions of the tumor volume remain necessary with the investigated tools. The use of the recently published Dosimetrically Informed Volume Edition tool for adaptive radiotherapy (DIVE-ART) (36) may prove effective in accelerating the edition process. It indicates regions of the replanning image that are most relevant to edit from a dosimetric point of view.

Regarding the evaluation of the different optimization strategies, we have observed that in addition to doubling the adaptation time, at minimum, the two template-based strategies (DEF-INIT and DL-INIT) do not achieve significantly lower target DVH metrics error as

compared to the DM-based strategies (DL-DEF and DL-DL) ($p > 0.05$). Furthermore, these template-based strategies fail to achieve acceptable target coverage for some outlier patients. For the OARs, we do not believe that the significantly higher errors observed with the DM-based strategies, in comparison with the template-based strategies, are clinically relevant, being less than 1.3Gy.

Notably, the DL-DEF strategy outperformed the DL-DL method on average. This finding contrasts with previous conclusions drawn by Borderias et al. (20), where the DL-predicted dose exhibited superior target coverage compared to the deformed dose. However, in the past years, dose deformation has been widely used for dose accumulation (37–39). In their paper, Zhang et al. (40) explored the impact of different DIR algorithms and settings available in RayStation. They demonstrated that refining the implementation of the ANACONDA DIR algorithm, by using more ROIs as guidance for the dose deformation (similar to what was done to deform the target from controlling DL-predicted controlling ROIs in this work) could further mitigate average dosimetric variations. Although we have chosen not to use such controlling ROIs to facilitate a clear distinction between our delineation and optimization strategies, the DL-DEF approach yielded satisfactory results, achieving acceptable target DVH metrics across nearly all patients when evaluated against auto-segmented or groundtruth contours, with the exception of the iCTV D98% in the worst-case scenario. Besides, DL-DEF exhibited a passing rate (iCTV D98% over 94%) over 90% when evaluated on the deformed iCTV against 2, 4 and 5mm and a passing rate of 88% with a 7mm setup, surpassing DL-DL (passing rate ranging from 88% to 53% when evaluated against 2 and 7mm, respectively). Nonetheless, we observed an outlier in the dosimetric evaluation of the DL-DEF strategy, for which perturbed scenarios appeared to be more effectively managed by the DL-DL approach. This discrepancy could be attributed to the intrinsic properties of both the deformation and prediction algorithms. Specifically, while the patient-specific nature of the DL-DEF algorithm may perform better in generic cases with minimal anatomical deformation, the population-based DL prediction algorithm might demonstrate superior performance in handling complex cases, given that it was trained on a large set of anatomies. The disparity with the conclusion drawn by Borderias et al. (20) may arise due to the substantial anatomical variations typically encountered in head and neck patients (41–43). Moreover, the non-diffeomorphic nature of some anatomical variations (e.g., apparition of gas bubbles in the gastro-intestinal tract) prevents the ANACONDA DIR algorithm from being used appropriately to deform the dose on patients with such variations (26,44).

Although DL-DL exhibits lower robustness as compared to DL-DEF, we believe it offers significant promise for improvements and has larger potential to exploit daily anatomy to reduce the dose to OARs or improve target coverage. In contrast to DL-DEF, where the dose is simply deformed from the initial plan and therefore misses out the sometimes favorable anatomy changes, the DL-DL will predict a lower dose to the OARs and try to mimic this lower dose when these OARs appear further away from the target volume in the rCT. Besides, in addition to enlarging the training dataset, which can enhance the model's performance in various anatomical variations, refining the DM algorithm used for optimization could also increase the performance of the two DM-based strategies. In an example of how DM algorithms can guide optimization toward high-quality plans, Borderias et al. (20) compared a non-adapted planning strategy with fully automated OAPT strategies based on DM for patients with head and neck cancer. They found that all implemented

OAPT strategies provide closer dose volume histograms (DVH) metrics to the initial plan than the plan without adaptation. In this work, we use a similar DM model, consisting of a voxel-wise least-squared error minimization problem coupled with two conventional radiotherapy objective functions on the tumor volume, to guide the optimization process towards automatically generated dose distributions. However, the selection of this optimization model must be carried out as carefully as the selection of the dose generation technique (45). This gives further possibilities for our DM-based optimization strategies refinement and further investigation should be done before clinical implementation is possible.

This study presents some limitations. First, the strategies were implemented on repeated CT scans while cone-beam CT (CBCT) remains the preferred modality for daily imaging, due to its lower cost and compact nature. However, assuming that similar photon-machine mounted CBCT quality could be achieved for protons (46,47), we expect that a generalization of our study to CBCT is feasible. Another important limitation of this study is that we assume the use of a slow CT, where the clinician does not have the time to delineate the target on all phases before reconstructing the iCTV. Instead, we directly deform the iCTV from the average pCT, which may introduce substantial segmentation errors (48).

Additionally, the timings reported in this work, necessary to complete the full proposed online adaptive workflows, are significantly influenced by the chosen number of iterations. For the template-based optimization strategies, we have set 200 iterations following a clinical protocol. For the dose-mimicking based strategies, the iteration number was reduced to 60. This decision was informed by several trials and observation of iterations counts required to achieve the optimization cost plateau. However, a comprehensive investigation of different iteration counts for both optimization approaches is essential to fully validate the timing and dosimetric expectations of the proposed OAPT strategies. Besides, we could use a reduced scenarios set while optimizing a plan in an adaptive setting, by using only the extreme setup shifts. After recording the optimization timings with reduced scenario sets, we observed a median reduction time of 400 seconds for the template-based optimization strategies and 180 seconds for the DM-based optimization strategy, while still achieving comparable plan quality. However, this was not investigated further and clinical acceptability of those plans still needs to be evaluated.

To conclude, this work demonstrates the feasibility of integrating four distinct online adaptive proton therapy strategies within a commercially available treatment planning system. Through comprehensive time and dosimetric evaluations, we have provided valuable insights into the performance of these strategies. While our findings present high similarity metrics with manual contours for most structures, the dosimetric impact of automated delineation errors remains significant for the target volume. Hopefully, the availability of numerous alternative commercial solutions offers promise for enhancing or replacing our delineation approaches. We have also highlighted the reliability of the mimicked dose deformation technique, which presents a good trade-off between time and dosimetric error considerations. It accurately restores important clinical objectives while maintaining a median adaptation time of merely 10 minutes. The deformed dose followed by a dose-mimicking step effectively recovered most of the dosimetric objectives from the manual adaptation and offers the benefit of enhanced interpretability compared to a black box deep learning model. However, for patients presenting more complex anatomy variations or

favorable anatomical changes, we recommend using a population-based method such as our deep learning based dose prediction algorithm followed by dose mimicking.

**References**


1. Lomax AJ, Boehringer T, Coray A, Egger E, Goitein G, Grossmann M, et al. Intensity modulated proton therapy: A clinical example. Med Phys. 2001 Mar;28(3):317–24.
2. Sterpin E, Barragan A, Souris K, Lee JA. Planification de traitement robuste en protonthérapie. Cancer/Radiothérapie. 2016 Oct;20(6–7):523–9.
3. Thomas M, Defraene G, Lambrecht M, Deng W, Moons J, Nafteux P, et al. NTCP model for postoperative complications and one-year mortality after trimodality treatment in oesophageal cancer. Radiother Oncol. 2019 Dec 1;141:33–40.
4. Beukema JC, Kawaguchi Y, Sijtsema NM, Zhai TT, Langendijk JA, Van Dijk LV, et al. Can we safely reduce the radiation dose to the heart while compromising the dose to the lungs in oesophageal cancer patients? Radiother Oncol. 2020 Aug;149:222–7.
5. Lin SH, Merrell KW, Shen J, Verma V, Correa AM, Wang L, et al. Multi-institutional analysis of radiation modality use and postoperative outcomes of neoadjuvant chemoradiation for esophageal cancer. Radiother Oncol. 2017 Jun;123(3):376–81.
6. Paganetti H, Botas P, Sharp GC, Winey B. Adaptive proton therapy. Phys Med Biol. 2021 Nov;66(22):22TR01.
7. Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A. Online daily adaptive proton therapy. Br J Radiol. 2020 Mar;93(1107):20190594.
8. Isgum I, Staring M, Rutten A, Prokop M, Viergever MA, Van Ginneken B. Multi-Atlas-Based Segmentation With Local Decision Fusion—Application to Cardiac and Aortic Segmentation in CT Scans. IEEE Trans Med Imaging. 2009 Jul;28(7):1000–10.
9. Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat Oncol. 2013 Jun 26;8(1):154.
10. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021 Feb;18(2):203–11.
11. Matoska T, Patel M, Liu H, Beriwal S. Review of Deep Learning Based Autosegmentation for Clinical Target Volume: Current Status and Future Directions. Adv Radiat Oncol. 2024 May;9(5):101470.
12. Zhang F, Wang Q, Li H. Automatic Segmentation of the Gross Target Volume in Non-Small Cell Lung Cancer Using a Modified Version of ResNet. Technol Cancer Res Treat. 2020 Jan 1;19:1533033820947484.
13. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med Phys. 2017;44(12):6377–89.
14. Ma CY, Zhou JY, Xu XT, Guo J, Han MF, Gao YZ, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. J Appl Clin Med Phys. 2022;23(2):e13470.
15. Jin D, Guo D, Ho TY, Harrison AP, Xiao J, Tseng C kan, et al. DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med Image Anal. 2021 Feb 1;68:101909.
16. Thomas M, Mortensen HR, Hoffmann L, Møller DS, Troost EGC, Muijs CT, et al. Proposal for the delineation of neoadjuvant target volumes in oesophageal cancer. Radiother Oncol. 2021 Mar 1;156:102–12.
17. Smolders A, Choulilitsa E, Czerska K, Bizzocchi N, Krcek R, Lomax A, et al. Dosimetric comparison of autocontouring techniques for online adaptive proton therapy. Phys Med Biol. 2023 Aug;68(17):175006.
18. Archambault Y, Boylan C, Bullock D, Morgas T, Peltola J, Ruokokoski E, et al. Making On-line adaptive radiotherapy possible using artificial intelligence and machine learning

for efficient daily re-planning. Med Phys Int. 2020;8(2):77–86.
19. Petersson K, Nilsson P, Engström P, Knöös T, Ceberg C. Evaluation of dual-arc VMAT radiotherapy treatment plans automatically generated via dose mimicking. Acta Oncol. 2016 Apr 2;55(4):523–5.
20. Borderias-Villarroel E, Fredriksson A, Cvilic S, Di Perri D, Longton E, Pierrard J, et al. Dose mimicking based strategies for online adaptive proton therapy of head and neck cancer. Phys Med Biol. 2023 May 2;68(10).
21. Eriksson O, Zhang T. Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking. Med Phys. 2022;49(6):3564–73.
22. Borderias-Villarroel E, Huet Dastarac M, Barragán-Montero AM, Helander R, Holmstrom M, Geets X, et al. Machine learning-based automatic proton therapy planning: Impact of post-processing and dose-mimicking in plan robustness. Med Phys. 2023;50(7):4480–90.
23. 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 Oct;24(10):e14065.
24. Draguet C, Barragán-Montero AM, Vera MC, Thomas M, Populaire P, Defraene G, et al. Automated clinical decision support system with deep learning dose prediction and NTCP models to evaluate treatment complications in patients with esophageal cancer. Radiother Oncol. 2022 Nov;176:101–7.
25. University of Aarhus. PROton Versus Photon Therapy for Esophageal Cancer - a Trimodality Strategy (PROTECT) A Multicenter International Randomized Phase III Study of Neoadjuvant Proton Versus Photon Chemoradiotherapy in Locally Advanced Esophageal Cancer [Internet]. clinicaltrials.gov; 2022 Nov [cited 2023 May 24]. Report No.: NCT05055648. Available from: https://clinicaltrials.gov/ct2/show/NCT05055648
26. Weistrand O, Svensson S. The ANACONDA algorithm for deformable image registration in radiotherapy. Med Phys. 2015 Jan;42(1):40–53.
27. Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. Phys Med Biol. 2012 Jun 7;57(11):R99–117.
28. Visser S, O. Ribeiro C, Dieters M, Mul VE, Niezink AGH, Van Der Schaaf A, et al. Robustness assessment of clinical adaptive proton and photon radiotherapy for oesophageal cancer in the model-based approach. Radiother Oncol. 2022 Dec;177:197–204.
29. Van Herk M, Remeijer P, Rasch C, Lebesque JV. The probability of correct target dosage: dose-population histograms for deriving treatment margins in radiotherapy. Int J Radiat Oncol. 2000 Jul;47(4):1121–35.
30. Pukala J, Johnson PB, Shah AP, Langen KM, Bova FJ, Staton RJ, et al. Benchmarking of five commercial deformable image registration algorithms for head and neck patients. J Appl Clin Med Phys. 2016;17(3):25–40.
31. Kadoya N, Nakajima Y, Saito M, Miyabe Y, Kurooka M, Kito S, et al. Multi-institutional Validation Study of Commercially Available Deformable Image Registration Software for Thoracic Images. Int J Radiat Oncol. 2016 Oct 1;96(2):422–31.
32. Barragán-Montero AM, Nguyen D, Lu W, Lin MH, Norouzi-Kandalan R, Geets X, 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):3679–91.
33. Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiother Oncol. 2019 Jun 1;135:130–40.
34. Sherer MV, Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol. 2021 Jul;160:185–91.
35. Mackay K, Bernstein D, Glocker B, Kamnitsas K, Taylor A. A Review of the Metrics Used to Assess Auto-Contouring Systems in Radiotherapy. Clin Oncol. 2023 Jun

1;35(6):354–69.
36. Roberfroid B, Lee JA, Geets X, Sterpin E, Barragán-Montero AM. DIVE-ART: A tool to guide clinicians towards dosimetrically informed volume editions of automatically segmented volumes in adaptive radiation therapy. Radiother Oncol. 2024 Mar;192:110108.
37. Yeo UJ, Taylor ML, Supple JR, Smith RL, Dunn L, Kron T, et al. Is it sensible to "deform" dose? 3D experimental validation of dose-warping. Med Phys. 2012;39(8):5065–72.
38. Nobnop W, Chitapanarux I, Neamin H, Wanwilairat S, Lorvidhaya V, Sanghangthum T. Evaluation of deformable image registration (DIR) methods for dose accumulation in nasopharyngeal cancer patients during radiotherapy. Radiol Oncol. 2017 Sep 15;51(4):438–46.
39. Chetty IJ, Rosu-Bubulac M. Deformable Registration for Dose Accumulation. Semin Radiat Oncol. 2019 Jul 1;29(3):198–208.
40. Zhang L, Wang Z, Shi C, Long T, Xu XG. The impact of robustness of deformable image registration on contour propagation and dose accumulation for head and neck adaptive radiotherapy. J Appl Clin Med Phys. 2018 Jul;19(4):185–94.
41. Barker JL, Garden AS, Ang KK, O'Daniel JC, Wang H, Court LE, et al. Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system. Int J Radiat Oncol. 2004 Jul 15;59(4):960–70.
42. Robar JL, Day A, Clancey J, Kelly R, Yewondwossen M, Hollenhorst H, et al. Spatial and Dosimetric Variability of Organs at Risk in Head-and-Neck Intensity-Modulated Radiotherapy. Int J Radiat Oncol. 2007 Jul 15;68(4):1121–30.
43. Height R, Khoo V, Lawford C, Cox J, Joon DL, Rolfo A, et al. The dosimetric consequences of anatomic changes in head and neck radiotherapy patients. J Med Imaging Radiat Oncol. 2010;54(5):497–504.
44. Arsigny V, Commowick O, Pennec X, Ayache N. A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: Larsen R, Nielsen M, Sporring J, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. Berlin, Heidelberg: Springer; 2006. p. 924–31.
45. Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. The importance of evaluating the complete automated knowledge-based planning pipeline. Phys Med. 2020 Apr;72:73–9.
46. Landry G, Hua C ho. Current state and future applications of radiological image guidance for particle therapy. Med Phys. 2018;45(11):e1086–95.
47. Lustermans D, Fonseca GP, Taasti VT, van de Schoot A, Petit SF, van Elmpt WJC, et al. Image quality evaluation of a new high-performance ring-gantry cone-beam computed tomography imager. Phys Med Biol [Internet]. 2024 [cited 2024 Apr 30]; Available from: http://iopscience.iop.org/article/10.1088/1361-6560/ad3cb0
48. Nakamura M, Narita Y, Matsuo Y, Narabayashi M, Nakata M, Yano S, et al. Geometrical differences in target volumes between slow CT and 4D CT imaging in stereotactic body radiotherapy for lung tumors in the upper and middle lobe. Med Phys. 2008;35(9):4142–8.