Correlations between contouring similarity metrics and simulated treatment outcome for prostate radiotherapy

D Roach[1,2], M G Jameson[2,3], J A Dowling[4], M A Ebert[5,6,9], P B Greer[7,8], A M Kennedy[5], S Watt[3], and L C Holloway[1,2,3,9]

[1]South Western Sydney Clinical School, University of New South Wales, Sydney, Australia
[2]Ingham Institute for Applied Medical Research, Sydney, Australia
[3]Liverpool Cancer Therapy Centre, Liverpool Hospital, Sydney, Australia
[4]Australian e-Health Research Centre, CSIRO, Royal Brisbane Hospital, Australia
[5]Sir Charles Gairdner Hospital, Nedlands, Australia
[6]School of Physics, Faculty of Science, University of Western Australia, Crawley, Australia
[7]Calvary Mater Newcastle Hospital, Newcastle, Australia
[8]School of Mathematical and Physical Sciences, University of Newcastle, Newcastle, Australia
[9]Centre for Medical Radiation Physics, University of Wollongong, Wollongong, Australia

E-mail: d.roach@student.unsw.edu.au

Abstract

Many similarity metrics exist for inter-observer contouring variation studies, however no correlation between metric choice and prostate cancer radiotherapy dosimetry has been explored. These correlations were investigated in this study. Two separate trials were undertaken, the first a thirty-five patient cohort with three observers, the second a five patient dataset with ten observers. Clinical and planning target volumes (CTV and PTV), rectum, and bladder were independently contoured by all observers in each trial. Structures were contoured on T2-weighted MRI and transferred onto CT following rigid registration for treatment planning in the first trial. Structures were contoured directly on CT in the second trial. STAPLE and majority voting volumes were generated as reference gold standard volumes for each structure for the two trials respectively. VMAT treatment plans (78 Gy to PTV) were simulated for observer and gold standard volumes, and dosimetry assessed using multiple radiobiological metrics. Correlations between contouring similarity metrics and dosimetry were calculated using Spearman's rank correlation coefficient.

No correlations were observed between contouring similarity metrics and dosimetry for CTV within either trial. Volume similarity correlated most strongly with radiobiological metrics for PTV in both trials, including $TCP_{Poisson}$ (ρ = 0.57, 0.65), TCP_{Logit} (ρ = 0.39, 0.62), and EUD (ρ = 0.43, 0.61) for each respective trial. Rectum and bladder metric correlations displayed no consistency for the two trials.

PTV volume similarity was found to significantly correlate with rectum normal tissue complication probability (NTCP) (ρ = 0.33, 0.48). Minimal to no correlations with dosimetry were observed for overlap or boundary contouring metrics. Future inter-observer contouring variation studies for prostate cancer should incorporate volume similarity to provide additional insights into dosimetry during analysis.

Keywords: prostate, inter-observer variability, contouring, similarity metrics, VMAT

1. Introduction

Inter- and intra-observer contouring variability remains one of the largest sources of uncertainty in radiotherapy (Weiss and Hess, 2003; Van Dyk *et al.*, 2013), with poor contouring significantly impacting the quality of treatment and patient outcome in clinical trials (Peters *et al.*, 2010). Uncertainties in contouring for prostate cancer radiotherapy have been attributed to imaging modality (Rasch *et al.*, 1999; Dubois *et al.*, 1998), observer training (Khoo *et al.*, 2012), and an over-cautiousness of clinicians limiting rectal tissue volume contoured by trimming the planning target volume (PTV) (Gao *et al.*, 2007). Implementation of clinical protocols and utilisation of MRI reduces inter-observer contouring variability (Rasch *et al.*, 1999; Dubois *et al.*, 1998; Mitchell *et al.*, 2009), although only limited agreement amongst observers is reached (Ost *et al.*, 2011).

The prevalence of inter-observer contouring variability for prostate cancer has been thoroughly investigated, however few studies have additionally assessed the impact of this variability on dosimetry (Vinod *et al.*, 2016a). Of the studies that have investigated this, only neighbouring organ-at-risk (OAR) dosimetry has been evaluated (Livsey *et al.*, 2004; Mitchell *et al.*, 2009; Foppiano *et al.*, 2003; Perna *et al.*, 2011). Consequently, no study has assessed the impact of inter-observer contouring variability on target volume dosimetry for prostate radiotherapy. Additionally, studies to date utilised 3D-CRT, whereas most prostate cancer treatments now employ more modern techniques such as Intensity Modulated Radiotherapy (IMRT), Volumetric Modulated Arc Therapy (VMAT), or Stereotactic Body Radiotherapy (SBRT). These techniques generate tighter dose distributions matching the radiotherapy target volumes, and have been shown to reduce dose and subsequent toxicities to OARs (Al-Mamgani *et al.*, 2009; Palma *et al.*, 2008; Quan *et al.*, 2012). However, due to the steep dose gradients produced by these techniques, poorly contoured target volumes could result in larger impacts on dosimetry than has been observed in previous studies.

Similarity metrics are utilised to quantify contouring variations; however, no consensus exists over the choice of metric to incorporate during an investigation (Jameson *et al.*, 2010; Fotina *et al.*, 2012). This restricts comparisons being made between studies, as similarity metrics cited may poorly correlate with one another (Sharp *et al.*, 2014). A combination of boundary and volume metrics is recommended (Fotina *et al.*, 2012), however these metric choices may have little to no correlation with dosimetry. Studies investigating non-small cell lung cancers (Jameson *et al.*, 2014) and head and neck cancers (Beasley *et al.*, 2016) found that commonly utilised overlap metrics conformity index and DSC displayed weak or no correlations with simulated treatment outcome respectively. The aim of this study was to evaluate correlations between contouring similarity metrics and dosimetry for prostate cancer planned for VMAT radiotherapy.

2. Materials and Methods

2.1. Patient datasets and contouring

Two patient datasets were utilised for this study. The first trial incorporated forty-two patients from a prior study containing pre-treatment CT and MRI scans for localised prostate radiotherapy (Dowling *et al.*, 2015). Three observers (two experienced radiation oncologists, one experienced research radiation therapist) independently contoured clinical target volume (CTV) (ICRU, 2010), rectum, and bladder on T2-weighted MRI based on trial contouring protocol using Eclipse™ treatment planning software (Varian Medical Systems, Palo Alto, CA, USA). MRI scans

were rigidly registered to CT with respect to gold fiducial markers implanted in the prostate, and CTV, rectum, and bladder contours were transferred to CT for treatment planning.

The second trial utilised a five patient subset from the previous trial, with patients selected based on prior clustering of post registration intensity based image similarities of potential atlas images with respect to the larger RADAR patient dataset (Trans-Tasman Radiation Oncology Group (TROG), 2005) using affinity propagation (Frey and Dueck, 2007; Kennedy et al., 2016). This work was to be used during additional atlas based segmentation analyses. Ten observers across four treatment centres (two medical physicists, one radiation therapist, one radiographer, and six radiation oncologists) contoured CTV, rectum, and bladder on CT based on trial contouring protocol, with five observers using Eclipse™ (Varian Medical Systems, Palo Alto, CA) and five observers using Pinnacle^3® (Philips Healthcare, Best, Netherlands) treatment planning software. A uniform 7 mm margin was applied to the CTV in both trials to define the PTV (ICRU, 2010). Following contouring, DICOM structure files were returned and imported into Pinnacle^3® for treatment planning.

2.2. Gold standard volumes

In the absence of pathological information, "gold standard" reference volumes were estimated for each structure from observer contours. The first trial utilised the Simultaneous Truth And Performance Level Estimation (STAPLE) algorithm to generate these volumes (Warfield *et al.*, 2004). For the second trial, the large number of observers resulted in overlapping CTV and rectum STAPLE volumes that were deemed inappropriate for the study. Consequently, a majority vote was used to define the gold standard volumes for each structure in this trial. The impact the choice of gold standard volume has on analysis was investigated by a supplementary study of the first trial dataset, whereby observer contours were iteratively designated as additional gold standard volumes. Gold standard volumes were created within MilxView, an open-sourced image manipulation and processing platform developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) biomedical informatics group (Burdett *et al.*, 2010).

2.3. Treatment Planning

The initial trial contained three patient datasets with corrupt DICOM structure set files, while an additional four patients were removed due to incorrect CTV delineation by a single observer (three

Table 1. Contouring similarity and radiobiological metrics.

Contouring Similarity Metric	Radiobiological Metric
Dice Similarity Coefficient (DSC)	Tumour Control Probability - Poisson Model [a]
Volume Similarity	Tumour Control Probability – Logit Model [a]
Average Relative Volume Difference	Normal Tissue Complication Probability [b]
Sensitivity	Equivalent Uniform Dose (EUD)
Specificity	Minimum Dose [a]
C-Factor (Popovic *et al.*, 2007)	Mean Dose
Mean Absolute Surface Distance	Maximum Dose [b]
95% Hausdorff Distance	Isodose Volumes (IsoX)
Centroid Distances	Dose Volume levels (DX)
	Dose homogeneity [a]

[a] CTV, PTV only

[b] Bladder, Rectum only

115

Table 2. Contouring similarity metric derivations.

Metric	Equations and Derivations
Dice Similarity Coefficient (DSC)	$\mathrm{DSC} = \frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Volume Similarity	$\mathrm{VOLSIM} = \frac{Y\text{-}X}{(X+Y)/2}$
True Positive (Negative)	Number of voxels lying within (outside) the observer contour that also lie within (outside) the gold standard contour
False Positive (Negative)	Number of voxels lying within (outside) the observer contour that lie outside (within) the gold standard contour
Sensitivity	$p = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
Specificity	$q = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$
Mean Absolute Surface Distance	$\mathrm{MASD} = \frac{1}{N_{X_S}+N_{Y_S}} \left(\sum_{x \in X_S} \min_{y \in Y_S} d(x,y) + \sum_{y \in Y_S} \min_{x \in X_S} d(y,x) \right)$
95% Hausdorff Distance	$\mathrm{HD}_{\mathrm{asym}}(X_S, Y_S) = 95\text{th} \underset{x \in X_S}{\text{percentile}} \left(\min_{y \in Y_S} d(x,y) \right)$
	$\mathrm{HD}(X_S, Y_S) = \max(\mathrm{HD}_{\mathrm{asym}}(X_S, Y_S), \mathrm{HD}_{\mathrm{asym}}(Y_S, X_S))$
Centroid (Euclidean)	Euclidean distance between centre-of-mass for gold standard and observer
Centroid (Sagittal Plane)	Distance in sagittal plane between centre-of-mass for gold standard and observer
Centroid (Coronal Plane)	Distance in coronal plane between centre-of-mass for gold standard and observer
Centroid (Axial Plane)	Distance in axial plane between centre-of-mass for gold standard and observer
Absolute Relative Volume Difference	$\mathrm{aRVD} = \left\lvert 100 \times \left(\frac{\lvert X \rvert}{\lvert Y \rvert} \text{-} 1 \right) \right\rvert$
C-Factor (Popovic *et al.*, 2007)	$d = \frac{2p(1\text{-}q)}{p+(1\text{-}q)} + \frac{2(1\text{-}p)q}{(1\text{-}p)+q}$
	$\mathrm{C} = \begin{cases} d, p \geq q \;\wedge\; p > 1\text{-}q \\ -d, p < q \;\&\wedge\; p > 1\text{-}q \\ \text{undefined}, p \leq 1\text{-}q \end{cases}$

X: Number of voxels within gold standard contour

Y: Number of voxels within observer contour

X_S: Surface points of gold standard contour

Y_S: Surface points of observer contour

N_{X_S}: Number of surface points of gold standard contour

N_{Y_S}: Number of surface points of observer contour

$d(x,y)$: Euclidean distance from point x to point y

included seminal vesicles within the CTV, one patient had prostate bed contoured). This resulted in thirty-five patient datasets being imported into Pinnacle^3® for the first trial, and five patient datasets imported during the second trial.

VMAT treatment plans (78 Gy to PTV) incorporating gold standard contours were initially generated for each patient using Pinnacle^3®'s Autoplanning module, and were assessed for quality by an experienced radiation therapist. Treatment plans considered clinically unacceptable had dose objectives manually adjusted and were resimulated, until all gold standard treatment plans were accepted in line with department prostate planning protocol (supplementary table 1). VMAT treatment plans for each set of observer contours were subsequently generated using dose objectives matching each patient's gold standard treatment plan.

2.4. Contouring Similarity Metrics

Structure DICOM files were exported from Pinnacle^3®, and converted into NifTI files within MilxView (Burdett *et al.*, 2010). Volumetric, statistical, and boundary similarity metrics, summarised in table 1, were calculated with respect to the corresponding gold standard contour for bladder, rectum, CTV, and PTV. Derivations for contouring similarity metrics are included in table 2.

2.5. Dosimetry and radiobiological analysis

Dose-volume histograms (DVHs) of gold standard volumes for all treatment plans were exported from Pinnacle^3®, with radiobiological metrics for target volumes and OARs (table 1) calculated using in-house developed software Comp Plan (Holloway *et al.*, 2012). The difference between metrics calculated from gold standard and observer treatment plans provided a measure of the impact on dosimetry due to observer contouring variability.

2.6. Statistical Considerations

All statistical analysis was completed within MATLAB R2015b (The Mathworks Inc., Natick, MA). Spearman's non-parametric rank correlation coefficient (ρ) was used to assess correlations between contouring similarity and radiobiological metrics. As 12 contouring similarity metrics were included in this study, and either 12 (CTV, PTV) or 16 (bladder, rectum) radiobiological metrics considered, Bonferroni corrections of 144 for CTV and PTV, and 192 for bladder and rectum were applied. Subsequently, p-values of $p < 0.00035$ and $p < 0.00026$ respectively were now considered significant. With 105 observer treatment plans, $\alpha = 0.05$ (two-sided) and $1 - \beta = 0.9$, the initial trial possessed the statistical power to detect correlations of $|\rho| > 0.3$. The second trial, with 50 observer treatment plans, could detect correlations of $|\rho| > 0.4$. Table 3 outlines the general interpretation of the strength of Spearman's correlations within biomedical sciences, where the sign of ρ signifies the direction of the correlation (Hinkle *et al.*, 2003).

Final analysis of the contour datasets was then undertaken as detailed by Fotina *et al.*, where the intra-class correlation coefficient (ICC) was utilised to assess whether each trial possessed the number of observers required to result in a minimum acceptable level of study reliability (Fotina *et al.*, 2012).

Table 3. Strength of spearman's correlation ρ

Spearman's $\lvert\rho\rvert$	**Strength of correlation**
0.90 – 1.00	Very strong correlation
0.70 – 0.90	Strong correlation
0.50 – 0.70	Moderate correlation
0.30 – 0.50	Weak correlation
0.00 – 0.30	Negligible correlation

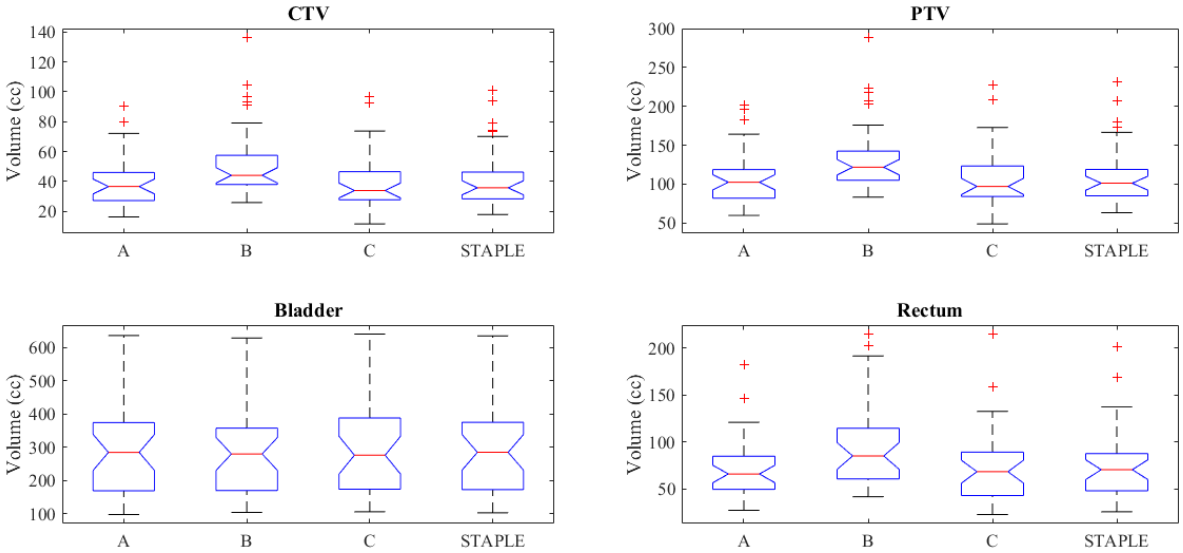

Figure 1. Observer and STAPLE volume spread for each structure across all 35 patients for trial 1. Notch box-plots show a significant difference in CTV and PTV median volumes by observer B compared to observers A and C. Differences between observer median rectum volumes were found to be not statistically significant.

3. Results

3.1. Trial 1 Results

Observer and STAPLE volumes for CTV, PTV, bladder, and rectum are plotted in figure 1. Mean, standard deviation (SD), and coefficient of variation (COV) of these volumes are summarised in table 4. Observer B consistently contoured larger CTV (+ 11.03 cc), PTV (+ 22.44 cc), and rectum (+ 19.09 cc), although only CTV and PTV were considered statistically significant (figure 1). Minimal variations were seen between observers A, C, and STAPLE volumes for CTV, PTV, and rectum, and between all observer and STAPLE bladder volumes. Additional descriptive statistics can be found in supplementary tables 2 and 3. Figure 2 plots DSC for observer contours with respect to the

Table 4. Mean, standard deviation, and coefficient of variation of structure volumes across trial 1 patient cohort.

Observer		CTV	PTV	Bladder	Rectum
A	Mean (cc)	40.13	107.53	289.98	71.50
	SD (cc)	19.32	37.97	141.24	33.50
	COV	0.48	0.35	0.49	0.47
B	Mean (cc)	53.74	135.07	285.72	95.51
	SD (cc)	24.66	45.01	138.09	46.38
	COV	0.46	0.33	0.48	0.49
C	Mean (cc)	40.74	109.59	297.99	73.35
	SD (cc)	21.17	41.54	144.15	40.17
	COV	0.52	0.38	0.48	0.55
STAPLE	Mean (cc)	42.71	112.63	291.63	76.42
	SD (cc)	21.30	40.98	141.51	39.05
	COV	0.50	0.36	0.49	0.51

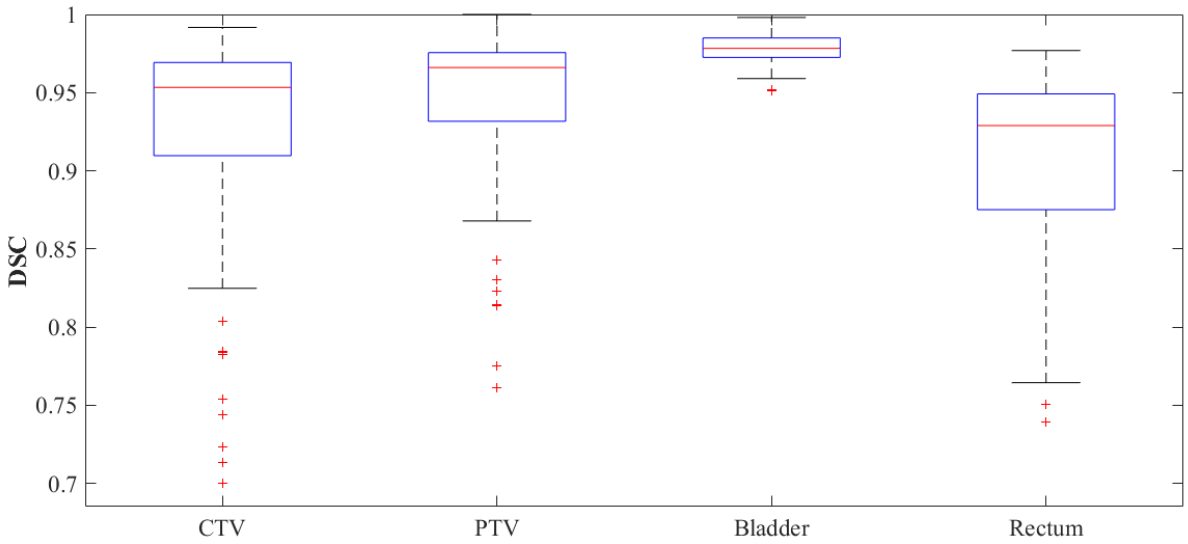

Figure 2. Spread in observer Dice Similarity Coefficient (DSC) for CTV, PTV, bladder, and rectum across all patients for trial 1.

corresponding STAPLE contour. The poorest overlapping PTV contour (DSC = 0.7612) is outlined purple in figure 3.

Mean differences between observer and STAPLE plan dosimetry for selected radiobiological metrics are shown in table 5. Negligible variations in CTV dosimetry were observed. Significant variations in dosimetry were seen for PTV, with several observer treatment plans failing to adequately treat the STAPLE volume (D95 < 76.2 Gy, supplementary table 1). One such treatment plan utilised observer C's contours (orange) in figure 4, where observer C's smaller PTV resulted in a significant portion of the STAPLE PTV receiving inadequate dose.

Correlations between contouring similarity and radiobiological metrics were investigated using Spearman's ρ, with complete lists of correlations found in supplementary tables 4 – 7. No significant correlations were observed for CTV and bladder. Significant correlations for PTV and rectum are displayed in figure 5. Volume similarity exhibited the strongest correlation to dosimetry for both PTV and rectum. The strongest correlating radiobiological metrics for PTV volume similarity were minimum dose (ρ = 0.67), $TCP_{Poisson}$ (ρ = 0.57), dose homogeneity (ρ = -0.52), EUD (ρ = 0.43), and TCP_{Logit} (ρ = 0.39). The strongest correlating radiobiological metrics for rectum volume similarity were mean dose (ρ = 0.45) and EUD (ρ = 0.43), with similar correlations for absolute relative volume difference also recorded (ρ = 0.46 and 0.38 respectively). No significant correlations between contouring similarity metrics and maximum dose or NTCP for the rectum were observed.

Sensitivity and specificity correlations (table 2) were comparable to volume similarity for rectum, but weaker and correlated with fewer radiobiological metrics for PTV. Unsurprisingly the C-Factor, described as a trade-off between sensitivity and specificity (Popovic *et al.*, 2007), exhibited similar correlations to these metrics. Significant correlations were observed between differences in centroid in the sagittal plane and minimum dose (ρ = 0.45) for PTV. Overlap metrics (DSC) and

Table 5. Mean differences between observer and STAPLE plan dosimetry.

	Δ $TCP_{Poisson}$		**Δ TCP_{Logit}**		**Δ NTCP**		**Δ EUD**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CTV	-0.0001	0.0007	-0.0009	0.0062	--	--	-0.07	0.46
PTV	-0.0545	0.2103	-0.0035	0.0107	--	--	-0.74	3.39
Bladder	--	--	--	--	0.0001	0.0006	0.86	1.95
Rectum	--	--	--	--	0.0042	0.0121	0.45	1.24

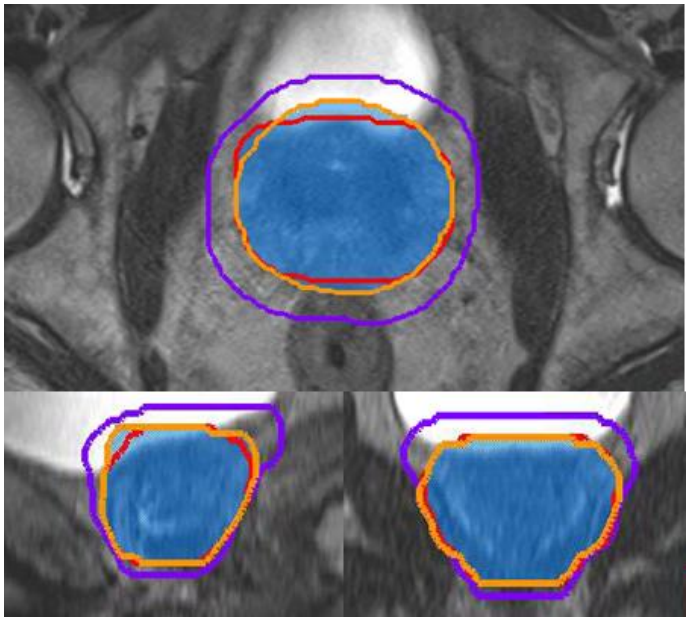

Figure 3. Patient 32 containing the poorest overlapping PTV contour (DSC = 0.7612) on T2 MR. Observer A (red), B (purple), and C (orange) contours are outlined on (clockwise from top) transverse, coronal, and sagittal images. STAPLE volume is shaded light blue.

205

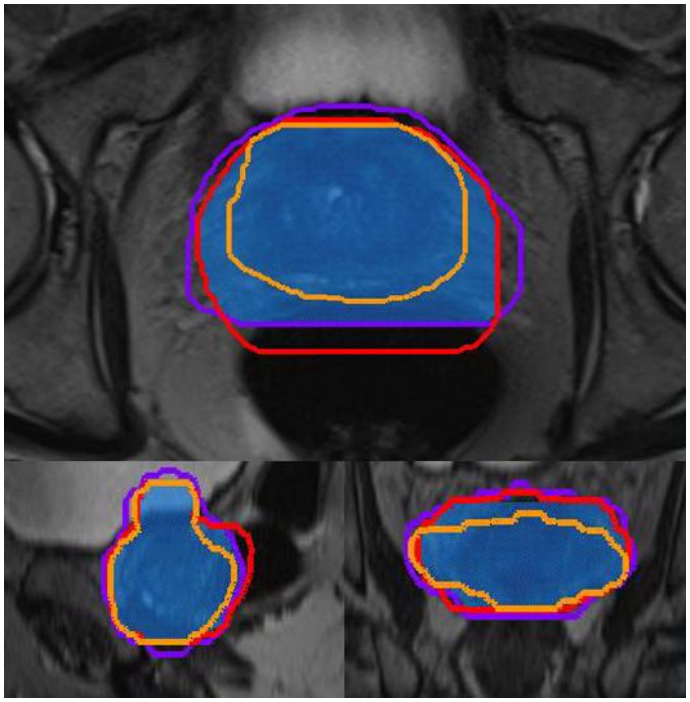

Figure 4. Patient 1 PTV contours on T2 MR. All observer contours recorded DSC > 0.9 with respect to the STAPLE volume shaded in light blue. However, due to significant portions of Observer C's (orange) PTV failing to include the STAPLE volume, insufficient dose was delivered to the STAPLE

210 PTV for these treatment plans.

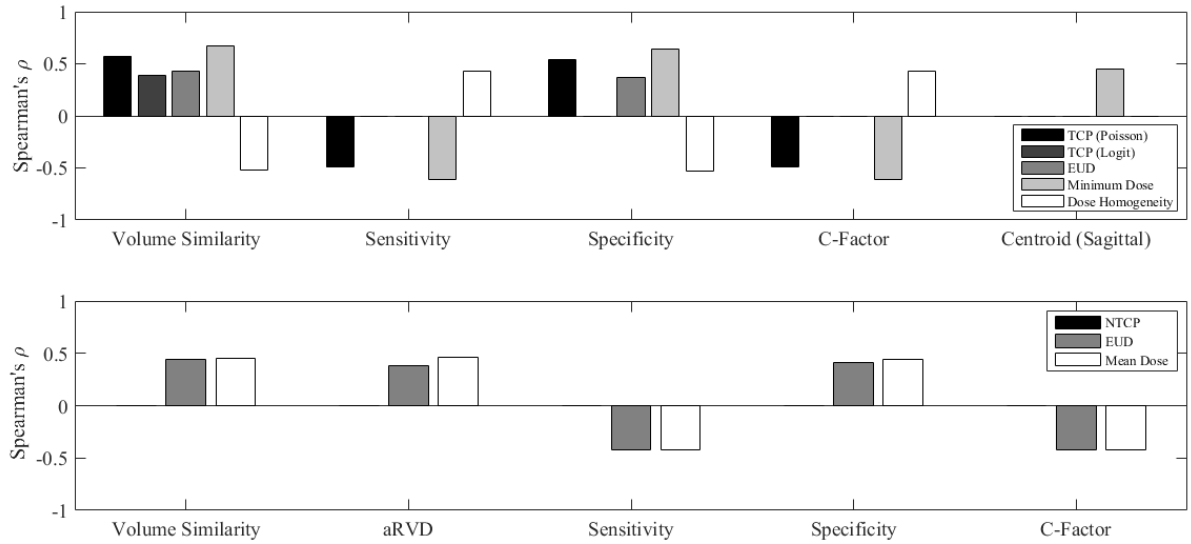

Figure 5. Significant spearman correlations for PTV (top image, $p < 0.00035$) and rectum (bottom image, $p < 0.00026$) for trial 1. Volume similarity, sensitivity, specificity and C-Factor significantly correlated with a range of radiobiological metrics for both structures. Most correlations identified were weak, although $TCP_{Poisson}$, minimum dose, and dose homogeneity showed moderate correlations with volume similarity, sensitivity, and specificity for PTV.

boundary metrics (mean absolute surface distance and 95% Hausdorff distance) showed no significant correlations with dosimetry for PTV, while DSC only showed minimal correlations (ρ = -0.35 – 0.38) with multiple rectum isodose curves (supplementary table 5). Analysis of additional gold standard volumes found the same similarity metrics exhibiting the strongest correlations with dosimetry (supplementary material tables 8 – 11).

Correlations between PTV contouring similarity metrics and OAR dosimetry were also investigated. No significant correlations were observed for bladder dosimetry. Figure 6 plots PTV volume similarity against rectum NTCP, where a weak correlation (ρ = 0.33) was observed.

3.1. Trial 2 Results

Figure 7 plots CTV, PTV, bladder and rectum volumes for trial 2 datasets, while spread in DSC is

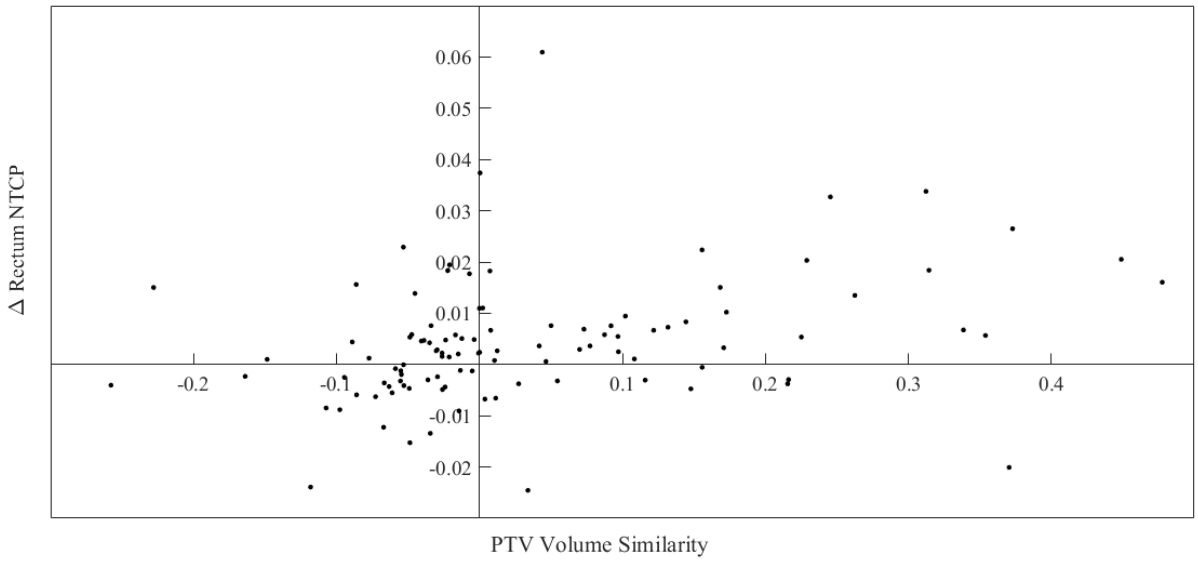

Figure 6. PTV volume similarity vs. Rectum NTCP, ρ = 0.33.

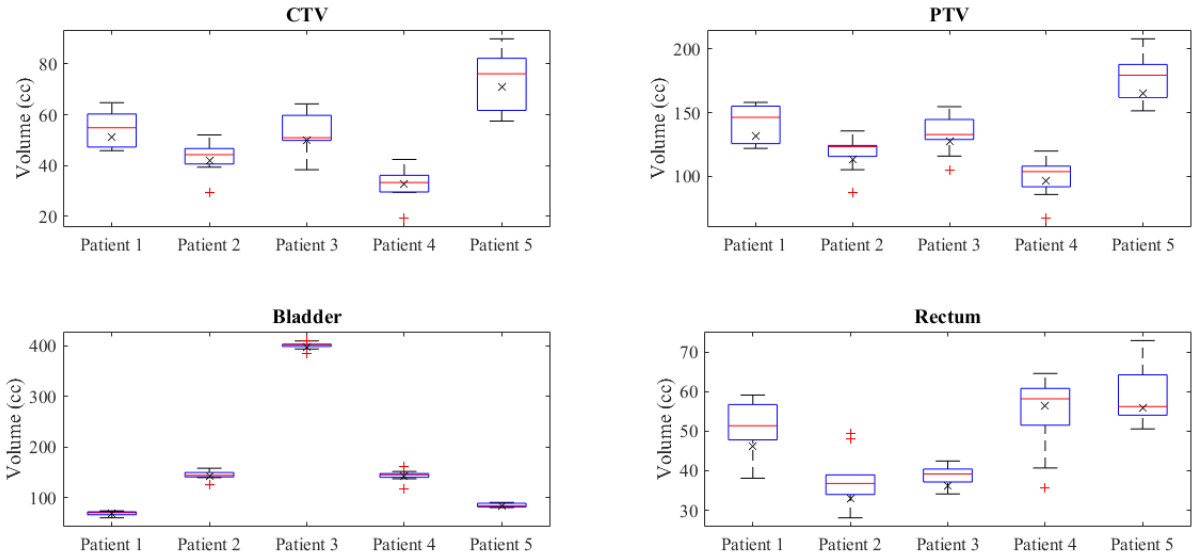

Figure 7. Variations in contoured volumes for each structure across the five patients in trial 2. Gold standard volumes for each structure are shown as a black cross. Due to the use of a majority vote, each gold standard volume is by definition smaller than the median volume for each structure.

illustrated in figure 8. Mean, standard deviation, and coefficient of variation for these volumes are again summarised in table 6, with additional descriptive statistics found in supplementary tables 12 and 13. Differences between observer and majority voting plan dosimetry are shown in table 7. Patient 4 recorded the smallest mean CTV volume (32.88 cc), while patient 5 had the largest mean CTV (74.02 cc).

Similar trends were observed for PTV volumes (patient 4: 99.61 cc, patient 5: 177.41 cc). Differences between mean rectum volumes between patients was less pronounced, ranging from patient 2 (37.69 cc) through to patient 5 (58.94 cc). Bladder volumes varied significantly between patients, ranging from patient 1 (69.27 cc) through to patient 3 (400.13 cc). Figure 9 shows all observer PTV contours for patient 4, as well as the majority vote PTV shaded in light blue. Observer 1 (green contour) recorded the poorest overlap (DSC = 0.8167) of all PTV volumes.

Table 7 again shows significant differences in PTV dosimetry, due to multiple observer PTV contours failing to adequately cover the majority vote PTV. Mean rectum dosimetry was marginally decreased for observer treatment plans, while mean bladder dosimetry improved compared to the gold standard treatment plan. Observer A's PTV from figure 9 is an example of a PTV failing to treat the majority vote PTV, with the corresponding treatment plan shown in figure 10. Complete lists of correlations between contouring similarity metrics and radiobiological metrics are given in supplementary tables 14 - 17.

No significant correlations between contouring similarity metrics and dosimetry were observed for either CTV or rectum. Bladder correlations were quite variable, a shifting of the coronal plane correlated moderately with NTCP (ρ = -0.51) and multiple isodose curves. Additionally, both

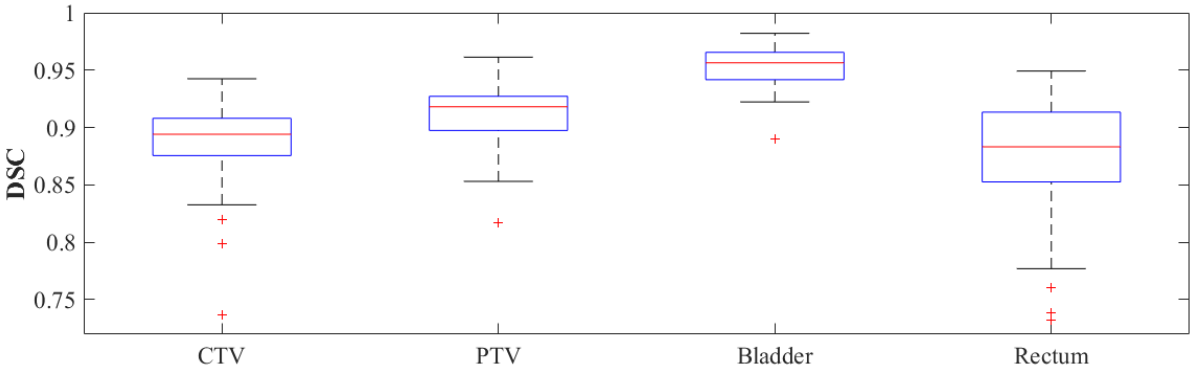

Figure 8. Spread in observer DSC for CTV, PTV, bladder, and rectum across all patients for trial 2.

DSC (ρ = 0.62) and aRVD (ρ = -0.54) were found to correlate with the maximum dose delivered to the bladder.

Specificity displayed the strongest correlations with dosimetry for PTV, with strong correlations observed for $TCP_{Poisson}$ (ρ = 0.81), TCP_{Logit} (ρ = 0.87), EUD (ρ = 0.85), and dose homogeneity (ρ = -0.82). Comparatively, while multiple radiobiological correlations for sensitivity were significant, they were only moderate in strength (ρ = -0.57, -0.52, and 0.60 for TCP_{Logit}, EUD, and dose homogeneity respectively). C-Factor correlations therefore matched the poorer performing sensitivity correlations. Volume similarity again correlated strongly with a range of radiobiological metrics for PTV, with the strongest correlations being dose homogeneity (ρ = -0.76), TCP_{Logit} (ρ = 0.70), EUD (ρ = 0.67), $TCP_{Poisson}$ (ρ = 0.65), and minimum dose (ρ = 0.61). These correlations are illustrated in figure 11.

Correlations between PTV volume similarity and rectum NTCP were again investigated, where a stronger (yet still weak) correlation of ρ = 0.48 was observed.

4. Discussion

Following treatment plan simulation, it was found that volumetric metrics (volume similarity, sensitivity, and specificity) assessing inter-observer contouring variation for PTV routinely correlated significantly with multiple radiobiological metrics. This was assessed using two separate trials, the first comprising of a small number of observers and large number of patients, the second with a large

Table 6. Mean, standard deviation, and coefficient of variation of observer structure volumes for trial 2 patients.

	Patient	CTV	PTV	Bladder	Rectum
1	Mean (cc)	54.43	141.99	69.27	51.13
	SD (cc)	6.58	14.42	4.13	6.20
	COV	0.12	0.10	0.06	0.12
	Majority vote (cc)	51.32	131.79	68.24	46.23
2	Mean (cc)	43.38	118.88	144.81	37.69
	SD (cc)	6.13	13.51	9.03	6.81
	COV	0.14	0.11	0.06	0.18
	Majority vote (cc)	42.13	113.44	143.33	32.89
3	Mean (cc)	52.64	133.17	400.13	38.61
	SD (cc)	7.96	15.37	7.38	2.78
	COV	0.15	0.12	0.01	0.07
	Majority vote (cc)	49.92	127.59	396.46	36.19
4	Mean (cc)	32.88	99.61	143.21	54.79
	SD (cc)	6.59	15.68	12.08	9.95
	COV	0.20	0.16	0.08	0.18
	Majority vote (cc)	32.81	96.41	142.10	56.35
5	Mean (cc)	74.02	177.41	84.91	58.94
	SD (cc)	11.14	20.12	3.82	6.82
	COV	0.15	0.11	0.05	0.12
	Majority vote (cc)	70.75	165.25	84.11	55.96

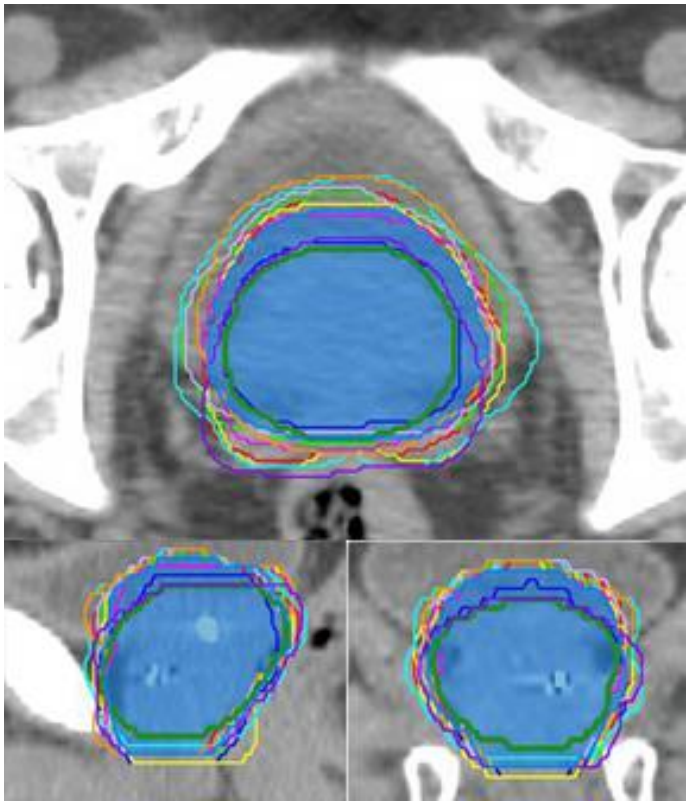

Figure 9. Patient 4 PTV contours on (clockwise from top) transverse, coronal, and sagittal images. Observer A (highlighted dark green) displayed the poorest overlapping PTV with respect to the majority vote PTV (shaded light blue), with a DSC of 0.8167.

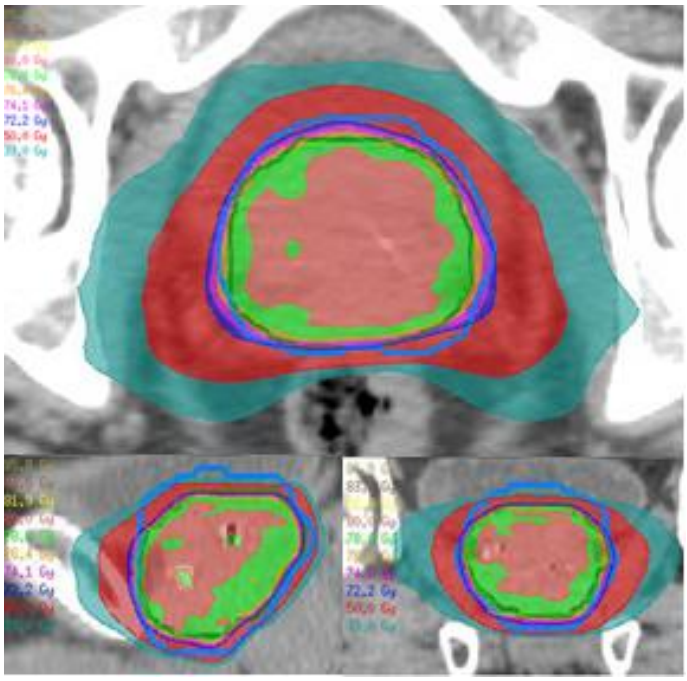

Figure 10. Patient 4 dose distribution derived from observer A's PTV contour (figure 9). The majority vote PTV is outlined in light blue, while the 78 Gy, 50 Gy, and 39 Gy isodose lines are shaded light green, red, and light blue respectively. This treatment plan resulted in zero tumour control probability for the majority vote PTV. It can clearly be seen on the sagittal and coronal slices that significant portions of the majority vote PTV were under-dosed during treatment planning.

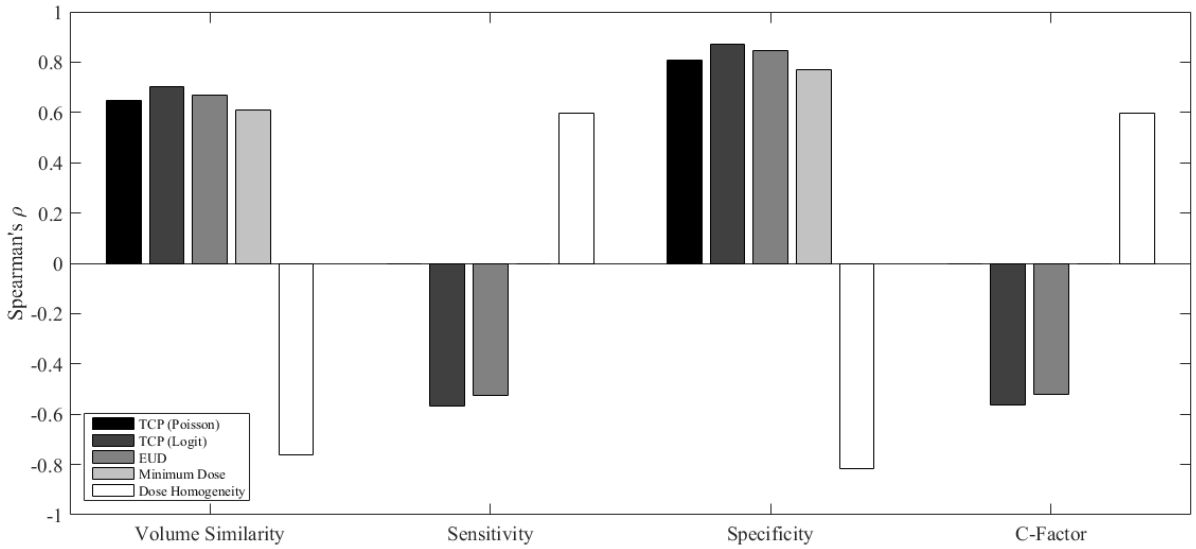

Figure 11. Significant spearman correlations for PTV with $p < 0.00035$ for trial 2. Correlations for PTV were much stronger than those observed in trial 1 (figure 5), ranging from moderate (sensitivity, C-Factor) to strong (volume similarity, specificity).

number of observers and small number of patients, ensuring the study captured the breadth of contouring variation that would be observed clinically. Additionally, variations in volume similarity when assessing PTV contours were found to moderately correlate with rectum NTCP. This study provides evidence linking inter-observer contouring variation metrics with dosimetry and treatment plan quality for prostate cancer patients.

Vinod *et al.* cited 16 studies investigating inter-observer contouring variation for GU structures, identifying the most popular similarity metrics for these studies as volume, surface, overlap, and centre-of-mass respectively (Vinod *et al.*, 2016a). Only a few studies included an assessment on dosimetry (Livsey *et al.*, 2004; Mitchell *et al.*, 2009; Foppiano *et al.*, 2003; Perna *et al.*, 2011), with these studies investigating impact on organ-at-risk dosimetry opposed to treatment target volumes. As stated by Vinod *et al.*, an assessment of contouring variation based solely on geometry may have little to no clinical significance, especially where no correlations between contouring metrics and dosimetry have been performed. Additionally, commonly cited spatial overlap metrics conformity index and DSC have been shown to display only minimal to no significant correlations with dosimetry for non-small cell lung cancers (Jameson *et al.*, 2014) and head and neck cancers (Beasley *et al.*, 2016) respectively. As no study correlating contouring similarity metrics and dosimetry had been performed for prostate radiotherapy, this study was undertaken to bring clinical relevancy and allow insights into target volume and organ-at-risk dosimetry to other inter-observer contouring variation studies.

Two trials were incorporated in this study; the first involving a small number of observers with many patients, the second a large number of observers with a small number of patients. This multiple trial analysis allowed for a technique to determine whether the calculated correlations were invariant to the trial setup. Fotina *et al.* introduced a method for calculating the minimum number of

Table 7. Mean differences between trial 2 observer and majority vote plan dosimetry

	Δ $TCP_{Poisson}$		**Δ TCP_{Logit}**		**Δ NTCP**		**Δ EUD**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CTV	0.0000	0.0001	-0.0014	0.0035	--	--	-0.11	0.27
PTV	-0.3498	0.4547	-0.0150	0.0380	--	--	-4.34	11.35
Bladder	--	--	--	--	0.0003	0.0009	0.44	2.75
Rectum	--	--	--	--	-0.0022	0.0248	-0.20	2.23

observers required within an inter-observer contouring study in order for a minimal acceptable level of study reliability to be reached (Fotina *et al.*, 2012). This is derived from the intraclass correlation coefficient (ICC), and suggests that a minimum value of 0.8 for the ICC is required. Table 8 lists the calculated ICC and corresponding minimum number of observers required for each structure across both trials. It was shown that each trial satisfied this criterion, with the three observers used within the first trial considered sufficient. It should be noted that, due to the initial derivation of ICC requiring knowledge of the number of observers, the derived number of observers is meant only to assess whether the trial's observer numbers are sufficient to produce reliable results. It does not give the recommended number of observers required for contouring analysis within the study.

No similarity metrics assessing CTV contouring variations correlated significantly with any radiobiological metric across both trials. As department treatment planning required uniform PTV dose distributions, variations in CTV dose distributions were negligible (tables 5 and 7), suggesting that the PTV margins applied adequately accounted for inter-observer CTV contouring variation. Within the initial trial substantial agreement between observer bladder contours was observed (figures 1 and 2), resulting in minimal variations in bladder dosimetry between observer treatment plans. Consequently, the insufficient spread in data prevented significant correlations for bladder being observed. Only when additional observers were included during the second trial were correlations between contouring variation metrics and dosimetry observed.

Curiously, additional observers within the second trial resulted in the loss of significance in correlations between contouring and radiobiological metrics for the rectum. A possible explanation could be the wider range of slices contoured by observers when defining the superior portion of the rectum during the second trial. Within the first trial all three observers contoured the rectum on approximately equivalent slices. Consequently, the variations in contouring were more significant within each slice, and hence within regions where higher dose would have been administered. As these correlations were only weak (table 3), they were subsequently masked when larger contouring variations were introduced during the second trial due to the variable number of slices contoured, coinciding with a region receiving less dose.

DSC and 95% Hausdorff distance are commonly quoted contouring similarity metrics within the literature, however in this study these metrics exhibited little or no significant correlation with dosimetry for all structures. An explanation for this is the inability for these metrics to differentiate between observer contours that lie either within or outside the gold standard volume. Figure 4 illustrates this. For this patient, all three observer PTV contours recorded DSC scores greater than 0.9, typically considered as excellent overlap (Zijdenbos *et al.*, 1994). Adequate treatment plans for the STAPLE target volumes were simulated based on observer A and B contours. However, due to observer C's PTV lying entirely within the STAPLE volume, significant regions of the STAPLE PTV received inadequate dose coverage during treatment planning. Consequently, markedly different treatment outcomes were simulated between the observer treatment plans, despite nearly identical DSC values being recorded. This is also observed in figures 9 and 10, where the smaller PTV from observer 1 resulted in inadequate coverage of the gold standard PTV. Volume similarity, sensitivity, and specificity possess the ability to differentiate between contours lying within or outside the gold standard volume, consequently these similarity metrics correlated with target volume dosimetry.

A concern when utilising volume similarity may be the lack of spatial information this metric provides, such that two hypothetical observer contours could each have equal volume, but share no

Table 8. Intraclass correlation coefficients and minimum required number of observers for study reliability

	Trial 1		**Trial 2**	
	ICC	Observers	ICC	Observers
CTV	0.9362	3 (2.45)	0.9752	1 (0.54)
PTV	0.9345	3 (2.99)	0.9719	1 (0.66)
Bladder	0.9969	1 (0.03)	0.9979	1 (0.03)
Rectum	0.9345	2 (1.82)	0.9495	1 (0.87)

spatial overlap. However, assuming adequate scan resolution, this hypothetical should never occur between two experienced observers for structures considered in this study. Inspection of the patient datasets found that variations in contouring for rectum, bladder, and CTV were found around the perimeter of the contours, never in the localisation of the structures. Another point of contention within this study was the choice of either STAPLE or majority vote as reference gold standard volumes for each trial. Additional analysis involving multiple gold standard volumes utilising the trial 1 dataset produced equivalent correlations to those seen in trial 1 (see supplementary tables 8 - 11).

Larger PTV volumes compromises dose sparing to nearby OARs, with a statistically significant correlation between PTV volume similarity and rectum NTCP observed in both trials (ρ = 0.33, 0.48), and shown for trial 1 in figure 6. A previous study investigating the impact of CTV and PTV contouring variations on rectum dosimetry was unable to observe statistically significant correlations (Livsey *et al.*, 2004). This discrepancy may be due to their study utilising 3D-CRT treatment plans, opposed to VMAT treatment planning incorporated in this study. VMAT treatment plans generate steeper dose gradients around the target volumes, increasing the sensitivity of target volume dosimetry to inter-observer contouring variations (also shown in figure 10). It is anticipated that future studies investigating SBRT treatment plans, where higher dose fractions are delivered, may yield stronger correlations than were observed in this study.

Within the initial trial observer B consistently contoured significantly larger CTV and PTV volumes, resulting in this observer's bias being over-represented during analysis. When investigating the cause of this discrepancy, it was found that observer A and C had prior experiences working with one another on prostate delineation projects. Consequently, both observers interpreted and contoured structures in this study in a similar fashion. This illustrates the importance of peer review during contouring for clinical trials, although as of yet no study has investigated the statistical impact of peer review on inter-observer contouring variation (Vinod *et al.*, 2016b). Contouring is routinely regarded as one of the highest priorities for peer review during a clinical trial (Marks *et al.*, 2013; Brundage *et al.*, 2013), with surveys showing up to 59% of observer contours undergo peer review (Hoopes *et al.*, 2015). For these reasons, it was decided that a second trial arm should be undertaken, to ensure that the range of contouring variation observed could be assumed to be representative of those found in general clinical practice.

The second trial arm included contours from multiple volunteers across a range of treatment centres. The large number of observers was required for an ongoing study, however included medical physicists and radiographers who lacked experience in contouring. As poor clinical training within this study could produce results not relevant to standard clinical practise, additional analysis utilising only the radiation oncologist and radiotherapist contours was performed. Adjusted CTV, PTV, bladder and rectum mean volumes (52.38 cc, 135.83 cc, 168.50 cc, 48.98 cc) and DSC (0.8976, 0.9190, 0.9561, 0.8807) across the five patients were comparable to the complete dataset (51.85 cc, 134.92 cc, 168.98 cc, 48.10 cc and 0.8872, 0.9118, 0.9530, 0.8724 respectively). Wilcoxon rank sum testing revealed no significant difference ($p < 0.05$) between either dataset. Additional analysis of correlations using this subset resulted in near equivalent correlations for PTV between volume similarity, sensitivity, specificity, and the C-Factor with multiple radiobiological metrics (supplementary tables 18 – 21). It can therefore be concluded that the lack of experience of some observers did not impact the findings of this study.

Additionally, as contouring within the second trial arm was performed on two separate platforms (Eclipse™ and Pinnacle^3®), there was potential for bias to be introduced within the study. Analysis of CTV, PTV, bladder, and rectum volumes and DSC overlap for Eclipse users (52.11 cc, 137.49 cc, 170.91 cc, 47.62 cc and 0.9013, 0.9228, 0.9554, 0.8913 respectively) compared to Pinnacle users (51.60 cc, 132.45 cc, 167.13 cc, 48.56 cc and 0.8737, 0.9013, 0.9507, 0.8543 respectively) revealed slightly improved overlap for CTV, PTV and Rectum from Eclipse users. Wilcoxon rank sum testing confirmed that distribution of DSC values for these structures differed significantly ($p < 0.05$) between the two treatment planning systems. While this confirmed a bias in spatial overlap dependent on treatment planning system (and consequently, treatment centre), additional analysis where either only Pinnacle or Eclipse observer contours were investigated again revealed identical correlations between contouring and radiobiological metrics that were statistically significant. Consequently, this bias was deemed by the authors not to be impactful to the studies aims.

Treatment plan generation was another potential study limitation, as poor quality treatment planning could mask the impact inter-observer contouring variability has on dosimetry. By utilising the Autoplanning module within Pinnacle^3®, and having treatment plans subsequently assessed by an experienced radiation therapist, treatment plans of similar quality satisfying department planning protocol were ensured across all observer contour sets.

Finally, it should be noted that a key difference between the two trial arms was the difference in scans used by observers for contouring. Trial 1 required contouring on T2-weighted MR scans, that were subsequently registered and fused to planning CT, while trial 2 required contouring to be performed directly on CT. However, as gold standard volumes were constructed from observer contours, differences in CTV and PTV volume typically observed between CT and MR scans in prostate contouring studies (Debois *et al.*, 1999) would be reflected in differences in gold standard volumes between the trials. As the study was investigating differences between observer contours with respect to these gold standard volumes, the impact of differing CTV and PTV volumes between the trials were compensated for during the analysis. Additionally, figures 2 and 6 reveal the spread in DSC across all structures for both trials, where similar trends were observed. These DSC values lie within the range typically seen within clinical trials (Sharp *et al.*, 2014), consequently the different imaging modality used by each trial was again felt not to be impactful to the study.

It is important to consider the clinical impact, and not just effect on dosimetry, that inter-observer contouring variations are responsible for. Incorrect contouring has been shown to be a significant cause of poor quality radiotherapy delivered in both pancreatic (Abrams *et al.*, 2012) and head and neck cancer trials (Peters *et al.*, 2010). With the latter it was found that poor quality treatment plans, from which 25% were attributed to poor quality contouring, were responsible for up to 20% reduction in both locoregional failure-free control and overall patient survival. Additional meta-analyses of clinical trials across multiple treatment sites concluded that radiotherapy protocol deviations were associated with both reduced treatment efficacy and patient survival rates (Ohri *et al.*, 2013; Weber *et al.*, 2012). In each meta-analysis, poor quality contouring was again concluded to be a significant factor for protocol deviations.

Additionally, the use of large rectangular treatment fields for prostate radiotherapy resulted in fewer clinical failures compared to the tightly conformal fields used in intensity-modulated radiotherapy and volumetric modulated arc therapy (Heemsbergen *et al.*, 2013). It was concluded that increased dose delivered to regions just outside the defined prostate, where subclinical spread of disease could be present, were partly responsible for this increase in treatment efficacy. Consequently, the accuracy of contouring is paramount in ensuring adequate dose coverage across the entire treatment volume for radiotherapy, and an assessment of accuracy based on treatment efficacy is required to relate contouring variations with clinical relevancy. Thus, this study is not just of interest dosimetrically, but has direct clinical relevance in improving the largest source of uncertainty in radiation treatment (Weiss and Hess, 2003).

As well as inter-observer contouring studies, contouring similarity metrics are routinely utilised during atlas development and validation (Acosta *et al.*, 2014). Automatically contoured structures generated by an atlas for a query patient must be assessed for accuracy and precision, which usually occurs using the clinician's original contour and commonly utilised similarity metrics DSC and Hausdorff distance. A recent proof of concept study found that inclusion of dosimetry assessment during automatic contouring of the prostate improved rectum dose sparing and dosimetry (Chang *et al.*, 2017). Importantly, these improvements came despite the automatically contoured structures recording nearly identical DSC and Hausdorff Distances to contours generated with no consideration of dosimetry. While these results were used to validate their automatic contouring algorithm, it also suggested that these metrics were unable to differentiate between contoured structures with varying dosimetry.

Volume similarity, sensitivity, and specificity significantly correlated with PTV dosimetry across both trials for prostate cancer radiotherapy. Conversely, correlations for rectum and bladder were only observed during trials 1 and 2 respectively. Due to correlations for these structures being only weak to moderate in magnitude, not being replicable across both trials, and due to no correlations for CTV being observed in either trial, a combination of contouring similarity metrics should still be cited during future inter-observer contouring variation studies. Fotina *et al.* recommend reporting a combination of overlap and statistical measures of agreement during analysis (Fotina *et al.*, 2012).

This study additionally advocates the reporting of volume similarity, as this would allow inter-observer contouring variation studies for prostate cancer radiotherapy to provide additional information on PTV dosimetry.

5. Conclusion

This study is the first to show statistically significant correlations between inter-observer contouring variations for prostate cancer radiotherapy and simulated patient dosimetry. Multiple significant correlations were observed between volume similarity, sensitivity, and sensitivity, and PTV dosimetry during both trials. Correlations between contouring similarity metrics and bladder and rectum dosimetry across the two trials were more variable, however variations in contouring PTV significantly correlated with differences in rectum dosimetry. No significant correlations between contouring similarity metrics and CTV dosimetry were observed. This study will greatly enhance future inter-observer contouring variation studies for prostate cancer radiotherapy, guiding contouring similarity metric choice to allow for insights into dosimetry and clinical relevancy during analysis.

Conflict of interest

None

Acknowledgement

The authors would like to thank Dr Wei Xuan for statistical advice, and Kirrily Cloak and Rohan Gray for treatment planning feedback. Additional thanks is given to Michelle Krawiec, Robba Rai, Prof Jim Denham, Dr Jeremiah De Leon, Dr Karen Lim, Dr Megan Berry, Dr Rohen White, Prof Sean Bydder, Dr Hendrick Tan, Dr Jeremy Croker, Dr Alycea McGrath, Dr Robert Jan Smeenk, and Dr John Matthews. The project was funded by NHMRC project grant number 1077788.

Appendix A. Supplementary Material

References

Abrams R A, Winter K A, Regine W F, Safran H, Hoffman J P, Lustig R, Konski A A, Benson A B, Macdonald J S and Rich T A 2012 Failure to adhere to protocol specified radiation therapy guidelines was associated with decreased survival in RTOG 9704—a phase III trial of adjuvant chemotherapy and chemoradiotherapy for patients with resected adenocarcinoma of the pancreas *Int. J. Radiat. Oncol. Biol. Phys.* **82** 809-16

Acosta O, Dowling J, Drean G, Simon A, De Crevoisier R and Haigron P 2014 *Abdomen and Thoracic Imaging*: Springer) pp 623-56

Al-Mamgani A, Heemsbergen W D, Peeters S T and Lebesque J V 2009 Role of intensity-modulated radiotherapy in reducing toxicity in dose escalation for localized prostate cancer *Int. J. Radiat. Oncol. Biol. Phys.* **73** 685-91

Beasley W J, McWilliam A, Aitkenhead A, Mackay R I and Rowbottom C G 2016 The suitability of common metrics for assessing parotid and larynx autosegmentation accuracy *J. Appl. Clin. Med. Phys.* **17**

Brundage M, Foxcroft S, McGowan T, Gutierrez E, Sharpe M and Warde P 2013 A survey of radiation treatment planning peer-review activities in a provincial radiation oncology programme: current practice and future directions *BMJ open* **3** e003241

Burdett N, Fripp J, Bourgeat P and Salvado O 2010 *E-Health*: Springer) pp 177-86

Chang J, Tian Z, Lu W, Gu X, Chen M and Jiang S B 2017 A novel geometry-dosimetry label fusion method in multi-atlas segmentation for radiotherapy: a proof-of-concept study *Phys. Med. Biol.* **62** 3656

Debois M, Oyen R, Maes F, Verswijvel G, Gatti G, Bosmans H, Feron M, Bellon E, Kutcher G and Van Poppel H 1999 The contribution of magnetic resonance imaging to the three-dimensional treatment planning of localized prostate cancer *Int. J. Radiat. Oncol. Biol. Phys.* **45** 857-65

Dowling J A, Sun J, Pichler P, Rivest-Hénault D, Ghose S, Richardson H, Wratten C, Martin J, Arm J and Best L 2015 Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences *Int. J. Radiat. Oncol. Biol. Phys.* **93** 1144-53

Dubois D F, Prestidge B R, Hotchkiss L A, Prete J J and Bice Jr W S 1998 Intraobserver and interobserver variability of MR imaging-and CT-derived prostate volumes after transperineal interstitial permanent prostate brachytherapy *Radiology* **207** 785-9

Foppiano F, Fiorino C, Frezza G, Greco C, Valdagni R and Radiotherapy A N W G o P 2003 The impact of contouring uncertainty on rectal 3D dose–volume data: Results of a dummy run in a multicenter trial (AIROPROS01–02) *Int. J. Radiat. Oncol. Biol. Phys.* **57** 573-9

Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R and Georg D 2012 Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy *Strahlenther. Onkol.* **188** 160-7

Frey B J and Dueck D 2007 Clustering by passing messages between data points *Science* **315** 972-6

Gao Z, Wilkins D, Eapen L, Morash C, Wassef Y and Gerig L 2007 A study of prostate delineation referenced against a gold standard created from the visible human data *Radiother. Oncol.* **85** 239-46

Heemsbergen W D, Al-Mamgani A, Witte M G, van Herk M and Lebesque J V 2013 Radiotherapy with rectangular fields is associated with fewer clinical failures than conformal fields in the high-risk prostate cancer subgroup: results from a randomized trial *Radiother. Oncol.* **107** 134-9

Hinkle D E, Wiersma W and Jurs S G 2003 Applied statistics for the behavioral sciences

Holloway L C, Miller J-A, Kumar S, Whelan B M and Vinod S K 2012 Comp Plan: A computer program to generate dose and radiobiological metrics from dose-volume histogram files *Med. Dosim.* **37** 305-9

Hoopes D J, Johnstone P A, Chapin P S, Kabban C M S, Lee W R, Chen A B, Fraass B A, Skinner W J and Marks L B 2015 Practice patterns for peer review in radiation oncology *Practical radiation oncology* **5** 32-8

ICRU 2010 ICRU Report 83: Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT) *Journal of the ICRU* **10**

Jameson M G, Holloway L C, Vial P J, Vinod S K and Metcalfe P E 2010 A review of methods of analysis in contouring studies for radiation oncology *J. Med. Imaging Radiat. Oncol.* **54** 401-10

Jameson M G, Kumar S, Vinod S K, Metcalfe P E and Holloway L C 2014 Correlation of contouring variation with modeled outcome for conformal non-small cell lung cancer radiotherapy *Radiother. Oncol.* **112** 332-6

Kennedy A M, Dowling J A, Ghose S, Rivest-Hénault D and Ebert M A 2016 Cluster Based selection of CT exemplars from a clinical dataset for inter-patient registration and dose mapping. Poster in: *International Conference on the use of Computers in Radiation Therapy,* London

Khoo E L, Schick K, Plank A W, Poulsen M, Wong W W, Middleton M and Martin J M 2012 Prostate contouring variation: can it be fixed? *Int. J. Radiat. Oncol. Biol. Phys.* **82** 1923-9

Livsey J E, Wylie J P, Swindell R, Khoo V S, Cowan R A and Logue J P 2004 Do differences in target volume definition in prostate cancer lead to clinically relevant differences in normal tissue toxicity? *Int. J. Radiat. Oncol. Biol. Phys.* **60** 1076-81

Marks L B, Adams R D, Pawlicki T, Blumberg A L, Hoopes D, Brundage M D and Fraass B A 2013 Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary *Practical radiation oncology* **3** 149-56

Mitchell D M, Perry L, Smith S, Elliott T, Wylie J P, Cowan R A, Livsey J E and Logue J P 2009 Assessing the effect of a contouring protocol on postprostatectomy radiotherapy clinical target volumes and interphysician variation *Int. J. Radiat. Oncol. Biol. Phys.* **75** 990-3

Ohri N, Shen X, Dicker A P, Doyle L A, Harrison A S and Showalter T N 2013 Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials *J. Natl. Cancer Inst.* djt001

Ost P, De Meerleer G, Vercauteren T, De Gersem W, Veldeman L, Vandecasteele K, Fonteyne V and Villeirs G 2011 Delineation of the postprostatectomy prostate bed using computed tomography: interobserver variability following the EORTC delineation guidelines *Int. J. Radiat. Oncol. Biol. Phys.* **81** e143-e9

Palma D, Vollans E, James K, Nakano S, Moiseenko V, Shaffer R, McKenzie M, Morris J and Otto K 2008 Volumetric modulated arc therapy for delivery of prostate radiotherapy: comparison with intensity-modulated radiotherapy and three-dimensional conformal radiotherapy *Int. J. Radiat. Oncol. Biol. Phys.* **72** 996-1001

Perna L, Cozzarini C, Maggiulli E, Fellin G, Rancati T, Valdagni R, Vavassori V, Villa S and Fiorino C 2011 Inter-observer variability in contouring the penile bulb on CT images for prostate cancer treatment planning *Radiat. Oncol.* **6** 1

Peters L J, O'Sullivan B, Giralt J, Fitzgerald T J, Trotti A, Bernier J, Bourhis J, Yuen K, Fisher R and Rischin D 2010 Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02 *J. Clin. Oncol.* **28** 2996-3001

Popovic A, de la Fuente M, Engelhardt M and Radermacher K 2007 Statistical validation metric for accuracy assessment in medical image segmentation *Int. J. Comput. Assist. Radiol. Surg.* **2** 169-81

Quan E M, Li X, Li Y, Wang X, Kudchadker R J, Johnson J L, Kuban D A, Lee A K and Zhang X 2012 A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment *Int. J. Radiat. Oncol. Biol. Phys.* **83** 1169-78

Rasch C, Barillot I, Remeijer P, Touw A, van Herk M and Lebesque J V 1999 Definition of the prostate in CT and MRI: a multi-observer study *Int. J. Radiat. Oncol. Biol. Phys.* **43** 57-66

Sharp G, Fritscher K D, Pekar V, Peroni M, Shusharina N, Veeraraghavan H and Yang J 2014 Vision 20/20: perspectives on automated image segmentation for radiotherapy *Med. Phys.* **41** 050902

Trans-Tasman Radiation Oncology Group (TROG) 2005 RADAR Trial - Randomised Androgen Deprivation and Radiotherapy.

Van Dyk J, Batista J and Bauman G S 2013 Accuracy and uncertainty considerations in modern radiation oncology *The Modern Technology of Radiation Oncology* **3** 361-412

Vinod S K, Jameson M G, Min M and Holloway L C 2016a Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies *Radiother. Oncol.*

Vinod S K, Min M, Jameson M G and Holloway L C 2016b A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology *J. Med. Imaging Radiat. Oncol.* **60** 393-406

Warfield S K, Zou K H and Wells W M 2004 Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation *Medical Imaging, IEEE Transactions on* **23** 903-21

Weber D C, Tomsej M, Melidis C and Hurkmans C W 2012 QA makes a clinical trial stronger: evidence-based medicine in radiation therapy *Radiother. Oncol.* **105** 4-8

Weiss E and Hess C F 2003 The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy *Strahlenther. Onkol.* **179** 21-30

615 Zijdenbos A P, Dawant B M, Margolin R and Palmer A C 1994 Morphometric analysis of white matter lesions in MR images: method and validation *Medical Imaging, IEEE Transactions on* **13** 716-24