cancers

MDPI

Article

Clinical Use of a Commercial Artificial Intelligence-Based Software for Autocontouring in Radiation Therapy: Geometric Performance and Dosimetric Impact

S M Hasibul Hoque [1], Giovanni Pirrone [1], Fabio Matrone [2], Alessandra Donofrio [2], Giuseppe Fanetti [2], Angela Caroli [2], Rahnuma Shahrin Rista [1], Roberto Bortolus [2], Michele Avanzo [1], Annalisa Drigo [1] and Paola Chiovati [1,*]

[1] Medical Physics Department, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, 33081 Aviano, Italy; smhasibul.hoque@cro.it (S.M.H.H.); giovanni.pirrone@cro.it (G.P.); rahnumashahrin.rista@cro.it (R.S.R.); mavanzo@cro.it (M.A.); adrigo@cro.it (A.D.)

[2] Radiation Oncology Department, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, 33081 Aviano, Italy; fabio.matrone@cro.it (F.M.); alessandra.donofrio@cro.it (A.D.); giuseppe.fanetti@cro.it (G.F.); angela.caroli@cro.it (A.C.); rbortolus@cro.it (R.B.)

* Correspondence: pchiovati@cro.it

Simple Summary: Auto contouring driven by artificial intelligence can improve the workflow of radiotherapy by accelerating the contouring process. However, quality assurance of artificial intelligence-based tools is necessary for ensuring safety and efficacy in a clinical practice. In this study investigated the geometric accuracy of structural contours created by a commercial software for autocontouring based on artificial intelligence using well established metrics. In particular, the impact on the radiotherapy treatment plan quality from the adoption of artificial intelligence generated contours was investigated. Our results show that the combination of automatically generated contours and careful review by a clinical radiation oncologist results in time saving without affecting the quality of treatment plan. In conclusion, after quality checks that involve both geometric accuracy as well as dosimetric impact, contouring based on *AI* can be safely adopted in clinical practice.

Abstract: Purpose: When autocontouring based on artificial intelligence (*AI*) is used in the radiotherapy (*RT*) workflow, the contours are reviewed and eventually adjusted by a radiation oncologist before an RT treatment plan is generated, with the purpose of improving dosimetry and reducing both interobserver variability and time for contouring. The purpose of this study was to evaluate the results of application of a commercial AI-based autocontouring for *RT*, assessing both geometric accuracies and the influence on optimized dose from automatically generated contours after review by human operator. Materials and Methods: A commercial autocontouring system was applied to a retrospective database of 40 patients, of which 20 were treated with radiotherapy for prostate cancer (PCa) and 20 for head and neck cancer (*HNC*). Contours resulting from *AI* were compared against *AI* contours reviewed by human operator and human-only contours using Dice similarity coefficient (*DSC*), Hausdorff distance (*HD*), and relative volume difference (*RVD*). Dosimetric indices such as D_{mean}, $D_{0.03cc}$, and normalized plan quality metrics were used to compare dose distributions from RT plans generated from structure sets contoured by humans assisted by *AI* against plans from manual contours. The reduction in contouring time obtained by using automated tools was also assessed. A Wilcoxon rank sum test was computed to assess the significance of differences. Interobserver variability of the comparison of manual vs. AI-assisted contours was also assessed among two radiation oncologists for PCa. Results: For PCa, AI-assisted segmentation showed good agreement with expert radiation oncologist structures with average *DSC* among patients ≥ 0.7 for all structures, and minimal radiation oncology adjustment of structures (*DSC* of adjusted versus *AI* structures ≥ 0.91). For *HNC*, results of comparison between manual and *AI* contouring varied considerably e.g., 0.77 for oral cavity and 0.11–0.13 for brachial plexus, but again, adjustment was generally minimal (*DSC* of adjusted against *AI* contours 0.97 for oral cavity, 0.92–0.93 for brachial plexus). The difference in dose for the target and organs at risk were not statistically significant

check for updates

Citation: Hoque, S.M.H.; Pirrone, G.; Matrone, F.; Donofrio, A.; Fanetti, G.; Caroli, A.; Rista, R.S.; Bortolus, R.; Avanzo, M.; Drigo, A.; et al. Clinical Use of a Commercial Artificial Intelligence-Based Software for Autocontouring in Radiation Therapy: Geometric Performance and Dosimetric Impact. *Cancers* **2023**, *15*, 5735. https://doi.org/10.3390/cancers15245735

Academic Editor: Hans Christiansen

Received: 11 October 2023
Revised: 30 November 2023
Accepted: 1 December 2023
Published: 7 December 2023

between human and AI-assisted, with the only exceptions of $D_{0.03cc}$ to the anal canal and D_{mean} to the brachial plexus. The observed average differences in plan quality for PCa and *HNC* cases were 8% and 6.7%, respectively. The dose parameter changes due to interobserver variability in PCa were small, with the exception of the anal canal, where large dose variations were observed. The reduction in time required for contouring was 72% for PCa and 84% for *HNC*. Conclusions: When an autocontouring system is used in combination with human review, the time of the RT workflow is significantly reduced without affecting dose distribution and plan quality.

Keywords: autocontouring; radiotherapy; artificial intelligence; time savings; dosimetry

1. Introduction

Radiation therapy (RT) is considered as an alternative to surgery for early-stage cancer, whereas locally advanced cancer is mostly treated in conjunction with surgery and systemic radiation therapies according to patient's age and comorbidities [1–3]. Improper delineation of the target volume and organs at risk (OARs) can affect the quality of dose distribution designed during planning of the RT treatment. As a consequence, inadequate target coverage or normal tissue sparing may occur, resulting in a reduced tumor control or an increased probability of side effects [4]. Traditionally, tumor volumes and OARs are manually contoured by radiation oncologists. This is a laborious procedure that is subject to both intra- and interobserver variability [5]. In this scenario, automatic contouring methods can minimize the clinical workload as well as improving reproducibility of RT. In the contouring workflow, the automatic contours depict a starting point, which is reviewed and, if necessary, manually edited before being sent to the treatment planning system.

Atlas-based contouring [6], statistical models of shape and appearance [7], artificial intelligence-based methods [8], and hybrid strategies are a few examples of the automated contouring techniques that have been introduced and developed with promising outcomes. The spread of artificial intelligence (*AI*) is impacting the workflow of RT treatment in several scenarios [9], and AI-based autocontouring software has been developed and made available to oncologists to optimize the contouring process [10]. A question that arises is whether the automated contours are of sufficient quality for clinical use, which can be answered only after effective validation, that is, evaluation of accuracy and reliability. The existing literature indicates that contour evaluation is performed mostly at the geometric level [11–13] using common geometric metrics, including moment-based methods, overlap metrics, and distance-based measures [14]. However, geometrical metrics alone do not necessarily reflect the actual clinical impact of the contour differences [11–13]. Treatment dosimetry, plan quality, and associated clinical decision-making processes are directly influenced by accuracy of contoured regions, and the impact of the geometric agreement into the dose domain and plan quality remains to be fully investigated [14–17].

Research Objectives

With the capability of automatically providing contours that can be used to generate clinically acceptable plans, commercial tools for automated segmentation can reduce treatment planning time substantially. The objective of this study was to investigate the accuracy of structure contours generated by commercial autocontouring software. Also, we wanted to investigate the dose distributions of the treatment plans generated from autocontoured structure sets.

2. Materials and Methods

2.1. Patient Data

After approval from the institutional review board of Centro di Riferimento Oncologico (CRO), 40 patients treated at CRO Aviano from September 2017 to June 2022 were selected retrospectively for this study. A total of 20 had been treated for prostate cancer (PCa) and

20 for head and neck cancer (*HNC*). The PCa patients' preparation before CT acquisition included full bladder and empty rectum. Patients with bilateral hip implants or rectal spacer were not included in the study. Patients with *HNC* cancer required no preparation but were immobilized with a thermoplastic mask. No contrast was administered for any patient before CT image acquisitions.

Patients CTs for planning of the treatment were acquired using a 90 cm wide bore Toshiba Aquilion 16 CT simulator with 5 mm slice thickness for PCa and 2 mm for *HNC*. Images were reconstructed using FC13 reconstruction algorithm having a 256 × 256 matrix.

2.2. Contouring Workflows

Target volume delineations are ruled by international guidelines and scientific associations recommendations. The contoured structures for PCa patients included the entire prostate and its capsule, which represent the clinical target volume (CTV) as well as the organs at risk. The planning target volume (*PTV*) was created by expanding CTV by 5 mm margin in all directions except 3 mm posterior. For *HNC*, organs at risk were contoured automatically, while the CTV was not automatically contoured.

Contoured structures, excluding the *PTV*, were generated for each patient using three methods, as follows:

- Manual contouring (C_{man}). Contours were delineated by a radiation oncologist with at least ten years of experience, also using semiautomated tools like flood fill and interpolation, within the integrated ARIA and Eclipse TPS systems (version: 16.1; Varian Medical Systems, Inc., NewYork, CA, USA) [18] and following the institutional guidelines [19–21]. These contours were assumed as the ground truth structures.
- Fully automated contouring based on artificial intelligence (C_{AI}). These were automatically created using a research version of Limbus Contour (version: 1.0.18; Limbus AI Inc., Regina, SK, Canada) [22] software. Limbus Contour (LC) employs organ-specific deep convolutional neural network models on the basis of a U-net architecture [23], which were trained on CT images from the Cancer Imaging Archive public database [24]. Following the creation of contours, LC applies a number of postprocessing techniques including as outlier removal, slice interpolation, z-plane cutoffs, and contour smoothing [23]. Contouring the structure set on a patient required up to 7 min on 3.2 GHz Intel Pentium CPU G3420; 4 GB RAM; 64-bit Operating System. The Limbus-generated RT structure sets were exported as DICOM files into the Varian Eclipse workstation for revision and validation.
- automatically generated contours reviewed and eventually adjusted by radiation oncologist ($C_{AI,adj}$). Expert ROs reviewed and, if necessary, modified the C_{AI} using the Eclipse contouring application in accordance with institutional consensus guidelines for target volume and OAR contours. In order to keep ROs blinded to the original contours' creation, only C_{AI} contours were visible to them during revision time.

For better reproducibility, manual contouring was performed by one radiation oncologist for each treated site and interobserver variability was measured between two operators for PCa (Section 2.9).

2.3. Treatment Planning and Delivery

The radiotherapy plans that had been previously delivered to the patients were assumed as the reference for dose comparison. Structure sets for these treatments were manually contoured by radiation oncologists (ROs) using the institutional protocol for Radiotherapy Oncology Group. Treatment planning was performed using dose prescription and constraints for planning shown in Tables 1 and 2.

Table 1. Dose constraints used for planning of PCa treatment.

Prostate (6000 cGy × 20 fx)	
PTV	D_{min} > 5700 cGy D_{max} < 6420 cGy
Rectum	$V_{4500\ cGy}$ < 15% $V_{2800\ cGy}$ < 35% $V_{900\ cGy}$ < 80%
Bladder	$V_{4500\ cGy}$ < 25% $V_{2800\ cGy}$ < 50%
Anal canal	$V_{4500\ cGy}$ < 15% $V_{2800\ cGy}$ < 35% $V_{900\ cGy}$ < 80%
Femoral head	$V_{3100\ cGy}$ < 1%

Table 2. Dose constraints used for planning of *HNC* treatment.

H&N (7095 cGy × 33 fx)	
PTV7095	D_{min} > 6953 cGy D_{max} < 7591 cGy
PTV6270	D_{min} > 5956 cGy
PTV5610	D_{min} > 5329 cGy
Brain	$V_{6000\ cGy}$ < 3%
Brainstem	D_{max} < 5000 cGy
Cochlea	D_{mean} < 4500 cGy
Spinal cord	D_{max} < 3600 cGy
PRV_Spinal cord	D_{max} < 4000 cGy
Oral cavity	D_{mean} < 3000 cGy $V_{3000\ cGy}$ < 73% $V_{4000\ cGy}$ < 20%
Ipsilateral parotid	D_{mean} < 2600 cGy
Contralateral parotid	D_{mean} < 2000 cGy
Ipsilateral submandibular gland	D_{mean} < 5000 cGy
Contralateral submandibular gland	D_{mean} < 3900 cGy
Mandible	$V_{5000\ cGy}$ < 31%
Arytenoid cartilage	$V_{5000\ cGy}$ < 50%
Constrictor muscle	$V_{5000\ cGy}$ < 70%
Constrictor muscle-PTV	D_{mean} < 5000 cGy $V_{5000\ cGy}$ < 31% $V_{5000\ cGy}$ < 31 cc
Thyroid	D_{mean} < 4500 cGy $V_{4000\ cGy}$ < 50% $V_{3000\ cGy}$ < 60%
Brachial plexus	$V_{6000\ cGy}$ < 0.1 cc
Esophagus	$V_{3500\ cGy}$ < 50% $V_{5000\ cGy}$ < 40% $V_{7000\ cGy}$ < 20%

PCa treatments were delivered using the volumetric modulated arc therapy (VMAT) technique with one or two 18 MV full coplanar arcs, 600 MU/min maximum dose rate, and a prescribed dose of 60 Gy in 20 fractions each of 3 Gy. *HNC* patients received intensity-modulated radiation therapy (IMRT) treatments with nine 6 MV photon beam fields, a maximum dose rate of 300 monitor units (MU) per minute, and a prescribed dose of 70.95 Gy in 33 fractions of 2.15 Gy each. These plans were generated using the Eclipse planning system (Varian Medical). Dose calculations were performed using the anisotropic analytical algorithm (AAA) with a grid resolution of 2.5 mm [25]. The treatment schedule consisted of 5 daily fractions per week. The treatments were administered using a Varian TrueBeam or Trilogy linear accelerator. A cone-beam computed tomography image was acquired at the beginning of each treatment session for image-guided RT [26]. For evaluation of dose difference due to autocontouring in the planning workflow, treatment plans were exactly same for the structure set $C_{AI,adj}$ following the same planning and optimization procedure as for the plans clinically used. Plans were exported for analysis in RT DICOM formats from the treatment planning system. DICOM files were transferred to a high-performance computer interface for analysis with homemade MATLAB scripts.

2.4. Qualitative Assessment of Automated Contouring

An experienced clinician assessed the C_{AI} for each patient using a four-point Likert scale, shown in Table 3, to evaluate qualitatively the automated contouring process. As such a test aims at distinguishing between *AI* and human operator, this is sometimes referred to as the Turing test [27,28].

Table 3. Scoring values for qualitative assessment of AI-generated contours.

Likert Scale for Each Patient		
• Severe correction	: Require correction	→ Large and obvious errors
• Medium correction	: Require correction	→ Minor errors that need a small amount of editing
• Slight correction	: Accepted	→ Minor errors, but these are clinically not significant
• No correction	: Accepted	→ Contour is very precise

2.5. Geometric Evaluation

For target and OAR structures, comparisons between C_{man} and C_{AI} contours before and after the physician review ($C_{AI,adj}$) were compared with these metrics described herein. For comparing AI- versus human-generated contours, we used different types of geometrical metrics that are based on distance between surfaces, size of overlapping volumes, and difference in size [29].

Dice similarity coefficient (*DSC*) provides a measure of the volumetric overlap of two contours of a structure with a score range from 0 (no overlay) to 1 (total overlay) [30]:

$$DSC\left(C_{man}, C_{AI,adj}\right) = \frac{2\left|C_{man} \cap C_{AI,adj}\right|}{|C_{man}| + \left|C_{AI,adj}\right|} \quad (1)$$

Hausdorff distance (*HD*) is a bidirectional measure of distance between contour surfaces [30]. This metric calculates the distance to the closest point in both directions, from contour C_{man} to contour $C_{AI,adj}$ and vice versa, to figure out the largest surface-to-surface separation between two contours.

$$HD\left(C_{man}, C_{AI,adj}\right) = max\left[h\left(C_{man}, C_{AI,adj}\right), h\left(C_{AI,adj}, C_{man}\right)\right] \quad (2)$$

where $h\ (C_{man}, C_{AI,adj})$ represents the Euclidean distance between a and b voxels corresponding to the C_{man} and $C_{AI}/C_{AI,adj}$ contours, respectively, and the formula is:

$$h\left(C_{man}, C_{AI,adj}\right) = max_{a \in C_{man}} min_{b \in C_{AI,adj}} ||a - b|| \quad (3)$$

Relative volume difference (*RVD*), also known as relative absolute volume difference, describes the size difference between the regions:

$$RVD = \left| \frac{\left|V_{AI,adj}\right| - |V_{man}|}{|V_{man}|} \right| \quad (4)$$

where V_{man} and $V_{AI,adj}$ represents the absolute volume corresponding to the C_{man} and $C_{AI,adj}$ contours, respectively.

2.6. Evaluation of Dose Differences

To assess the potential impact of *AI* on dosimetry, we calculated the difference in dose indexes among plans as:

$$\Delta D_X = \frac{(D_X)_{AI,adj} - (D_X)_{man}}{(D_X)_{man}} \quad (5)$$

where $(D_X)_{man}$ and $(D_X)_{Ai,adj}$ referred to dose parameters for C_{man} and $C_{AI,adj}$ contours, respectively. And X represents the dose metrics such as D_{min}, D_{mean}, and $D_{0.03cc}$.

Dose distribution to the organs-at-risk (OAR) doses were evaluated using D_{mean} (mean dose) and the highest dose encompassing $0.03cc$, $D_{0.03cc}$ [31].

Homogeneity Index (*HI*) was utilized to evaluate the dose uniformity within the PTV. *HI* was assessed using the following formula [32]:

$$HI = \frac{(D_{2\%} - D_{98\%})}{D_{50\%}} \quad (6)$$

Where $D_{2\%}$ (near maximum dose), $D_{98\%}$ (near minimum dose), and $D_{50\%}$ represent the minimum dose covering 2%, 98%, and 50% of the target volume, respectively. Formula for *HI* comparison between plans for C_{man} and $C_{AI,adj}$ contours was-

$$\Delta HI = \frac{\left|HI_{AI,adj} - HI_{man}\right|}{|HI_{man}|} \quad (7)$$

where HI_{man} and $HI_{AI,adj}$ represent the *HI* for C_{man} and $C_{AI,adj}$ contours, respectively.

Conformity Index (*CI*) was used to obtain a quantitative evaluation of the *PTV* coverage by the prescribed dose. *CI* was evaluated using the following equation [33]:

$$CI = \frac{(TV_{PV})^2}{V_{PTV} \times V_{TV}} \quad (8)$$

where V_{TV} indicates the volume that receives 95% of the prescribed dose, V_{PTV} represents the *PTV* volume, and TV_{PV} is the *PTV* volume inside the V_{TV}. *CI* comparison between C_{man} and $C_{AI,adj}$ contours plan was carried out by using the formula below:

$$\Delta CI = \frac{\left|CI_{AI,adj} - CI_{man}\right|}{|CI_{man}|} \quad (9)$$

where CI_{man} and $CI_{AI,adj}$ indicates the C_{man} and $C_{AI,adj}$ contours *CI*, respectively.

2.7. Normalized Plan Quality Metric

The plan quality metric (*PQM*) framework was designed to establish a standardized approach for assessing how well a particular treatment plan achieves specific dose volume objectives that serve as a hypothetical "virtual physician" [34]. A *PQM* scorecard is often created for every objective which assigns a score based on how effectively the objective is achieved by a particular plan. To enable meaningful comparisons across our study cases, we utilized the normalized *PQM* (*nPQM*) score, which divides the *PQM* score by the peak score achievable by the plan of a certain district (PQM_{max}) and scales to the percentage. The formula used for the normalized plan quality metric was:

$$nPQM = \frac{PQM}{PQM_{max}} \times 100 \quad (10)$$

The *PQM* scorecard to be used for analysis in this trial is shown in Tables 4 and 5. To calculate the score for a particular objective, there are two different types of functions: threshold and linear score. The threshold score's function awards no points if the objective is not achieved and the maximum number of points for the accomplished objective. The linear score's function makes use of two thresholds. Maximum points are awarded if the plan satisfies the constraint's "ideal threshold" and no points are assigned if it does not exceed the constraint's "minimally acceptable threshold". Using the value of the dose-volume statistic, linear interpolation between the two thresholds is used to calculate the number of scores awarded if the objective is between the two thresholds.

Table 4. *PQM* of PCa treatment plans.

Structure	Constraint	Function	Thresholds	Max Score
PTV	D99%	Linear	>6000 cGY (100%)	5
			>5700 cGy (95%)	4
	D0.1 cc	Linear	<6300 cGy(105%)	5
			<6420 cGy(107%)	4
Rectum	V4500 cGy	Threshold	<15%	3
	V2800 cGy	Threshold	<35%	3
	V900 cGy	Threshold	<80%	3
Bladder	V4500 cGy	Threshold	<25%	3
	V2800 cGy	Threshold	<50%	3
Anal canal	V4500 cGy	Threshold	<15%	3
	V2800 cGy	Threshold	<35%	3
	V900 cGy	Threshold	<80%	3
Femur Right	V3500 cGy	Threshold	<1%	2
Femur Left	V3500 cGy	Threshold	<1%	2

2.8. Evaluation of Contouring Time

The amount of time required for contouring was measured in order to estimate the increase in performance made possible with autocontouring. Since in clinical practice AI-generated contours need always to be reviewed and eventually modified by a radiation oncologist, we measured the reduction in contouring time from autocontouring as

$$\text{Reduction in time} = \frac{T(C_{man}) - T(C_{AI,adj})}{T(C_{man})} \quad (11)$$

Table 5. *PQM* of *HNC* treatment plans.

Structure	Constraint	Function	Thresholds	Max Score
BrachialPlexus_C	D_{max}	Threshold	<6000 cGy	3
BrachialPlexus_O	D_{max}	Threshold	<6270 cGy	3
Brain	$V_{6000\ cGy}$	Threshold	<3%	3
Brainstem	D_{max}	Threshold	<5000 cGy	5
Chiasm	D_{max}	Threshold	<5400 cGy	5
Cochlea_Contralateral	D_{max}	Threshold	<1000 cGy	2
Cochlea_Ipsilateral	D_{max}	Threshold	<3500 cGy	2
Esophagus	$V_{3500\ cGy}$	Threshold	<50%	1
	$V_{5000\ cGy}$	Threshold	<40%	1
	$V_{7000\ cGy}$	Threshold	<20%	1
Eye_L	D_{max}	Threshold	<4500 cGy	5
Eye_R	D_{max}	Threshold	<4500 cGy	5
Larynx	D_{mean}	Threshold	<4000 cGy	2
	$V_{5000\ cGy}$	Threshold	<27%	2
Lens_L	D_{max}	Threshold	<400 cGy	4
Lens_R	D_{max}	Threshold	<400 cGy	4
Mandible	$V_{5000\ cGy}$	Threshold	<31%	3
OpticNerve_L	D_{max}	Threshold	<5400 cGy	5
OpticNerve_R	D_{max}	Threshold	<5400 cGy	5
Oral Cavity	D_{mean}	Threshold	<3000 cGy	3
	$V_{3000\ cGy}$	Threshold	<73%	2
	$V_{4000\ cGy}$	Threshold	<20%	2
Parotid_Contralateral	D_{mean}	Threshold	<2000 cGy	2
	D_{median}	Threshold	<2000 cGy	4
Parotid_Ipsilatateral	D_{mean}	Threshold	<2600 cGy	2
	D_{median}	Threshold	<2600 cGy	4
PharynxConst	$V_{5000\ cGy}$	Threshold	<70%	2
Pituitary	D_{max}	Threshold	<5000 cGy	5
SpinalCord	D_{max}	Threshold	<4000 cGy	5
Submandibular_Co	D_{mean}	Threshold	<3900 cGy	3
Submandibular_Ho	D_{mean}	Threshold	<5000 cGy	3
Thyroid	D_{mean}	Threshold	<4500 cGy	1
	$V_{4000\ cGy}$	Threshold	<50%	1
	$V_{3000\ cGy}$	Threshold	<60%	1

2.9. Interobserver Variability

The interobserver variability was assessed by comparing autogenerated structures reviewed and adjusted by two different operators. The geometric differences were calculated by assessing *DSC*, *HD*, and *RVD* among AI-assisted contours performed by the two operators:

$$DSC(1,2) = \frac{2\left|C_{AI,adj1} \cap C_{AI,adj2}\right|}{\left|C_{AI,adj1}\right| + \left|C_{AI,adj2}\right|} \quad (12)$$

where $C_{AI,adj1}$ is the contour generated by *AI* and adjusted by operator 1.

Dosimetric evaluation was performed using the same methods previously described. For instance, the interobserver variability for D_{min} to an organ at risk was calculated as

$$\Delta D_{min} = \frac{|D_{min,1} - D_{min,2}|}{|D_{min,2}|} \quad (13)$$

where $D_{min,1}$ and $D_{min,2}$ are the minimum doses to an organ at risk generated using *AI* and adjusted by operators 1 and 2, respectively.

2.10. Data Analysis

For geometrical and dosimetric evaluation, we developed an in-house script in MATLAB version R2021a (The MathWorks, Inc, Boston, MA, USA) [35] to compare structure sets and treatment plans for both the automatic and the manually edited contours as shown in Figure 1. Wilcoxon rank sum test was employed to perform dosimetric comparisons to determine if there were any significant differences between the individual OARs in each arm in terms of the D_{min}, D_{mean}, and $D_{0.03cc}$ doses based on the reference dose distribution, and significant differences for the *PTV* doses in terms of *HI* and *CI* were assessed with the alpha (α) value 0.05 for 95% *CI*.

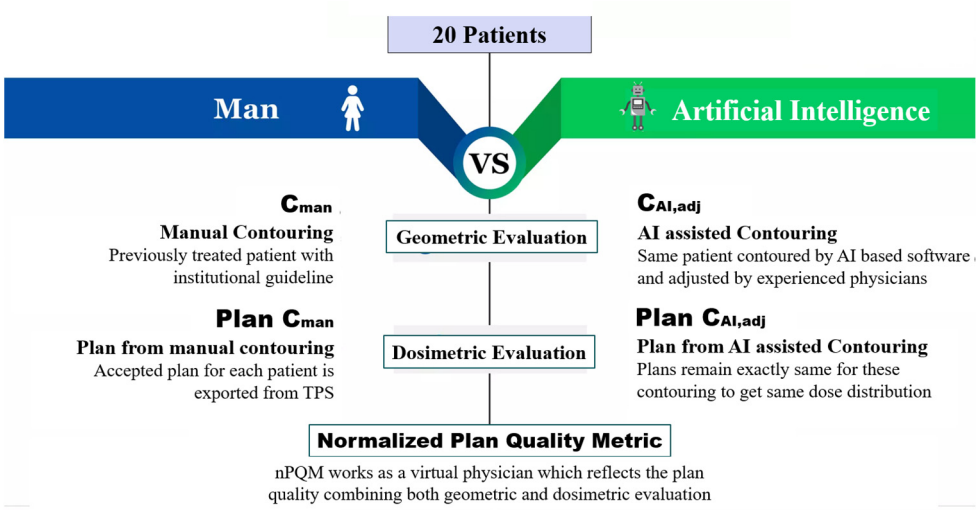


Figure 1. Comparison procedures for contouring evaluation.

3. Results

3.1. Qualitative Assessment of Automated Contouring

Results of the quality assessment for the PCa and *HNC* contours are shown in Figures 2 and 3, respectively.

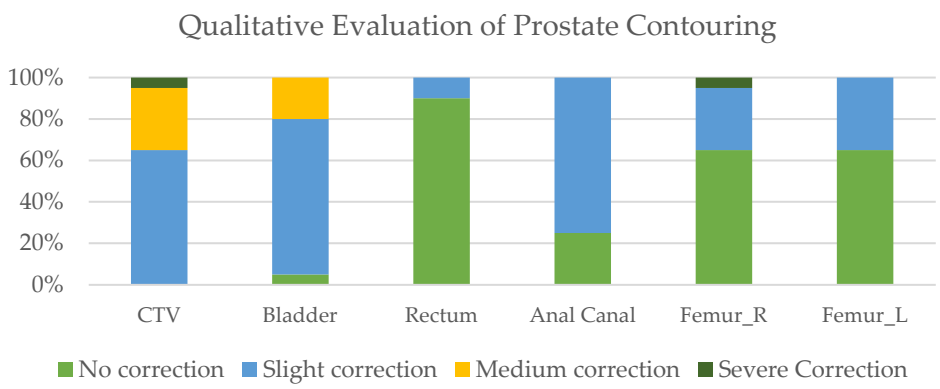


Figure 2. Evaluation of AI-based contouring for PCa.

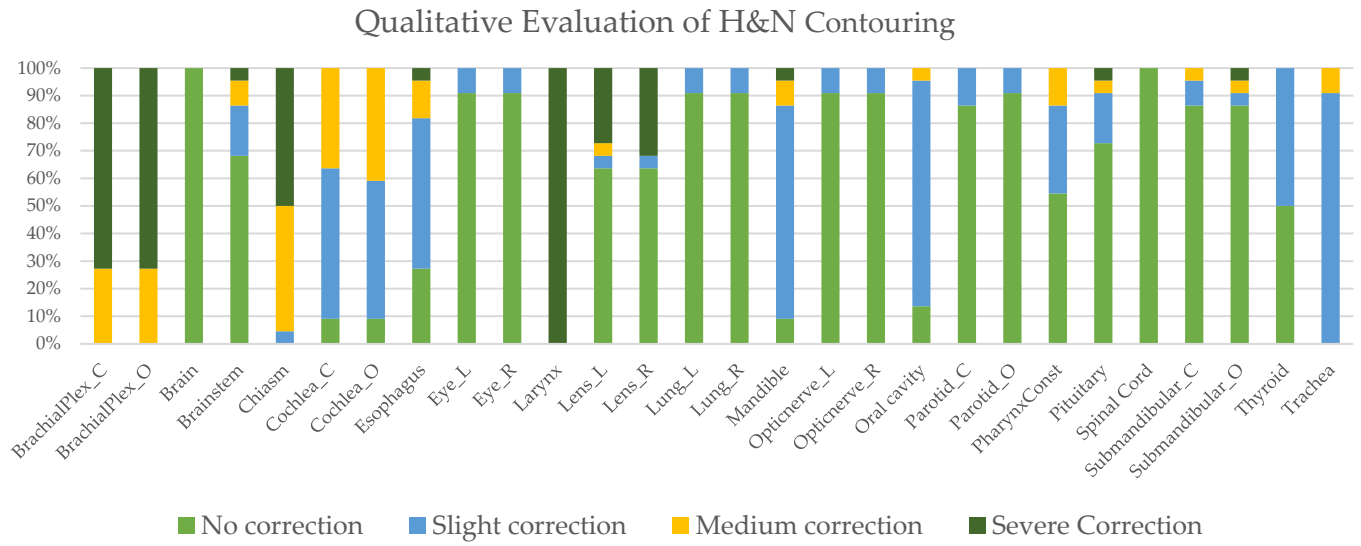


Figure 3. Physician assessment of AI-based contouring for *HNC*.

3.2. Geometric Comparison

Tables 6 and 7 show the differences among structures contoured with different modalities. The highest average *DSC* values were observed for the bladder and rectum, followed by the anal canal and prostate. The values of the average *HD* were 4.19 mm, 2.85 mm, and 1.08 mm for prostate, bladder, and rectum, respectively. The values of the *RVD* showed the same trend: 0.08, 0.02, 0.01 for prostate, bladder, and rectum, respectively.

Table 6. Summary of *DSC*, *HD*, and *RVD* values measured before and after physician contours, (C_{AI} vs. $C_{AI,adj}$) for PCa cases.

	DSC		***HD* (mm)**		***RVD***	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**Mean (SD)**	**Median (Range)**
CTV	0.93 (0.07)	0.96 (0.68–0.99)	4.19 (3.31)	3.00 (1.0–11.09)	0.08 (0.11)	0.03 (0.0–0.46)
Rectum	0.99 (0.00)	0.99 (0.99–1.0)	1.08 (0.20)	1.00 (1.0–1.73)	0.01 (0.0)	0.01 (0.0–0.02)
Bladder	0.99 (0.01)	0.99 (0.97–1.0)	2.85 (2.54)	1.21 (1.0–9.22)	0.02 (0.02)	0.02 (0.0–0.06)
Anal Canal	0.91 (0.06)	0.89 (0.81–1.0)	2.44 (1.36)	2.00 (1.0–6.08)	0.16 (0.11)	0.20 (0.0–0.32)
Left Femur	0.98 (0.02)	1.00 (0.94–1.0)	2.23 (1.80)	1.00 (1.0–5.66)	0.03 (0.04)	0.00 (0.0–0.11)
Right Femur	0.97 (0.08)	1.00 (0.62–1.0)	3.34 (6.46)	1.00 (1.0–30.15)	0.03 (0.04)	0.00 (0.0–0.15)

Table 7. Summary of geometric difference metrics measured between C_{man} and $C_{AI,adj}$ contours measured with different metrics for PCa cases.

	DSC		***HD* (mm)**		***RVD***	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**Mean (SD)**	**Median (Range)**
PTV	0.80 (0.07)	0.80 (0.67–0.91)	15.38 (9.20)	15.39 (3.16–37.7)	0.12 (0.10)	0.10 (0.01–0.35)
Rectum	0.83 (0.07)	0.85 (0.60–0.89)	14.20 (20.2)	8.62 (2.24–96.30)	0.15 (0.17)	0.09 (0.01–0.64)
Bladder	0.94 (0.01)	0.95 (0.90–0.97)	4.11 (2.49)	3.39 (2.0–13.19)	0.03 (0.03)	0.01 (0.0–0.12)
Anal Canal	0.70 (0.08)	0.70 (0.54–0.90)	5.43 (2.26)	4.47 (3.0–11.40)	0.33 (0.27)	0.30 (0.03–1.29)
Left Femur	0.78 (0.16)	0.83 (0.47–0.93)	18.06 (16.4)	12.64 (3.16–54.0)	0.50 (0.69)	0.18 (0.02–1.98)
Right Femur	0.78 (0.16)	0.84 (0.45–0.94)	17.98 (16.8)	13.01 (2.24–54.3)	0.49 (0.65)	0.20 (0.02–1.78)

Figure 4 shows *DSC*, *HD*, and *RVD* scores for PCa cases. For PCa, large variabilities in terms of *DSC* and *RVD* were observed for anal canal and both femur heads in comparison between C_{man} and $C_{AI,adj}$ contouring. As for *HD* values, a wide range of values was reported for femur heads.

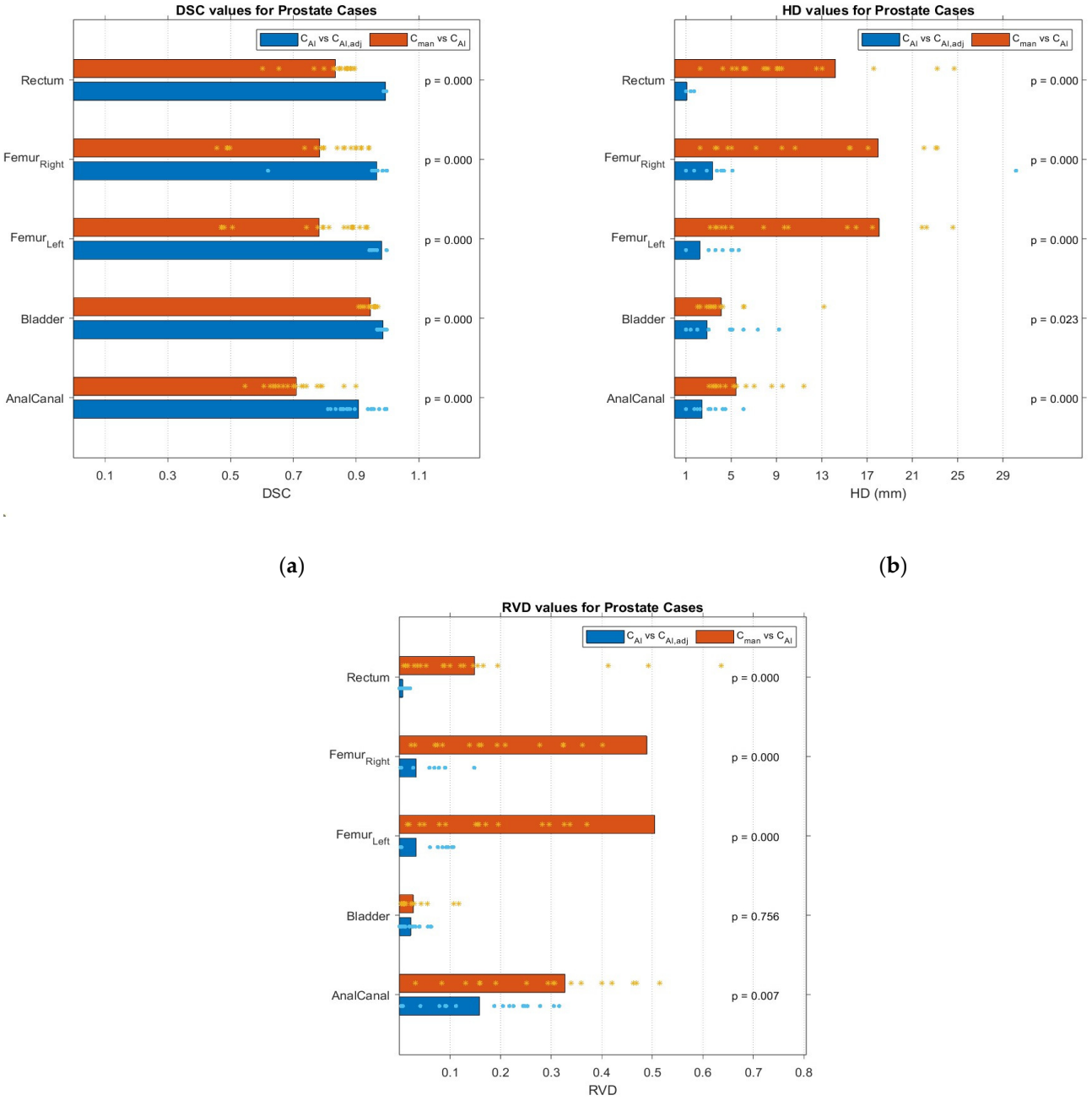


Figure 4. Geometric evaluation results: (**a**) *DSC*, (**b**) *HD* in mm, and (**c**) *RVD*, for $C_{AI,adj}$ contours in comparison with both C_{AI} and C_{man} contours of PCa cases. Each * represents a value.

Table 8 provides a complete list of *DSC*, *HD*, and *RVD* values for *HNC* contours. The brain, mandible, parotids, and thyroid showed a high level of correlation with average *DSC* scores of 1.00, 0.98, 0.99, and 0.94, and average *HD* scores of 0.65 mm, 8.13 mm, 1.50 mm, and 9.58 mm, respectively, between the contours of before (C_{AI}) and after physician review ($C_{AI,adj}$). *RVD* values were generally close to 0.

Table 8. Summary of *DSC*, *HD*, and *RVD* values measured before and after physician contours (C_{AI} vs. $C_{AI,adj}$) for *HNC* cases.

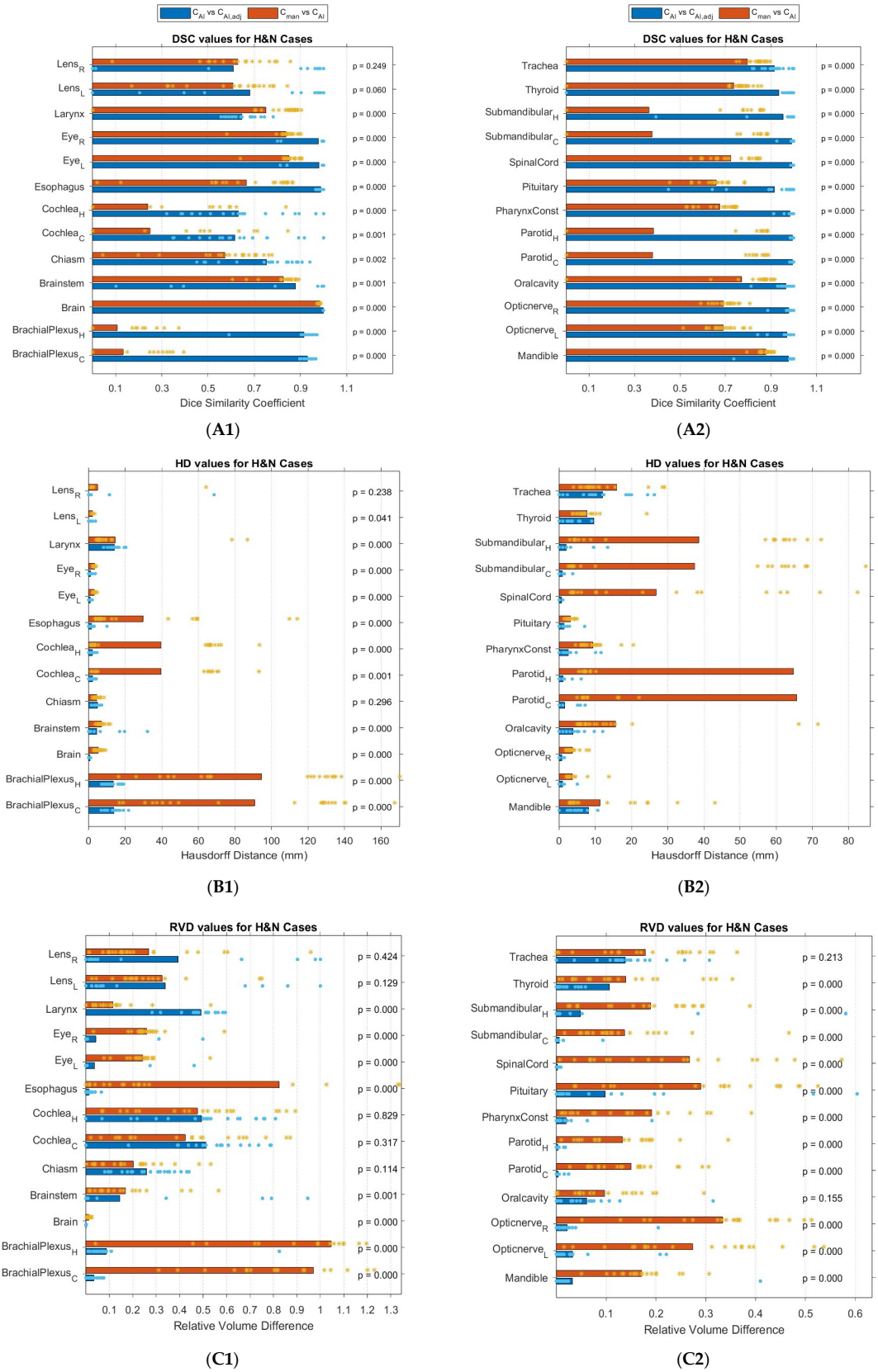
	DSC		***HD* (mm)**		***RVD***	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Contralateral Brachial Plexus	0.93 (0.02)	0.93 (0.90–0.97)	13.87 (3.80)	14.59 (6.78–21.8)	0.03 (0.02)	0.03 (0.00–0.08)
Ipsilateral Brachial Plexus	0.92 (0.08)	0.93 (0.59–0.97)	13.52 (4.12)	13.28 (7.07–19.1)	0.09 (0.18)	0.05 (0.01–0.82)
Brain	1.00 (0.00)	1.00 (1.00–1.00)	00.65 (0.49)	01.00 (0.00–1.00)	0.00 (0.00)	0.00 (0.00–0.00)
Brainstem	0.88 (0.27)	1.00 (0.10–1.00)	04.35 (8.54)	1.00 (0.00–32.16)	0.15 (0.31)	0.00 (0.00–0.95)
Chiasm	0.75 (0.15)	0.78 (0.45–0.94)	04.82 (1.40)	05.10 (1.41–7.00)	0.26 (0.12)	0.26 (0.08–0.44)
Contralateral Cochlea	0.62 (0.19)	0.60 (0.35–1.00)	02.21 (1.06)	02.24 (0.00–4.12)	0.51 (0.23)	0.57 (0.00–0.79)
Ipsilateral Cochlea	0.63 (0.20)	0.63 (0.32–1.00)	02.03 (0.96)	01.87 (0.00–4.58)	0.49 (0.24)	0.52 (0.00–0.81)
Esophagus	0.99 (0.01)	0.99 (0.97–1.00)	01.67 (2.12)	01.00 (0.00–10.0)	0.01 (0.02)	0.01 (0.00–0.07)
Left Eye	0.98 (0.05)	0.99 (0.81–1.00)	00.85 (0.59)	01.00 (0.00–2.00)	0.04 (0.12)	0.00 (0.00–0.46)
Right Eye	0.98 (0.06)	0.99 (0.80–1.00)	00.93 (0.82)	01.00 (0.00–3.61)	0.04 (0.13)	0.00 (0.00–0.50)
Larynx	0.65 (0.06)	0.63 (0.56–0.78)	14.11 (2.79)	13.83 (8.12–20.3)	0.49 (0.08)	0.51 (0.28–0.60)
Left Lens	0.68 (0.42)	0.93 (0.00–1.00)	01.00 (1.00)	01.00 (0.00–3.74)	0.34 (0.43)	0.06 (0.00–1.00)
Right Lens	0.61 (0.47)	0.95 (0.00–1.00)	5.96 (17.54)	01.00 (0.00–68.6)	0.39 (0.47)	0.05 (0.00–1.00)
Mandible	0.98 (0.06)	0.99 (0.74–1.00)	8.13 (19.10)	3.32 (0.00–88.21)	0.03 (0.09)	0.01 (0.00–0.41)
Left Optic Nerve	0.97 (0.04)	0.98 (0.84–1.00)	00.99 (1.06)	01.00 (0.00–5.00)	0.03 (0.06)	0.01 (0.00–0.22)
Right Optic Nerve	0.98 (0.03)	0.98 (0.89–1.00)	00.72 (0.49)	01.00 (0.00–1.41)	0.02 (0.04)	0.01 (0.00–0.20)
Oral Cavity	0.97 (0.04)	0.98 (0.81–1.00)	03.82 (3.07)	3.50 (0.00–12.04)	0.06 (0.07)	0.04 (0.00–0.31)
Contralateral Parotid	0.99 (0.00)	0.99 (0.98–1.00)	01.50 (2.02)	01.00 (0.00–7.14)	0.00 (0.01)	0.00 (0.00–0.02)
Ipsilateral Parotid	0.99 (0.00)	0.99 (0.98–1.00)	01.13 (1.43)	01.00 (0.00–6.08)	0.00 (0.00)	0.00 (0.00–0.02)
Pharynx Constrictor Muscle	0.98 (0.02)	0.99 (0.91–1.00)	02.57 (3.03)	1.21 (0.00–11.58)	0.02 (0.04)	0.01 (0.00–0.19)
Pituitary	0.92 (0.15)	0.97 (0.45–1.00)	01.33 (1.50)	01.00 (0.00–7.07)	0.10 (0.17)	0.01 (0.00–0.60)
Spinal Cord	0.99 (0.01)	0.99 (0.99–1.00)	00.65 (0.49)	01.00 (0.00–1.00)	0.00 (0.00)	0.00 (0.00–0.01)
Contralateral Submandibular	0.99 (0.02)	0.99 (0.93–1.00)	00.86 (0.84)	01.00 (0.00–3.74)	0.01 (0.02)	0.00 (0.00–0.09)
Ipsilateral Submandibular	0.95 (0.14)	0.99 (0.39–1.00)	01.91 (3.42)	1.00 (0.00–13.42)	0.05 (0.14)	0.00 (0.00–0.58)
Thyroid	0.94 (0.22)	0.99 (0.00–1.00)	9.58 (31.67)	01.62 (0.0–143.8)	0.11 (0.41)	0.01 (0.00–1.84)
Trachea	0.92 (0.05)	0.92 (0.82–1.00)	12.04 (7.25)	11.05 (0.0–26.29)	0.14 (0.08)	0.14 (0.00–0.31)

Figure 5 shows the geometric evaluation results of C_{AI} contour and C_{man} contour both compared with $C_{AI,adj}$ contours. Autocontouring resulted in similar results for the brain, brainstem, mandible, and eyes (*DSC* > 0.83). For the brachial plexuses, parotids, cochlea and submandibular glands, there was a significant difference between the C_{man} contour and $C_{AI,adj}$ contour, while other OARs had better performance in terms of *DSC*, *HD*, and *RVD*.

Table 9 summarizes the *DSC*, *HD*, and *RVD* values calculated between the C_{man} and $C_{AI,adj}$ contours. The worst metrics were found in smaller structures such as lenses, while larger structures including the brain, mandible, eyes, and trachea showed a high level of correlation, with average *DSC* around 80% and lower *HD* and *RVD* values.

3.3. PTV Evaluation

Figure 6 shows the geometric and dosimetric comparison of the C_{man} plan compared with $C_{AI,adj}$ for prostate *PTV* in terms of *DSC* and *RVD* scores and differences in *HI* and *CI*.



(**A1**) (**A2**)

(**B1**) (**B2**)

(**C1**) (**C2**)

Figure 5. Geometric evaluation results: (**A1**,**A2**) *DSC*, (**B1**,**B2**) *HD* in mm, and (**C1**,**C2**) *RVD*, for $C_{AI,adj}$ contours in comparison with both C_{AI} and C_{man} contour of *HNC* cases. Each * represents a value.

Table 9. Summary of geometric difference metrics measured between C_{man} and $C_{AI,adj}$ contours measured with different metrics for *HNC* cases.

	DSC		*HD* (mm)		*RVD*	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Contralateral Brachial Plexus	0.13 (0.16)	0.00 (0.00–0.39)	90.86 (50.80)	119.9 (16.9–167)	0.97 (0.46)	0.87 (0.31–2.08)
Ipsilateral Brachial Plexus	0.11 (0.13)	0.00 (0.00–0.37)	94.65 (46.10)	120.6 (16.4–170)	1.05 (0.43)	1.07 (0.04–1.80)
Brain	0.98 (0.00)	0.99 (0.97–0.99)	05.20 (01.90)	04.79 (2.80–9.22)	0.01 (0.01)	00.01 (0.0–0.03)
Brainstem	0.83 (0.08)	0.85 (0.60–0.89)	06.99 (02.30)	6.44 (3.50–11.31)	0.17 (0.15)	0.12 (0.01–0.57)
Chiasm	0.57 (0.19)	0.63 (0.04–0.78)	04.50 (01.70)	04.12 (2.20–8.31)	0.20 (0.15)	00.20 (0.0–0.53)
Contralateral Cochlea	0.25 (0.31)	0.00 (0.00–0.85)	39.55 (34.70)	64.07 (1.4–93.09)	0.42 (0.27)	0.45 (0.02–0.87)
Ipsilateral Cochlea	0.24 (0.29)	0.00 (0.00–0.84)	39.66 (34.90)	64.55 (1.00–93.3)	0.48 (0.26)	0.51 (0.07–0.89)
Esophagus	0.67 (0.24)	0.79 (0.02–0.87)	29.76 (35.40)	8.33 (3.60–114.2)	5.62 (21.33)	0.22 (0.02–95.9)
Left Eye	0.85 (0.06)	0.85 (0.64–0.91)	02.97 (00.60)	03.00 (2.00–4.58)	0.24 (0.09)	0.25 (0.08–0.53)
Right Eye	0.84 (0.07)	0.85 (0.58–0.90)	03.17 (00.60)	03.08 (2.40–4.24)	0.26 (0.10)	0.25 (0.03–0.59)
Larynx	0.75 (0.26)	0.85 (0.00–0.90)	14.50 (23.50)	06.20 (4.0–86.98)	11.93 (36.3)	0.11 (0.03–121)
Lens_L	0.61 (0.19)	0.69 (0.17–0.84)	02.11 (00.60)	02.00 (1.40–3.16)	0.33 (0.32)	0.22 (0.03–1.41)
Lens_R	0.63 (0.17)	0.65 (0.08–0.86)	04.87 (14.00)	01.73 (1.0–64.33)	0.27 (0.23)	0.18 (0.02–0.96)
Mandible	0.88 (0.03)	0.88 (0.79–0.91)	11.29 (12.10)	04.12 (2.8–43.12)	0.17 (0.06)	0.17 (0.05–0.31)
Left Optic Nerve	0.69 (0.07)	0.69 (0.51–0.81)	03.68 (0270)	02.83 (2.0–13.78)	0.27 (0.15)	0.27 (0.02–0.54)
Right Optic Nerve	0.69 (0.06)	0.70 (0.59–0.81)	03.73 (01.90)	03.08 (1.70–8.31)	0.33 (0.12)	0.36 (0.05–0.51)
Oral Cavity	0.77 (0.27)	0.87 (0.00–0.92)	15.56 (18.70)	9.09 (5.0–71.510)	4.48 (15.36)	0.12 (0.02–66.7)
Contralateral Parotid	0.38 (0.43)	0.00 (0.00–0.89)	65.68 (52.30)	94.79 (4.9–147.8)	0.15 (0.08)	0.15 (0.03–0.31)
Ipsilateral Parotid	0.38 (0.44)	0.00 (0.00–0.89)	64.74 (54.10)	96.12 (5.4–148.1)	0.13 (0.08)	0.14 (0.01–0.35)
Pharynx Constrictor Muscle	0.67 (0.08)	0.71 (0.53–0.75)	09.34 (03.80)	9.06 (4.60–20.45)	0.19 (0.24)	00.14 (0.01–1.1)
Pituitary	0.66 (0.08)	0.67 (0.45–0.79)	03.08 (00.90)	03.00 (1.40–5.00)	0.29 (0.17)	0.33 (0.04–0.52)
Spinal Cord	0.72 (0.10)	0.71 (0.55–0.85)	26.80 (26.90)	11.63 (3.2–82.49)	0.27 (0.22)	0.23 (0.01–0.91)
Contralateral Submandibular	0.38 (0.43)	0.00 (0.00–0.89)	37.42 (31.20)	56.35 (2.5–84.73)	0.14 (0.10)	0.11 (0.02–0.47)
Ipsilateral Submandibular	0.36 (0.41)	0.00 (0.00–0.87)	38.61 (31.00)	58.19 (3.0–88.42)	0.19 (0.10)	0.20 (0.02–0.39)
Thyroid	0.74 (0.18)	0.77 (0.00–0.86)	07.68 (04.47)	06.63 (3.6–24.19)	0.14 (0.11)	0.11 (0.01–0.35)
Trachea	0.80 (0.19)	0.84 (0.00–0.90)	15.95 (18.90)	09.43 (4.0–90.63)	7.20 (31.34)	0.22 (0.0–140.3)

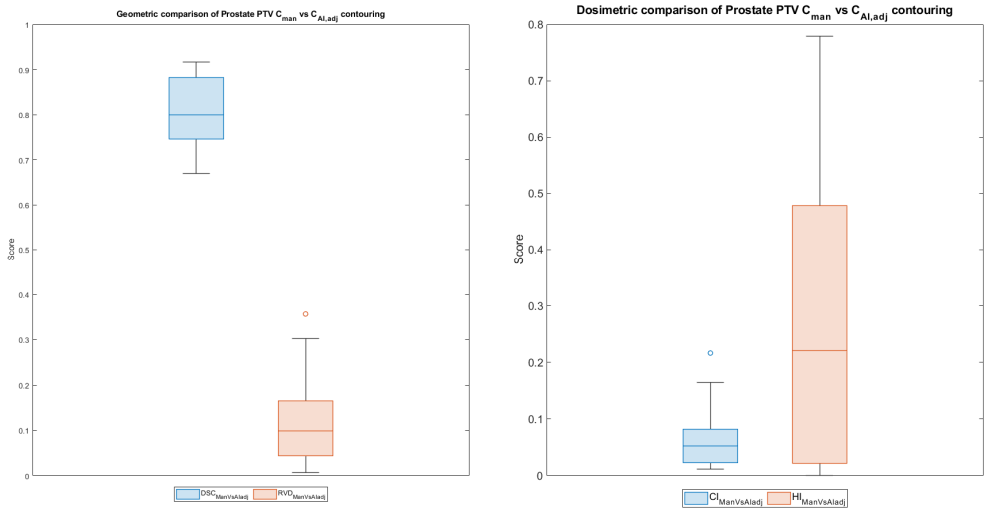
3.4. Dosimetric Comparison

Differences in D_{min}, D_{mean}, and $D_{0.03cc}$ between manual and AI-assisted are shown in Figure 7.

The quantitative results of the dosimetric comparisons of plans with the C_{man} plan compared with $C_{AI,adj}$ contours are summarized in Table 10. No significant dose differences were measured between manual and autocontour workflows, except the anal canal for PCa cases.

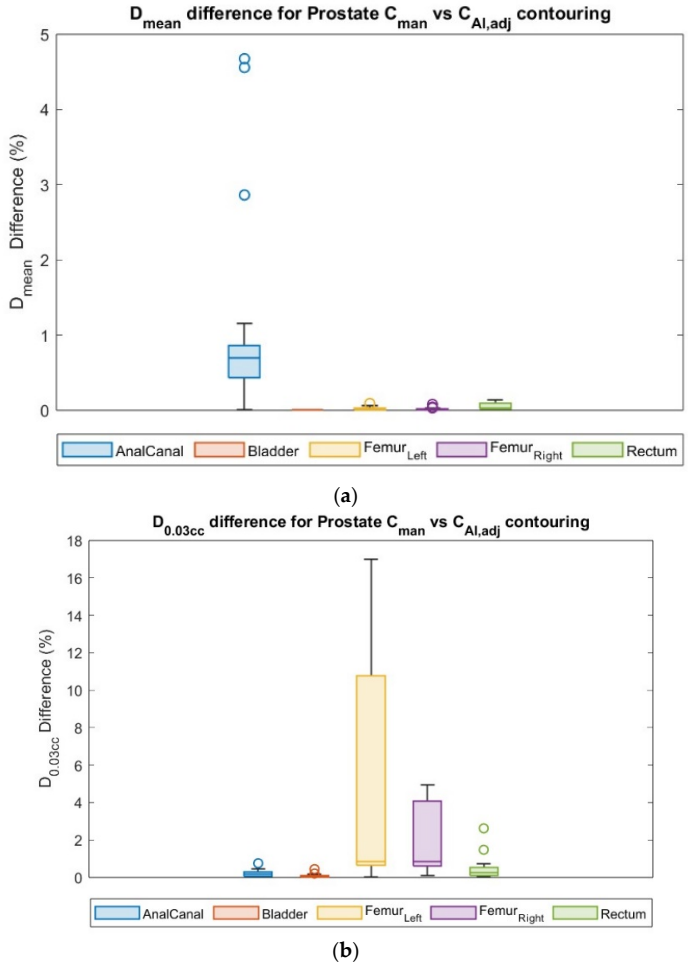
Table 10. Relative differences in D_{mean} and $D_{0.03cc}$ values measured between C_{man} and $C_{AI,adj}$ contours for PCa cases.

	ΔD_{mean}		$\Delta D_{0.03cc}$	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Rectum	0.13 (0.17)	0.07 (0.0–0.73)	0.05 (0.05)	0.03 (0.0–0.13)
Bladder	0.03 (0.02)	0.02 (0.0–0.07)	0.00 (0.00)	0.00 (0.0–0.01)
Anal canal	0.49 (0.28)	0.48 (0.03–1.22)	1.11 (1.33)	0.69 (0.0–4.67)
Femur Left	0.21 (0.16)	0.19 (0.0–0.53)	0.02 (0.02)	0.01 (0.0–0.09)
Femur Right	0.21 (0.16)	0.21 (0.0–0.54)	0.01 (0.02)	0.01 (0.0–0.08)



(**a**) (**b**)

Figure 6. (**a**) Geometric evaluation by *DSC* and *RVD* and (**b**) dosimetric evaluation in terms of homogeneity index and conformity index of prostate PTV. Each circle symbol represents a value outside the standard deviation. Each circle symbol represents a value outside the standard deviation.

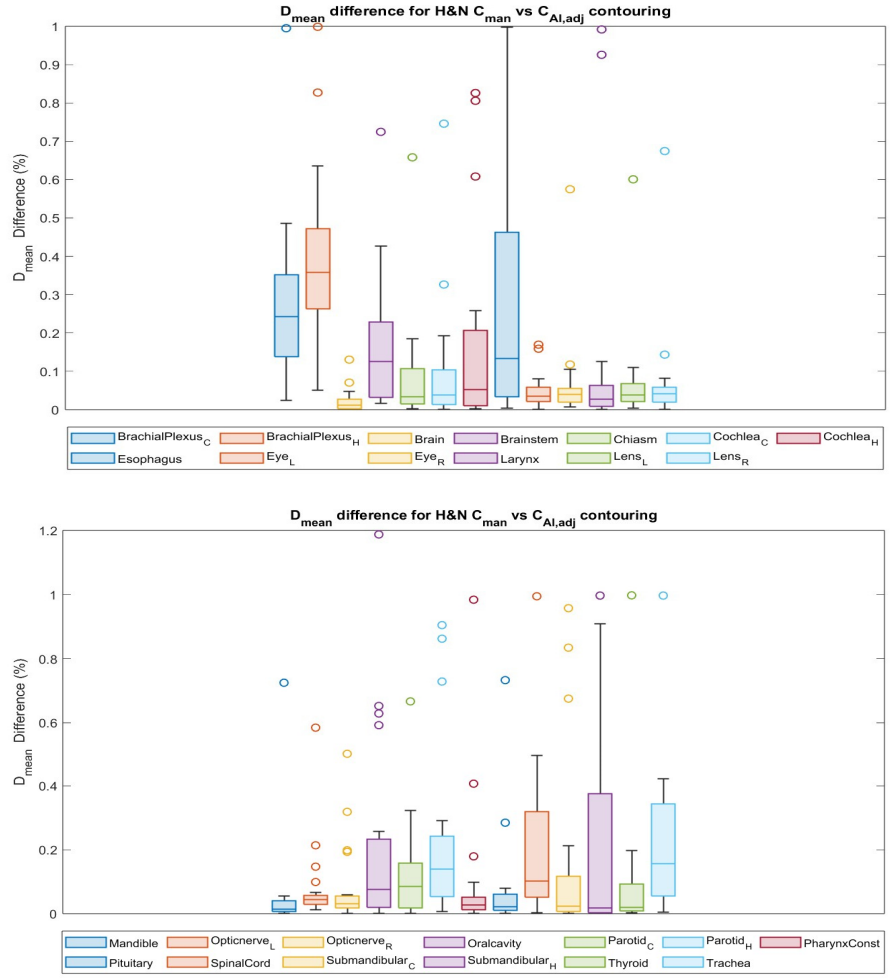


(**a**)

(**b**)

Figure 7. Dosimetric evaluation results: (**a**) relative difference in mean dose (D_{mean}), and (**b**) relative difference in dose of 0.03*cc* volume ($D_{0.03cc}$), for plan form C_{man} contour in comparison with $C_{AI,adj}$ contours of PCa cases. Each circle symbol represents a value outside the standard deviation.

Differences in D_{mean} to OARs for *HNC* between C_{man} and $C_{AI,adj}$ contours are shown in Figure 8a, where the esophagus exhibited relatively large variations in ΔD_{mean}. $D_{0.03cc}$ to the eyes and cochleas had a difference of a maximum of 13% between the C_{man} and $C_{AI,adj}$ plans (Figure 8b), while other OARs showed <10% differences, except for the constraint of contralateral brachial plexus and brainstem. Each circle symbol represents a value outside the standard deviation.



(**a**)

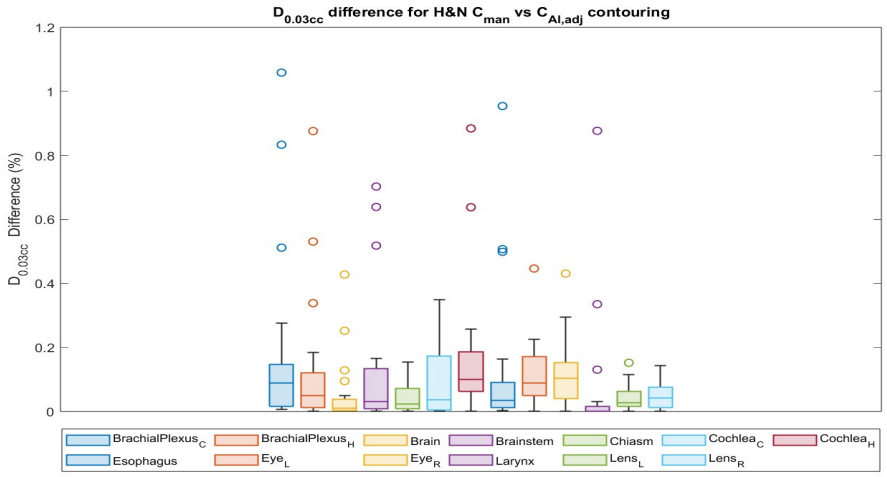
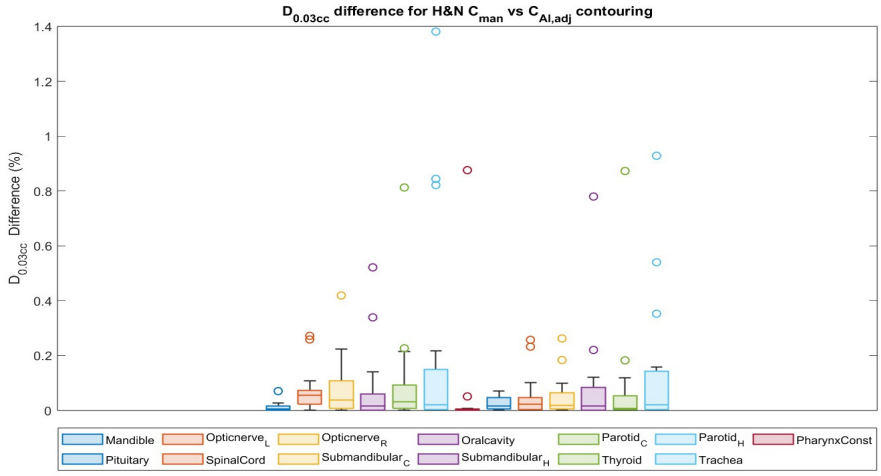


Figure 8. *Cont.*



(**b**)

Figure 8. Dosimetric evaluation results: (**a**) relative difference in mean dose (D_{mean}), and (**b**) relative difference in dose of 0.03*cc* volume ($D_{0.03cc}$), for plan form C_{man} contour in comparison with $C_{AI,adj}$ contours of *HNC* cases. Each circle symbol represents a value outside the standard deviation.

The dosimetric parameters of the *HNC* patients are listed in Table 11. The largest differences were seen in both brachial plexuses D_{min} and D_{mean}, with differences up to 82% and 35%, respectively, between the C_{man} and $C_{AI,adj}$ pairs. The differences were relatively smaller for other OARs between the C_{man} and $C_{AI,adj}$ contour plan pairs.

Table 11. Summary of relative differences in D_{min}, D_{mean}, and $D_{0.03cc}$ values for C_{man} and $C_{AI,adj}$ contours for *HNC* cases.

	ΔD_{mean}		$\Delta D_{0.03cc}$	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Contralateral Brachial Plexus	0.25 (0.13)	0.24 (0.02–0.49)	0.12 (0.19)	0.08 (0.01–0.83)
Ipsilateral Brachial Plexus	0.35 (0.15)	0.35 (0.05–0.64)	0.07 (0.09)	0.02 (0.00–0.34)
Brain	0.02 (0.03)	0.01 (0.00–0.13)	0.06 (0.11)	0.02 (0.00–0.43)
Brainstem	0.14 (0.11)	0.13 (0.02–0.43)	0.14 (0.23)	0.04 (0.00–0.70)
Chiasm	0.05 (0.05)	0.03 (0.00–0.13)	0.05 (0.05)	0.04 (0.00–0.15)
Contralateral Cochlea	0.08 (0.09)	0.04 (0.00–0.33)	0.11 (0.11)	0.08 (0.00–0.35)
Ipsilateral Cochlea	0.11 (0.20)	0.04 (0.00–0.83)	0.16 (0.19)	0.10 (0.02–0.88)
Esophagus	0.18 (0.21)	0.08 (0.00–0.64)	0.09 (0.16)	0.03 (0.00–0.51)
Left Eye	0.04 (0.04)	0.03 (0.00–0.16)	0.13 (0.11)	0.09 (0.02–0.45)
Right Eye	0.04 (0.03)	0.03 (0.01–0.12)	0.13 (0.11)	0.10 (0.03–0.43)
Larynx	0.09 (0.23)	0.03 (0.00–0.93)	0.07 (0.23)	0.00 (0.00–0.88)
Left Lens	0.04 (0.03)	0.04 (0.00–0.11)	0.05 (0.04)	0.04 (0.00–0.15)
Right Lens	0.04 (0.03)	0.04 (0.00–0.14)	0.05 (0.04)	0.05 (0.00–0.14)
Mandible	0.02 (0.02)	0.01 (0.00–0.06)	0.01 (0.02)	0.00 (0.00–0.17)
Left Optic Nerve	0.05 (0.03)	0.04 (0.01–0.15)	0.07 (0.08)	0.06 (0.01–0.27)
Right Optic Nerve	0.05 (0.08)	0.03 (0.00–0.32)	0.07 (0.10)	0.04 (0.00–0.42)
Oral Cavity	0.11 (0.15)	0.04 (0.00–0.59)	0.05 (0.09)	0.02 (0.00–0.34)
Contralateral Parotid	0.10 (0.09)	0.10 (0.01–0.32)	0.11 (0.19)	0.05 (0.00–0.81)
Ipsilateral Parotid	0.15 (0.17)	0.09 (0.01–0.73)	0.10 (0.20)	0.02 (0.00–0.82)
Pharynx Constrictor Muscle	0.06 (0.10)	0.03 (0.00–0.41)	0.00 (0.01)	0.00 (0.00–0.05)
Pituitary	0.04 (0.07)	0.02 (0.00–0.28)	0.02 (0.02)	0.02 (0.00–0.07)

Table 11. *Cont.*

	ΔD_{mean}		$\Delta D_{0.03cc}$	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Spinal Cord	0.17 (0.17)	0.10 (0.00–0.50)	0.04 (0.06)	0.02 (0.00–0.26)
Contralateral Submandibular Gland	0.13 (0.24)	0.03 (0.00–0.83)	0.05 (0.07)	0.02 (0.00–0.26)
Ipsilateral Submandibular Gland	0.14 (0.20)	0.01 (0.00–0.64)	0.04 (0.06)	0.01 (0.00–0.22)
Thyroid	0.05 (0.06)	0.02 (0.00–0.20)	0.03 (0.05)	0.01 (0.00–0.18)
Trachea	0.17 (0.15)	0.13 (0.00–0.42)	0.05 (0.09)	0.02 (0.00–0.35)

Differences between the achieved dosimetric parameters for PCa planning were not significant according to the Wilcoxon test, with the exception of $D_{0.03cc}$ for the anal canal. Brachial plexuses showed significant differences in terms D_{mean}. The statistical analysis results are shown in Table 12.

Table 12. Statistical test results for D_{min}, D_{mean}, and $D_{0.03cc}$ values measured between the plans generated from C_{man} and $C_{AI,adj}$ contours of PCa and *HNC* cases.

Study Site	**OAR**	ΔD_{mean}	$\Delta D_{0.03cc}$
Prostate	Rectum	0.64	0.35
	Bladder	0.90	0.37
	Anal Canal	0.11	0.04
	Left Femur	0.47	0.93
	Right Femur_R	0.23	0.82
Head and Neck	Contralateral Brachial Plexus	0.02	0.22
	Ipsilateral Brachial Plexus	0.00	0.41
	Brain	0.95	0.79
	Brainstem	0.92	0.84
	Chiasm	0.90	0.74
	Contralateral Cochlea	0.82	0.58
	Ipsilateral Cochlea	0.74	0.97
	Esophagus	0.34	0.66
	Left Eye	0.71	0.90
	Right Eye	0.86	1.00
	Larynx	0.51	0.47
	Left Lens	0.90	0.51
	Right Lens	0.95	0.52
	Mandible	0.92	0.94
	Left Optic Nerve	0.84	0.88
	Right Optic Nerve	0.80	0.79
	Oralcavity	0.52	0.39
	Parotid_C	0.86	0.42
	Parotid_H	0.30	0.44
	PharynxConst	0.99	0.92
	Pituitary	0.84	0.82
	SpinalCord	0.34	0.78
	Contralateral Submandibular gland	0.12	0.22
	Ipsilateral Submandibular gland	0.17	0.37
	Thyroid	0.95	0.66
	Trachea	0.92	1.00

3.5. nPQM Comparison

nPQM revealed that all the plans optimized from $C_{AI,adj}$ were considered equivalent to C_{man}, with only few plans deemed as inferior to the clinical plan but clinically acceptable. Table 13 summarizes the difference in plan quality for all study sites.

Table 13. Relative difference in normalized plan quality metric between treatment plans with C_{man} and $C_{AI,adj}$ contours.

Study Site	Mean (SD)	Median (Range)
Prostate	0.080 (0.097)	0.032 (0.00–0.276)
Head and Neck	0.067 (0.057)	0.054 (0.00–0.173)

3.6. Time Savings

Table 14 reports the average times required for contouring over all test subjects with different methods in absolute and percentage units of time savings.

Table 14. Time savings using AI-assisted autocontouring for study sites.

Study Site	T_{man}	$T_{AI,adj}$	Time Savings	Saved Time (%)
Prostate	23 min	6 min 25 s	16 min 35 s	72%
Head and Neck	2 h 30 min	23 min 35 s	2 h 6 min 25 s	84%

3.7. Interobserver Variability

The qualitative test results showed no significant difference between the two observers. Time saving percentages varied among ROs (from 64% to 72% and 16 to 19 min for PCa, respectively). Only 2% variation was observed in *nPQM*. A detail geometric differences are shown in Table 15.

Table 15. Interobserver variability in terms of *DSC*, *HD* and *RVD* values measured between $C_{AI,adj}$ performed by two independent physicians for PCa cases.

	DSC		*HD* (mm)		*RVD*	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**Mean (SD)**	**Median (Range)**
PTV	0.91 (0.06)	0.94 (0.80–0.98)	5.87 (4.70)	3.74 (1.41–18.14)	0.09 (0.09)	0.04 (0.0–0.26)
Rectum	0.97 (0.03)	0.98 (0.90–1.00)	4.76 (5.79)	3.08 (0.0–23.73)	0.06 (0.06)	0.04 (0.0–0.21)
Bladder	0.98 (0.01)	0.98 (0.95–1.00)	3.10 (3.17)	2.00 (1.0–14.28)	0.02 (0.02)	0.02 (0.0–0.08)
Anal Canal	0.91 (0.05)	0.92 (0.83–0.99)	3.11 (1.53)	2.34 (1.0–6.08)	0.12 (0.09)	0.10 (0.01–0.29)
Femur Left	0.98 (0.02)	0.99 (0.95–1.00)	2.02 (1.71)	02.12 (0.0–5.0)	0.03 (0.04)	0.02 (0.0–0.11)
Femur Right	0.98 (0.03)	0.99 (0.87–1.00)	2.53 (3.29)	1.87 (0.0–14.59)	0.03 (0.04)	0.02 (0.0–0.12)

As shown in Figure 9, the plans with $C_{AI,adj}$ resulted in anal canal coverage that largely differed from the manual contour plan. No significant geometric differences were found for *DSC* and *RVD* by comparison of both RO-reviewed contours with the C_{man} contour. Table 16 tabulates the difference in dosimetric parameters for observer variability.

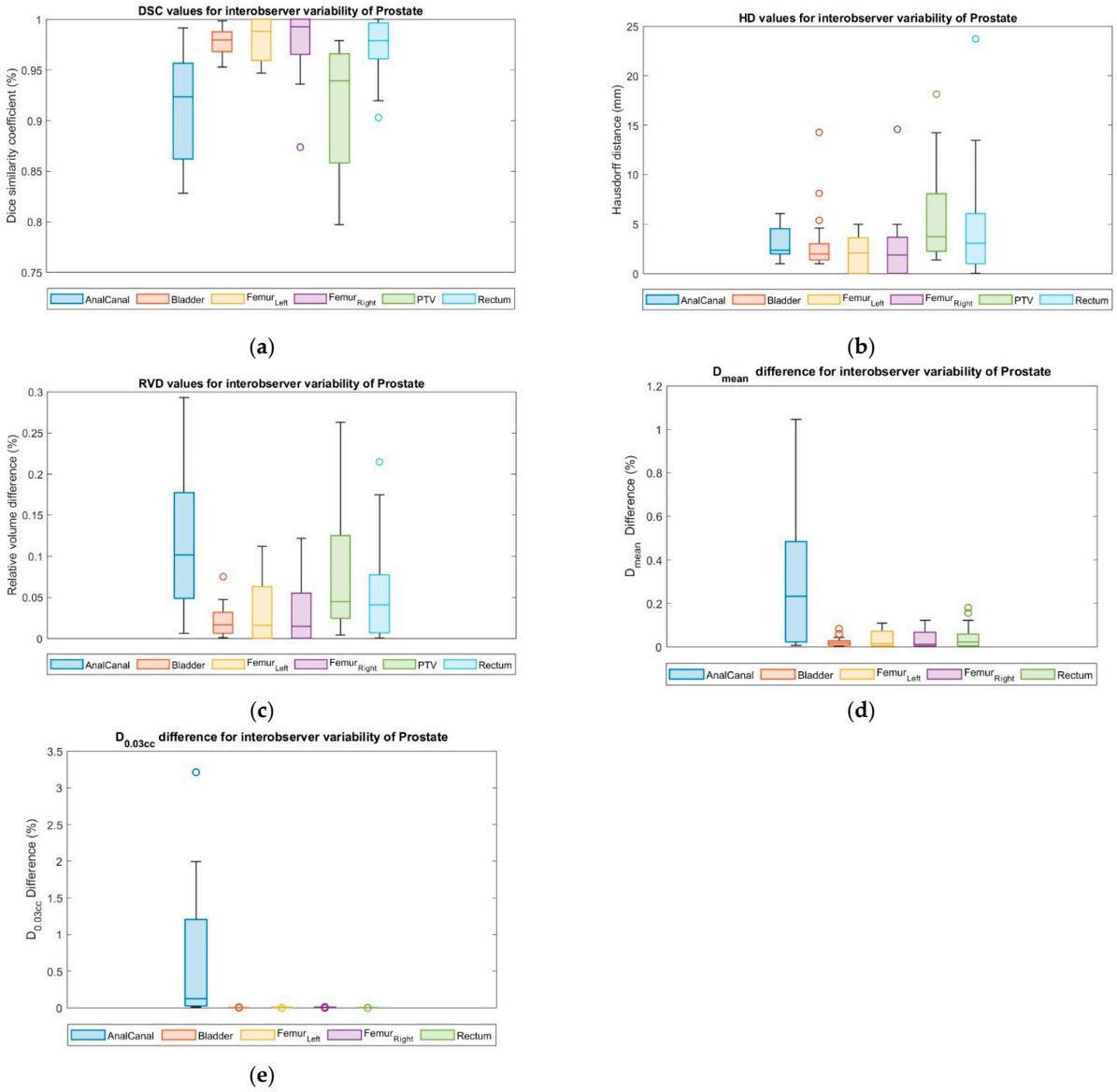


Figure 9. Geometric evaluation results: (**a**) *DSC*, (**b**) *HD* in mm, and (**c**) *RVD* and dosimetric evaluation results; (**d**) relative difference in mean dose (D_{mean}) and (**e**) relative difference in dose of 0.03*cc* volume ($D_{0.03cc}$) of interobserver variability. Each circle symbol represents a value outside the standard deviation.

Table 16. Summary of relative differences in D_{mean} and $D_{0.03cc}$ values measured between $C_{AI,adj}$ performed by two different radiation oncologists.

	ΔD_{mean}		$\Delta D_{0.03cc}$	
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**
Rectum	0.04 (0.05)	0.02 (0.0–0.18)	0.00 (0.00)	0.00 (0.0–0.00)
Bladder	0.02 (0.02)	0.02 (0.0–0.08)	0.00 (0.00)	0.00 (0.0–0.01)
Anal Canal	0.28 (0.31)	0.23 (0.0–1.05)	0.65 (0.91)	0.12 (0.0–3.21)
Femur Left	0.04 (0.04)	0.01 (0.0–0.11)	0.00 (0.00)	0.00 (0.0–0.00)
Femur Right	0.04 (0.04)	0.01 (0.0–0.12)	0.00 (0.00)	0.00 (0.0–0.01)

4. Discussion

Since automatic segmentation tools have become a more efficient alternative to expert manual segmentation, it is important that these applications undergo a thorough review, as the full responsibility of the use of *AI* falls to humans [36]. In particular, the medical physicists have the responsibility of a thorough quality assurance [37] and the radiation oncologist has clinical responsibility of the resulting contours. The purpose of this work was to explore the potential advantages of including an artificial intelligence-based autocontouring system in a clinical pathway in terms of time saving, contour generation accuracy, and radiotherapy plan quality obtained from such reviewed structure sets. The analysis was performed on a dataset of 40 cancer patients equally distributed for PCa and *HNC*.

As the majority of reported evaluation metrics in the literature are based on geometric metrics [38], and usually evaluate autocontouring without human intervention, we compared both the geometric and dosimetric plan quality performance of the autocontouring software (version: 1.0.18; Limbus AI Inc, Regina, SK, Canada), after physician validation and adjustment, against manual contours.

The first results of this work clearly indicate that with the aid of an AI-based autocontouring system, 72% and 84% of contouring time can be saved for PCa and *HNC* cases, respectively. More time saving is possible by implementing a fully integrated system that automatically detects the CT image by predefined protocol and contour structures, eliminating manual export/import function. Moreover, the geometric accuracy reached by Limbus *AI* showed a high compliance with the contours used in the clinical routine. The target and OARs of PCa patients were segmented to high geometric precision, with *DSC* between C_{man} and $C_{AI,adj} \geq 0.7$. The anal canal contours had the largest differences, with an average value of *DSC* (0.70) as well as a 30% difference in volume between the C_{man} and $C_{AI,adj}$ contours.

In comparison to AI-based $C_{AI,adj}$ contours, most of the structures of *HNC* cases, including the brain, mandible, eyes, and optic nerves, had a high degree of geometric correlation (*DSC* > 0.98, *HD* < 3.32 mm, and *RVD* near to 0). However, there were also structures with low *DSC*, such as the brachial plexus (*DSC* = 0.11–0.13), leading to a large variety of results, which is consistent with the previous literature [6,11]. The institutional recommendation to contour a larger larynx, for instance, may result in a poorer geometric correlation of this OAR. Moreover, for this study, the autocontouring software and the oncologist utilized only CT images without contrast enhancement for contouring and revision, while normally, ROs register MRI images to CT images for contouring the OARs.

In principle, the accuracy of contouring has a direct influence on plan optimization, and hence the assessment and decision-making process for treatment plans. As a result, the focus of this study was to determine whether $C_{AI,adj}$ contours could provide equivalent dosimetric findings to C_{man} contours when examined using dosimetric parameters. The prostate *PTV* conformity index showed nearly no change in dosimetric analysis; however, there was a 22% difference in *HI*. Although this study did not contain target volume auto-segmentation, we exclusively examined prostate *PTV* for observation. The modest dosimetric variation in *PTV* might be attributed mostly to the expertise and different approach to planning by various medical physicists.

The greatest notable dose difference for PCa OARs' dose-volume metrics was in the anal canal for the C_{man} vs. $C_{AI,adj}$ contour plan, whilst other OARs maintained almost the same dose distribution. Femurs indicated slightly higher mean dose, which might be attributed to volume variance in the femur segmentation. In terms of *HNC* cases, both brachial plexuses showed a greater divergence in the mean dose for the $C_{AI,adj}$ contours as compared to the C_{man} contours. Otherwise, no significant differences in dose-volume metrics were discovered for those plans. Dosimetric disparities between the C_{man} and $C_{AI,adj}$ contour plans, on the other hand, were minimal for organs such as the cochlea, parotids, and submandibular glands. Only for the brachial plexus were mean dose differences statistically significant; otherwise, the Wilcoxon rank sum tests failed to identify a significant

difference in the achieved dosimetric parameters between these plan pairs, implying that the $C_{AI,adj}$-generated plans perform similarly to the C_{man} contour in the dose optimization and evaluation process for *HNC* planning.

The complex interplay between structure geometry and dose distribution is reflected in the discrepancy between geometric and dosimetric performance. In addition to geometric accuracy, spatial dose distribution and steepness of dose gradients also affect dosimetry performance. Even if there is a significant difference in the dosimetric metrics between the C_{man} and $C_{AI,adj}$ contours for a structure located far away from the high-dose zone, their absolute dosimetric values may be too small to have an impact on plan assessment and decision making. Furthermore, depending on whether it extracts point or volume-based dosimetry, each dosimetric parameter (i.e., maximum, mean, or volume-based parameter) has a distinct reliance and sensitivity to geometric change. For example, when the size of a structure varies in a high-dose gradient zone, the maximum dose may fluctuate more than the mean dose [4]. Overall, the complex interplay between structure geometry and dose distribution suggests that employing a commercial autosegmentation system that was not trained on local data necessitates further examination that includes both geometric and dosimetric analysis. This critical situation highlights the significance of adopting normalized plan quality metrics as a virtual physician that integrates both geometry and dosimetry assessment. The overall plan quality of PCa and *HNC* cases with the $C_{AI,adj}$ contour changed by 8.0% and 6.7%, respectively, when compared to the reference plan that was in a relatively acceptable range.

Interobserver variability analysis was conducted for PCa cases, where the geometric and dosimetric data acquired using each of the studied delineations by two ROs and the manual one was analyzed. Time savings and acceptance of AI-driven contours are approximately the same for both ROs. Except for the anal canal contour, there was a good correlation of geometric metrics ($DSC > 0.92$, $HD < 3.74$ mm, and $RVD < 0.04$) between two ROs. There was also a large dose variation ($D_{0.03cc}$ was 12% and D_{mean} was 23%) for the anal canal, despite the fact that the dose parameters for other OARs were identically matched between ROs. The overall normalized plan quality variation was 2% between ROs, whereas the difference between the C_{man} and $C_{AI,adj}$ contour plan was 3.2%, suggesting that a standard starting point of contouring can reduce interobserver variability.

We considered manually delineated contours of CRO Aviano as the gold standard in this research. This is not to claim that manual delineation is "better" or "accurate" than AI-based delineation. Experts favored autosegmented contours over manual delineation for specific structures in our ongoing evaluation study. Manual delineation provides a clinically acceptable and recognized contour quality, implying some clinical expertise or local institution practices. As Limbus software (version: 1.0.18; Limbus AI Inc., Regina, SK, Canada) was trained using universal structure sets, software using local institutional datasets can lessen discrepancies because there are always some variances in practice between institutions.

This study has some limitations. Even if the selected cases for each district resulted in a homogeneous dataset, only a subgroup of the patients in this research were evaluated for dosimetry. Although it clearly highlighted the disparity between geometric metrics and dosimetry performance, further research including a wider pool of patient samples will be advantageous in characterizing the dosimetry performance of each unique structure. The contouring was carried out retrospectively using CT images without contrast enhancement and without the registration of MRI and/or PET images, which is now strongly suggested for the contouring of not only treatment volumes, but also OARs in some pathologies. For a more in-depth examination, research registering CT autocontouring with MRI and/or PET images might be a feasible option. To obtain a more complete scenario on how the performance of the Limbus autocontouring system affects the contouring procedure, a comparison with other similar software should be performed. Finally, the 5 mm CT slice thickness in the prostate patients, which is standard practice in our institution, is a relatively large value used in prostates [39]. A change in slice thickness from 5 to 3 mm has been

shown to affect only the volume of the bladder significantly [40]. However, this should not affect the main conclusions of the present study, as the slice thickness was always consistent during the comparison among *AI* and humans in the prostate patients. Despite its limitations, this study offers a proof-of-concept methodology to investigate the impact of including in the RT workflow an autocontouring software.

5. Conclusions

In the contouring process, human assessment is required due to the lack of absolute dependability of automatic segmentation. Nonetheless, providing an approach that has the potential to speed up the contouring process in the vast majority of cases would be an improvement over present clinical practice.

The clinical acceptability and efficacy of the AI-driven approach are dependent on the structural segmentation for the site, and clinical criteria stringency, as demonstrated by the cancer sites. The varying performance of $C_{AI,adj}$ contours across structure sets suggests a different approach, in which automatic segmentation is used to generate a subset of contours where AI consistently performs well, and clinical effort is reserved for the complement subset, which may be more sensitive and subject to significantly larger error or variation.

Dose parameter analysis revealed that treatment plans optimized using AI-generated contours did not result in statistically significant differences when examined using normalized plan quality metrics. The results show that plans based on automatically generated contours do not overdose nearby OARs. However, no statistically significant link between geometric and dosimetric metrics was found. The outcomes from dosimetric analysis and interobserver variability suggest that AI-based autocontouring may help to establish a standard starting point for radiation therapy treatment.

Author Contributions: Conceptualization, S.M.H.H. and P.C.; methodology, S.M.H.H.; software, S.M.H.H. and G.P.; formal analysis, G.P. and M.A.; investigation, A.D. (Alessandra Donofrio) and F.M.; resources, G.F., A.C. and R.B.; writing—original draft preparation, S.M.H.H.; writing—review and editing, A.D. (Annalisa Drigo), R.S.R. and M.A.; supervision, P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Italian Ministry of Health (Ricerca Corrente) (no grant number provided). The authors would also like to acknowledge the ACC reti 2021—RCR WP12.

Institutional Review Board Statement: The studies involving human participants were reviewed and approved by Comitato Etico Unico Regionale—CEUR Friuli Venezia Giulia, Azienda Regionale di Coordinamento per la Salute (ARCS), via Pozzuolo n. 330—33100 Udine (palazzina B).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request due to privacy/ethical restrictions.

Acknowledgments: We are grateful to "Limbus AI Inc." and "Dosimetrica" for providing a research version of the Limbus AI software.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barton, M.B.; Jacob, S.; Shafiq, J.; Wong, K.; Thompson, S.R.; Hanna, T.P.; Delaney, G.P. Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012. *Radiother. Oncol.* **2014**, *112*, 140–144. [CrossRef]
2. Barton, M.B.; Frommer, M.; Shafiq, J. Role of radiotherapy in cancer control in low-income and middle-income countries. *Lancet Oncol.* **2006**, *7*, 584–595. [CrossRef] [PubMed]
3. Cancer Today. Available online: http://gco.iarc.fr/today/home (accessed on 5 November 2022).
4. Cao, M.; Stiehl, B.; Yu, V.Y.; Sheng, K.; Kishan, A.U.; Chin, R.K.; Yang, Y.; Ruan, D. Analysis of Geometric Performance and Dosimetric Impact of Using Automatic Contour Segmentation for Radiotherapy Planning. *Front. Oncol.* **2020**, *10*, 1762. Available online: https://www.frontiersin.org/articles/10.3389/fonc.2020.01762 (accessed on 30 November 2023). [CrossRef] [PubMed]

5. Jameson, M.G.; Holloway, L.C.; Vial, P.J.; Vinod, S.K.; Metcalfe, P.E. A review of methods of analysis in contouring studies for radiation oncology. *J. Med. Imaging Radiat. Oncol.* **2010**, *54*, 401–410. [CrossRef] [PubMed]
6. Kim, N.; Chang, J.S.; Kim, Y.B.; Kim, J.S. Atlas-based auto-segmentation for postoperative radiotherapy planning in endometrial and cervical cancers. *Radiat. Oncol.* **2020**, *15*, 106. [CrossRef] [PubMed]
7. Tong, N.; Gou, S.; Yang, S.; Cao, M.; Sheng, K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med. Phys.* **2019**, *46*, 2669–2682. [CrossRef] [PubMed]
8. Jackson, P.; Kron, T.; Hardcastle, N. A future of automated image contouring with machine learning in radiation therapy. *J. Med. Radiat. Sci.* **2019**, *66*, 223–225. [CrossRef] [PubMed]
9. Wong, J.; Huang, V.; Wells, D.; Giambattista, J.; Giambattista, J.; Kolbeck, C.; Otto, K.; Saibishkumar, E.P.; Alexander, A. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: A workflow study at two cancer centers. *Radiat. Oncol.* **2021**, *16*, 101. [CrossRef]
10. Brouwer, C.L.; Dinkla, A.M.; Vandewinckele, L.; Crijns, W.; Claessens, M.; Verellen, D.; van Elmpt, W. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Phys. Imaging Radiat. Oncol.* **2020**, *16*, 144–148. [CrossRef]
11. Casati, M.; Piffer, S.; Calusi, S.; Marrazzo, L.; Simontacchi, G.; Di Cataldo, V.; Greto, D.; Desideri, I.; Vernaleone, M.; Francolini, G.; et al. Clinical validation of an automatic atlas-based segmentation tool for male pelvis CT images. *J. Appl. Clin. Med. Phys.* **2022**, *23*, e13507. [CrossRef]
12. D'aviero, A.; Re, A.; Catucci, F.; Piccari, D.; Votta, C.; Piro, D.; Piras, A.; Di Dio, C.; Iezzi, M.; Preziosi, F.; et al. Clinical Validation of a Deep-Learning Segmentation Software in Head and Neck: An Early Analysis in a Developing Radiation Oncology Center. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9057. [CrossRef] [PubMed]
13. Zabel, W.J.; Conway, J.L.; Gladwish, A.; Skliarenko, J.; Didiodato, G.; Goorts-Matthews, L.; Michalak, A.; Reistetter, S.; King, J.; Nakonechny, K.; et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. *Pract. Radiat. Oncol.* **2021**, *11*, e80–e89. [CrossRef] [PubMed]
14. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef] [PubMed]
15. Voet, P.W.; Dirkx, M.L.; Teguh, D.N.; Hoogeman, M.S.; Levendag, P.C.; Heijmen, B.J. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother. Oncol.* **2011**, *98*, 373–377. [CrossRef] [PubMed]
16. Vinod, S.K.; Jameson, M.G.; Min, M.; Holloway, L.C. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother. Oncol.* **2016**, *121*, 169–179. [CrossRef] [PubMed]
17. Sharp, G.; Fritscher, K.D.; Pekar, V.; Peroni, M.; Shusharina, N.; Veeraraghavan, H.; Yang, J. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med. Phys.* **2014**, *41*, 050902. [CrossRef]
18. Eclipse | Varian. Available online: https://www.varian.com/products/radiotherapy/treatment-planning/eclipse (accessed on 12 November 2022).
19. Offersen, B.V.; Boersma, L.J.; Kirkove, C.; Hol, S.; Aznar, M.C.; Sola, A.B.; Kirova, Y.M.; Pignol, J.-P.; Remouchamps, V.; Verhoeven, K.; et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiother. Oncol.* **2015**, *114*, 3–10. [CrossRef]
20. Salembier, C.; Villeirs, G.; De Bari, B.; Hoskin, P.; Pieters, B.R.; Van Vulpen, M.; Khoo, V.; Henry, A.; Bossi, A.; De Meerleer, G.; 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. [CrossRef]
21. Grégoire, V.; Evans, M.; Le, Q.-T.; Bourhis, J.; Budach, V.; Chen, A.; Eisbruch, A.; Feng, M.; Giralt, J.; Gupta, T.; et al. Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. *Radiother. Oncol.* **2018**, *126*, 3–24. [CrossRef]
22. Limbus AI-Automatic Contouring for Radiation Therapy, Limbus AI. Available online: https://limbus.ai/ (accessed on 12 November 2022).
23. Wong, J.; Fong, A.; McVicar, N.; Smith, S.; Giambattista, J.; Wells, D.; Kolbeck, C.; Giambattista, J.; Gondara, L.; Alexander, A. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother. Oncol.* **2020**, *144*, 152–158. [CrossRef]
24. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [CrossRef]
25. Avanzo, M.; Pirrone, G.; Vinante, L.; Caroli, A.; Stancanello, J.; Drigo, A.; Massarut, S.; Mileto, M.; Urbani, M.; Trovo, M.; et al. Electron Density and Biologically Effective Dose (BED) Radiomics-Based Machine Learning Models to Predict Late Radiation-Induced Subcutaneous Fibrosis. *Front. Oncol.* **2020**, *10*, 490. [CrossRef]
26. Avanzo, M.; Chiovati, P.; Boz, G.; Sartor, G.; Dozza, F.; Capra, E. Image-guided volumetric arc radiotherapy of pancreatic cancer with simultaneous integrated boost: Optimization strategies and dosimetric results. *Phys. Medica* **2016**, *32*, 169–175. [CrossRef]

27. Baroudi, H.; Brock, K.K.; Cao, W.; Chen, X.; Chung, C.; Court, L.E.; El Basha, M.D.; Farhat, M.; Gay, S.; Gronberg, M.P.; et al. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? *Diagnostics* **2023**, *13*, 667. [CrossRef]
28. Gooding, M.J.; Smith, A.J.; Tariq, M.; Aljabar, P.; Peressutti, D.; van der Stoep, J.; Reymen, B.; Emans, D.; Hattu, D.; van Loon, J.; et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. *Med. Phys.* **2018**, *45*, 5105–5115. [CrossRef]
29. Yeghiazaryan, V.; Voiculescu, I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J. Med. Imaging* **2018**, *5*, 015006. [CrossRef] [PubMed]
30. StructSeg2019-Grand Challenge, Grand-Challenge.org. Available online: https://structseg2019.grand-challenge.org/Evaluation/ (accessed on 30 November 2022).
31. Mayo, C.S.; Moran, J.M.; Bosch, W.; Xiao, Y.; McNutt, T.; Popple, R.; Michalski, J.; Feng, M.; Marks, L.B.; Fuller, C.D.; et al. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int. J. Radiat. Oncol.* **2017**, *100*, 1057–1066. [CrossRef] [PubMed]
32. Doses, A. 3. Special Considerations Regarding Absorbed-Dose and Dose–Volume Prescribing and Reporting in IMRT. *J. ICRU* **2010**, *10*, 27–40. [CrossRef]
33. Feuvret, L.; Noël, G.; Mazeron, J.-J.; Bey, P. Conformity index: A review. *Int. J. Radiat. Oncol.* **2006**, *64*, 333–342. [CrossRef] [PubMed]
34. Nelms, B.E.; Robinson, G.; Markham, J.; Velasco, K.; Boyd, S.; Narayan, S.; Wheeler, J.; Sobczak, M.L. Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems. *Pract. Radiat. Oncol.* **2012**, *2*, 296–305. [CrossRef] [PubMed]
35. MATLAB-MathWorks. Available online: https://ww2.mathworks.cn/en/products/matlab.html (accessed on 30 November 2022).
36. Avanzo, M.; Trianni, A.; Botta, F.; Talamonti, C.; Stasi, M.; Iori, M. Artificial Intelligence and the Medical Physicist: Welcome to the Machine. *Appl. Sci.* **2021**, *11*, 1691. [CrossRef]
37. Zanca, F.; Brusasco, C.; Pesapane, F.; Kwade, Z.; Beckers, R.; Avanzo, M. Regulatory Aspects of the Use of Artificial Intelligence Medical Software. *Semin. Radiat. Oncol.* **2022**, *32*, 432–441. [CrossRef] [PubMed]
38. 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*, 354–369. [CrossRef] [PubMed]
39. Kim, Y.; Hsu, I.J.; Lessard, E.; Pouliot, J.; Vujic, J. Dose uncertainty due to computed tomography (CT) slice thickness in CT-based high dose rate brachytherapy of the prostate cancer. *Med. Phys.* **2004**, *31*, 2543–2548. [CrossRef] [PubMed]
40. Berthelet, E.; Liu, M.; Truong, P.; Czaykowski, P.; Kalach, N.; Yu, C.; Patterson, K.; Currie, T.; Kristensen, S.; Kwan, W.; et al. CT slice index and thickness: Impact on organ contouring in radiation treatment planning for prostate cancer. *J. Appl. Clin. Med. Phys.* **2003**, *4*, 365–373. [CrossRef]