Physics and Imaging in Radiation Oncology 32 (2024) 100646

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

Autodelineation methods in a simulated fully automated proton therapy workflow for esophageal cancer

Pieter Populaire [a,b], Beatrice Marini [c,d], Kenneth Poels [b], Stina Svensson [e], Edmond Sterpin [a,f], Albin Fredriksson [e], Karin Haustermans [a,b,*]

[a] KU Leuven, Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium
[b] University Hospital Leuven, Department of Radiation Oncology, Leuven, Belgium
[c] Humanitas University, Department of Biomedical Sciences, Milan, Italy
[d] Humanitas Research Hospital IRCCS, Department of Radiotherapy and Radiosurgery, Milan, Italy
[e] RaySearch Laboratories AB, Stockholm, Sweden
[f] Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, UCLouvain, Belgium

ARTICLE INFO

Keywords:
Contouring
Adaptive radiotherapy
Treatment planning
Artificial intellegence
Proton therapy
Oesophageal cancer

ABSTRACT

Background and purpose: Proton Online Adaptive RadioTherapy (ProtOnART) harnesses the dosimetric advantage of protons and immediately acts upon anatomical changes. Here, we simulate the clinical application of delineation and planning within a ProtOnART-workflow for esophageal cancer. We aim to identify the most appropriate technique for autodelineation and evaluate full automation by replanning on autodelineated contours.
Materials and methods: We evaluated 15 patients who started treatment between 11-2022 and 01-2024, undergoing baseline and three repeat computed tomography (CT) scans in treatment position. Quantitative and qualitative evaluations compared different autodelineation methods. For Organs-at-risk (OAR) deep learning segmentation (DLS), rigid and deformable propagation from baseline to repeat CT-scans were considered. For the clinical target volume (CTV), rigid and three deformable propagation methods (default, heart as controlling structure and with focus region) were evaluated. Adaptive treatment plans with 7 mm (ATP_{7mm}) and 3 mm (ATP_{3mm}) setup robustness were generated using best-performing autodelineated contours. Clinical acceptance of ATPs was evaluated using goals encompassing ground-truth CTV-coverage and OAR-dose.
Results: Deformation was preferred for autodelineation of heart, lungs and spinal cord. DLS was preferred for all other OARs. For CTV, deformation with focus region was the preferred method although the difference with other deformation methods was small. Nominal ATPs passed evaluation goals for 87 % of ATP_{7mm} and 67 % of ATP_{3mm}. This dropped to respectively 2 % and 29 % after robust evaluation. Insufficient CTV-coverage was the main reason for ATP-rejection.
Conclusion: Autodelineation aids a ProtOnART-workflow for esophageal cancer. Currently available tools regularly require manual annotations to generate clinically acceptable ATPs.

1. Introduction

Esophageal cancer (EC) is commonly treated with radiation therapy (RT), either as part of neoadjuvant or definitive treatment. The former is the treatment of choice in locally advanced EC, as part of trimodality treatment comprising of neo-adjuvant chemoradiotherapy followed by surgery [1]. Unfortunately RT also poses risks of side effects, impacting quality of life and increasing morbidity and mortality risks. Herein, dose to neighbouring organs-at-risk (OAR), i.e. heart and lungs, is thought to play a major role [2,3]. Proton therapy (PT), may reduce these risks compared to regular photon-based radiotherapy (XT), due to its beneficial dose distribution [4–6]. Phase II trial data already support this, while phase III trials are currently recruiting patients [7–9]. Although the finite range of the proton beam results in a favourable dose distribution, it can also result in significant dose perturbations if, e.g., the patient's anatomy changes. This explains the need for a thorough follow-up of the patients' internal anatomy during treatment via plan recalculations and, if necessary, plan adaptations [10]. About half of all patients need at least one replanning throughout their PT-treatment course, more often than with XT [11]. Adaptation is usually performed

* Corresponding author at: KU Leuven, Department of Oncology, Laboratory of Experimental Radiotherapy, Leuven, Belgium.
E-mail address: karin.haustermans@uzleuven.be (K. Haustermans).

https://doi.org/10.1016/j.phro.2024.100646
Received 5 August 2024; Received in revised form 6 September 2024; Accepted 10 September 2024
Available online 14 September 2024

'offline', when the patient is off the treatment couch, using a new CT-simulation scan and subsequent treatment replanning. However, advanced imaging techniques combined with enhanced treatment planning (TPS) and delivery systems, could allow for online treatment adaptation, when the patient is on the treatment couch. Proton Online Adaptive RadioTherapy (ProtOnART) would immediately enact upon mentioned impactful anatomical changes by reoptimization of the treatment plan [12]. However, clinical implementation of ProtOnART is still lacking, except for early applications [13]. ProtOnART could use the superior physical properties of PT while also reducing uncertainties. Consequently, this would lower the dose to the OARs, especially in areas difficult to treat, like for EC.

While daily ProtOnART appears to benefit EC-treatment, it presents practical challenges. Image acquisition, delineation, planning and quality assurance (QA) protocols need to be performed while the patient is on-couch [12]. Fast plan adaptation and acceptance is needed to limit the overall treatment time for three reasons: (1) on the per-patient level, it will limit the chance of slow intra-fraction movement; (2) for patients themselves, limiting the on-couch time will increase comfort and (3) from an organizational point of view, it limits the time usage of PT-treatment facilities that are relatively scarce.

Delineation is among the most time-consuming steps in this workflow, especially for large target volumes as is the case for EC [14,15]. Today, several methods of automation are available, which both save time and reduce inter- and intra-observer variability [16–18]. In this analysis, we explored the autodelineation methods available in a commercially available TPS for PT. We aimed to identify the best method for each structure, based on quantitative and qualitative evaluations. The preferred methods were then used to investigate the feasibility of a fully automated delineation and planning EC-ProtOnART-workflow.

2. Materials and methods

2.1. Patient population

For this analysis, 60 simulation CT-scans of 15 patients (4 per patient) were used. Patients were included in a clinical trial (**Supplementary Table 1**) at our centre and gave prior written informed consent to use their data for translational research. The study was approved by our Ethical Board with reference number S65789. All 15 patients commenced neoadjuvant treatment between November 2022 and January 2024 for locally advanced EC. Each underwent a baseline planning CT-scan (CT1) and three repeated simulation CT-scans in treatment position after one, two and three weeks (CT2, CT3 and CT4 respectively), mimicking an in-room CT ProtOnART-worklow. All CTs were performed on Siemens SOMATOM Drive (Siemens Healthineers, Germany), acquired at 2 mm slice thickness and encompassed the whole lungs, heart and –depending on the target location– the whole liver and kidneys. On all CTs (CT1-CT4), delineation of the clinical target volume (CTV) and OARs was executed by a radiation oncologist and supervised by a senior radiation oncologist.

2.2. Delineation

Delineation on all scans followed the guidelines outlined in the protocol of the PROTECT-trial [9]. This meant the use consensus guidelines of Thomas et al. for CTV-delineation and ASTRO consensus guidelines for delineation of the heart [19,20]. Other OARs comprised of left and right lung (Lung_L, Lung_R), left and right kidney (Kidney_L, Kidney_R), spinal cord (Spinal_cord), stomach, liver and spleen. After review, these contours were labelled 'ground truth' (GT). For this analysis, autodelineations were generated on CT2 through CT4 in RayStation version 2023B (Raysearch, Sweden) using different methods as described below.

For OARs, three autodelineation methods were tested: contour propagation through rigid image registration (RIR), deformable image registration (DIR) and deep learning segmentation (DLS). Image registration mimicked the clinically used matching protocols to account for positioning errors. RIR was done using the intensity-based algorithm available in RayStation. For DIR, the ANACONDA method in RayStation was used with intensity information only (no controlling structures) [21]. Propagations via DIR and RIR were performed from baseline (GT-contours on CT1) to the three adaptive scenarios (CT2 through CT4). For DLS, the clinically released models were used [22–24].

Evaluation of the OARs was done both quantitatively and qualitatively [25]. For the quantitative analysis, Dice similarity coefficients (DICE), mean Hausdorff Distance (HD) and mean distance to agreement (meanDTA) were calculated based on the GT-contours [25,26]. The ideal values for these parameters are DICE=1, meanDTA=0, and HD=0, representing perfect overlap, perfect alignment, and no distance between corresponding points on the two contours, respectively. For the qualitative analysis, contours were independently scored by two radiation oncologists (BM and PP) using a four-point Likert scale (Accept, Minor variation, Major variation or Reject, key in **Supplementary Table 2**). The four-point Likert scale was dichotomized in Accept or Minor variation versus Major variation or Reject accounting for inter-observer variability and creating two distinct subgroups that also encompass expected annotation times (respectively <90 and ≥ 90 s). In case of disagreement, cases were discussed until consensus was reached.

For CTV-autodelineation, four propagation methods were investigated: RIR and three different DIR-methods. For RIR, the same method as for OAR propagation was used. For DIR, we used the ANACONDA method with (DIR_def) intensity information only (no controlling structures); (DIR_focus) intensity information only but within a focus region; and (DIR_ctrl) intensity information and the GT-heart contour as a controlling structure [21,27]. As focus region in DIR_focus, the GT-CTV on CT1 expanded with 3 cm was used to include the relevant parts of the neighbouring structures such as the lungs, spine and heart. Quantitative analysis was performed using the same metrics as for OAR. Qualitative analysis was performed by the same radiation oncologists, evaluating the three DIR-methods. The same four-point Likert scale, used for OARs, was applied (**Supplementary Table 2**). Additionally, the different DIR-methods were ranked from best to worst.

The results of the quantitative evaluation were compared using t-tests, significance level alpha = 0.05. Qualitative evaluations were analysed using Chi Squared (group size > 5) or Fisher exact tests (group size ≤ 5), significance level alpha = 0.05. Given each patient contributes the same number of samples (three CT scans), any possible intra-patient correlation of results is assumed to be consistently distributed across all patients, preserving balance in the analysis. Statistics were performed in IBM SPSS Statistics version 29 [28].

2.3. Treatment planning

Based on the best-scoring autodelineated contours of OARs and CTV, adaptive treatment planning was performed in RayStation using predetermined objectives and constraints (**Supplementary Table 3**). Adaptive treatment plans (ATP) were created on all repeated simulation CTs after rigidly transferring the isocenter. All plans were created for an IBA ProteusOne using 3-beam setup (left-posterior-oblique, posterior and right-posterior-oblique). Treatment plans were optimized based on class-solution objectives for CTV and OARs using maximum 250 iterations (or until cost function <10E-8) with spot filtering after 30 iterations. The planned dose was calculated with Monte-Carlo dose engine (v5.2). Calculation grid-size was set to 2.5x2.5x2.5 mm^3. Two different uncertainty levels were used for adaptive replanning, an isotropic error of 7 mm (ATP_{7mm}) and 3 mm (ATP_{3mm}) for setup robustness and 3 % range uncertainty, reflecting respectively the currently clinically used robustness parameters without online adaptation and a proposed reduced-robustness scenario that does not take positioning errors or inter-fraction motion into account.

ATPs, based on the autodelineated contours, were then evaluated on the GT-contours using the plan evaluation goals (**Supplementary Table 4**). Evaluation took place on the nominal plan as well as considering a 7 mm for ATP_{7mm} or 3 mm for ATP_{3mm} isotropic margin and 3 % range uncertainty for robustness. Evaluation goals were evaluated as passed or failed, both nominally and robustly. ATPs were labelled as clinically accepted when fulfilling the evaluation goals versus rejected when this was not the case.

3. Results

3.1. Delineation

In the OAR, qualitative (Table 1) and quantitative (Fig. 1, detailed data in **Supplementary Table 5**) evaluations indicated that for autodelineation a combination of DIR and DLS is preferred. Heart and lung autodelineation using DIR and DLS showed comparable, limited need for time-consuming major annotations. Conversely, RIR often required major annotations. DIR had the largest DICE for both heart and lungs. For the spinal cord, all DIR autodelineations required less than 90 s annotation time. Similarly, DIR had the largest DICE. For autodelineation of the kidneys, liver, spleen and stomach, DLS outperformed the other methods, both in qualitative and quantitative assessment.

Table 1

Qualitative analysis of different methods for autodelineating organs-at-risk. Dichotomized 4-point Likert scale as detailed in **Supplementary Table 2** is used.

Structure	Method	Score	
		Accept/Minor % (n)	Major/Reject % (n)
Heart	RIR	4 % (2)	96 % (43)
	DIR	**78 % (35)***	22 % (10)
	DLS	**71 % (32)***	29 % (13)
Kidney_L	RIR	11 % (5)	89 % (40)
	DIR	20 % (9)	80 % (36)
	DLS	**98 % (44)* †**	2 % (1)
Kidney_R	RIR	16 % (7)	84 % (38)
	DIR	24 % (11)	76 % (34)
	DLS	**96 % (43)* †**	4 % (2)
Liver	RIR	2 % (1)	98 % (44)
	DIR	9 % (4)	91 % (41)
	DLS	**89 % (40)* †**	11 % (5)
Lung_L	RIR	16 % (7)	84 % (38)
	DIR	**98 % (44)***	2 % (1)
	DLS	**98 % (44)***	2 % (1)
Lung_R	RIR	11 % (5)	89 % (40)
	DIR	**98 % (44)***	2 % (1)
	DLS	**100 % (45)***	0 % (0)
Spinal_cord	RIR	16 % (7)	84 % (38)
	DIR	**100 % (45)***^	0 % (0)
	DLS	**82 % (37)***	18 % (8)
Spleen	RIR	9 % (4)	91 % (41)
	DIR	**44 % (20)***	56 % (25)
	DLS	**96 % (43)* †**	4 % (2)
Stomach	RIR	0 % (0)	100 % (45)
	DIR	0 % (0)	100 % (45)
	DLS	**56 % (25)* †**	44 % (20)

RIR=Rigid Image Registration, DIR=Deformable Image Registration, DLS=Deep Learning Segmentation

* Statistically significantly larger proportion accepted/minor than for RIR with $p < 0.001$.

† Statistically significantly larger proportion accepted/minor than for DIR with $p < 0.001$.

^ Statistically significantly larger proportion accepted/minor than for DLS with $p < 0.010$.

For CTV autodelineation, quantitative analysis (Fig. 2**,** detailed data in **Supplementary Table 6**) ruled out RIR as a potential candidate. DIR_focus had the most favourable metrics compared to the other DIR methods, though these differences were not statistically significant. Qualitatively comparing the DIR methods (Table 2), no statistically significant differences were identified using the 4-point Likert scale: all methods resulted in a major need for annotation in approximately 40 % of the cases. Using the ranking system, DIR_focus was found to result in the most favourable (best) contour in about 60 % ($p < 0.001$ compared to DIR_def) of the cases and the least favourable (worst) contour in only 2 % of the cases ($p < 0.001$ compared to both DIR_def and DIR_ctrl). DIR_focus was therefore identified as the preferred propagation method. Moreover, this method is slightly faster to compute and does not require annotation of the heart contour as a controlling structure prior to propagation, unlike DIR_ctrl.

Table 3 summarises the best-performing autodelineated contours.

3.2. ATP evaluation

Percentages of passed evaluation metrics of the ATPs (**Supplementary Table 4**) on the GT-contours are displayed in Table 4. Insufficient CTV coverage was the only reason for nominal ATP-rejection. Lowering the CTV-$V_{95\%}$ evaluation criterion from 99 % (objective used in plan optimization) to 98 % or 97 % resulted in a larger plan acceptance rate. Dose to spinal cord was never a reason for plan rejection. Body dose only resulted in ATP_{7mm}-rejection after robust evaluation (three out of 45 plans (7 %)). Using CTV-$V_{95\%}$>98 % as evaluation criterion, in the 45 nominal plans distributed over 15 patients, ATP_{7mm}-rejections occurred for three patients, in a total of six plans (13 %), while ATP_{3mm}-rejections of the plan occurred for eight patients, in a total of 15 plans (33 %). After robust evaluation, only one ATP_{7mm} (2 %) and 13 ATP_{3mm} (29 %) passed the CTV-$V_{95\%}$>98 % criterion, graphical CTV-evaluation per CT is displayed in **Supplementary Fig. 1**. Predicting nominal ATP-rejection based on qualitative evaluation of delineation (major variations) proved inaccurate, with sensitivity and specificity of respectively 0.50 and 0.59 for ATP_{7mm} and respectively 0.60 and 0.67 for ATP_{3mm}. After robust evaluation, sensitivity and specificity of ATP_{3mm}-rejections were 0.53 and 0.85 respectively. Sensitivity and specificity of ATP_{7mm}-rejections after robustness evaluation were not considered given only 1/45 plans passed. While not evaluation metrics, OAR-planning objectives and constraints (**Supplementary Table 3**) were met for all nominal ATPs.

4. Discussion

While ProtOnART could mitigate this risk of CTV underdosage due to geometric uncertainties, its clinical implementation still poses significant challenges. We evaluated different autodelineation methods for CTV and OARs as the first step in a ProtOnART-workflow, based on 45 adaptive scenarios in 15 patients. Our analyses showed that a combination of DLS and DIR is most useful for adaptive autodelineation. However, our findings indicate manual annotations remain necessary as using the autodelineated contours often failed to produce clinically acceptable ATPs.

To our knowledge, we are the first group who explored the feasibility of a ProtOnART-workflow for EC. ProtOnART delineation in the thoracic region was already discussed by Smolders et al., quantitatively evaluating five patients with non-small cell lung cancer [29]. Their findings cannot, however, be extrapolated to EC given the different CTV and position for EC within the body. Additionally, delineation within a

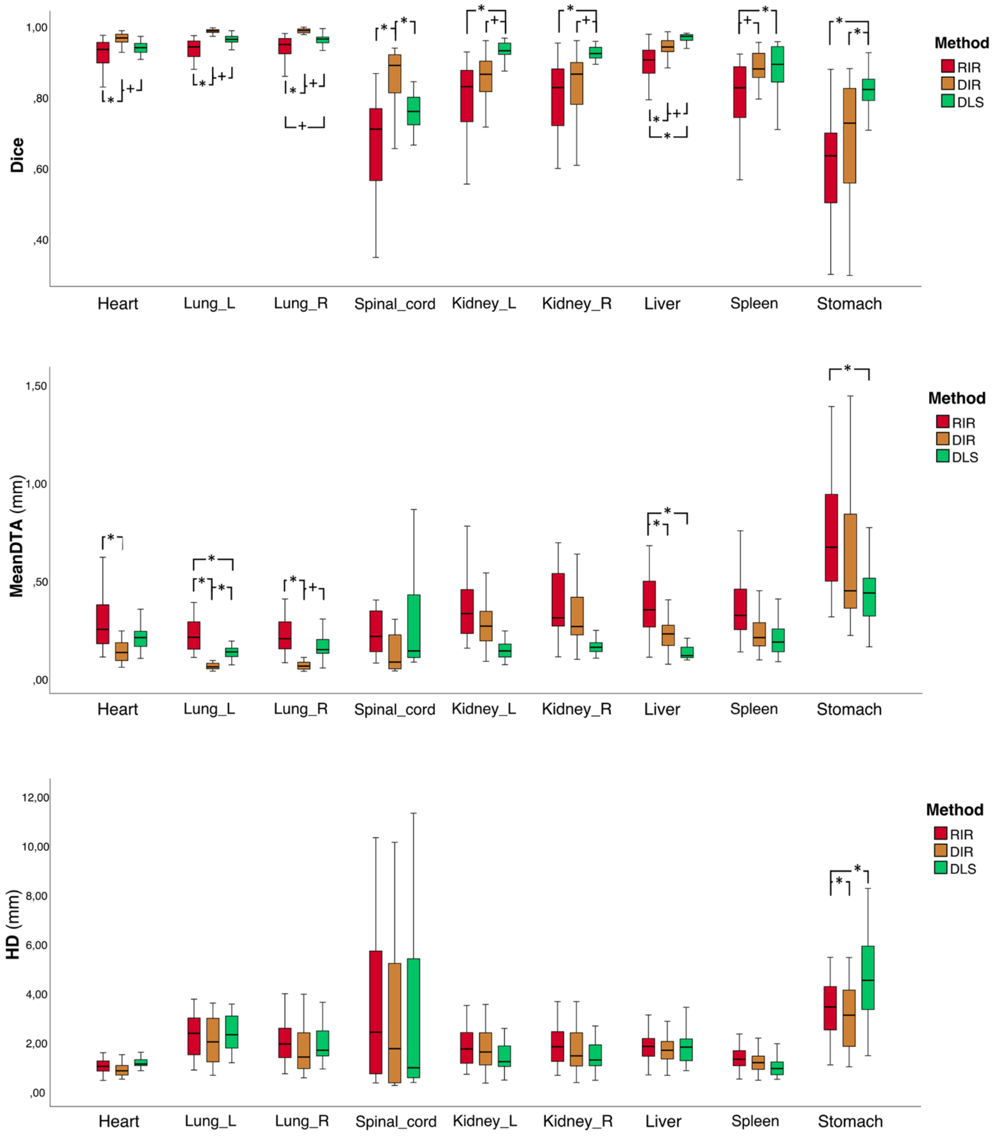

Fig. 1. Quantitative analysis of autodelineation of organs-at-risk Quantitative analysis of autodelineation of respective organs-at-risk comparing validated ground truth contours with rigid image registration (RIR), deformable image registration (DIR) or deep learning segmentation (DLS). Metrics used are Dice similarity coefficient (DICE, first plane), mean distance to agreement (meanDTA, second plane) and Hausdorff distance (HD, third plane). Based on these findings, DLS appears to be the preferred method for autodelineation for all organs-at-risk, except for spinal cord, lung and heart where DIR is preferred. Significance: * indicated p < 0.005; + indicates p < 0.050.

ProtOnART-workflow will require a clinical approval, expressing the need for qualitative assessments that remain the gold standard [30,31]. Using our findings, we also assessed whether these autodelineations could be used to generate clinically acceptable ATPs without further user input.

Apart for lungs, heart and spinal cord, DLS was the preferred method for autodelineation. For lungs and heart, DIR was slightly preferred, based on DICE and MeanDTA scores. For spinal cord, both qualitative and quantitative evaluations favoured DIR. For the EC-CTV, propagation using DIR_focus was preferred, though the difference compared to other DIR methods was small. About 40 % of the propagated CTVs required major annotations prior to clinical approval that is an inherent part of an online adaptive workflow. Boekhoff et al. reported that online contour adaptation takes close to 20 min using Elekta Unity MR-Linac [15]. Improving guideline-based EC-CTV autodelineation will likely reduce the need for major annotations and is therefore a priority [20].

Rejections of fully automated ATPs were mainly due to GT-CTV coverage, especially when accounting for uncertainties by robust evaluation. This indicates the importance of improving autodelineation of the EC-CTV. Some rejections, however, could also be attributable to small differences that can be considered within the scope of interobserver variability. Even after introduction of consensus guidelines, interobserver variability results in DICE scores and HDs in the vicinity of 0.85 and 2.5 mm respectively that are similar in order of magnitude to our automated propagations (average DICE and HD of 0.89 and 1.21 mm respectively) [20]. Future tools could aid clinicians in identifying only the dosimetrically relevant regions for annotation [32]. This likely contributed to the larger number of rejections in GT-CTV-$V_{95\%}$>99 % when compared to GT-CTV-$V_{95\%}$>98 % or GT-CTV-$V_{95\%}$>97 % and in ATP_{3mm} when compared to ATP_{7mm}. Acknowledging the impact of interobserver variability while still assuring adequate coverage of the target volume, we proposed to reduce the GT-CTV-$V_{95\%}$ objective to 97 or 98 % in evaluation, as opposed to 99 % used in optimisation. This is also how the need for (offline) replanning based on weekly surveillance repeat CTs is identified within the PROTECT-trial [9]. Additionally, robust evaluations were performed using the isotropic shifts of 7 mm

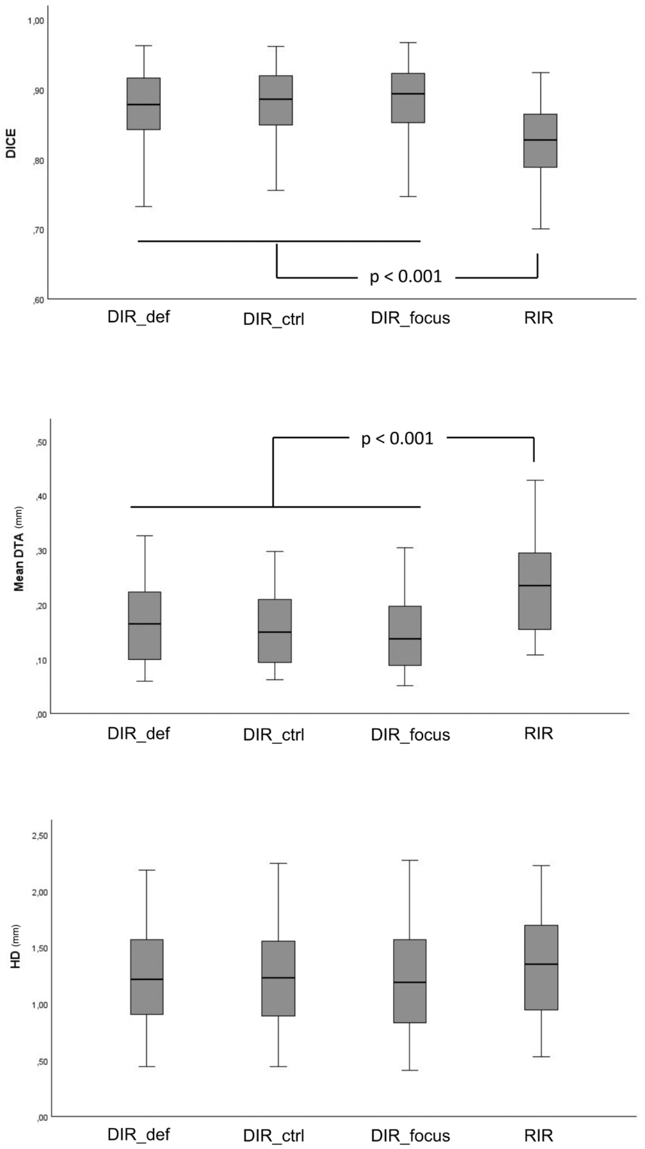

Fig. 2. Quantitative analysis of autodelineation of clinical target volume Quantitative analysis of autodelineation of the clinical target volume comparing validated ground truth contours with rigid image registration (RIR), deformable image registration (DIR) with the default ANACONDA algorithm (DIR_def), the heart as a controlling structure (DIR_ctrl) and a focus region (DIR_focus). Metrics used are Dice similarity coefficient (DICE, first plane), mean distance to agreement (meanDTA, second plane) and Hausdorff distance (HD, third plane). All DIR-methods perform better than RIR for DICE and MeanDTA. In this quantitative analysis, no statistically significant differences between the DIR-methods can be demonstrated.

and 3 mm for respectively ATP_{7mm} and ATP_{3mm} which could potentially be reduced within a ProtOnART-workflow. Research guiding online adaptive planning optimization and evaluation parameters is needed. Our findings show the limitations (limited sensitivity and specificity) of a qualitative contour evaluation in identifying ATP-rejections. This highlights the difficulty in determining when to adapt a PT-plan. Rather, in absence of clear and accurate triggers, efficient day-to-day adaptation is a better alternative as is used in commercially available systems for XT with CBCT-guidance (Ethos, Varian Medical Systems, Palo Alto, CA, USA) and MR-guidance (Unity, Elekta AB, Stockholm, Sweden).

Our analysis has some limitations. First, in this ProtOnART-simulation we retrospectively used data from three weekly simulation CT-scans. Day-to-day variations could be less pronounced and result in favourable propagation of contours. The repeated ATP-rejections in the same patient indicate anatomical shifts in CT2-4 when compared to the baseline treatment plan (CT1). In practice, one would change the reference CT, likely reducing ATP-rejections. Furthermore, ProtOnART is ideally CBCT-based, compared to our simulation implying use of in-room CT. Nonetheless, clinical application is still lacking given concerns about accuracy [12]. Second, we identified one single autodelineation method per structure to be used while other methods could be favoured in select cases. For example, one in three patients did not have DIR_focus as preferred propagation method for the CTV. However, nor the 4-point Likert scale for qualitative evaluation, nor the quantitative evaluation suggested other propagation methods would significantly improve results. Furthermore, within ProtOnArt, there is no (computational) time nor resources to generate and evaluate multiple variations of the same structure. We did not evaluate time parameters during our analyses as data would not accurately represent clinical application in the, at time of writing, absence of a publicly available system for ProtOnART to test. However, our qualitative analysis used a 90-second expected annotation cutoff to determine major variations. Nearly half of the autodelineated CTVs surpassed this threshold, confirming delineation to be a significant bottleneck [12]. For planning, timing will depend on computational power and uncertainty parameters used [33]. The extra time for ProtOnART in the thoracic region should be limited to 10–15 min [34], including integration between imaging, TPS and treatment delivery system as well as QA (not evaluated in our analysis). Third and finally, we only evaluated tools available in one commercial system (RayStation) as this would be capable to perform the ProtOnART-workflow. Other automation tools could deliver different conclusions.

In conclusion, we have offered the necessary input needed for delineation and planning within a clinically applicable ProtOnART-workflow for EC. Our findings underline that clinical and dosimetric evaluations are both needed within this workflow and full automation is currently not accurate enough. We believe an executable EC-ProtOnART-workflow is feasible. Nonetheless, improvements for (CTV-)autodelineation, finetuning of the planning/QA steps and integration of software and hardware are required to speed up the workflow and allow clinical implementation. ProtOnART will require close collaboration between radiation oncologists, radiation therapists and medical physics experts to provide the needed care for this innovative treatment concept.

CRediT authorship contribution statement

Pieter Populaire: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Beatrice Marini:** Formal analysis, Investigation, Writing – original draft. **Kenneth Poels:** Formal analysis, Investigation, Methodology, Validation, Writing – review & editing. **Stina Svensson:** Methodology, Resources, Software, Writing – review & editing. **Edmond Sterpin:** Conceptualization, Project administration, Supervision, Writing – review & editing. **Albin Fredriksson:** Resources, Software, Writing – review & editing. **Karin Haustermans:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stina Svensson: employee of RaySearch Laboratories. Albin Fredriksson: employee of RaySearch Laboratories. Pieter Populaire: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Beatrice Marini: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Kenneth Poels: no known competing financial interests or personal relationships that could have appeared to influence the work reported in

Table 2
Qualitative analysis of autodelineation of clinical target volume (CTV) using different methods. Dichotomized 4-point Likert scale as detailed in **Supplementary Table 2** is used, as well and ranking from “best” to “worst”.

		Method					
		DIR_def		DIR_ctrl		DIR_focus	
		Percentage	(n)	Percentage	(n)	Percentage	(n)
4 Point Likert Scale	Minor	58 %	(26)	56 %	(25)	58 %	(26)
	Major	42 %	(19)	44 %	(20)	42 %	(19)
Ranking	Best	4 %	(2)	36 %	(16)	**60 %***	(27)
	Middle	36 %	(16)	27 %	(12)	38 %	(17)
	Worst	60 %	(27)	38 %	(17)	**2 %**†	(1)

RIR=Rigid image registration
DIR_def = Deformable image registration with the default ANACONDA algorithm.
DIR_ctrl = Deformable image registration with the heart as a controlling structure.
DIR_focus = Deformable image registration with a focus region
* Statistically significantly larger proportion than for DIR_def with $p < 0.001$.
† Statistically significantly smaller proportion than for DIR_def and DIR_ctrl with $p < 0.001$.

Table 3
Summary of preferred methods for autodelineation to be used in a proton online adaptive radiotherapy workflow for esophageal cancer treatment.

Structure	Method
CTVtotal	Intensity-driven deformable image registration with focus region (CTV+3 cm)
Heart	Deformable image registration
Lungs (Lung_L+Lung_R)	Deformable image registration
Kidneys (Kidney_L+Kidney_R)	Deep learning segmentation
Liver	Deep learning segmentation
Spleen	Deep learning segmentation
Stomach	Deep learning segmentation
Spinal Cord	Deformable image registration

Table 4
Percentage of plans optimized using autodelineated contours passing evaluation clinical goals and constraints based on ground truth structures. Dosimetric evaluation of ground-truth contours of adaptive treatment plans (ATP) based on uncorrected autodelineations with a 7 mm (ATP_{7mm}) or 3 mm (ATP_{3mm}) setup error. Robust evaluation was performed using isotropic 7 mm (ATP_{7mm}) and 3 mm (ATP_{3mm}) shifts and 3 % range uncertainty. The nominal plan and worst-case scenario of robust evaluation are reviewed.

Evaluation goals	Success rate ATP_{7mm} % (n/45)		Success rate ATP_{3mm} % (n/45)	
	Nominal	Worst-case robust	Nominal	Worst-case robust
CTV-$V_{95\%} > 99$ %	78 % (35/45)	0 % (0/45)	47 % (21/45)	2 % (1/45)
CTV-$V_{95\%} > 98$ %	87 % (39/45)	2 % (1/45)	67 % (30/45)	29 % (13/45)
CTV-$V_{95\%} > 97$ %	96 % (43/45)	29 % (13/45)	73 % (33/45)	40 % (18/45)
Body $D_{1cm3} <$ 110 %	100 % (45/45)	93 % (42/45)	100 % (45/45)	100 % (45/45)
Body $D_{5cm3} <$ 107 %	100 % (45/45)	93 % (42/45)	100 % (45/45)	100 % (45/45)
Spinal_cord $D_{0.05cm3} < 50$ Gy	100 % (45/45)	100 % (45/45)	100 % (45/45)	100 % (45/45)

this paper. Stina Svensson: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Edmond Sterpin: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Albin Fredriksson: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Karin Haustermans: no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2024.100646.

References

[1] Eyck BM, van Lanschot JJB, Hulshof MCCM, van der Wilk BJ, Shapiro J, van Hagen P, et al. Ten-year outcome of neoadjuvant chemoradiotherapy plus surgery for esophageal cancer: the randomized controlled CROSS trial. J Clin Oncol 2021; 39:1995–2004. https://doi.org/10.1200/JCO.20.03614.
[2] Wang J, Wei C, Tucker SL, Myles B, Palmer M, Hofstetter WL, et al. Predictors of postoperative complications after trimodality therapy for esophageal cancer. Int J Radiat Oncol Biol Phys 2013;86:885–91. https://doi.org/10.1016/J.IJROBP.2013.04.006.
[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;141:33–40. https://doi.org/10.1016/J.RADONC.2019.09.015.
[4] Chuong MD, Hallemeier CL, Jabbour SK, Yu J, Badiyan S, Merrell KW, et al. Improving outcomes for esophageal cancer using proton beam therapy. Int J Radiat Oncol Biol Phys 2016;95:488–97. https://doi.org/10.1016/J.IJROBP.2015.11.043.
[5] Wang X, Hobbs B, Gandhi SJ, Muijs CT, Langendijk JA, Lin SH. Current status and application of proton therapy for esophageal cancer. Radiother Oncol 2021;164: 27–36. https://doi.org/10.1016/J.RADONC.2021.09.004.
[6] Gergelis KR, Jethwa KR, Tryggestad EJ, Ashman JB, Haddock MG, Hallemeier CL. Proton beam radiotherapy for esophagus cancer: state of the art. J Thorac Dis 2020; 12:7002. https://doi.org/10.21037/JTD-2019-CPTN-06.
[7] Lin SH, Hobbs BP, Verma V, Tidwell RS, Smith GL, Lei X, et al. Randomized phase IIB trial of proton beam therapy versus intensity-modulated radiation therapy for locally advanced esophageal cancer. J Clin Oncol 2020;38:1569–78. https://doi.org/10.1200/JCO.19.02503.
[8] NCT03801876 | Comparing Proton Therapy to Photon Radiation Therapy for Esophageal Cancer | ClinicalTrials.gov n.d. https://clinicaltrials.gov/study/NCT03801876?term=NRG&cond=Esophageal%20Cancer&rank=2 (accessed April 22, 2024).
[9] Mortensen HR, Populaire P, Hoffmann L, Moeller DS, Appelt A, Nafteux P, et al. Proton versus photon therapy for esophageal cancer - A trimodality strategy (PROTECT) NCT050555648: a multicenter international randomized phase III study of neoadjuvant proton versus photon chemoradiotherapy in locally advanced esophageal cancer. Radiother Oncol 2024;190:109980. https://doi.org/10.1016/J.RADONC.2023.109980.
[10] Møller DS, Alber M, Nordsmark M, Nyeng TB, Lutz CM, Hoffmann L. Validation of a robust strategy for proton spot scanning for oesophageal cancer in the presence of anatomical changes. Radiother Oncol 2019;131:174–8. https://doi.org/10.1016/J.RADONC.2018.09.018.
[11] Visser S, Ribeiro CO, Dieters M, Mul VE, Niezink A, van de Schaaf A, et al. Robustness assessment of clinical adaptive proton and photon radiotherapy for oesophageal cancer in the model-based approach. Radiother Oncol 2022;177: 197–204. https://doi.org/10.1016/J.RADONC.2022.11.001.
[12] Albertini F, Matter M, Nenoff L, Zhang Y, Lomax A. Online daily adaptive proton therapy. Br J Radiol 2020;93:20190594. https://doi.org/10.1259/BJR.20190594.
[13] Bobić M, Choulilitsa E, Lee H, Czerska K, Christensen JB, Mayor A, et al. Multi-institutional experimental validation of online adaptive proton therapy workflows. Phys Med Biol 2024;69:165021. https://doi.org/10.1088/1361-6560/AD6527.

[14] Lavrova E, Garrett MD, Wang YF, Chin C, Elliston C, Savacool M, et al. Adaptive radiation therapy: a review of CT-based techniques. Radiol Imaging Cancer 2023;5:e230011.

[15] Boekhoff MR, Bouwmans R, Doornaert PAH, Intven MPW, Lagendijk JJW, van Lier ALHMW, et al. Clinical implementation and feasibility of long-course fractionated MR-guided chemoradiotherapy for patients with esophageal cancer: An R-IDEAL stage 1b/2a evaluation of technical innovation. Clin Transl Radiat Oncol 2022;34:82–9. https://doi.org/10.1016/J.CTRO.2022.03.008.

[16] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol 2019;29:185–97. https://doi.org/10.1016/J.SEMRADONC.2019.02.001.

[17] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41:050902. https://doi.org/10.1118/1.4871620.

[18] Mao W, Riess J, Kim J, Vance S, Chetty IJ, Movsas B, et al. Evaluation of auto-contouring and dose distributions for online adaptive radiation therapy of patients with locally advanced lung cancers. Pract Radiat Oncol 2022;12:e329–38. https://doi.org/10.1016/J.PRRO.2021.12.017.

[19] Feng M, Moran JM, Koelling T, Chughtai A, Chan JL, Freedman L, et al. Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer. Int J Radiat Oncol Biol Phys 2011;79:10–8. https://doi.org/10.1016/J.IJROBP.2009.10.058.

[20] 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;156:102–12. https://doi.org/10.1016/J.RADONC.2020.11.032.

[21] Weistrand O, Svensson S. The ANACONDA algorithm for deformable image registration in radiotherapy. Med Phys 2015;42:40–53. https://doi.org/10.1118/1.4894702.

[22] Ng CKC, Leung VWS, Hung RHM. Clinical evaluation of deep learning and atlas-based auto-contouring for head and neck radiation therapy. Appl Sci 2022;12:11681. https://doi.org/10.3390/APP122211681.

[23] Doolan PJ, Charalambous S, Roussakis Y, Leczynski A, Peratikou M, Benjamin M, et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. Front Oncol 2023;13:1213068. https://doi.org/10.3389/FONC.2023.1213068.

[24] Almberg SS, Lervåg C, Frengen J, Eidem M, Abramova TM, Nordstrand CS, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol 2022;173:62–8. https://doi.org/10.1016/J.RADONC.2022.05.018.

[25] 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;35:354–69. https://doi.org/10.1016/J.CLON.2023.01.016.

[26] Hammers JE, Pirozzi S, Lindsay D, Kaidar-Person O, Tan X, Chen RC, et al. Evaluation of a commercial DIR platform for contour propagation in prostate cancer patients treated with IMRT/VMAT. J Appl Clin Med Phys 2020;21:14–25. https://doi.org/10.1002/ACM2.12787.

[27] Polo AL, Nix M, Thompson C, O'Hara CJ, Entwisle J, Murray L, et al. Improving hybrid image and structure-based deformable image registration for large internal deformations. Phys Med Biol 2024;69:095011. https://doi.org/10.1088/1361-6560/AD3723.

[28] IBM Corp. IBM SPSS Statistics for Windows, Version 28.0 2021.

[29] 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;68:175006. https://doi.org/10.1088/1361-6560/ACE307.

[30] 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;160:185–91. https://doi.org/10.1016/J.RADONC.2021.05.003.

[31] 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. https://doi.org/10.1016/J.PHRO.2019.12.001.

[32] 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;192:110108. https://doi.org/10.1016/J.RADONC.2024.110108.

[33] Buti G, Souris K, Barragán Montero AM, Cohilis M, Lee JA, Sterpin E. Accelerated robust optimization algorithm for proton therapy treatment planning. Med Phys 2020;47:2746–54. https://doi.org/10.1002/MP.14132.

[34] Borderías-Villarroel E, Barragán-Montero A, Sterpin E. Time is NTCP: should we maximize patient throughput or perform online adaptation on proton therapy systems? Radiother Oncol 2024;198:110389. https://doi.org/10.1016/J.RADONC.2024.110389.