Physics and Imaging in Radiation Oncology 28 (2023) 100494

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

# A geometry and dose-volume based performance monitoring of artificial intelligence models in radiotherapy treatment planning for prostate cancer

Geert De Kerf [a,*], Michaël Claessens [a,b], Fadoua Raouassi [a], Carole Mercier [a,b], Daan Stas [a,c], Piet Ost [a,b], Piet Dirix [a,b], Dirk Verellen [a,b]

[a] Department of Radiation Oncology, Iridium Netwerk, Wilrijk (Antwerp), Belgium
[b] Centre for Oncological Research (CORE), Integrated Personalized and Precision Oncology Network (IPPON), University of Antwerp, Antwerp, Belgium
[c] Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

ARTICLE INFO

*Keywords:*
Performance monitoring
Artificial intelligence
SBRT prostate
Deep Learning Segmentation
Deep Learning Planning
Clinical metrics

ABSTRACT

*Background and Purpose:* Clinical Artificial Intelligence (AI) implementations lack ground-truth when applied on real-world data. This study investigated how combined geometrical and dose-volume metrics can be used as performance monitoring tools to detect clinically relevant candidates for model retraining.
*Materials and Methods:* Fifty patients were analyzed for both AI-segmentation and planning. For AI-segmentation, geometrical (Standard Surface Dice 3 mm and Local Surface Dice 3 mm) and dose-volume based parameters were calculated for two organs (bladder and anorectum) to compare AI output against the clinically corrected structure. A Local Surface Dice was introduced to detect geometrical changes in the vicinity of the target volumes, while an Absolute Dose Difference (ADD) evaluation increased focus on dose-volume related changes. AI-planning performance was evaluated using clinical goal analysis in combination with volume and target overlap metrics.
*Results:* The Local Surface Dice reported equal or lower values compared to the Standard Surface Dice (anorectum: (0.93 ± 0.11) vs (0.98 ± 0.04); bladder: (0.97 ± 0.06) vs (0.98 ± 0.04)). The ADD metric showed a difference of (0.9 ± 0.8)Gy for the anorectum $D_{1cm^3}$. The bladder $D_{5cm^3}$ reported a difference of (0.7 ± 1.5)Gy. Mandatory clinical goals were fulfilled in 90 % of the DLP plans.
*Conclusions:* Combining dose-volume and geometrical metrics allowed detection of clinically relevant changes, applied to both auto-segmentation and auto-planning output and the Local Surface Dice was more sensitive to local changes compared to the Standard Surface Dice. This monitoring is able to evaluate AI behavior in clinical practice and allows candidate selection for active learning.

## 1. Introduction

Every step in the radiation therapy (RT) workflow has been exposed to Artificial Intelligence (AI) solutions and the benefits have been demonstrated [1–4]. However, most studies report on retrospective or simulated evaluations, which does not represent a clinical setting [5].

To date, auto-segmentation tools contour a variety of structures on both CT and MR with high accuracy [6–10]. Some papers even report feasibility of treatment plan creation based on AI generated structures [11]. However, AI-segmented regions of interests (ROIs) close to the target volume still need verification [12,13]. Small user-adjustments also remain necessary to ensure quality and guideline compliance, mainly in case of changing guidelines or image acquisition protocols [14].

Automated treatment plan generation in general improves planning efficiency and reduces plan quality variability [15]. Different approaches are available and able to create treatment plans for a variety of pathologies and dose prescriptions [16–19]. Despite these benefits, the main challenges remain safe implementation, ongoing maintenance of an AI model and adaptation to changing workflows and procedures [20]. Stereotactic body radiation therapy (SBRT) treatments, delivering high doses in a limited number of fractions, require even more stringent quality assurance (QA) measures [21].

Guidelines for safe AI implementation and QA were published [22] and the success rate of AI in RT will largely depend on the interpretability and data-model dependency [23]. Also, data standardization and

* Corresponding author.
*E-mail address:* geert.dekerf@gza.be (G. De Kerf).

https://doi.org/10.1016/j.phro.2023.100494
Received 6 June 2023; Received in revised form 20 September 2023; Accepted 20 September 2023
Available online 23 September 2023

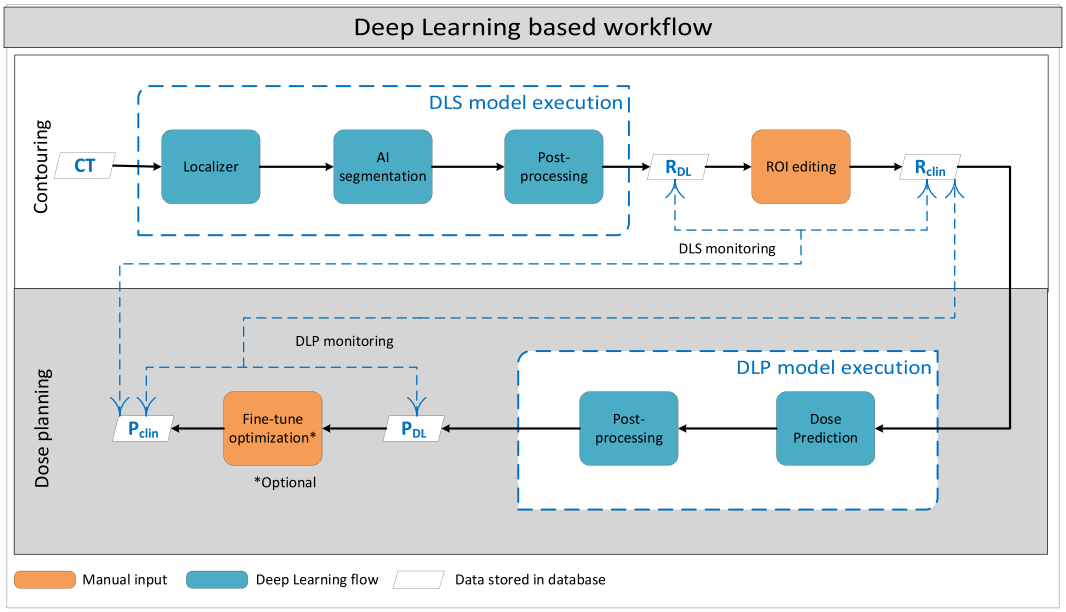


**Fig. 1.** Schematic overview of a deep learning guided contouring and dose planning workflow. The input is the CT image on which the DLS model generates AI contours $R_{DL}$, followed by a radiation oncologist's review resulting in contours $R_{clin}$. Based on $R_{clin}$, the DLP model creates plan $P_{DL}$ which can be further optimized by a planner leading to plan $P_{clin}$ within the fine-tune optimization step. All datasets ($R_{DL}$, $R_{clin}$, $P_{DL}$, $P_{clin}$) are stored in the database and were used for DLS and DLP monitoring.

integration into existing clinical workflows remains a challenge [24]. Validation and commissioning reports help documenting intended use and limitations of a model. Reviewing automated plans using checklists will improve patient safety as well [25]. However, automated and independent QA of the AI models is preferable as more efficiency in the radiotherapy workflow is needed [26]. Uncertainty is embedded in AI models and continuously monitoring the behavior on unseen data can make AI even more efficient by guiding users towards cases that need attention [27].

Continuously monitoring AI output in clinical routine benefits from the large number of patients, but also requires new, combined metrics, both geometrical and dose-volume related, to detect the clinical relevance of anomalous output and to allow the selection of appropriate datasets for further optimization of AI models within the framework of expert-augmented machine learning [28].

This study investigated how problem-specific clinical knowledge from experts can be automatically extracted from an AI supported workflow using a combined analysis of both geometrical and dose-volume metrics. Performance of both segmentation and planning models, clinically implemented for SBRT prostate treatments, was evaluated. In addition, metrics with specific thresholds were proposed to detect outliers for model retraining.

## 2. Material and Methods

### 2.1. Patient data

Fifty SBRT prostate cancer patients were simulated on a Brilliance - Big Bore (Philips, The Netherlands) or a SOMATOM Go-Sim CT scanner (Siemens, Germany). All auto-segmentations and automated treatment plans were created in our treatment planning system (TPS) (RayStation 11B, RaySearch Laboratories AB, Sweden) and all AI models were commercially available and trained, validated and tested by the company. The AI-segmentation model was initially trained on local data of the Iridium Netwerk (Belgium), and data of the University Health Network (UHN) (Canada) was used for AI-planning model training. All patients were treated at the Iridium Netwerk following the PACE Trial guidelines (ClinicalTrials.gov Identifier: NCT01584258) in five fractions, received a dose of 7.25 Gy per fraction to the PTV, which was a uniform expansion of the CTV of 5 mm, except in the posterior direction where a margin of 3 mm was used. The training dataset of UHN was based on a uniform PTV expansion of 7 mm. Due to de-identified nature of the data, the need for ethics review was waived for this retrospective study.

### 2.2. Clinical AI models

The Deep Learning Segmentation (DLS) model (RSL Male Pelvic CT (v1.0.0), RaySearch Laboratories AB, Sweden) segmented five different organs (prostate (as CTV), anorectum, bladder and left and right femoral head) and used the simulation CT as input and outputted deep learning ROIs ($R_{DL}$) based on the ACROP and RTOG guidelines [29,30]. In clinical routine, contours were reviewed by a radiation oncologist (RO) and corrected when needed, providing planners with clinically approved ROIs ($R_{clin}$) (Fig. 1). The CTV generated by the model differed from the clinically used CTV on two accounts: inclusion of seminal vesicles and the use of MRI. As described in the PACE Trial guidelines, part of the seminal vesicles will be included in the CTV, depending on the patient's risk group (low, intermediate or high). Additionally, MRI imaging was used in clinical routine to determine the anatomical borders of the prostate, which impacts the CTV contour. Therefore, the CTV contour was not further evaluated.

The Deep Learning Planning (DLP) model configuration (RSL-Prostate-3625-SBRT (v3.0.0), RaySearch Laboratories AB, Sweden) consisted of a two-step approach: prediction/mimicking of a 3D-dose distribution based on input ROIs, followed by a post-processing optimization. The latter facilitated model sharing between clinics as prediction/mimicking highly relies on the training dataset, the post-processing allowed additional output tuning based on local, clinical needs. A deep learning generated plan ($P_{DL}$) was the result of running the model (Fig. 1). An optional 'Fine-tune optimization' step allowed manual tweaking of the treatment plan to fulfill most of the patient specific requirements, if appropriate. This final plan was the clinical plan ($P_{clin}$) which was used for treatment delivery.

Prior to clinical use, DLS and DLP model commissioning was performed on five patients. For DLS, all $R_{DL}$'s were evaluated by an

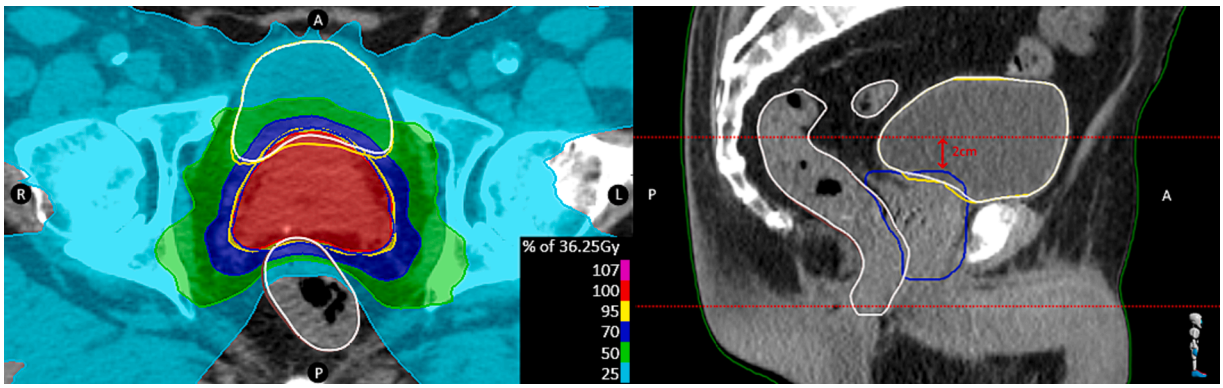


**Fig. 2.** Typical output of DLS and DLP models. White, DLS generated ROIs are plotted next to the clinical correct yellow bladder and brown anorectum. On the sagittal plane, the dashed red lines mark the boundaries for the Local Surface Dice calculation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Mandatory and optimal ROI dose constraints for SBRT.

| | Dose volume constraints | | |
|---|---|---|---|
| | Dose (Gy) | Minimum volume (%) | |
| Target | | Mandatory | Optimal |
| PTV | 36.25 | 90 | 95 |
| | Dose (Gy) | Maximum volume (% or $cm^3$) | |
| OAR | | Mandatory | Optimal |
| Rectum | 18.1 | 50 % | |
| | 29.0 | 20 % | |
| | 36.0 | 2 $cm^3$ | 1 $cm^3$ |
| Bladder | 18.1 | 40 % | |
| | 37.0 | 10 $cm^3$ | 5 $cm^3$ |
| Femoral heads | 14.5 | – | 5 % |
| Bowel | 18.1 | 5 $cm^3$ | |
| | 30.0 | 1 $cm^3$ | |
| Penile bulb | 29.5 | – | 50 % |
| Urethra | 42.0 | – | 50 % |

experienced radiation oncologist (RO) and scored based on quality and timesaving. For DLP, the model's output $P_{DL}$ was retrospectively compared against manually created, clinically used plans and quality was scored using the Plan Quality Index (PQI) [31]. To assess treatment machine deliverability, patient specific plan QA was performed using the SunCHECK platform (Sun Nuclear Corporation, Melbourne, USA).

### 2.3. Auto-segmentation performance monitoring

Continuous monitoring compared $R_{DL}$ with $R_{clin}$ using both geometrical and dose-volume parameters on bladder (#50) and anorectum (#50) ROIs. The geometrical performance of the model was assessed via the Standard Surface Dice 3 mm and a new Local Surface Dice 3 mm. Both metrics were calculated based on following definition of a Surface Dice with a tolerance of 3 mm:

$$Surface\_Dice(S_{DLS}, S_{clin}, 3mm) = \frac{|S_{DLS} \cap B^{3mm}_{clin}| + |S_{clin} \cap B^{3mm}_{DLS}|}{|S_{DLS}| + |S_{clin}|}$$ where $S_{DLS}$ and $S_{clin}$ were the surfaces of $R_{DL}$ and $R_{clin}$ and $B^{3mm}_{DLS}$ and $B^{3mm}_{clin}$ were the tolerance regions, 3 mm in- and outside the corresponding surface $S$ [32]. The Standard Surface Dice used the entire ROI volumes as input for $S_{DLS}$ and $S_{clin}$. On the other hand, the Local Surface Dice only evaluated differences between $S'_{DLS}$ and $S'_{clin}$ where $S'$ is the surface $S$ cropped at the transversal slices 2 cm away from the target, as visualized in the sagittal plane of Fig. 2 and this distance was chosen based on used thresholds in the online adaptive setting [33,34]. The Standard Surface Dice already provided clinically relevant and quantitative measures [35,36], the local variation further increased the focus on the clinically relevant changes in the vicinity of the target. Dose-volume impact of the applied changes to $R_{DL}$ was investigated by evaluating different dose volume constraints for both $R_{DL}$ and $R_{clin}$ on the clinically approved dose distribution $P_{clin}$. The absolute dose difference (ADD) at different clinically relevant dose levels, as listed in Table 1, were reported next to the average dose ($D_{average}$) and the near maximum dose ($D_{0.03cm^3}$). For every dose volume constraint, the absolute dose differences ADD = $|D_{clin} - D_{DLS}|$ were calculated as a change in a contour might have an impact on the reported dose value. Segmented organs were categorized into four groups, based on an ADD of 100 cGy and a mean Local Surface Dice minus one Standard Deviation (SD). Zone 1 resembled patients with a high Local Dice score and an ADD less than 100 cGy, while in Zone 2 the reported ADD became larger than 100 cGy. Zone 3 contains patients with a Local Dice score that was one SD less than the mean Local Dice. In Zone 4 both reported values showed larger deviations. Patient with hip prostheses were marked with a red border.

### 2.4. Auto-planning performance monitoring

In clinical routine, $P_{DL}$ was autonomously created via the DLP process and a consistent shape of the dose distribution (Fig. 2) was obtained for all fifty patients. $P_{clin}$ underwent minor adaptations from scaling to the dose prescription to further optimize the plan for better fulfilling the clinical goals or finding a different balance between OAR sparing and target coverage. To quantify the differences between $P_{DL}$ and $P_{clin}$, clinical goals of Table 1 were evaluated on both dose distributions against $R_{clin}$ volumes. Additionally, spatial information, such as ROI volumes and overlap ratios between ROI and PTV, were reported as the dose prediction method uses ROIs as input to the model. Geometrical metrics related to the PTV margin are standardized using:

$$x' = \frac{x - \text{average}(X)}{\text{stdev}(X)}$$

with $x$ the PTV volumes, or the overlap volume between the PTV and an organ, and $X$ the corresponding values of all fifty plans. The applied scaling allowed comparing volume-based metrics between the training set and the local data in case different PTV margins were used. All these parameters were calculated for both the autonomously created $P_{DL}$'s and clinically approved $P_{clin}$'s.

Target volume reported standardized PTV volumes because of the difference in PTV margin and the standardized overlap volume considered the intersection with both bladder and anorectum. Model's performance was classified based on the optimal clinical goal analysis for both $P_{DL}$ and $P_{clin}$ and four different groups could be distinguished: green dots resemble patient ROIs for which $P_{DL}$ did achieve the optimal dose constraint and not further optimization of $P_{clin}$ was needed ($P_{DL}$ was selected for treatment). In case $P_{DL}$ did not pass the clinical goal, but $P_{clin}$ did, the plan was marked orange. Patient ROIs for which both $P_{DL}$

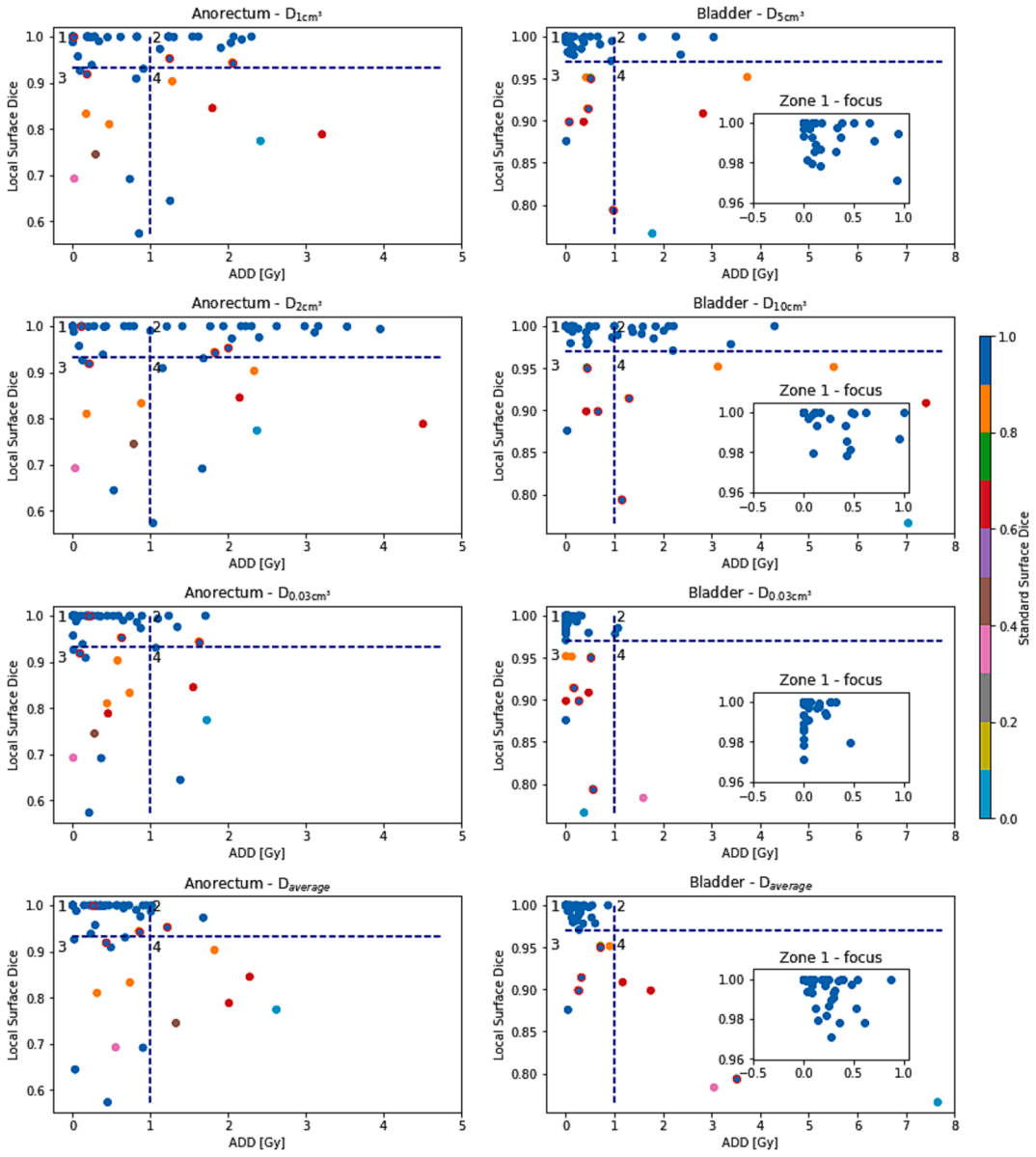


**Fig. 3.** On the x-axis, Absolute Dose Difference (ADD) for anorectum and bladder were reported for different critical volumes. On the y-axis, the Local Surface Dice is plotted and the color of each dot resembles the Standard Surface Dice as shown by the color bar. The dashed lines resemble the ADD of 1 Gy and the first standard deviation of the Local Surface Dice. Each graph is divided into four zones, reflecting perfect DLS segmentations in zone 1, minor corrected DLS ROIs with large dose-volume impact (zone 2), major corrected DLS ROIs outside the high dose region and less clinical impact in zone 3 and zone 4 shows major corrections with large difference in ADD.

and $P_{clin}$ were not able to fulfill the most stringent clinical goal were marked as red. Finally, if a different compromise between organ sparing and target coverage was preferred a dot can turn black in case $P_{DL}$ did pass the clinical goal, but $P_{clin}$ did not.

To compare reported means, a T-test for two independent samples was performed in Python 3.9 using the SciPy package.

## 3. Results

### 3.1. Model commissioning

The commissioning phase on five patients revealed necessity of no (cf. femoral heads) to minor (cf. bladder, anorectum) corrections with high timesaving for DLS. For DLP, a significantly ($p < 0.001$) better sparing of the anorectum (V18.1 Gy = (23.9 ± 7)% for the manual plan and (15.3 ± 6)% for $P_{DL}$) is reflected in a better PQI for $P_{DL}$ (65.4 %) compared to manual planning (74.4 %). No other significant clinical goals differences were detected. In general, the number of Monitor Units increases when comparing manual plans against $P_{DL}$: (1963 ± 192) and (2680 ± 200) ($p < 0.001$) respectively. The gamma pass rate (3 %2mm) remained identical: (100 ± 0.1)% for manual plans and (99.9 ± 0.2)% for $P_{DL}$.

### 3.2. Auto-segmentation – Performance monitoring

The Standard Surface Dice 3 mm and the Local Surface Dice 3 mm reported values of (0.98 ± 0.04) and (0.93 ± 0.11) ($p = 0.006$) respectively for the anorectum, and (0.98 ± 0.04) and (0.97 ± 0.06) ($p = 0.3$) for bladder. The anorectum was evaluated against four different ADD's: $D_{0.03cm^3}$ (0.5 ± 0.5)Gy, $D_{1cm^3}$ (0.9 ± 0.8)Gy, $D_{2cm^3}$ (1.3 ± 1.2)Gy and $D_{average}$ (0.6 ± 0.6)Gy ($p < 0.001$). Bladder reported ADD's of (0.2 ± 0.3)Gy,(0.7 ± 1.5)Gy, (1.5 ± 2.3)Gy and (0.6 ± 1.2)Gy ($p < 0.001$)

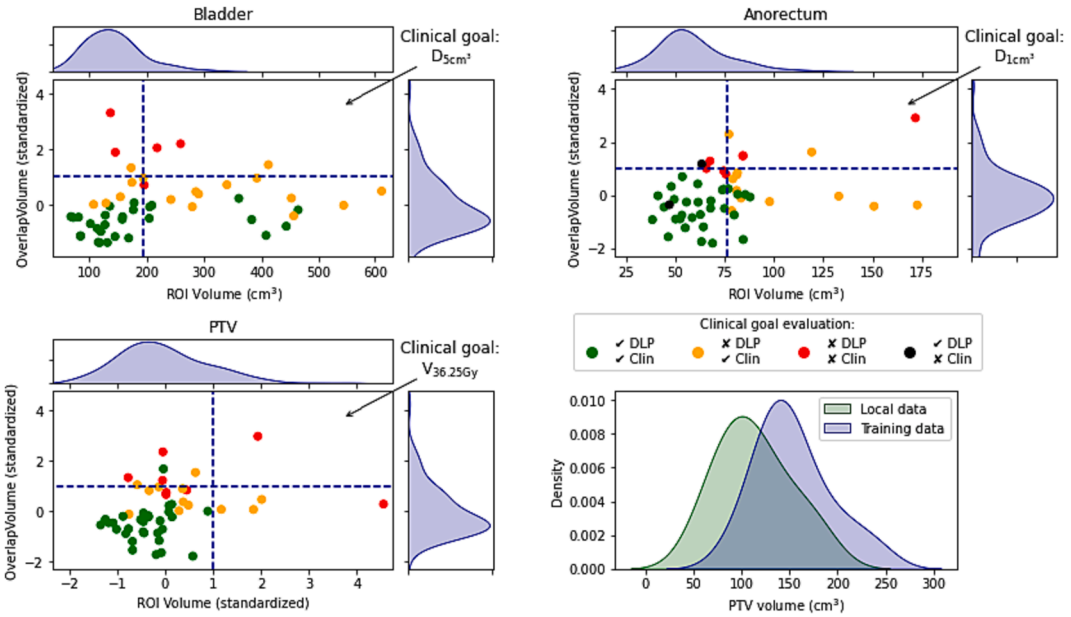


**Fig. 4.** Clinical goal analysis comparing DLP output against the clinical plan for both bladder ($D5cm^3 < 37$ Gy), anorectum ($D1cm^3 < 36$ Gy) and PTV ($V36.25\ Gy > 95$ %) and every colored dot resembles a patient ROI belonging to one of the four clinical goal evaluation groups. The blue distribution plots above and to the right of each scatter plot show the corresponding data of the training data. The dashed blue lines show the first standard deviation of this training data. The lower left graph visualizes the difference in PTV volume between local and training data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for respectively $D_{0.03cm^3}$, $D_{5cm^3}$, $D_{10cm^3}$ and $D_{average}$. All information is visualized in Fig. 3.

### 3.3. Auto-planning – Performance monitoring

Clinical goal evaluation revealed that $P_{DL}$ did fulfill all mandatory clinical goals (Table 1) in 90 % of the cases, while all where fulfilled in $P_{clin}$. The success rate dropped to 32 % and 74 % respectively for $P_{DL}$ and $P_{clin}$ when the optimal dose constraints were evaluated. The geometrical analysis showed significant volume differences between the training and the local dataset for respectively PTV ($(154 \pm 45)cm^3$; $(120 \pm 50)cm^3$ with $p < 0.001$), bladder ($(141 \pm 51)cm^3$; $(226 \pm 136)cm^3$ with $p < 0.001$) and rectum ($(58 \pm 18)cm^3$; $(76 \pm 29)cm^3$ with $p < 0.001$).

Fig. 4 also reveals clinical data points that were outside the first SD of the training data.

## 4. Discussion

Model commissioning proved its advantage with respect to both time and quality of models' output. Nevertheless, the results largely depended on the quality and variety of the commissioning test set and even large retrospective studies are not representative and not able to provide a full estimate of the models' clinical behavior [5]. Also, retrospective analysis often compared against ground truth datasets for both segmentation and planning models [37–39]. So, proper performance monitoring tools are appropriate and must deal with variations inherent to a clinical workflow such as, due to timesaving, mainly relevant segmentation errors will be corrected in the vicinity of the target volumes as this will impact the dose distribution and the lack of ground truth data [12]. For DLP, different PTV margins compromise direct comparison between training and local datasets. Despite the feasibility to automatically create deliverable treatment plans, detecting sub-optimal created plans remains a challenge [40].

DLS performance monitoring must be able to detect local and certainly small changes with a large potential impact on the dose distribution. Fig. 3 shows that the Local Surface Dice is more sensitive compared to the Standard Surface Dice in cases where the Standard Dice is high (>0.9), but the Local Dice score is low (<0.7) and even significantly different in case of rectum analysis. Additionally, the ADD improves revealing cases with a high Local Surface Dice (>1SD) but with large (>1Gy) difference in reported dose. Literature already showed the dose-volume impact, but either a new plan was optimized, introducing additional uncertainty [13], or only ROIs were reported further away from high doses [41]. By comparing both $R_{DL}$ and $R_{clin}$ to the approved dose distribution of $P_{clin}$, no plan optimization bias was introduced. The combination of both geometrical and dose-volume metrics proves its advantage as the Local Surface Dice is sensitive to detect large deviations close to the target volumes (Fig. 3, zone 3 and 4) and the ADD can detect small changes in the vicinity of the high dose region (Fig. 3, zone 2). The bladder in Fig. 2 is a zone 2 example. Deviations in zone 3 can be categorized as clinically less relevant with no major impact on the dose distribution, while the opposite is true in zone 4. Fig. 3 also shows that $D_{0.03cm^3}$ and $D_{average}$ are more robust as a dose-volume metric compared to the $D_{1cm^3}$, $D_{2cm^3}$, $D_{5cm^3}$ or $D_{10cm^3}$ as the mean ADD is higher for the latter and differ significantly. Too small volumes ($D_{0.03cm^3}$) are not sensitive enough in case a ROI shifts in a homogenous dose distribution. Further, average doses report on too large volumes, which decreases sensitivity as well. In addition to previous papers, a minimal critical volume ($>1cm^3$) is recommended in case of dose-volume analysis of AI segmented contours. A different dose-volume analysis can be obtained by creating a treatment plan directly on $R_{DL}$ to investigate how DLS does impact the dose distribution. As this approach inherently increases the user interaction, we currently only evaluated the ADD parameter.

Femoral head and prostate segmentations were not reported. The former was never adapted manually, leading to a Standard Surface Dice score of 1 for all femoral heads, both left and right. The latter was impacted by the local delineation guideline of the prostate contour and the superposition of Gaussian distributions when plotting the clinical PTV volume data also reveals the different, risk group dependent, PTV definitions in Fig. 4. The different PTV margin necessitated data standardization, the risk-group dependent delineations made it impossible to monitor the quality of the prostate segmentation.

DLP monitoring showed a lower mean PTV volume in the clinical data compared to the training data, which originates from the smaller

PTV margin and it impacts the target-ROI overlap volume. As $P_{DL}$ did fulfill almost all mandatory clinical goals, comparison of the plan quality between $P_{clin}$ and $P_{dlp}$ did focus on the fulfillment of the optimal dose constraints. To date, dose normalization can differ between $P_{clin}$ and $P_{dlp}$ which may impact the dose comparison, but has the advantage that $P_{dlp}$ is created fully autonomously. Fig. 4 shows the impact of both organ and overlap volume on the clinical goal fulfillment. Larger overlap volumes between OAR and target lead to a more complex plan optimization problem as reflected by the red dots for which even $P_{clin}$ could not fulfill the clinical goal. This increased complexity dominates the incapability of automatically creating a clinically acceptable $P_{dlp}$ and user interaction is needed to find the optimal balance between organ sparing and target coverage. It is unclear if adding these plans to the training dataset will improve model's output. In case of increasing ROI volume, the dose prediction might be impacted as large ROI volumes are not represented in the training dataset and most of the orange dots have a volume that exceeds the first SD of the training dataset. Consequently, these plans are interesting candidates to add to the training dataset to improve performance and robustness of the model. Volume differences between local data and training data might be introduced by different bladder and anorectum preparations between the different centers. This observation strengthens the need for OOD detection to better understand the behavior of AI models and at the same time the robustness of the dose prediction is proven as DLP hardly generates plans causing major violations.

In conclusion, combining geometrical and dose-volume metrics is of added value when monitoring the performance of both AI based segmentation and planning models. For segmentation, the Local Surface Dice is more sensitive compared to the Standard Surface Dice and detection of changes in the vicinity of the target volume can further improve in combination with ADD analysis if a critical volume of at least 1 $cm^3$ is used. For planning, reporting volumes in combination with clinical goal analysis enables OOD detection. Future work might explore analyzing dose-volume metrics of plans created directly on $R_{DL}$ in an automated way as well as improving the comparison between $P_{dlp}$ and $P_{clin}$ by taking into account the dose normalization.

## CRediT authorship contribution statement

**Geert De Kerf:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Michaël Claessens:** Conceptualization, Writing – original draft, Writing – review & editing. **Fadoua Raouassi:** Conceptualization. **Carole Mercier:** Validation, Writing – review & editing. **Daan Stas:** Validation, Writing – review & editing. **Piet Ost:** Validation, Writing – review & editing. **Piet Dirix:** Validation, Writing – review & editing. **Dirk Verellen:** Writing – review & editing, Supervision.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] Boon I, Au Yong T, Boon C. Assessing the role of artificial intelligence (AI) in clinical oncology: utility of machine learning in radiotherapy target volume delineation. Medicines 2018;5:131. https://doi.org/10.3390/medicines5040131.

[2] Wang C, Zhu X, Hong JC, Zheng D. Artificial intelligence in radiotherapy treatment planning: present and future. Technol Cancer Res Treat 2019;18:1–11. https://doi.org/10.1177/1533033819873922.

[3] Bijman R, Sharfo AW, Rossi L, Breedveld S, Heijmen B. Pre-clinical validation of a novel system for fully-automated treatment planning. Radiother Oncol 2021;158:253–61. https://doi.org/10.1016/j.radonc.2021.03.003.

[4] Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. Phys Imaging Radiat Oncol 2021;19:13–24. https://doi.org/10.1016/j.phro.2021.05.007.

[5] McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005. https://doi.org/10.1038/s41591-021-01359-w.

[6] Savenije MHF, Maspero M, Sikkes GG, Der Voort V, Van Zyp JRN, Alexis AN, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiat Oncol 2020;15:1–12. https://doi.org/10.1186/s13014-020-01528-0.

[7] Cha E, Elguindi S, Onochie I, Gorovets D, Deasy JO, Zelefsky M, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiother Oncol 2021;159:1–7. https://doi.org/10.1016/j.radonc.2021.02.040.

[8] Wong J, Huang V, Wells D, Giambattista J, Giambattista J, Kolbeck C, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. Radiat Oncol 2021;16:1–10. https://doi.org/10.1186/s13014-021-01831-4.

[9] Robert C, Munoz A, Moreau D, Mazurier J, Sidorski G, Gasnier A, et al. Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers. Cancer/Radiotherapie 2021;25:607–16. https://doi.org/10.1016/j.canrad.2021.06.023.

[10] Almeida G, Tavares JMRS. Deep Learning in Radiation Oncology Treatment Planning for Prostate Cancer: A Systematic Review. J Med Syst 2020:44. https://doi.org/10.1007/s10916-020-01641-3.

[11] Duan J, Bernard M, Downes L, Willows B, Feng X, Mourad WF, et al. Evaluating the clinical acceptability of deep learning contours of prostate and organs-at-risk in an automated prostate treatment planning process. Med Phys 2022;49:2570–81. https://doi.org/10.1002/mp.15525.

[12] Vaassen F, Hazelaar C, Canters R, Peeters S, Petit S, van Elmpt W. The impact of organ-at-risk contour variations on automatically generated treatment plans for NSCLC. Radiother Oncol 2021;163:136–42. https://doi.org/10.1016/j.radonc.2021.08.014.

[13] Johnston N, De Rycke J, Lievens Y, van Eijkeren M, Aelterman J, Vandersmissen E, et al. Dose-volume-based evaluation of convolutional neural network-based auto-segmentation of thoracic organs at risk. Phys Imaging Radiat Oncol 2022;23:109–17. https://doi.org/10.1016/j.phro.2022.07.004.

[14] Vaassen F, Boukerroui D, Looney P, Canters R, Verhoeven K, Peeters S, et al. Real-world analysis of manual editing of deep learning contouring in the thorax region. Phys Imaging Radiat Oncol 2022;22:104–10. https://doi.org/10.1016/j.phro.2022.04.008.

[15] Panettieri V, Ball D, Chapman A, Cristofaro N, Gawthrop J, Griffin P, et al. Development of a multicentre automated model to reduce planning variability in radiotherapy of prostate cancer. Phys Imaging Radiat Oncol 2019;11:34–40. https://doi.org/10.1016/j.phro.2019.07.005.

[16] McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. Phys Med Biol 2017;62:5926–44. https://doi.org/10.1088/1361-6560/aa71f8.

[17] Smith A, Granatowicz A, Stoltenberg C, Wang S, Liang X, Enke CA, et al. Can the Student Outperform the Master? A Plan Comparison Between Pinnacle Auto-Planning and Eclipse knowledge-Based RapidPlan Following a Prostate-Bed Plan Competition. Technol Cancer Res Treat 2019;18:1–8. https://doi.org/10.1177/1533033819851763.

[18] De Roover R, Crijns W, Poels K, Dewit B, Draulans C, Haustermans K, et al. Automated treatment planning of prostate stereotactic body radiotherapy with focal boosting on a fast-rotating O-ring linac: Plan quality comparison with C-arm linacs. J Appl Clin Med Phys 2021;22:59–72. https://doi.org/10.1002/acm2.13345.

[19] Nawa K, Haga A, Nomoto A, Sarmiento RA, Shiraishi K, Yamashita H, et al. Evaluation of a commercial automatic treatment planning system for prostate cancers. Med Dosim 2017;42:203–9. https://doi.org/10.1016/j.meddos.2017.03.004.

[20] Moore KL. Automated Radiotherapy Treatment Planning. Semin Radiat Oncol 2019;29:209–18. https://doi.org/10.1016/j.semradonc.2019.02.003.

[21] Mancosu P, Lambri N, Castiglioni I, Dei D, Iori M, Loiacono D, et al. Applications of artificial intelligence in stereotactic body radiation therapy. Phys Med Biol 2022;67:16TR01. https://doi.org/10.1088/1361-6560/ac7e18.

[22] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[23] Barragan-Montero A, Bibal A, Dastarac MH, Draguet C, Valdes G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Phys Med Biol 2022:67. https://doi.org/10.1088/1361-6560/ac678a.

[24] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6. https://doi.org/10.1038/s41591-018-0307-0.The.

[25] Nealon KA, Court LE, Douglas RJ, Zhang L, Han EY. Development and validation of a checklist for use with automatically generated radiotherapy plans. J Appl Clin Med Phys 2022:1–7. https://doi.org/10.1002/acm2.13694.

[26] Claessens M, Vanreusel V, De Kerf G, Mollaert I, Lofman F, Gooding MJ, et al. Machine learning-based detection of aberrant deep learning segmentations of target and organs at risk for prostate radiotherapy using a secondary segmentation algorithm. Phys Med Biol 2022:67. https://doi.org/10.1088/1361-6560/ac6fad.

[27] van Rooij W, Verbakel WF, Slotman BJ, Dahele M. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. Adv. Radiat Oncol 2021:6. https://doi.org/10.1016/j.adro.2021.100658.
[28] Gennatas ED, Friedman JH, Ungar LH, Pirracchio R, Eaton E, Reichmann LG, et al. Expert-augmented machine learning. PNAS 2020;117:4571–7. https://doi.org/10.1073/pnas.1906831117.
[29] Salembier C, Villeirs G, De Bari B, Hoskin P, Pieters BR, Van Vulpen M, et al. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. Radiother Oncol 2018;127:49–61. https://doi.org/10.1016/j.radonc.2018.01.014.
[30] Gay H, Barthold H, O'Meara E, Bosch W, El Naqa I, Al-Lozi R, et al. Pelvic Normal Tissue Contouring Guidelines for Radiation Therapy: A Radiation Therapy Oncology Group Consensus Panel Atlas. Int J Radiat Oncol Biol Phys 2012:83. https://doi.org/10.1016/j.ijrobp.2012.01.023.Pelvic.
[31] Leung LHT, Kan MWK, Cheng ACK, Wong WKH, Yau CC. A new dose-volume-based Plan Quality Index for IMRT plan comparison. Radiother Oncol 2007;85:407–17. https://doi.org/10.1016/j.radonc.2007.10.018.
[32] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, de Fauw J, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. J Med Internet Res 2021:23. https://doi.org/10.2196/26151.
[33] Bohoudi O, Bruynzeel AME, Senan S, Cuijpers JP, Slotman BJ, Lagerwaard FJ, et al. Fast and robust online adaptive planning in stereotactic MR-guided adaptive radiation therapy (SMART) for pancreatic cancer. Radiother Oncol 2017;125:439–44. https://doi.org/10.1016/j.radonc.2017.07.028.
[34] Lamb J, Cao M, Kishan A, Agazaryan N, Thomas DH, Shaverdian N, et al. Online Adaptive Radiation Therapy: Implementation of a New Process of Care. Cureus 2017:9. https://doi.org/10.7759/cureus.1618.
[35] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.
[36] Kiser KJ, Barman A, Stieb S, Fuller CD, Giancardo L. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. J Digit Imaging 2021;34:541–53. https://doi.org/10.1007/s10278-021-00460-3.
[37] Kiljunen T, Akram S, Niemelä J, Löyttyniemi E, Seppälä J, Heikkilä J, et al. A deep learning-based automated CT segmentation of prostate cancer anatomy for radiation therapy planning-a retrospective multicenter study. Diagnostics 2020;10:959. https://doi.org/10.3390/diagnostics10110959.
[38] Kusters M, Miki K, Bouwmans L, Bzdusek K, van Kollenburg P, Smeenk RJ, et al. Evaluation of two independent dose prediction methods to personalize the automated radiotherapy planning process for prostate cancer. Phys Imaging Radiat Oncol 2022;21:24–9. https://doi.org/10.1016/j.phro.2022.01.006.
[39] Lempart M, Benedek H, Nilsson M, Eliasson N, Bäck S, Munck af Rosenschöld P, et al. Volumetric modulated arc therapy dose prediction and deliverable treatment plan generation for prostate cancer patients using a densely connected deep learning model. Phys Imaging. Radiat Oncol 2021;19:112–9. https://doi.org/10.1016/j.phro.2021.07.008.
[40] Wortel G, Eekhout D, Lamers E, van der Bel R, Kiers K, Wiersma T, et al. Characterization of automatic treatment planning approaches in radiotherapy. Phys Imaging Radiat Oncol 2021;19:60–5. https://doi.org/10.1016/j.phro.2021.07.003.
[41] Zhu J, Chen X, Yang B, Bi N, Zhang T, Men K, et al. Evaluation of Automatic Segmentation Model With Dosimetric Metrics for Radiotherapy of Esophageal Cancer. Front Oncol 2020;10:1–9. https://doi.org/10.3389/fonc.2020.564737.