Novel Methodology to assess the Effect of Contouring Variation on Treatment Outcome

DOI:
10.1002/mp.14865

Document Version
Accepted author manuscript

Novel Methodology to assess the Effect of Contouring Variation on Treatment Outcome

Running title: Contour variation and treatment outcome

Authors:

Alexander Jenkins[#]**,**
Thomas Soares Mullen[+#]**,**
Corinne Johnson-Hart[&]**,**
Andrew Green,
Alan McWilliam,
Marianne Aznar,
Marcel van Herk and
Eliana Vasquez Osorio

Affiliations:

- AJ and TSM: University of Manchester, School of Physics & Astronomy - Faculty of Science and Engineering, Manchester, M13 9PL, United Kingdom
- CJH, AG, AM, MA, MvH, EVO: University of Manchester, Division of Cancer Studies - School of Medical Sciences - Faculty of Biology- Medicine and Health, Manchester, M20 4BX, United Kingdom

Footnotes:

[+] Current address: Champalimaud Research, Champalimaud Centre for the Unknown, Avenida Brasília, Doca de Pedrouços, Lisboa 1400-038, Portugal

[&] Current address: Christie Medical Physics and Engineering, The Christie NHS Foundation Trust, Wilmslow Road, Manchester, M20 4BX

[#] These authors contributed equally to this work

Corresponding author:
Dr Eliana Vasquez Osorio
email: eliana.vasquezosorio@manchester.ac.uk

Division of Cancer Sciences | Faculty of Biology, Medicine and Health | The University of Manchester
Department 58 The Christie | Wilmslow Road | Manchester M20 4BX | Tel (+44) 161 918 7480

Abstract

Purpose: Contouring variation is one of the largest systematic uncertainties in radiotherapy, yet its effect on clinical outcome has never been analysed quantitatively. We propose a novel, robust methodology to locally quantify target contour variation in a large patient cohort and find where this variation correlates with treatment outcome. We demonstrate its use on biochemical recurrence for prostate cancer patients.

Method: We propose to compare each patient's target contours to a *consistent* and *unbiased* reference. This reference was created by auto-contouring each patient's target using an externally trained deep learning algorithm. Local contour deviation measured from the reference to the manual contour were projected to a common frame of reference, creating *contour deviation maps* for each patient. By stacking the contour deviation maps, time to event was modelled pixel-wise using a multivariate Cox proportional hazards model (CPHM). Hazard ratio (HR) maps for each covariate were created, and regions of significance found using cluster-based permutation testing on the z-statistics.

This methodology was applied to Clinical Target Volume (CTV) contours, containing only the prostate gland, from 232 intermediate- and high-risk prostate cancer patients. The reference contours were created using ADMIRE® v3.4 (Elekta AB, Sweden). Local contour deviations were computed in a spherical coordinate frame, where differences between reference and clinical contours were projected in a 2D map corresponding to sampling across the coronal and transverse angles every 3^{o}. Time to biochemical recurrence was modelled using the pixel-wise CPHM analysis accounting for contour deviation, patient age, Gleason score and treated CTV volume.

Results: We successfully applied the proposed methodology to a large patient cohort containing data from 232 patients. In this patient cohort, our analysis highlighted regions where the contour variation was related to biochemical recurrence, producing expected and unexpected results: 1) the interface between prostate-bladder and prostate-seminal vesicle interfaces where increase of the manual contour relative to the reference was related to a *reduction* of risk of biochemical recurrence by 4-8% per mm and 2) the prostate's right, anterior and posterior regions where an increase of the manual contour relative to the reference contours was related to an *increase* of risk of biochemical recurrence by 8-24% per mm.

Conclusion: We proposed and successfully applied a novel methodology to explore the correlation between contour variation and treatment outcome. We analysed the effect of contour deviation of the prostate CTV on biochemical recurrence for a cohort of more than 200 prostate cancer patients

while taking basic clinical variables into account. Applying this methodology to a larger dataset including additional clinically important covariates and externally validating it can more robustly identify regions where contour variation directly relates to treatment outcome. For example, in the prostate case we use to demonstrate our novel methodology, external validation will help confirm or reject the counter-intuitive results (larger contours resulting in higher risk). Ultimately, the results of this methodology could inform contouring protocols based on actual patient outcomes.

Keywords:

Contour variation, treatment outcome modelling, CPHM, Data mining

Introduction

Prostate cancer is the second most common cancer in men worldwide[1] and approximately 30-60% of cases will be treated with curative-intent radiotherapy[2,3]. Radiotherapy relies on accurate definition of the target volume, i.e. the region where the prescribed dose of radiation is to be delivered. However, target definition is subjective, and is known to vary between observers; known as inter-observer variability (IOV). Variation in contouring systematically deforms the planned dose distribution relative to the (unknown) true target, potentially delivering lower dose than required to the cancer cells and influencing the effectiveness of treatment[4].

Despite the known presence of IOV in target contours, as far as the authors are aware, its effect on clinical outcomes has never been analysed quantitatively. Typically, IOV is quantified in studies where multiple observers contour the same structure on patient images[5–7]. As drawing these contours takes a long time and requires expertise, most studies include only a limited number of observers on a limited number of patient cases[7]. Furthermore, variations are often reduced to a single number, where all spatial information is lost; typically Dice similarity coefficient or Hausdorff distance[7]. For example, computing the Dice similarity coefficient of two structures involves only their respective volumes and overlap, this results in a single scalar value that contains no spatial information and it is impossible to infer how each spatial region contributed to the metric. These two drawbacks, a small cohort and the simplification of the contour differences into a single number, makes it impossible to effectively analyse the spatial effects of IOV on clinical outcome.

In recent years, automatic contouring of structures has been made possible using deep learning (DL) techniques[8]. For some organs, DL auto-contoured structures are of comparable quality to those drawn by an observer[9–12]. Additionally, advances in computational tools are also being used to automatize and improve clinical target volumes (CTV) beyond anatomical organs[13]. Contours produced by DL contouring tools can therefore be seen as being drawn by a virtual observer that is highly consistent and unbiased to data beyond the medical image being segmented.

In this study, we present a novel methodology to quantify the effect of contour deviations on clinical outcome. The methodology relies on quantifying local contour deviations by comparing the manually delineated contour with the DL generated contour, which is used as an arbitrary yet consistent reference. These local contour deviations are then analysed statistically to define regions where observer deviation correlates with outcome, taking clinical variables into account[14]. A major advantage of this approach is that instead of being restricted to a limited IOV study setting, it can exploit the information contained in large quantities of routine clinical data. As a first application

of this novel methodology, we analysed the effect of contour deviation of the prostate CTV on biochemical recurrence for a cohort of prostate cancer patients treated with radical radiotherapy.

Materials and methods

Patient dataset

247 intermediate- and high-risk prostate cancer patients, treated between 2007 and 2013 at a single institution (The Christie Hospital NHS Foundation Trust) with 57 Gy in 19 fractions of intensity-modulated radiotherapy were included in this study. Patients were setup via an empty bladder and rectum protocol, both for planning and treatment. Patients were followed-up for at least 4 years as standard of care and biochemical recurrence status, defined as a rise in the blood level of prostate-specific antigen (PSA) of 2 ng/ml above nadir after treatment, was stored for all patients. The characteristics of this cohort are summarised in Table 1. Each patient had one CTV contour encompassing the prostate gland only, defined by the treating oncologist. This contour shall be referred to as the *manual contour*. All data was collected from the ukCAT distributed learning database (ethics approval from the UK North West - Haydock Research Ethics Committee, reference number 17/NW/0060, local approval consent ukCAT ref. 2018-018).

For each patient, a DL auto-contour of the prostate gland was generated as a reference using the research version of ADMIRE® v3.4, (Elekta AB, Sweden).

Table 1: Characteristics of the cohort of prostate cancer patients, before and after refinement. The refined dataset includes only patients whose clinical target volume was limited to the prostate gland, and without anomalies in their contour deviation maps. For patients with a biochemical recurrence, the mean time to event was 4.80 years (0.85 - 8.46 years).

	Original ($n = 247$)		Refined ($n = 232$)	
Variable	№	%	№	%
Gleason score				
6	30	12	25	11
7	134	54	128	55
8	42	17	39	17
$9-10$	41	17	40	17
Age (years)				
< 65	82	33	75	33
$65-75$	147	60	140	60
> 75	18	7	17	7
Recurrence				
Yes	83	34	74	32
No	164	66	158	68

Quantifying local contour deviations

Figure 1 shows the steps for quantifying local contour deviations. The reference DL contour, and the manual contour are first triangulated into 3D surfaces using the marching cubes algorithm[15]. Then, using a similar approach as Remeijer et al.[16], local contouring deviation was measured in spherical polar coordinates. Contour deviation, δR, as a function of angles Θ (for the coronal plane) and φ (for the transverse plane) was defined as

$$\delta R(\theta, \phi) = |\overrightarrow{OM}| - |\overrightarrow{OR}| \qquad (1)$$

where distances $|\overrightarrow{OM}|$ and $|\overrightarrow{OR}|$ are quantified from the centre-of-mass of the DL reference contour (O) in the direction determined by Θ and φ (blue arrow in Figure 1b) to the point of intersection with the surface of the manual (M) and DL reference contour (R), respectively (Figure 1b). By sampling the coronal and transverse angles every 3°, contour deviation maps of 60x120 pixels were created for each patient (Figure 1c).

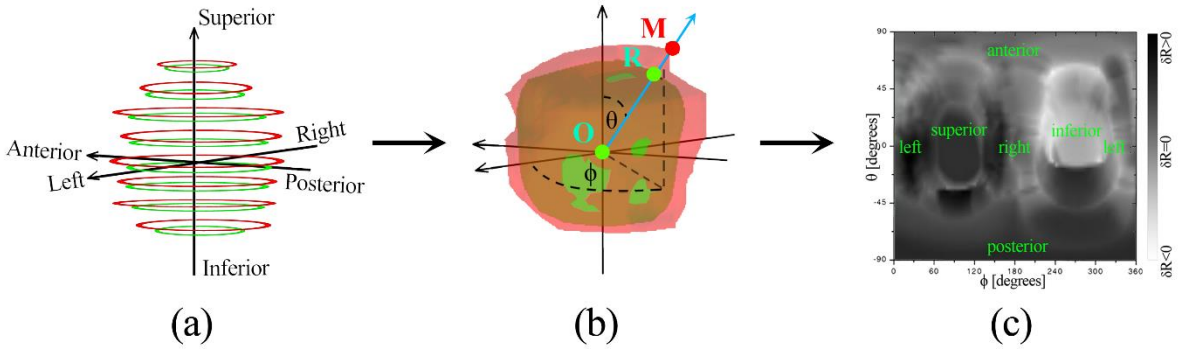


Figure 1: Method to obtain a contour deviation map for a given patient from the manual and the DL reference contours of the CTV, as shown in (a). Contours are each triangulated to form a surface and a spherical polar coordinate system, centred at the centre-of-mass of the auto-contour (O) is defined, as shown in (b). At each 3◦ × 3◦ angle, the difference between the distances from O to the manual contour and the DL reference contour δR(θ, φ) is calculated (Equation 1), to construct a single patient contour deviation map, shown in (c). Maps of hundreds of patients are next correlated with clinical outcome.

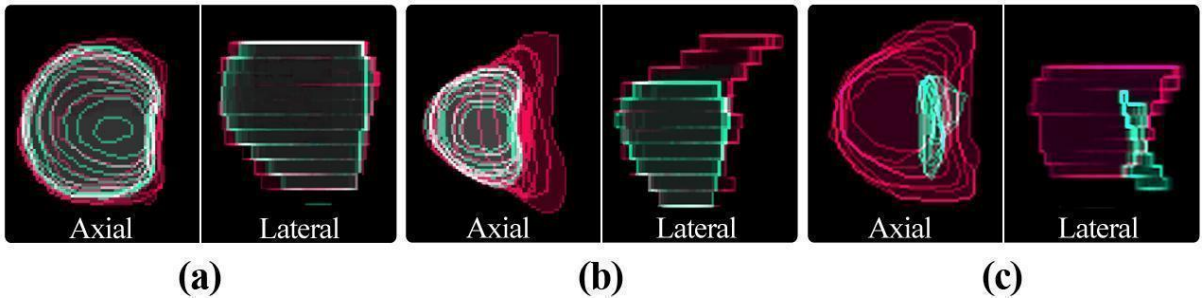


Figure 2: Examples of the axial and lateral projections of the DL auto-contour (blue) and the manual contour (red) created for each patient. These images were used to identify and remove patients from the analysis if their seminal vesicles are included in the manual contour, or if the DL contouring has failed. (a) A patient kept in the analysis with only their prostate in the manual contour. (b) A patient removed from the analysis with their prostate and seminal vesicles in the manual contour. (c) The only patient where DL contouring failed (due to artifacts caused by bilateral metal hips) was also removed.

Cohort refinement

To ensure consistency in the input data, we curated the patient data. First, the manual contour of each patient was visually verified to ensure the contour only contained the prostate gland, i.e. to ensure a consistent contouring protocol in the dataset. Secondly, the quality of the DL contour was visually verified. Patients with seminal vesicles included in the manual contour or where the DL contouring failed were removed from analysis (see Figure 2). Thirdly, the contour deviation map of each patient was visually verified to ensure intuitive deviation, (i.e. in the order of mm), in all directions. Patients with local anomalies, $\delta R(\theta, \phi)$, often in the order of 10 cm, were removed. The cause of this problem was traced back to holes in the triangulation of contours into surfaces (see Figure S1 in the supplementary materials). Visual verification was performed by AJ and TSM, to ensure consistency of contours rather than clinical correctness.

In addition to cohort refinement, we assessed the similarity between the DL contour and manual contour to identify if systematic differences exist across the cohort. We assessed this and report the following metrics: histogram of Dice similarity coefficient, scatter plot of DL vs. manual contour volumes, three scatter plots of the DL vs. manual contour x, y and z centre-of-mass coordinates respectively, and boxplots of δR values in each region from Figure 3.

Pixel-wise survival analysis

We assumed that each pixel in the contour deviation map referred to a consistent anatomical location[16]. For each pixel in the contour deviation maps, time to biochemical recurrence was modelled using a multivariable Cox Proportional Hazard Model (CPHM) accounting for contour deviation, patient age, Gleason score and manual CTV volume. Note that the last three variables

were constant for all pixels for a single patient. This methodology has been developed by Green, et al.[14], and in their publication details the development of the statistical technique. It is important to note that Green's implementation is validated with respect to the 'Survival' toolkit in R[17]; the R code for a single voxel is exemplified by figure S2 in the supplementary materials. By assembling the Hazard Ratios (HRs) of the 7200 CPHM in the 60x120 grid, HR maps for each variable were created. Regions of significance were found using the cluster-based permutation test on the z-statistics[18] (10^4 permutations). Ranges of HRs on each interface and within each prostate region are reported by extracting them using 8 rectangular regions of interest (see Figure 3).

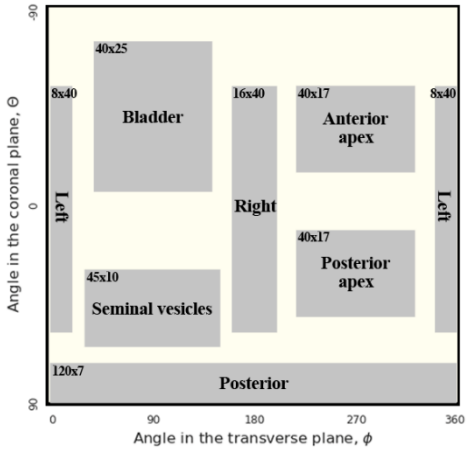


Figure 3: The shaded rectangular kernels shown here represent regions and interfaces of interest on the prostate, which were used to calculate the range of HRs for the confounding variables. The dimensions (x,y) of each kernel is denoted in its upper left and is measured in pixels.

Results

Cohort refinement

From the original 247 patients, 7 were removed as their manual contour included the seminal vesicles, 1 because of DL failure, and 9 were removed due to local anomalies in surface reconstruction, leaving 232 patients for analysis (Table 1, Figure 2).

When assessing the similarity between the DL contour and manual contour, we found that the Dice similarity coefficient for the contours was on average 0.81 (range 0.41-0.92) (see supplementary materials, Figure S3). A scatter plot of the DL vs. manual contour volume revealed that the manual contour has a consistently larger volume on average than the DL contour (see supplementary materials, Figure S4). Boxplots of δR values in each region from Figure 3, revealed that median δR values were positive and different across all regions; this confirmed our observation from Figure

S2, but also identified the posterior apex region as that with greatest systematic over-contouring (see supplementary materials, Figure S5). Scatter plots of the x, y and z centre-of-mass coordinates for the DL and manual contour revealed no systematic shifts in coordinates across the cohort (see supplementary materials, Figure S6).

Pixel-wise survival analysis

Figures 4 and 5 show the HR maps for the considered variables in the CPHM. The effect of contouring deviation, $\delta R(\Theta,\varphi)$, against biochemical recurrence, controlling for Gleason score, age, and manual CTV volume is shown in Figure 5. The contours encapsulate varying statistically significant regions of $HR < 1$ and $HR > 1$. This result shows that per mm increase of the manual contour relative to the DL reference in the prostate-bladder and -seminal vesicle interfaces, *reduces* the risk of biochemical recurrence by 4-8% ($p<0.05$). Conversely, per mm increase of the manual contour relative to the DL reference in the prostate's right, anterior and posterior regions, *increases* the risk of biochemical recurrence by 8-24% ($p<0.01$). Figure 5 shows HR maps for the controlled confounding variables. Patient age showed a significant relationship in the bladder and seminal vesicles interfaces, and the posterior and apical regions ($p<0.05$) as shown on Figure 5a. This implies an interaction between contour variation, age and biochemical recurrence. Manual CTV volume shows a significant relationship with biochemical recurrence throughout the prostate's superior ($p<0.001$), as shown on Figure 5b. When Gleason scores 7 and 8 are compared to Gleason score 9-10, as shown in figure 5c-d, all values in the HR maps are less than unity, as expected. Throughout Figure 5c all values are statistically significant ($p<0.05$), however only the region contoured is statistically significant for Figure 5d. This shows that patients with a Gleason score lower than 9-10 have a reduced relative risk of biochemical recurrence, which is an intuitive result, with some interaction with contour variation. Table 2 displays the range of HRs on each interface and within each prostate region, extracted using the rectangular regions defined in Figure 3.

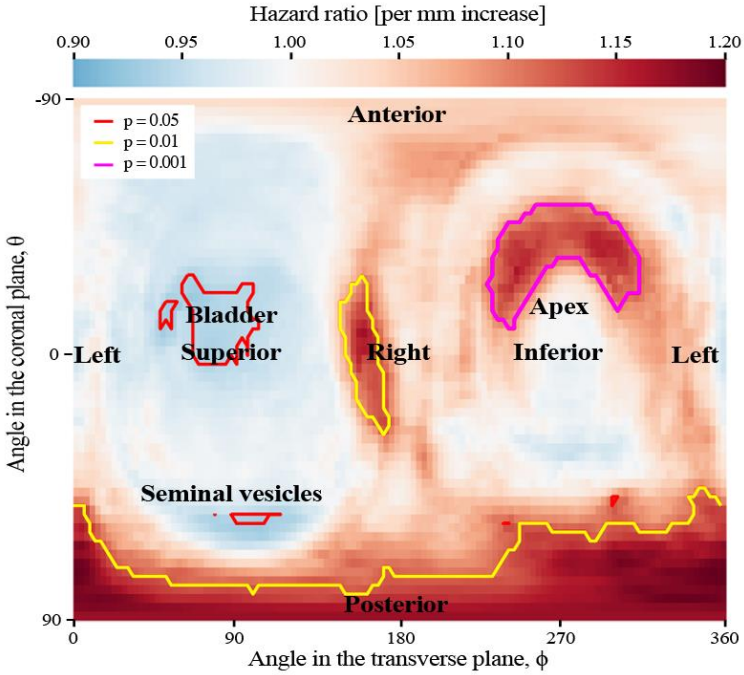


Figure 4: HR map for contouring deviation. Regions of significance are contoured. The map suggests that contouring larger volumes in the left region, bladder and seminal vesicle interfaces could lead to better biochemical control. This HR map is modulated by all variables included in the analysis.

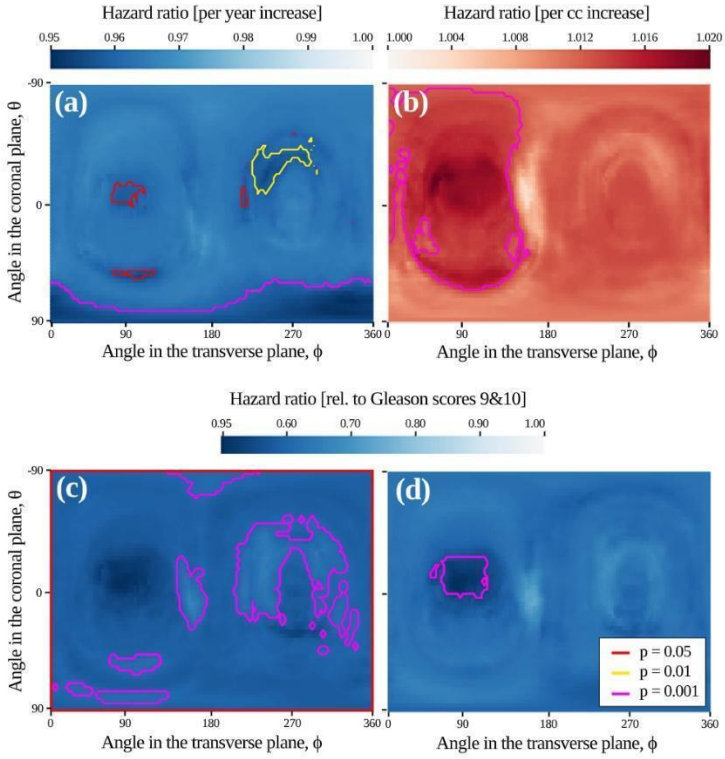


Figure 5: The HR maps of the risk of recurrence for the confounding variables modulated by the spatially varying contour deviation. (a) Age per year increase, (b) Manual CTV volume per cm^3 increase, (c) Gleason score 7 relative to 9-10, and (d) Gleason score 8 relative to 9-10. Notice that every HR map is based on modelling all variables included in the analysis.

Table 2: Range of HR for each covariate at different regions across the prostate, extracted from the HR maps shown in Figures 4 and 5, using the regions of interest defined in Figure 3.

Region	Prostate Volume (per cc)	Age (per year)	Gleason Score 7 (rel. 9-10)	Gleason Score 8 (rel. 9-10)	Contouring Variation (per mm)
Posterior	1.007-1.011	0.954-0.968	0.546-0.592	0.514-0.626	1.002-1.236
Right	1.003-1.011	0.963-0.974	0.544-0.645	0.538-0.717	0.994-1.201
Anterior Apex	1.007-1.011	0.958-0.986	0.552-0.636	0.542-0.666	0.987-1.190
Posterior Apex	1.009-1.011	0.962-0.970	0.504-0.613	0.527-0.649	0.948-1.151
Seminal Vesicles	1.008-1.013	0.958-0.969	0.547-0.600	0.560-0.608	0.916-1.078
Left	1.007-1.013	0.962-0.969	0.548-0.591	0.550-0.610	0.952-1.188
Bladder	1.004-1.016	0.962-0.971	0.476-0.598	0.454-0.628	0.915-1.088

Discussion

In this study we proposed a novel methodology to analyse the effect of contouring uncertainty on clinical outcome for a large cohort of patients. Here, we used our method on the prostate site and demonstrated how to identify regions where contour deviation and recurrence are correlated, by measuring the deviation of each patient's clinical contour relative to a highly consistent reference contour and applying pixel-wise CPHM, followed by permutation testing. It is important to notice that this methodology can be used to explore relationships to other outcomes as well.

After applying our proposed methodology to a cohort of 232 prostate cancer patients, we found regions where contour deviations were correlated with biochemical recurrence. In detail, we observed:

- A per mm increase of the manual contour relative to the DL reference in the prostate's bladder and seminal vesicle interfaces is related to a *reduction* of the risk of biochemical recurrence by 4-8% (p<0.05). This can be interpreted as having larger manual contours in these regions is associated with improved control.
- A per mm increase of the manual contour relative to the DL reference in the prostate's right, anterior and posterior regions, is related to an *increase of* the risk of biochemical recurrence by 8-24% (p<0.01). This means that larger manual contours in these regions are associated with poorer control.

To the best of our knowledge, this is the first investigation showing a direct effect of contouring variation on biochemical recurrence following prostate radiotherapy. For the pixel-wise CPHM analysis, we included the Gleason score at diagnosis, patient age at the start of treatment, manual CTV volume and contour deviation. However, the presence of the counter-intuitive observation made, that larger tumour coverage may increase the risk of recurrence, potentially invalidates the more logical one. This clearly points to the need to correct for additional confounding variables. Confounding variables missing from the analysis may influence the interpretation and significance of our findings. Such additional variables include the PSA level on diagnosis, spatial variation of the planned target dose, and the patient's rectum volume upon planning, which are all known to affect the risk of biochemical recurrence[19–21]. This analysis should therefore be repeated in a cohort, preferable larger, where these covariates are available. In addition, the pixel-wise CPHM needs to be internally and externally validated to ensure the observed HRs will generalise well to new patient data, from a variety of different populations. These additional steps will help turn our observations into conclusions, and ultimately translate these results into the clinic.

HR maps of the other covariates also had significant regions (figure 5), indicating that contour deviation is not the only variable affecting survival, which is not surprising. Interpretation of the HR maps is less intuitive than the contour deviation maps. The HR maps of the confounding variables will be approximately constant when there is no interaction between contouring deviation and the confounding variable. This is clearly the case for age. However, the HR maps of prostate volume shows different HR for different regions, which is logical because delineation deviation and volume change have the same effect – a motion of the delineated contours. The most interesting interaction is for Gleason scores (7 and 8 relative to 9-10) which both suggests an increase in risk of delineation deviation for higher Gleason scores. We also checked whether there was a correlation between Gleason score and manual CTV volume, however, this was not observed (see supplementary materials, Figure S7).

Classical contouring variation studies require the contours of multiple observers, often oncologists: an expensive resource. In a recent review of IOV, the number of patients range between 1 and 26 for prostate cancer cases[7]. Such a limited number of patients makes meaningful analysis of clinical outcomes impossible. In our novel approach, we use a DL auto-contour as a reference to quantify contour deviation for more than 200 patients. Despite the clinical correctness of DL auto-contours being debatable, it provides a consistent baseline to compare the manual (clinically used) contour. As DL only uses the image dataset as input, clinical circumstances cannot influence the results of the DL model, meaning that detected deviations are not confounded by such clinical variables. However, it is important to note that the DL model will consistently reproduce any bias present in the training data, which may be the reason of the DL contours are smaller than the manual on average. Recent developments are starting to handle CTV definition automatically[13] which, if implemented clinically, may improve consistency for future patients. For our work, we visually inspected the contours and removed failed DL contours, therefore minimising the impact of image quality on DL contouring. Re-training the DL model is beyond our reach as we used commercial tools. As discussed in the results section, we observed that the manual contour is consistently larger on average than the DL contour (see supplementary materials, Figure S4), and also found that the manual contour is systematically larger in all regions of interest from Figure 3 (see supplementary materials, Figure S5). This consistent difference does not affect the estimates resulting from our analysis: the estimate for δR for each Cox model (at each pixel) is determined from the spread of δR, rather than its absolute value. Having a reference structure that is consistently smaller will only affect the intercept term at each pixel from the Cox model, which is not of interest here.

For the current analysis, contour deviation maps were extracted using a spherical coordinate system centred on the DL reference contour. This assumes that the same spherical direction (i.e., Θ and φ) will capture the same anatomical prostate region for all patients. This assumption is likely valid for convex organs which shape and orientation is similar among patients, such as the prostate. A similar assumption was made by Witte et al.[19] for image-based data mining. However, further work on improving inter-patient mapping of structures to a common frame of reference could further improve the results and allow this methodology to be extended to non-convex structures.

For our current analysis, we explored the magnitude of δR and its relation to biochemical recurrence. As the contours of the prostate shape the region of high dose, we assumed that radial variations on these contours would indirectly relate to treatment failure. Our method could be adapted if directional variability is of interest, where instead of looking at the magnitude of δR, we could look at its vector components separately (e.g., right-left, anterior-posterior or inferior-superior components).

For the studied patient cohort, our method produced both expected and unexpected results when relating contour deviations to biochemical recurrence, for example larger contours around the seminal vesicles predict better control, but larger contours around the posterior region predict worse control. As such, these results should be interpreted with caution and extra analysis on an external and larger dataset should be performed before translating them into clinical practice. These results also highlight the need to include deviation cross-correlations introduced by observer contouring 'styles' into the analysis. The effect of these cross-correlations could be minimised by including a large number of observers to reduce the bias on contouring style. Alternatively, the observer identification could be added as a confounding categorical variable in the CPHM. In our case, individual observers could not be identified from the retrospective data and therefore, this effect could not be accounted for in our analysis. Other deviation correlations may be introduced by the observer's level of expertise and the individual interpretation of the local contouring protocol factors that could be included in future studies[22–24].

From the HR map for contouring deviation, as presented here, regions are identified where clinical contours could be altered in order to limit a patient's risk of biochemical recurrence following prostate radiotherapy. Again, we highlight that it is important to notice our methodology can be used to explore relationships between contouring and any outcome. As a result, competing risk models could be built on top of our methodology and used to highlight the regions where contouring deviation should be reduced to find the optimal therapeutic ratio. Thus, our developed methodology

could be used to better define protocols for contouring, and potentially improve patient outcomes, once applied to a data set where the aforementioned additional covariates are available, and an internal and external validation has been conducted. Furthermore, the methodology proposed here could be adapted to other radiotherapy treatment sites.

The primary goal of this manuscript was to introduce a methodology to explore the correlations between contour variation and outcome. The general framework followed by our method, i.e., image-based data-mining or voxel-based analysis, been used in neuroimaging for over a decade[25,26] and it has been successfully used to explore radiotherapy dose and outcome in several sites[19,27–29]. We refer the interested reader to the recent article by Palma et al. where a 'Cookbook' dedicated to this method for use in radiation oncology is presented[30] and a critical editorial on the importance of correct statistical analysis[31].

Conclusion

We have proposed a novel method to analyse the effect of contouring variation on clinical outcome for a large cohort of patients using deviations to a consistent virtual observer. We exemplify our methodology on a cohort of prostate cancer patients, and use time to biochemical recurrence following treatment as our outcome. Regions were identified in which contour deviations of the prostate relate to biochemical recurrence, with some expected and unexpected results (produced both expected and unexpected results). After including relevant covariates and validating results with an external dataset, results from this methodology could inform contouring protocols based on actual patient outcomes.

Acknowledgements

M.V.H. was supported by NIHR Manchester Biomedical Research Centre. Prostate Cancer UK (RIA15-ST2-031) and Cancer Research UK via funding to the Cancer Research Manchester Centre (C147/A25254) supported this work. M.A is also supported by Cancer Research UK (C8225/A21133).

Conflict of Interest Statement

The author(s) declare no conflict of interests.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi:10.3322/caac.21492
2. National Cancer Registration & Analysis Service, Cancer Research UK. *National Cancer Registration and Analysis Service Short Report: Chemotherapy, Radiotherapy and Surgical Tumour Resections in England: 2013-2014 (V2)*.; 2018. http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/main_cancer_treatments
3. Pascale M, Azinwi CN, Marongiu B, Pesce G, Stoffel F, Roggero E. The outcome of prostate cancer patients treated with curative intent strongly depends on survival after metastatic progression. *BMC Cancer*. 2017;17(1):651. doi:10.1186/s12885-017-3617-6
4. Nyholm T, Jonsson J, Söderström K, et al. Variability in prostate and seminal vesicle delineations defined on magnetic resonance images, a multi-observer, -center and -sequence study. *Radiat Oncol*. 2013;8(1). doi:10.1186/1748-717X-8-126
5. Fiorino C, Reni M, Bolognesi A, Cattaneo GM, Calandrino R. Intra- and inter-observer variability in contouring prostate and seminal vesicles: Implications for conformal treatment planning. *Radiother Oncol*. 1998;47(3):285-292. doi:10.1016/S0167-8140(98)00021-8
6. Villeirs GM, Van Vaerenbergh K, Vakaet L, et al. Interobserver delineation variation using CT versus combined CT + MRI in intensity-modulated radiotherapy for prostate cancer. *Strahlentherapie und Onkol*. 2005;181(7):424-430. doi:10.1007/s00066-005-1383-x
7. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol*. 2016;121(2):169-179. doi:10.1016/j.radonc.2016.09.009
8. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol*. 2019;29(3):185-197. doi:10.1016/j.semradonc.2019.02.001
9. McCarroll RE, Beadle BM, Balter PA, et al. Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: A step toward automated radiation treatment planning for low- And middle-income countries. *J Glob Oncol*. 2018;2018(4). doi:10.1200/JGO.18.00055
10. Zabel WJ, Conway JL, Gladwish A, et al. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. *Pract Radiat Oncol*. Published online 2020. doi:10.1016/j.prro.2020.05.013
11. Gooding MJ, Smith AJ, Tariq M, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. *Med Phys*. 2018;45(11):5105-5115. doi:10.1002/mp.13200
12. Almeida G, Tavares JMRS. Deep Learning in Radiation Oncology Treatment Planning for Prostate Cancer: A Systematic Review. *J Med Syst*. 2020;44(10). doi:10.1007/s10916-020-01641-3
13. Unkelbach J, Bortfeld T, Cardenas CE, et al. The role of computational methods for automating and improving clinical target volume definition. *Radiother Oncol*. Published online 2020. doi:10.1016/j.radonc.2020.10.002
14. Green A, Vasquez Osorio E, Aznar MC, McWilliam A, van Herk M. Image Based Data Mining Using Per-voxel Cox Regression. *Front Oncol*. 2020;10:1178.

doi:10.3389/fonc.2020.01178
15. Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987*. Vol 21. Association for Computing Machinery, Inc; 1987:163-169. doi:10.1145/37401.37422
16. Remeijer P, Rasch C, Lebesque J V., Van Herk M. A general methodology for three-dimensional analysis of variation in target volume delineation. *Med Phys*. 1999;26(6):931-940. doi:10.1118/1.598485
17. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer; 2000.
18. Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Trans Med Imaging*. 1999;18(1):32-42. doi:10.1109/42.750253
19. Witte MG, Heemsbergen WD, Bohoslavsky R, et al. Relating Dose Outside the Prostate With Freedom From Failure in the Dutch Trial 68 Gy vs. 78 Gy. *Int J Radiat Oncol Biol Phys*. 2010;77(1):131-138. doi:10.1016/j.ijrobp.2009.04.040
20. Ang M, Rajcic B, Foreman D, Moretti K, O'Callaghan ME. Men presenting with prostate-specific antigen (PSA) values of over 100 ng/mL. *BJU Int*. 2016;117:68-75. doi:10.1111/bju.13411
21. Heemsbergen WD, Hoogeman MS, Witte MG, Peeters STH, Incrocci L, Lebesque J V. Increased Risk of Biochemical and Clinical Failure for Prostate Patients with a Large Rectum at Radiotherapy Planning: Results from the Dutch Trial of 68 GY Versus 78 Gy. *Int J Radiat Oncol Biol Phys*. 2007;67(5):1418-1424. doi:10.1016/j.ijrobp.2006.11.014
22. Rasch C, Barillot I, Remeijer P, Touw A, Van Herk M, Lebesque J V. Definition of the prostate in CT and MRI: A multi-observer study. *Int J Radiat Oncol Biol Phys*. 1999;43(1):57-66. doi:10.1016/S0360-3016(98)00351-4
23. Altorjai G, Fotina I, Lütgendorf-Caucig C, et al. Cone-beam CT-based delineation of stereotactic lung targets: The influence of image modality and target size on interobserver variability. *Int J Radiat Oncol Biol Phys*. 2012;82(2). doi:10.1016/j.ijrobp.2011.03.042
24. Nicholls L, Gorayski P, Poulsen M, et al. Maintaining prostate contouring consistency following an educational intervention. *J Med Radiat Sci*. 2016;63(3):155-160. doi:10.1002/jmrs.168
25. Ashburner J, Friston KJ. Voxel-based morphometry - The methods. *Neuroimage*. 2000;11(6 I):805-821. doi:10.1006/nimg.2000.0582
26. Whitwell JL. Voxel-based morphometry: An automated technique for assessing structural changes in the brain. *J Neurosci*. 2009;29(31):9661-9664. doi:10.1523/JNEUROSCI.2160-09.2009
27. Palma G, Monti S, D'Avino V, et al. A Voxel-Based Approach to Explore Local Dose Differences Associated With Radiation-Induced Lung Damage. *Int J Radiat Oncol Biol Phys*. 2016;96(1):127-133. doi:10.1016/j.ijrobp.2016.04.033
28. McWilliam A, Kennedy J, Hodgson C, Vasquez Osorio E, Faivre-Finn C, van Herk M. Radiation dose to heart base linked with poorer survival in lung cancer patients. *Eur J Cancer*. 2017;85. doi:10.1016/j.ejca.2017.07.053
29. Guo Y, Jiang W, Lakshminarayanan P, et al. Spatial Radiation Dose Influence on Xerostomia Recovery and Its Comparison to Acute Incidence in Patients With Head and Neck Cancer. *Adv Radiat Oncol*. 2020;5(2):221-230. doi:10.1016/j.adro.2019.08.009

30. Palma G, Monti S, Cella L. Voxel-based analysis in radiation oncology: A methodological cookbook. *Phys Medica*. 2020;69:192-204. doi:10.1016/j.ejmp.2019.12.013
31. Shortall J, Palma G, Mistry H, et al. Flogging a dead salmon? Reduced dose posterior to prostate correlates with increased PSA progression in voxel-based analysis of 3 randomised phase 3 trials – Marcello et el. *Int J Radiat Oncol Biol Phys*. Published online 2021:(Accepted).

Supplementary materials accompanying the publication:

Novel Methodology to assess the Effect of Contouring Variation on Treatment Outcome

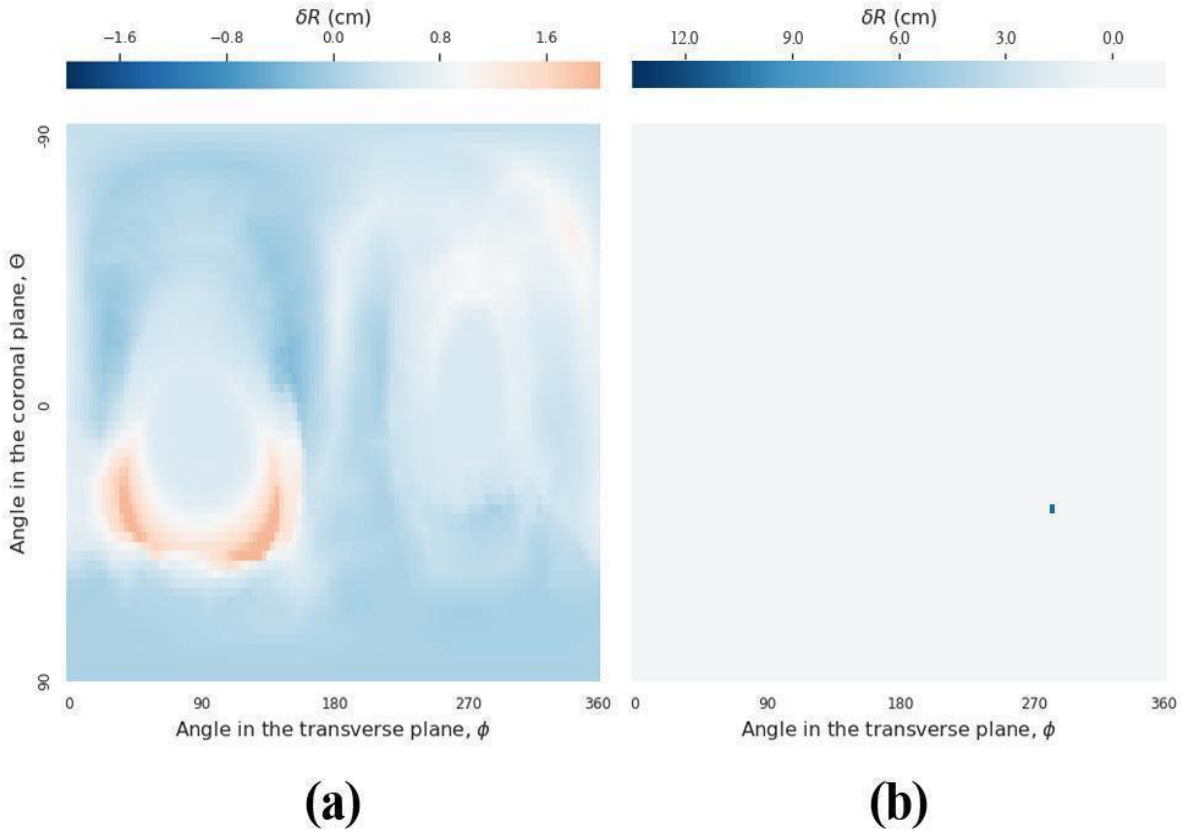


Figure S1: Each patient's deviation map was inspected to identify and remove patients with local contouring anomalies induced by triangulating each contour into a surface. (a) The deviation map of a patient, kept in the analysis. (b) The deviation map of a patient with a local anomaly of order 10 cm, removed from the analysis.

```
require("survival")
# apply cox proportional hazard model to a single ∂R pixel value
# with confounding variables: CTV volume, Age, Gleason score
# outcome metric: biochemical recurrence
model4onePixel <- coxph(relapse ~ deltaR4onePixel + age + ctvvol + gleason_score)
```

Figure S2: Example code snippet demonstrating the equivalent R code to perform pixel-wise cox proportional hazard model toolbox using the "survival" package.

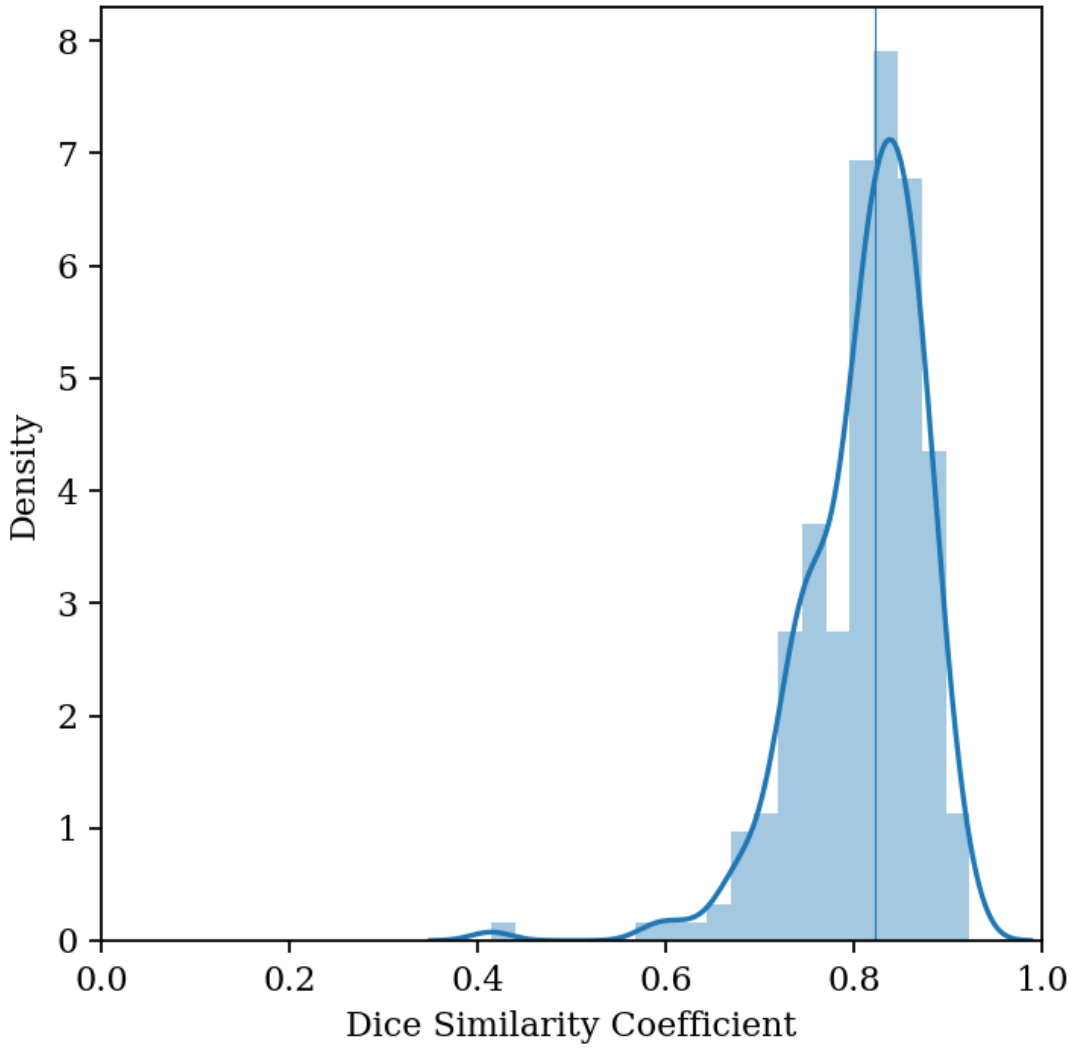


Figure S3: Distribution of Dice similarity coefficient across the entire patient cohort. Dice similarity coefficient measures the spatial overlap between the volume of the manual contours and the DL algorithm contours. The majority of Dice coefficients were greater than 0.8, suggesting high similarity between the DL and manual contours.

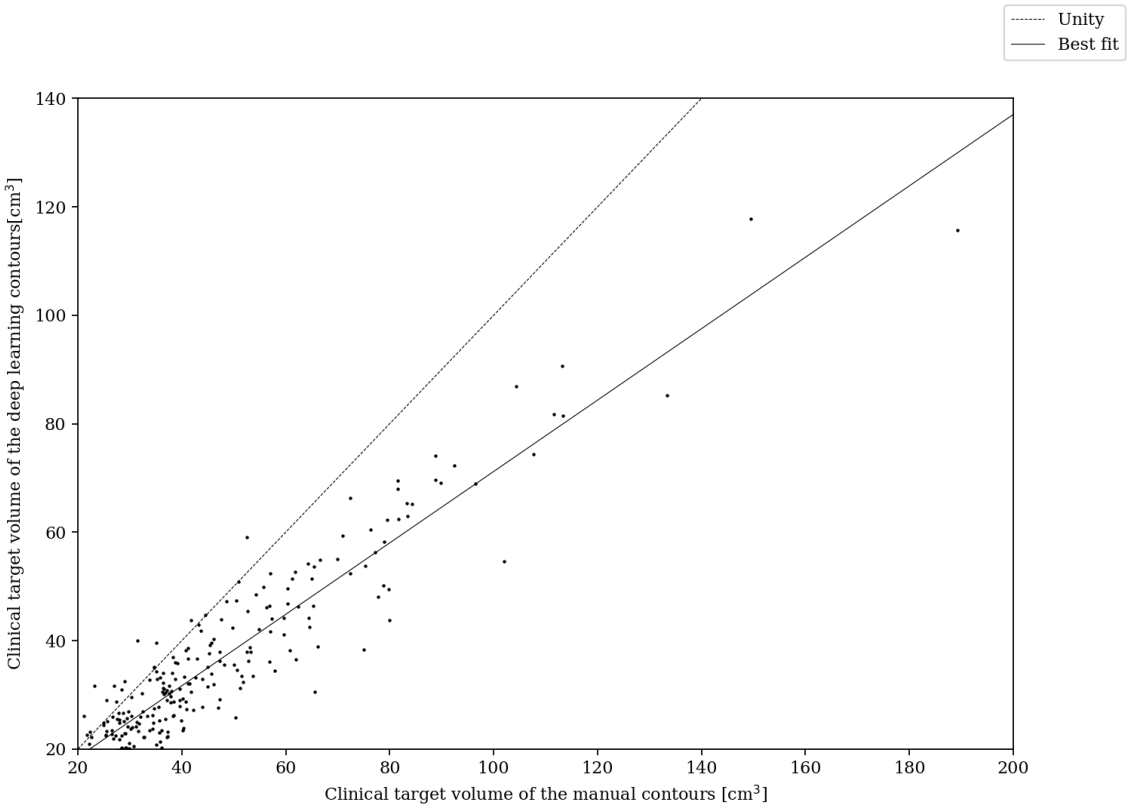


Figure S4: Scatter plot showing the volume of the manual contour against the volume of the deep learning algorithm. We observe a tendency of the datapoints to be below the line of unity (dashed line), meaning that the DL CTVs were smaller than the manual CTVs. The $y = 0.659x + 5.333$ *line is the fitted linear model (*$R^2 = 0.867$*).*

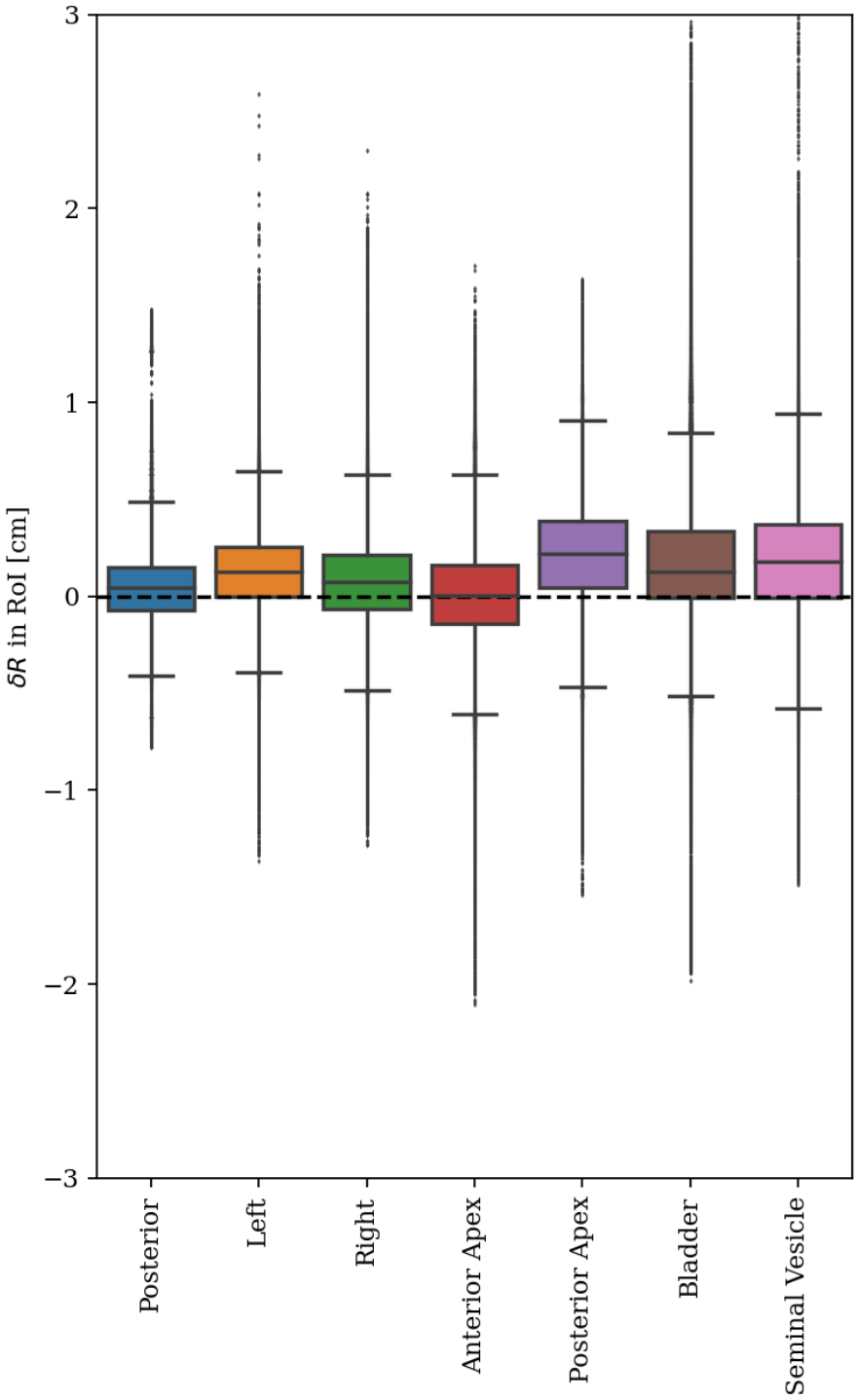
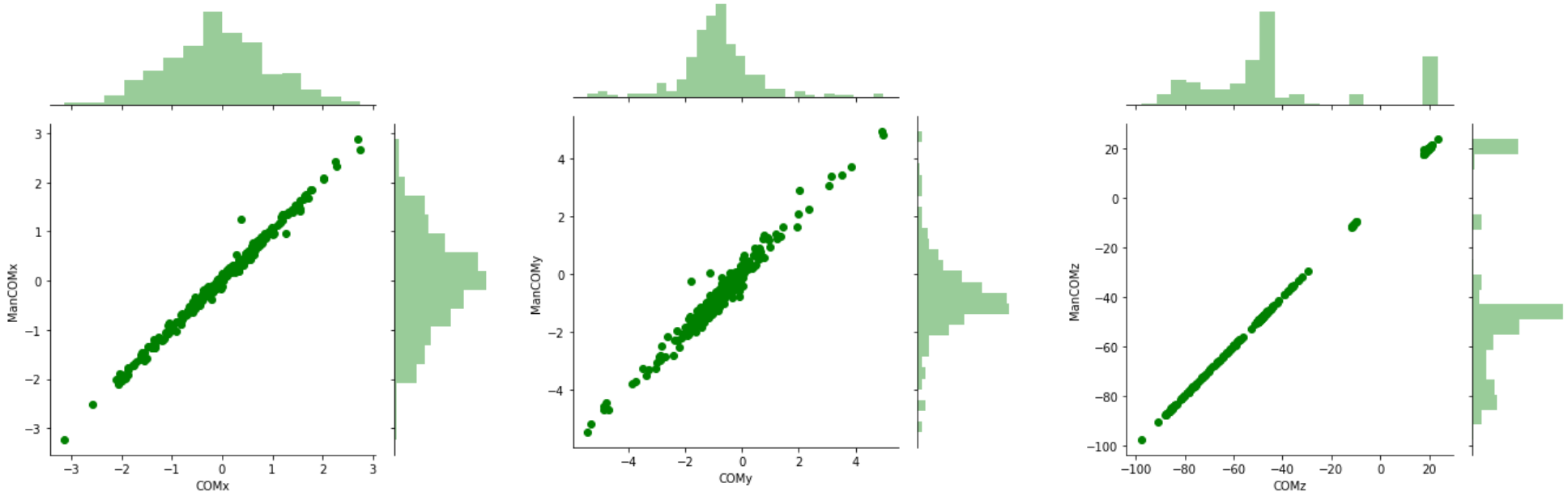


Figure S5: The distribution of ∂R pixel values for regions and interfaces of interest on the prostate, defined in Figure 3 in the main manuscript, indicating systematic differences across all regions. All median ∂R values are ≥ 0 *cm agreeing with Figure S4 showing that the DL segmentations are systematically smaller than the manual contours. One-way ANOVA test showed significant systematic differences over all regions of interest (F-score = 1.08e4, p < 10e-4). This systematic variation is expected as contour variation has been reported to be different in different regions.*



(a) (b) (c)

Figure S6: Scatter plots of the centre-of-mass of the manual contours (y-axis) against the DL contours (x-axis) in the x plane (right-left), y-plane and z-plane shown in (a), (b) and (c) respectively. The null hypothesis that distributions of x, y and z centre-of-mass coordinates between manual and DL contours have equal means, has been tested using a two-sample t-test. We found no significant evidence to suggest the means centre-of-mass coordinates are different in x (p=0.64), y (p=0.64), and z (p=0.98).

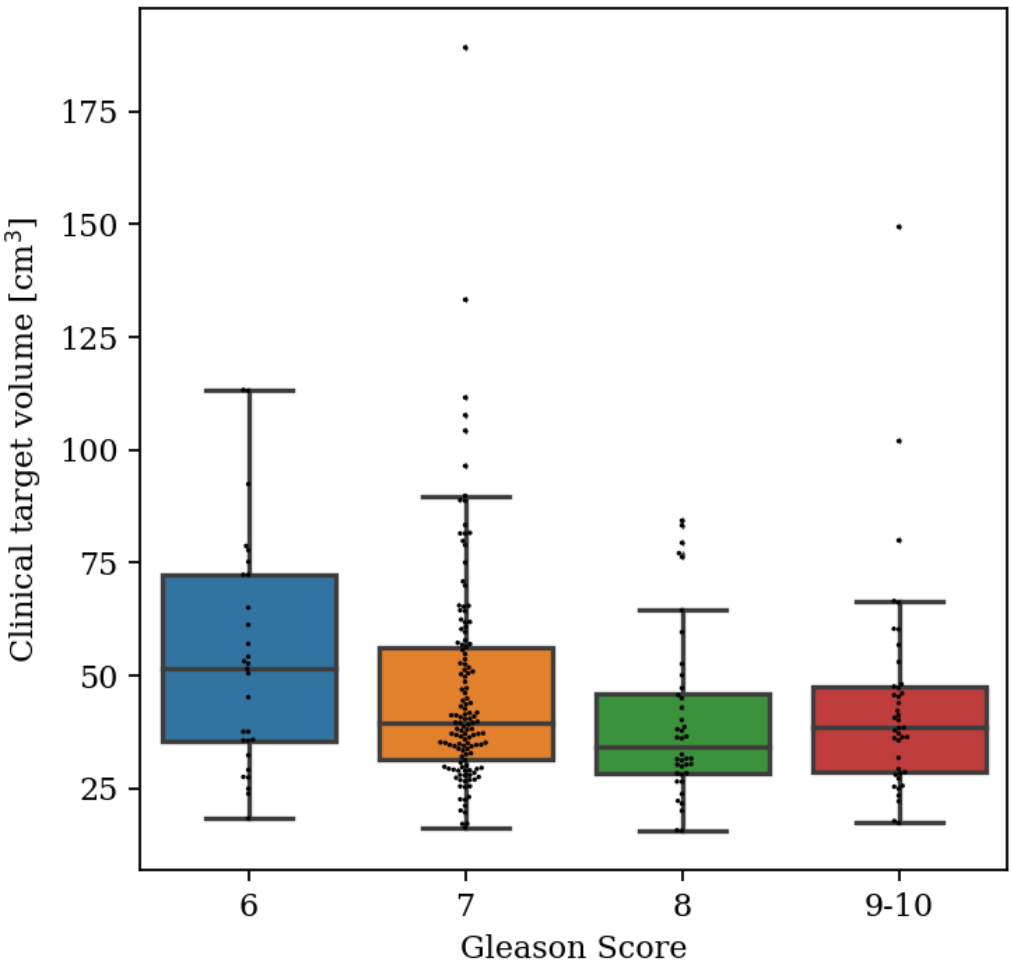


Figure S7: Distribution of the clinical target volume (CTV) for patients with different Gleason scores. We performed a one-way analysis of variation (ANOVA) to test if there were differences for the CTV patients with different Gleason score groups (F-score = 1.9196, p-value = 0.127). The insignificant score suggests that the Gleason score of the patient has an insignificant interaction with their defined CTV.