Novel dosimetric validation of a commercial CT scanner based deep learning automated contour solution for prostate radiotherapy

Salvatore Berenato a, Matthew Williams a, Owain Woodley a, Christian Möhler b, Elin Evans c, Anthony E. Millin a, Philip A. Wheeler b, c

a Velindre Cancer Centre, Radiotherapy Physics Department, Cardiff, Wales, United Kingdom
b Siemens Healthineers, Forchheim, Germany
c Velindre Cancer Centre, Medical Directorate, Cardiff, Wales, United Kingdom

ARTICLE INFO
Keywords:
AI
Automation
Prostate cancer
VMAT
IMRT
Delineation
Segmentation
CT

ABSTRACT
Purpose: OAR delineation accuracy influences: (i) a patient’s optimised dose distribution (PD), (ii) the reported doses (RD) presented at approval, which represent plan quality. This study utilised a novel dosimetric validation methodology, comprehensively evaluating a new CT-scanner-based AI contouring solution in terms of PD and RD within an automated planning workflow.

Methods: 20 prostate patients were selected to evaluate AI contouring for rectum, bladder, and proximal femurs. Five planning ‘pipelines’ were considered; three using AI contours with differing levels of manual editing (nominally none (AI Std), minor editing in specific regions (AI MinEd), and fully corrected (AI FullEd)). Remaining pipelines were manual delineations from two observers (MD Ob1, MD Ob2). Automated radiotherapy plans were generated for each pipeline. Geometric and dosimetric agreement of contour sets AI Std, AI MinEd, AI FullEd and MD Ob2 were evaluated against the reference set MD Ob1. Non-inferiority of AI pipelines was assessed, hypothesising that compared to MD Ob1, absolute deviations in metrics for AI contouring were no greater than that from MD Ob2.

Results: Compared to MD Ob1, organ delineation time was reduced by 24.9 min (96 %), 21.4 min (79 %) and 12.2 min (45 %) for AI Std, AI MinEd and AI FullEd respectively. All pipelines exhibited generally good dosimetric agreement with MD Ob1. For RD, median deviations were within ± 1.8 cm³, ± 1.7 % and ± 0.6 Gy for absolute volume, relative volume and mean dose metrics respectively. For PD, respective values were within ± 0.4 cm³, ± 0.5 % and ± 0.2 Gy. Statistically (p < 0.05), AI MinEd and AI FullEd were dosimetrically non-inferior to MD Ob2.

Conclusions: This novel dosimetric validation demonstrated that following targeted minor editing (AI MinEd), AI contours were dosimetrically non-inferior to manual delineations, reducing delineation time by 79 %.

1. Introduction

Delineation of organs at risk (OARs) is a critical component in modern inverse planned radiotherapy pathways and has two main purposes. Firstly, segmented OARs form a key input to the inverse optimiser, directly influencing the patient’s radiotherapy dose distribution and hence the treatment’s efficacy. Secondly, delineations are used to report OAR treatment doses, which when compared against established dose tolerances, guide the treating physician on the quality and safety of the plan prior to approval. Accurate delineation is therefore imperative to ensure (i) the patient’s dose (PD) is optimal, with efficacy maximised and (ii) the reported doses (RD) are accurate with unsafe treatments not erroneously approved. Fig. 1 illustrates the concept of PD and RD, and their clinical importance in relation to segmentation accuracy.

OAR segmentation is however time consuming and resource intensive. A potential solution to this problem is automated OAR segmentation via deep learning (DL), which has been shown to generate contours on computed tomography (CT) scans with good geometric correspondence to manual delineations [1]. However, whilst numerous studies have assessed the geometric accuracy of AI solutions, only a small handful have considered the dosimetric impact of delineation errors, with 3 identified for prostate cancer [2–4]. Importantly, within the literature none of the dosimetric validations utilise a methodology that
explicitly validates AI contours in terms of both PD and RD. Instead, dosimetric studies typically focus solely on the impact on PD [2–4] and do not consider the risks of erroneous plan approvals due to errors in RP. The clinical impact of contour deviations due to AI delineation has therefore not been fully explored or understood.

In terms of the AI solutions that have been validated, existing CT based DL approaches use reconstructed CT scans as the input [1]. However, radiotherapy CT scans are optimised for the treatment planning pathway and may not represent the ideal input for AI delineation algorithms. Recently a commercial AI segmentation solution that is natively integrated within a CT scanner has been released (DirectORGANS; Siemens Healthineers, Erlangen). The solution offers a potential advantage over existing approaches by automatically generating and utilising a dedicated, hidden CT reconstruction that is optimised and standardised for AI segmentation. DL contours are generated using the AI tailored reconstruction and automatically propagated to the reconstruction generated for the treatment planning process, and therefore are not influenced by the chosen reconstruction parameters.

The purpose of this study was to perform a comprehensive validation of this new scanner-based AI delineation software in the context of an automated treatment planning pipeline for extreme hypo-fractionated prostate radiotherapy (EHRT Prostate). Generating plans using a fully validated automated planning solution [5,6], five planning ‘pipelines’ were considered: three using AI contours with differing levels of manual editing (nominally none, minor and full), and two using manual delineations from different observers. The study used multiple levels of editing to investigate the possibility of focussing amendments to regions where accuracy was hypothesised to be most critical; providing a potentially efficient alternative to full amendment of all structures. Through using a novel dosimetric validation methodology, the study aimed to comprehensively compare each AI pipeline against manual methods in terms of timing, geometric agreement, and dosimetric impact for both PD and RD. The testing hypothesis was that compared to a reference observer, dosimetric and geometric differences introduced using an AI pipeline were no greater than that introduced by a second observer. The study therefore sought to provide sound evidence as to the non-inferiority of CT scanner generated AI delineations within an automated planning workflow. To our knowledge this work represents the first clinical study of this commercial tool for prostate cancer and the first evaluation of the dosimetric impact of contour deviations in terms of PD and RD.

2. Methods and materials

2.1. Auto-contouring software

The software evaluated in this study was DirectORGANS (version VA30), which is natively integrated into a SOMATOM go.Sim CT scanner (Siemens Healthineers, Erlangen). DirectORGANS utilises a dedicated CT reconstruction that is optimised in aspects such as slice thickness, metal artifact reduction and reconstruction filter to provide an optimal, standardised input to the DL models. The DL segmentation consists of two major steps: first, a characteristic landmark is detected for the target organ or group of organs using Deep Reinforcement Learning [7] to roughly locate the respective anatomy. The volume is then cropped to a volume of interest around the landmark, which is subsequently fed into a deep image-to-image network (DI2IN) for the actual segmentation step [8]. After inference on the dedicated reconstruction is completed, the resulting contours are mapped to the clinical reconstruction. All AI contour models were pre-installed standard releases, which aligned with RTOG’s consensus guidelines [9] and were trained independently to our institution. Relevant models were AnoRectumSig, bladder, left proximal femur (femur_L), and right proximal femur (femur_R). No rectum model was available therefore.

![Clinical Example](image)

Fig. 1. Definition of PD and RD, with different scenarios for a 40 Gy in 5 fraction prostate treatment demonstrating how delineation errors in the patient pathway impact PD and RD for a key dose metric (rectum V36Gy) and how that in turn impacts clinical decision making. For scenario 1, RD is extracted using ‘Rectum_Manual’ (as the planning contour) and PD extracted using ‘Rectum_Manual’ (as the best estimate contour). For scenario 2, RD is extracted using ‘Rectum_AI_Small’ (as the planning contour) and PD extracted using ‘Rectum_Manual’ (as the best estimate contour).
AnoRectumSig was manually segmented into its constituent organs (anus, rectum, sigmoid).

2.2. Patient selection and preparation

20 prostate patients treated between 1st April 2021 and 1st June 2021 were chronologically selected for this study according to local criteria for EHRTProstate (Inclusion criteria: low/intermediate risk, exclusion criteria: bilateral hip prostheses). All study patients were intermediate risk, with three having a single hip prosthesis.

Patients were scanned on a Siemens SOMATOM go.Sim at 120 kV using CAREDose4D automatic tube current modulation. Images for the treatment planning process were reconstructed with a 1 mm slice thickness, using the Qr40 reconstruction filter and SAFIRE Strength 3 iterative reconstruction, with AI contours automatically generated by the scanner. The AI contours and clinical images were transferred to the treatment planning system RayStation v6B (RaySearch, Stockholm), where the remainder of the study was performed. Prostate and 1 cm of proximal seminal vesicles were delineated as the clinical target volumes (CTV) by a specialist dosimetrist according to the PACEB trial protocol [10] and expanded by 5 mm (6 mm craniocaudally) to form the planning target volume (PTV). All AI contours were hidden during target delineation.

2.3. Contour generation

Rectum, bladder, femur L and femur R were evaluated in this study. All delineations followed RTOG’s consensus guidelines [9], in keeping with both local practice and UK clinical trials. Prior to the study, peer review sessions (on non-study patients) were performed to ensure delineation practice and interpretation of the guidelines was harmonised as far as possible across participants. During segmentation all target volumes were hidden to eliminate operator influence from target volume boundaries. Prostheses and bowel proximal to the PTV were delineated manually by each operator for the purposes of plan generation but were not included in any analysis.

Two sets of manually delineated OARs (MDStd, MDOb1) were independently generated for each study patient by two fully trained observers (Ob1, Ob2). For the AI contours, a preprocessing step was required whereby Ob1 defined the superior and inferior extent of the rectum (according to RTOG’s standard definition) with a custom RayStation script automatically deleting contours outside these boundaries, thereby extracting the rectum component from the AI AnoRectumSig contour. This preprocessed set represented the base AI contours (AStd). Using AStd as a starting point, two additional contour sets with different levels of manual editing were generated by Ob1. The first set (AStd) included minor editing that was either trivial to perform, or in areas where accuracy was hypothesised to be critical in the planning process. For rectum and bladder, contours were corrected in regions perceived to be abutting the prostate. Femoral contours were automatically cropped using a script to match the inferior boundary with the AStd rectum contour (as per RTOG guidelines). For the second contour set (AFullL, AFullR) contours were fully corrected according to current clinical requirements.

In summary, five contour sets were generated AStd, AStd, AFullL, AFullR, MDOb1 and MDOb2, with all interventions on AI sets performed by Ob1. Delineation was performed in controlled conditions, with timing data (including preprocessing steps) recorded for AStd, AStd, AFullL and AFullR.

2.4. Plan generation

Radiotherapy plans were generated for each contour set using an automated planning solution [11] that has been validated for prostate cancer [5] and implemented clinically for EHRTProstate patients [6]. Utilisation of automated planning ensured treatment plans were generated without operator bias. The clinical planning protocol (Table 1) was based on the PACEB clinical trial, delivering 36.25 Gy and 40 Gy in 5 fractions to the PTV and CTV respectively. The automated planning solution required an AutoPlan protocol as an input. The protocol used in this study is presented in Table 2 and consisted of a set of clinical goals at three differing priority levels. Priority one goals represented mandatory OAR tolerances that must be met, priority two goals defined target coverage and hotspots, and priority three goals represented all other trade-offs. For each clinical goal a weighting factor was assigned, alongside if the objective is ‘static’ or ‘dynamic’. Static objectives are selected for fixed clinical aims (e.g. PTV minimum dose > 36.25 Gy), whereas dynamic objectives are selected where the aim is to minimise dose (e.g. minimise Bladder V17.0 Gy). The automated planning solution followed a ‘Protocol Based Automatic Iterative Optimisation’ (PBAIO) methodology during plan generation where: (i) targets are retracted from priority one OARs; (ii) a set of optimisation objectives derived from the AutoPlan protocol are loaded into the planning system’s native optimiser; (iii) the optimisation process is started, with the weight and position of ‘dynamic’ objectives automatically updated at regular intervals to ensure OAR doses are minimised and competing trade-offs balanced in a consistent manner. Full details of the PBAIO algorithm are provided by Wheeler et al. [11].

Plans consisted of two full VMAT arcs using a Varian TrueBeamSTx (Varian Medical Systems, Palo Alto) machine model. The resultant plans for each contour set represented the output of a potential clinical planning ‘pipeline’, with the AIStd pipeline nominally representing a fully automated process.

2.5. Contour evaluation

In all geometric and dosimetric analysis MDOb1 was defined as the reference contour set. The geometric similarity of each contour set with MDOb1 was assessed using the Dice similarity coefficient (DSC) and mean surface distance (MSD) metrics. DSC was defined as D(X, Y) = 2X ∩ Y/(X + Y), where X and Y were the contours to be assessed. In line with Duan et al. [2], MSD was defined as the mean of the directed average Hausdorff distance (HD) for X vs Y and Y vs X, where HD X vs Y represents the mean distance of points in X to their closest points in Y.

For the dosimetric comparison, RD and PD metrics were extracted for all plans within a given patient, with each plan representing a particular contouring pipeline. In line with Fig. 1, RD was defined as DVH parameter data that would be reported in the patient records for a particular pipeline implementation (RDpipeline) and PD as the best estimate of the actual dose the patient would receive (PDpipeline). As with a

| Table 1 | Prostate EHRT planning protocol. |
| Target Objectives | | |
| ROI Name | Dose Parameter | Optimal | Actionable* |
| CTV | V40.0 Gy | ≥95% | ≥90% |
| PTV | V36.25 Gy | ≥95% | ≥90% |
| CI | – | – | <120% |

| OAR Objectives | | |
| ROI Name | Dose Parameter | Optimal | Actionable* |
| Rctum | V18.1 Gy | <50% | – |
| Rectum | V20.0 Gy | <20% | <20% |
| Rectum | V36.0 Gy | <1cc | <2cc |
| Bladder | V18.1 Gy | <40% | – |
| Bladder | V37.0 Gy | <10 cc | <15 cc |
| Proximal Femur | V14.5 Gy | <5cc | – |
| Bowel | V18.1 Gy | – | – |
| Bowel | V30.0 Gy | – | <1cc |

*Deviations from actionable planning goals are permissible if approved by the treating oncologist.

1 CI: (Body V36.25 Gy / PTV Volume)*100 %.
and PD DVH extraction methodology, in terms of the plan and contour set, for RD and the patient dose when optimised with the reference contour set similar study [2] we estimated PD hence final dose distribution, the difference (ΔPD) between PD(hence final dose distribution, the difference (ΔPD) between PD and MD) was determined. The data extraction methodology for RD and PD DVH was presented in Table 3.

To estimate the error in RD due to contour discrepancies (ΔRD), the difference between RDpipeline and the best estimate of the actual dose to the patient (PDpipeline) was calculated for each pipeline. Similarly, to estimate the contour set’s influence on the optimisation process and hence final dose distribution, the difference (ΔPD) between PDpipeline and the patient dose when optimised with the reference contour set (PDMD) was determined. ΔPD and ΔRD were calculated for all treatment protocol dose metrics (Table 1) and OAR mean doses. Note that assessment of ΔRD for target dose metrics was not applicable as a single set of target contours was used across all the pipelines and therefore ΔRD would equal zero.

For statistical analysis, the testing hypothesis was ‘when compared against the reference contours (MD081), absolute dosimetric and geometric differences introduced using an AI pipeline are no greater than that introduced by a 2nd observer (MD082)’. In this regard DSC, MSD, |ΔRD| and |ΔPD| for the AI pipelines were statistically compared to MD082 using two tailed Wilcoxon signed ranked tests. Absolute differences for dose metrics were selected to ensure significant results represented differences in the magnitude of deviations rather than the direction.

### 3. Results

A summary of the recorded OAR delineation times per pipeline is presented in Fig. 2. The total median delineation time for ALstd, ALminEd, ALfullEd and MD081 was 1.0 min, 5.4 min, 15.2 min and 25.9 min respectively. Compared to MD081, total median time savings for ALstd, ALminEd and ALfullEd were 24.9 min [range 21.5 to 44.9 min], 21.4 min [13.4 to 30.7 min] and 12.2 min [-6.1 to 24.9 min] respectively, representing median efficiency savings of 96 %, 79 % and 45 %. Savings were observed across each of the assessed OARs. For ALfullEd, delineation time was greater than Ob1 for three rectum (+4.3 min, +0.2 min, +0.9 min) and four bladder (+2.4 min, +0.8 min, +3.1 min, +2.2 min) contours. This was due to poorer AI performance attributed to either (i) prosthetic hip artifacts, or (ii) high proximity between bowel and bladder/rectum leading to minimal greyscale contrast at boundaries.

Box plots of the geometric contour comparison with MD081 are presented in Fig. 3. In terms of inter-observer variability, there was good agreement between MD082 and MD081 with median DSC scores of 0.915, 0.967, 0.968 and 0.967 for rectum, bladder, femur_R and femur_L respectively and median MSD scores ≤ 1 mm across all OARs. For the AI pipelines, ALstd was the worst performer with statistically (p < 0.05) poorer results than MD082 for DSC and MSD across all organs. Minor editing for rectum and bladder (ALminEd) led to marginal improvements, with DSC increased by < 0.01, however both DSC and MSD were still significantly (p < 0.05) poorer than MD082. In contrast, minor edits to femur_L and femur_R led to DSC and MSD values that were not

---

**Table 2**

Prostate EHRT AutoPlan Protocol.

### Priority 1: Primary OAR Goals

<table>
<thead>
<tr>
<th>ROI Name</th>
<th>Dose Parameter</th>
<th>Objective Type</th>
<th>Target (Gy)</th>
<th>WF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowel, 1 mm</td>
<td>Dmax</td>
<td>Static</td>
<td>30</td>
<td>1000.0</td>
</tr>
</tbody>
</table>

### Priority 2: Target Goals

<table>
<thead>
<tr>
<th>ROI Name</th>
<th>Dose Parameter</th>
<th>Objective Type</th>
<th>Target (Gy)</th>
<th>WF</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTV</td>
<td>Dmin</td>
<td>Static</td>
<td>40.7</td>
<td>250.0</td>
</tr>
<tr>
<td>PTV</td>
<td>Dmin</td>
<td>Static</td>
<td>36.25</td>
<td>250.0</td>
</tr>
<tr>
<td>PTV</td>
<td>Dmax</td>
<td>Static</td>
<td>37.83</td>
<td>250.0</td>
</tr>
</tbody>
</table>

### Priority 3: Trade-off Goals (Standard)

<table>
<thead>
<tr>
<th>ROI Name</th>
<th>Dose Parameter</th>
<th>Objective Type</th>
<th>Target (Gy, % cm²)</th>
<th>WF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectum, 5 mm</td>
<td>Dmean (Gy)</td>
<td>Dynamic</td>
<td>5.0</td>
<td>61.6</td>
</tr>
<tr>
<td>Rectum, 5 mm</td>
<td>Dmax (Gy)</td>
<td>Dynamic</td>
<td>32.0</td>
<td>11.7</td>
</tr>
<tr>
<td>Rectum, 5 mm</td>
<td>V28.0 Gy (%)</td>
<td>Dynamic</td>
<td>0.0</td>
<td>99.0</td>
</tr>
<tr>
<td>Rectum</td>
<td>V34.0 Gy (cm³)</td>
<td>Static</td>
<td>1.0</td>
<td>415.3</td>
</tr>
<tr>
<td>Bladder</td>
<td>Dmax (Gy)</td>
<td>Dynamic</td>
<td>32.0</td>
<td>6.9</td>
</tr>
<tr>
<td>Bladder</td>
<td>V17.0 Gy (%)</td>
<td>Dynamic</td>
<td>0.0</td>
<td>44.6</td>
</tr>
<tr>
<td>Bladder</td>
<td>V36.3 Gy (%)</td>
<td>Static</td>
<td>8.0</td>
<td>1500.0</td>
</tr>
<tr>
<td>Femur, L</td>
<td>Dmax (Gy)</td>
<td>Static</td>
<td>13.5</td>
<td>150.0</td>
</tr>
<tr>
<td>Femur, R</td>
<td>Dmax (Gy)</td>
<td>Static</td>
<td>13.5</td>
<td>150.0</td>
</tr>
<tr>
<td>Bowel</td>
<td>V24.0 Gy (%)</td>
<td>Dynamic</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Bowel</td>
<td>V30.4 Gy (%)</td>
<td>Dynamic</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Prosthesis, 5 mm</td>
<td>Dmax (Gy)</td>
<td>Static</td>
<td>4.0</td>
<td>77.0</td>
</tr>
</tbody>
</table>

### Priority 3: Trade-off Goals (Dose Fall Off)

<table>
<thead>
<tr>
<th>ROI Name</th>
<th>Fall Off Type</th>
<th>Objective Type</th>
<th>High Dose Level (Gy)</th>
<th>Low Dose Level (Gy)</th>
<th>Dose Gradient (%Pres cm⁻¹)</th>
<th>WF</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTV</td>
<td>Intra PTV Falloff</td>
<td>Dynamic</td>
<td>38.0</td>
<td>38.0</td>
<td>50.0</td>
<td>42.0</td>
</tr>
<tr>
<td>External</td>
<td>Normal Tissue Falloff</td>
<td>Dynamic</td>
<td>36.0</td>
<td>20.0</td>
<td>50.0</td>
<td>219.9</td>
</tr>
<tr>
<td>External</td>
<td>Normal Tissue Falloff</td>
<td>Static</td>
<td>80.0</td>
<td>26.0</td>
<td>80.0</td>
<td>20000.0.0</td>
</tr>
</tbody>
</table>

**Abbreviations:** %Pres = % of overall treatment prescription; WF = weighting factor; Dmin = minimum dose; Dmax = maximum dose; Dmean = mean dose

**Notes:** The WF and target values of ‘dynamic’ objectives are automatically updated during the optimisation process such that dose parameters are minimised.

---

**Table 3**

DVH extraction methodology, in terms of the plan and contour set, for RDpipeline and PDpipeline.

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>RDpipeline</th>
<th>PDpipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALstd</td>
<td>ALstd</td>
<td>ALstd</td>
</tr>
<tr>
<td>ALminEd</td>
<td>ALminEd</td>
<td>ALminEd</td>
</tr>
<tr>
<td>ALfullEd</td>
<td>ALfullEd</td>
<td>ALfullEd</td>
</tr>
<tr>
<td>MD082</td>
<td>MD082</td>
<td>MD082</td>
</tr>
<tr>
<td>MD081</td>
<td>MD081</td>
<td>MD081</td>
</tr>
</tbody>
</table>

---

S. Berenato et al.  
Physica Medica 122 (2024) 103339
significantly different to MD
Ob2. AI
FullEd led to further minor improvements in DSC (< 0.01) for bladder and both femurs, and a moderate improvement for rectum (0.915 vs 0.894). Compared to MD
Ob2, AI
FullEd was superior for both femurs, inferior for bladder, with no significant difference observed for rectum. Unlike AI
MinEd, AI
FullEd led to reductions in the spread of DSC and MSD, indicating an improved agreement at a per-patient level that was in line with MD
Ob2. Examples of typical and outlier cases are presented in Fig. 4. In general, areas of the poorest geometric agreement were due to prosthetic hip artefacts, highly atypical patient anatomy or high proximity between bowel and bladder/rectum.

Boxplots summarising the dosimetric evaluation (in terms of ΔRD and ΔPD) and results of the hypothesis testing are presented in Fig. 5. In terms of RD, except for AI
Std bladder V37.0 Gy where median ΔRD = 1.8 cm³, all pipelines exhibited good dosimetric agreement at the population level with median ΔRD values for all absolute volume, relative volume and mean dose metrics within ±1.0 cm³, ±1.7 % and ±0.6 Gy respectively. For PD, dosimetric agreement for OARs was substantially better than for RD, with median ΔPD values within ±0.4 cm³, ±0.4 % and ±0.2 Gy. In terms of target doses, again agreement was very good with median ΔPD for coverage, hotspot and conformity index (CI) metrics within ±0.5 %, ±0.2 Gy and ±0.8 % respectively. Hypothesis testing demonstrated that AI
Std was inferior to MD
Ob2, with |ΔRD| and |ΔPD| significantly (P < 0.05) different for 9/11 and 5/11 OAR metrics

Fig. 2. Manual contouring time required for AI
Std, AI
MinEd, AI
FullEd and MD
Ob1 delineation workflows. Timing for rectum AI
std included preprocessing step to extract rectum from AnoRectumSig. No manual contouring intervention was required for bladder AI
Std, femur AI
std and femur AI
MinEd. Stars indicate a statistically significant difference (p < 0.05) in delineation time between the AI contour set and MD
Ob1.

Fig. 3. Dice similarity coefficient (top) and mean surface distance (bottom) scores comparing AI
Std, AI
MinEd, AI
FullEd and MD
Ob2 contours to the reference contours MD
Ob1. Stars indicate a statistically significant difference (p < 0.05) in geometry metrics between the AI contour set and the second observer (MD
Ob2). Data is presented in tabular format in the supplementary table S1.
respectively. In contrast, \( A_{\text{MinEd}} \) and \( A_{\text{FullEd}} \) were considered non-inferior to MD\( _{\text{Std}} \) with \( A_{\text{MinEd}} \) V14.5 Gy metrics for femur\( _L \) and femur\( _R \) (\( \Delta \text{PD} \) only) statistically inferior but at clinically insignificant levels (< 0.1 %). On a per-patient basis, \( A_{\text{MinEd}} \) generally resulted in an increased spread in \( \Delta \text{RD} \) compared to MD\( _{\text{Std}} \). This was most noticeable for femur\( _L \) and femur\( _R \) mean dose, and bladder V37.0 Gy with respective ranges of [-0.6 to 0.3 GY], [-0.5 to 0.9 GY] and [-11.5 to 6.6 cm\(^2\)] compared to [-0.04 to 0.1 GY], [-0.04 to 0.2 GY] and [-5.5 to 4.4 cm\(^2\)] for MD\( _{\text{Std}} \). This increased spread was present but less pronounced for \( \Delta \text{PD} \). For \( A_{\text{MinEd}} \) and \( A_{\text{FullEd}} \), the spread of data was considered nominally equivalent to MD\( _{\text{Std}} \) further indicating parity between the three pipelines.

4. Discussion

In this study a commercial, CT scanner based, deep-learning auto-segmentation solution has been geometrically and dosimetrically validated for EHRT\(_{\text{Prostate}}\) using a novel dosimetric validation methodology. When compared to a 2nd observer, \( A_{\text{MinEd}} \) and \( A_{\text{FullEd}} \) contours were statistically inferior across the majority of geometry metrics. However, for \( A_{\text{MinEd}} \) no inferiority was observed in the more clinically relevant dosimetric validation, where the two pipelines (\( A_{\text{MinEd}} \) and MD\( _{\text{Std}} \)) were considered nominally equivalent for both RD and PD. This study therefore provides clear evidence that AI contours, once edited in CTV abutment regions and manually cropped in the superior/inferior aspect (rectum and femurs), are statistically non-inferior to manual delineations even for highly targeted EHRT\(_{\text{Prostate}}\). With total respective delineation times for \( A_{\text{MinEd}} \) and MD\( _{\text{Std}} \) of 5.4 mins and 25.9 mins, clinical implementation of AI segmentation in this manner would lead to substantial efficiency savings with no compromise to treatment accuracy or efficacy. It also represents a method with a considerably higher efficiency saving (vs. MD\( _{\text{Std}} \)) than the common AI implementation approach of \( A_{\text{FullEd}} \) (79 % vs. 45 %). It is important to note that blind adoption of such an approach may lead to increased clinical risk due to gross AI delineation errors not being corrected. Therefore, appropriate risk mitigation processes should accompany any implementation. For example, at our clinic, in addition to amendments in the target abutment region, all AI volumes are reviewed by a trained operator with perceived deviations > 5 mm always amended and contours fully amended if there are reasonable concerns their quality may impact dosimetry.

The study’s findings are based on the hypothesis testing comparing deviations in AI pipelines (vs Ob\(_1\)) to deviations between two observers (Ob\(_2\) vs Ob\(_1\)). The results of which are strongly influenced by the level of inter-observer variation (IOV) between Ob\(_1\) and Ob\(_2\). A high IOV would lead to a weak comparator, with AI pipelines more likely to be considered non-inferior to Ob\(_2\), the converse being true for a low IOV. By performing a peer review session prior to the study proper, we sought to reduce IOV to as low as reasonably practicable, thereby forming a strong comparator to the AI pipelines. Whilst this may be less representative of clinical practice, it ensured any conclusions of non-inferiority could be considered excellent, with respective DSC values for bladder, rectum and proximal femur of 0.97 (vs 0.96), 0.92 (vs 0.79) and 0.97 (vs 0.91) for femoral head IOV.

In terms of the geometric validation, \( A_{\text{MinEd}} \) yielded average DSC results that were highly consistent with the literature for rectum (0.89 vs [0.79–0.92]) and bladder (0.95 vs [0.93–0.97]) [1–3]. For proximal femur, reported studies tended to evaluate femoral head contours that exclude the neck, which whilst not directly equivalent were considered similar enough to act as reasonable comparators. In this regard, proximal femur DSC results were aligned with the upper end of published results (0.96 (rt) and 0.95 (lt) vs [0.68–0.97]) [1,2]. The geometric validation therefore provides supportive evidence that the unedited AI output (albeit including manual pre-processing to extract the rectum from AnoRectumSig) was in-line with the current state-of-the-art in auto-segmentation.

In comparison to Ob\(_2\), a range of statistically significant detractors in DSC and MSD were observed across the AI pipelines. However, these differences did not necessarily translate to statistically and clinically significant dosimetric differences, especially for PD. Most notably, for \( A_{\text{MinEd}} \) none of the significant geometric differences resulted in significant dosimetric differences. This weak relationship between geometry and dosimetry metrics is aligned with that observed for head and neck cancer [13] and highlights the importance of dosimetric assessments in the clinical validation of contouring methods.

Results of the dosimetric validation yielded some important observations. Firstly, all dosimetric deviations for proximal femur were small (< 1 Gy or 1 %) and not considered clinically significant. This indicates
that unedited proximal femur contours are suitable for direct clinical implementation. Secondly, delineation uncertainty in regions where bladder/rectum abutted the prostate, led in many cases to substantial deviations in high dose metrics ($\geq 36$ Gy) across both manual and AI pipelines. Delineations were performed solely using CT data and these results highlight important limitations in the appropriateness of this imaging modality for delineating boundaries accurately. Finally for PD, which directly affects treatment efficacy, delineation accuracy had minimal impact on low-mid dose metrics ($< 36.0$ Gy), with deviations in volume and mean doses across all patients less than $\pm 3\%$ and $\pm 2$ Gy respectively. For context, comparisons of manual planning vs. automated planning for prostate cancer have demonstrated substantially

Fig. 5. Dosimetric comparison of the AI_{Std}, AI_{MinEd}, AI_{FullEd} and MD_{Ob2} planning pipelines in terms of difference in the patient dose ($\Delta$PD: LHS) and reported dose ($\Delta$RD: RHS) to the reference pipeline (MD_{Ob1}). Stars indicate a statistically significant difference ($p < 0.05$) in $|\Delta$PD$|$ or $|\Delta$RD$|$ between the AI contour set and the second observer (MD_{Ob2}). Data is presented in tabular format in the supplementary table S2.
larger deviations, with mean rectum and bladder dose differences ranging between [-4 to 12] Gy and [-10 to 10] Gy respectively [14]. The same was not true for high dose metrics, where OAR delineation in the abutment regions influenced the compromise of target coverage during the plan optimisation. Results of this study therefore point to the insensitivity of delineation accuracy on plan optimisation, except in regions which influence target compromise. For RD, variation in delineation led to more significant deviations, with errors in volume and dose metrics of up to ± 8 % and ± 2.6 Gy respectively. For volume metrics, deviations at this level could substantially influence clinical decision-making during approval and therefore were considered of likely clinical relevance. These differing conclusions for PD and RD highlight the importance of our novel dosimetric validation approach in assessing both metrics to fully understand the influence of contour variations on the planning pathway in terms of both clinical efficacy (PD) and clinical decision making (RD).

In terms of previous studies, a significant number have assessed CT based AI OAR contouring for prostate cancer [8,12,15-21] in a recent systematic review [1]. However, we identified only three dosimetric evaluations [2-4], with two generating plans using unmodified fixed optimisation objectives [3,4], which is not representative of clinical practice. Duan et al [2] present the most relevant study, utilising Van’s RapidPlan to automatically generate patient tailored plans prescribed to 70 Gy in 25 fractions. They evaluated the impact on PD of an unedited AI contour pipeline versus a manual contour pipeline. Results demonstrated significant dosimetric differences for bladder doses only, which were reduced in the AI pipeline. No assessment of RD or IOV (from a dosimetry perspective) was performed. Our study therefore builds on this evidence base through a more comprehensive dosimetric evaluation that included: different levels of contour editing, statistical comparison against a 2nd observer, and analysis of both PD and RD. Furthermore, the EHRTProstate planning protocol we used represents a highly complex planning scenario, which will likely become the standard of care for many patients [10].

To our knowledge this work also presents the first clinical evaluation for prostate cancer of a commercial CT scanner-based AI contouring solution. By utilising CT reconstructions that are optimised and standardised for AI segmentation rather than for clinical purposes, there is a theoretical advantage over existing solutions. This is especially likely in applications where clinical imaging requirements (e.g. for image smoothness and slice thickness) are variable and contrary to optimal AI performance. Whilst a direct comparison of the two approaches was not performed, results from the geometric evaluation are supportive of this novel approach, with DSC metrics on a par with the top performing AI solutions in the literature. An additional benefit of the AI implementation is seamless integration into the radiotherapy pathway with structures autonomously segmented on the scanner prior to export. Our study evaluated independently trained, pre-installed ‘out-of-the-box’ models, which has the advantage that results should be highly generalisable, with low implementation barriers for centres with the same equipment. However, it also meant reliance on appropriate models being available. In this study, a pre-processing step was required to extract the rectum from the AnoRectumSig AI output, reducing the level of automation possible. Release of further models as the software matures should however alleviate these issues.

This study had several potential weaknesses. Firstly, utilisation of only two observers may not have adequately sampled the extent of IOV across the observer population. Interestingly in this case, by the nature of the hypothesis, the excellent agreement between O$_{\text{B}_1}$ and O$_{\text{B}_2}$ compared to other studies indicate that our results are more likely to be biasing against, rather than for, AI segmentation. Secondly, it is worth noting that in this study AI contours were generated by the scanner using CT reconstruction settings that were optimised for AI contouring, however the manual amendments/delineation were performed on standard clinical reconstructed images. This could theoretically provide AI delineation with an advantage over manual delineations, for example due to enhanced tissue boundary definition. However, in this study as the comparator (OB) was delineated on clinical images, any deviations in AI contouring due to improved tissue visibility would lead to lower geometric agreement scores and again bias results against AI contouring. Finally, in this study OARs were delineated with no knowledge of the target volume. Whilst this enabled timing and geometric analysis to be independent of the cancer site and therefore generalisable, it was not representative of clinical practice where OAR delineation is strongly guided by CTV contours. This has important consequences for practical clinical implementation, where an ‘approved’ target contour is generally available. Firstly, unlike in this study, editing the abutment region for bladder and rectum would be a trivial process, leading to meaningful reductions in delineation time with percentage efficiency gains highest for A$_{\text{final}}$. Secondly, the variation in high dose metrics across the pipelines would likely be reduced to negligible levels. It is reasonable to assume results of this study represents a ‘worst case’ implementation, with efficiency likely improved and variation reduced in clinical practice.

5. Conclusions

In summary, a commercial CT scanner-based AI delineation software has been comprehensively evaluated for prostate cancer using a novel dosimetric validation methodology. The pre-installed OAR models yielded AI contours of very good correspondence with manual contours, with results fully congruent with the current state-of-the-art in AI segmentation. Following limited edits in specific regions (e.g. target/OAR abutment), AI contours were dosimetrically non-inferior to manual delineations and reduced total delineation time for the evaluated set of OARs from 25.9 mins to 5.4 mins.

CRediT authorship contribution statement

SB: Data curation, Investigation, Formal Analysis, Software. MW: Conceptualization, Funding acquisition, Writing – review & editing, Investigation, Formal Analysis, Methodology, Supervision, Software. OW: Data curation, Writing – review & editing. CM: Conceptualization, Writing – review & editing, Methodology. EE: Writing – review & editing, Methodology. AEM: Conceptualization, Funding acquisition, Writing – review & editing, Methodology. PW: Conceptualization, Funding acquisition, Writing, Writing – review & editing, Investigation, Formal Analysis, Methodology, Supervision, Software.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christian Möhler is an employee of Siemens Healthineers. All other authors have no conflict of interest to disclose.

Acknowledgments

This research was funded by Velindre Cancer Centre’s Advancing Radiotherapy Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2024.103339.

References


