Supplementary material

Table S1: Average contouring times for prostate structures 3 centres using manual contouring and deep learning contouring (DLC). Standard errors used. P-value from paired T-test for centres 1 and 4and for unpaired T-test for centre 2.

	Average OAR contouring time (minutes)		
	Centre 1	**Centre 2**	**Centre 4**
Manual	15.0±0.7	12.5±0.5	31.1±2.4
DLC	13.0±1.3	11.5±0.3	16.5±2.0
Average time difference (negative is time saving)	-2.0±1.4	-1.0±0.6	14.7±3.1
p-value	0.17	0.09	<0.01

Table S2: Median dice similarity coefficient (DSC) for prostate organs at risk for 2 centres comparing manual contouring and deep learning contouring (DLC). Standard errors used.

Structure	**DSC**	
	Centre 1	**Centre 2**
Rectum	0.62±0.06	0.72±0.03
Bladder	0.83±0.02	0.93±0.02
Femoral Head Left	0.93±0.03	0.90±0.11
Femoral Head Right	0.90±0.04	0.92±0.01

*Table S3: Median distance to agreement (DTA) for prostate organs at risk 2 centres comparing manual contouring and deep learning contouring (DLC). Standard errors used. *One result excluded for error calculation due to DLC contour being entirely outside the manual contour.*

Structure	**DTA (mm)**	
	Centre 1	**Centre 2**
Rectum	4.6±0.7	4.7±1.0
Bladder	2.1±0.3	1.7±0.4
Femoral Head Left	1.5±0.6	1.6±0.2*
Femoral Head Right	2.2±0.6	1.2±0.2

Table S4: Average scores for prostate organs at risk from 2 centres using deep learning contouring (DLC). Standard errors used. Centres scored based on agreement to a clinical contour. centre 2 scored 1=good agreement, 2=very minor differences, 3=minor differences, 4=edits required, 5=moderate edits required, 6=major edits required, 7=gross error. Centre 4 scored 1=clinically acceptable, 2=clinically acceptable (but not quite as reviewing clinician would draw it), 3= requires minor adjustments to be clinically acceptable, 4=requires major adjustments to be clinically acceptable, 5= would be easier to start from scratch.

Structure	**Average Score**	
	Centre 2	**Centre 4**
Rectum	5.3±0.2	3.6±0.4
Bladder	4.5±0.2	3.6±0.4
Femoral Head Left	1.8±0.2	1.3±0.1
Femoral Head Right	1.5±0.1	1.2±0.1

Table S5: Average head and neck organ at risk contouring times for 4 centres using an existing clinical method and deep learning contouring (DLC). Standard errors used.

	Average OAR contouring time (minutes)			
	Centre 1	**Centre 2**	**Centre 3**	**Centre 4**
Existing Method	18.3±1.5	11.9±1.2	74.7±7.6	10.4±1.0
DLC	10.0±0.8	18.0±0.6	62.5±3.1	8.4±0.2
Average time difference (negative is time saving)	-8.3±1.7	+6.1±1.3	-12.2±8.2	-2.0±1.0
p-value	<0.01	0.01	0.54	0.279

Table S6: Median dice similarity coefficient (DSC) for head and neck organs at risk for 4 centres comparing existing clinical contouring and deep learning contouring (DLC). Standard errors used. – indicates the structure was not analysed by that centre

	DSC			
	Centre 1	**Centre 2**	**Centre 3**	**Centre 4**
Brainstem	0.78±0.03	0.81±0.01	-	0.82±0.02
Mandible	0.86±0.03	0.90±0.01	0.85±0.01	-
Parotid Left	0.71±0.05	0.82±0.02	0.77±0.03	0.80±0.01
Parotid Right	0.75±0.03	0.85±0.01	0.72±0.02	0.80±0.01
Spinal Cord	0.72±0.05	0.69±0.05	-	0.80±0.02
Submandibular Left	0.51±0.09	-	0.71±0.03	-
Submandibular Right	0.64±0.09	-	0.67±0.03	-
Larynx	0.66±0.06	-	-	-
Oral Cavity	0.77±0.03	-	-	-

Table S7: Median distance to agreement (DTA) for head and neck organs at risk 3 centres comparing the existing clinical contouring and deep learning contouring (DLC). Standard errors used. – indicates the structure was not analysed by that centre.

	DTA (mm)		
	Centre 1	**Centre 2**	**Centre 4**
Brainstem	2.7±0.4	1.6±0.2	1.5 ±0.2
Mandible	1.6±0.6	0.6±0.1	-
Parotid Left	3.3±0.9	2.2±0.6	2.2±0.1
Parotid Right	2.8±0.4	1.9±0.5	2.5±0.1
Spinal Cord	1.6±4.2	1.0±0.2	2.5±0.2
Submandibular Left	4.0±0.9	-	-
Submandibular Right	2.6±0.9	-	-
Larynx	3.5±4.1	-	-
Oral cavity	5.2±0.6	-	-

Table S8: Average scores for head and neck organs at risk from 3 centres using deep learning contouring (DLC). Standard errors used. Centres scored based on agreement to a clinical contour. Centres 1 and 2 scored 1=good agreement, 2=very minor differences, 3=minor differences, 4=edits required, 5=moderate edits required, 6=major edits required, 7=gross error. Centre 4 scored 1=clinically acceptable, 2=clinically acceptable (but not quite as reviewing clinician would draw it), 3= requires minor adjustments to be clinically acceptable, 4=requires major adjustments to be clinically acceptable, 5= would be easier to start from scratch.

	Centre 1	Centre 2	Centre 4
Brainstem	4.9	4.0±0.2	3.4±0.2
Mandible	3.0	3.9±0.2	-
Parotid Left	4.2	2.1±0.2	2.4±0.2
Parotid Right	3.9	2.1±0.2	2.5±0.2
Spinal Cord	3.1	2.3±0.2	2.4±0.2
Submandibular Left	4.7	-	-
Submandibular Right	5.7	-	-
Larynx	5.3	-	-
Oral Cavity	4.8	-	-

Table S9: Median dice similarity coefficient (DSC) from inter-observer comparisons using manual contouring and deep learning contouring (DLC) for head and neck organs at risk. P-values from a Wilcoxon signed rank test.

	Median DSC		p-value
	Manual	DLC	
Brainstem	0.78±0.03	0.93±0.01	<0.01
Larynx	0.79±0.02	0.83±0.01	0.01
Left SMG	0.80±0.02	0.89±0.04	<0.01
Left parotid	0.84±0.01	0.92±0.02	<0.01

Table S10: Median distance to agreement (DTA) from inter-observer comparisons using manual contouring and deep learning contouring (DLC) for head and neck organs at risk. P-values from a Wilcoxon signed rank test.

	Median DTA (mm)		p-value
	Manual	DLC	
Brainstem	2.2±0.8	0.9±0.1	<0.01
Larynx	1.9±0.2	1.8±0.1	0.699
Left SMG	1.1±0.1	0.7±0.1	<0.01
Left parotid	1.4±0.1	0.9±0.1	<0.01

Table S11: Mean dice similarity coefficient (DSC) for prostate organs at risk from the study. Standard errors used.

Structure	Centre 1	Centre 2
Rectum	0.58±0.05	0.71±0.02
Bladder	0.84±0.01	0.82±0.02
Femoral Head Left	0.90±0.08	0.82±0.01
Femoral Head Right	0.89±0.03	0.92±0.01