Impact of Contouring Variability on Dose-Volume Metrics used in Treatment Plan Optimization of Prostate IMRT

Arvand Barghi, Carol Johnson [2] , Andrew Warner [3] , Glenn Bauman [4] , Jerry Battista, George Rodrigues [5]

1. 2. Department of Medical Biophysics, Western University 3. Radiation Oncology, London Health Sciences Centre, London, Ontario, CA 4. Department of Radiation Oncology, London Regional Cancer Program, London Health Sciences Centre, London, ON, CANADA 5. Department of Radiation Oncology, London Regional Cancer Program, London, Ontario, CA; Schulich School of Medicine & Dentistry, Western University, London, Ontario, CA

✉ **Corresponding author:** George Rodrigues, george.rodrigues@lhsc.on.ca
Disclosures can be found in Additional Information at the end of the article

Abstract

Background and Purpose: Contouring variability remains a major source of uncertainty in radiotherapy treatment planning. The objective of this study was to identify the effect of contouring variability on dose-volume histogram (DVH) metrics used for treatment plan optimization of prostate IMRT.

Methods: A total of 25 observers were recruited to delineate the bladder, prostate, and rectum in a CT scan of low-risk prostate cancer. Dice similarity coefficients (DSC) were calculated between observer and an algorithmically-generated consensus contour. The observer contours were used to generate treatment plans and calculate DVH for each organ. The variance between DVH curves was calculated for the values $D_{95\%}$ for prostate, and $V_{65, 70, 75 \text{ Gy}}$ for bladder and rectum.

Results: DSC for the bladder, prostate, and rectum were 0.971 ± 0.007, 0.838 ± 0.067, 0.771 ± 0.124, respectively. DVH variance for all three structures was primarily driven by differences in prostate contouring. Variations in rectal contouring had important additional impacts only on rectal DVH. Bladder contouring variation had little impact on DVH metrics.

Conclusions: Although the rectum was the most inconsistently contoured structure, its variability did not impact DVH as much as prostate variability. It has been demonstrated that the dosimetric impact of contouring variability cannot be predicted solely with DSC.

Introduction

Modern radiation therapy delivers highly conformal dose distributions. With improving technological capability, an increased need for accurate delineation of the planning target volumes (PTV) and organs at risk (OAR) is required. Inter-clinician variability during contouring remains a substantial source of uncertainty. Contouring variability has therefore

become a focus of research, with emphasis on both PTV and OARs for many anatomical sites [1-4]. The impact of contour variability on dose-volume histogram (DVH) parameters used in treatment plan optimization has been less well-characterized.

DVH summarizes dose distribution across an entire organ or target volume of interest and is sensitive to variations in contouring. Previous work has identified DVH differences due to contouring variation in breast, oropharyngeal and prostate cancer [5-7], but these studies did not examine the independent contribution of contouring variability of individual structures to DVH variations. For example, it remains unclear how contouring variability, specifically of the bladder, affects the DVH of the prostate. The dose optimization algorithm seeks to deposit as much dose in the PTV, while minimizing dose deposition in OARs delineated. The interplay of PTV and OAR contours therefore requires consideration because treatment plan optimization is driven collectively by the competing parameters of dose to target and normal tissue toxicity.

We examined the effect of individual and combined contouring variability on OAR and PTV DVH for a low-risk prostate cancer case contoured by multiple observers. In doing so, we sought to isolate the effects of variability in contouring individual and grouped structures on the DVH for all the structures of interest and potentially identify sources where efforts to reduce variability in contouring may have the largest impact on plan optimization results as characterized by deviations in key DVH parameters.

Materials And Methods

Contouring collection and comparison of similarity

An online contouring challenge of the prostate, bladder, and rectum was completed on an anonymized low-risk prostate cancer abdominal-pelvic CT data set (120 kV$_P$, 512x512 pixels/slice, 3-mm slice thickness) DICOM-RT structure files containing 25 unique contours of prostate, bladder, and rectum were obtained using a multi-institutional online program (www.contouringchallenge.com). In agreeing to participate in the challenge, observers provided the requested contours and consented to the use of the contours for research analysis.

To compare the contours for individual observers, a "gold standard" contour is required (Figure 1). For this study, the gold standards used were consensus contours created for each of the three structures using the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [8] that estimates the true volume of a structure from a collection of observer contours.
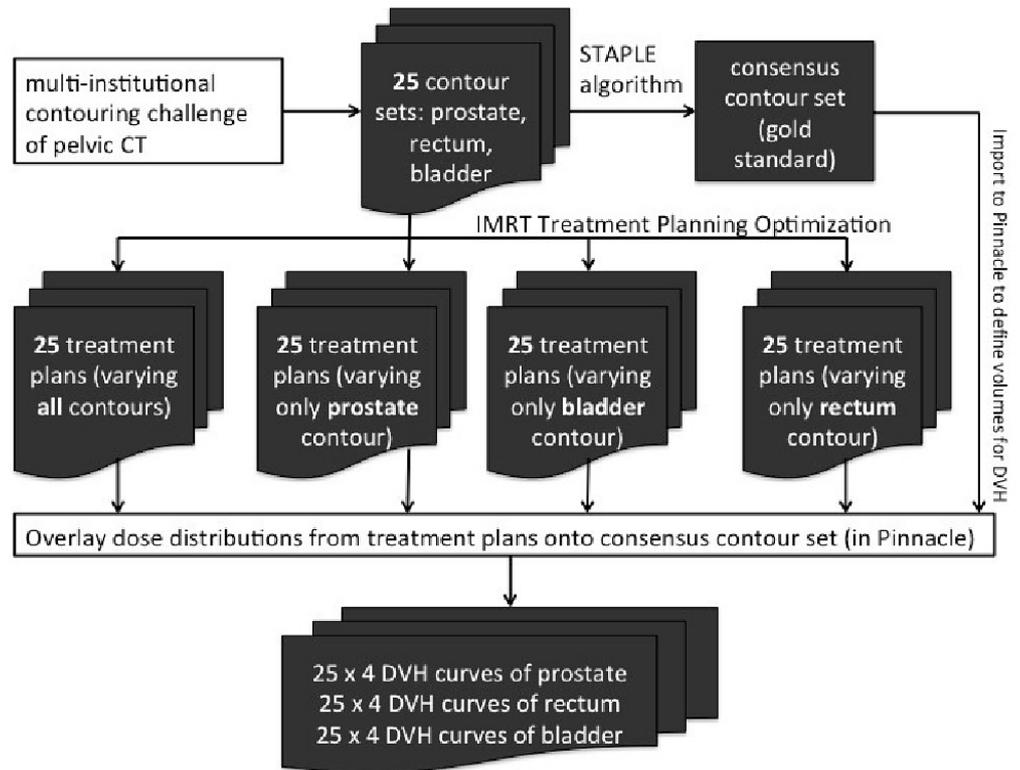
FIGURE 1: "Gold standard" contour

The 25 observer contours (prostate, rectum, bladder) were each compared against the corresponding consensus contour with StructSure software (StructSure TM, Standard Imaging Inc., Middletown, WI, USA). StructSure imports DICOM-RT structure files and performs measurements of similarity between a structure file identified as the "gold standard" and a comparison "test" structure file. For this study, we used the Dice similarity coefficient (DSC) [9] to compare observer contours against the consensus contour obtained by the STAPLE algorithm.

Creation of contour series isolating variability in a single structure

Differences in the DSC do not necessarily reflect potential differences in treatment planning dose effects.   An objective of this study was to determine the individual contribution of prostate, bladder, and rectum contouring variability to the DVH deviations of these three contours superimposed on the consensus dose distribution. For this purpose, we have designed four *contour series* for investigation, corresponding the variation of a different structure's contour. The first contour series includes the 25 contour sets of all observers (Vary All). The second series was created by importing only the 25 observer contours for the prostate combined with the STAPLE contours of the bladder and rectum (Vary Prostate Only). The third and fourth series were created by importing the observer contours for bladder (Vary Bladder Only) and for rectum (Vary Rectum Only) combined with STAPLE contours of the remaining two organs. These contour series were investigated with the treatment plan optimization protocol as outlined below.

Treatment plan optimization and DVH curves

The contour series were used in a treatment plan optimization protocol. To avoid bias, an automated class solution was applied to all contour sets in the four contour series described above using the scripting and plug-in utilities in Pinnacle software (version 8.1y). The IMRT plan was designed to use five fields of 18 MV x-rays targeting a PTV (GTV, plus a uniform margin of 10 mm except 7 mm posteriorly) with 76 Gy to the normalization point (isocentre), and a consistent set of objectives and weights. Using this solution, an optimized reference plan containing the STAPLE gold standard contours was generated and approved by two radiation oncologists, guided by RTOG-0415 standards (Radiation Therapy Oncology Group - www.rtog.com).

Using the same class solution developed for the reference plan based on STAPLE consensus contours, another set of treatment plans was generated for each test contour series. The resulting dose distributions were superimposed onto the STAPLE consensus contours to determine the dose that would have resulted in the "true" structures. To be clear, the dose distributions were optimized using test contours, but the dose-volume analysis was based on the "gold standard" consensus contour set. Potential geographic misses of the "true" target or dose spillage into normal structures were thus assayed. The DVH data for these standard structures subjected to test dose exposure were exported for analysis of variance.

Statistical analysis

For prostate, the variance (σ^2) of DVH curves between the 25 observers was characterized at the 95% dose level ($D_{95\%}$). Thus, the % volume was kept fixed, and the variance in Gy of the 25 observer DVH curves was measured (a horizontal line sampling the DVH curves). The curves were *a priori* sampled from $D_{92.5\%}$ to $D_{97.5\%}$ at increments of 0.2 %, resulting in 25 measurements of σ^2 from which the mean and standard deviation of σ^2 was calculated for the region $D_{95 + 2.5\%}$. For rectum and bladder, the dose is kept fixed, and σ^2 of the % volume between the 25 observer DVH curves was calculated (vertical line sampling). For rectum and bladder, σ^2 was *a priori* characterized at three different regions of the curve: $V_{65 + 2.5\,Gy}$, $V_{70 + 2.5\,Gy}$, $V_{75 + 2.5\,Gy}$ sampling at increments of 0.2 Gy.

To test for differences in variance (in volume or in dose depending on which structures were varied), the Paired T-test was used, comparing the following DVH pairings: All vs. Prostate Only, All vs. Bladder Only, All vs. Rectum Only. All statistical analysis was performed using SAS software (version 9.2), using two-sided statistical testing at the 5% significance level.

Results

Figure *2* shows a representative central slice of the CT study with all observer contours superimposed. The mean and standard deviation of DSC for observer contours compared to the respective STAPLE gold standard contours were 0.971 ± 0.007 (bladder), 0.838 ± 0.067 (prostate) and 0.771 ± 0.124 (rectum). Figures *3A*-*3D* illustrates the superimposed DVHs of treatment plans when varying all contours (A), as well as when varying the contours of prostate, bladder, and rectum individually (B-D). From this figure, qualitatively there is relatively high dispersion of DVH curves for all three structures when all contours are varied or the prostate contour only is varied. Dispersion in the DVH curves of the rectum also occurs when the rectum contours are varied. Little to no dispersion occurs when bladder contours are varied. Tables *1-3* quantitatively summarize the variance statistics of the relevant bladder (Table *1*), rectum (Table *2*), and prostate (Table *3*) DVH ranges.
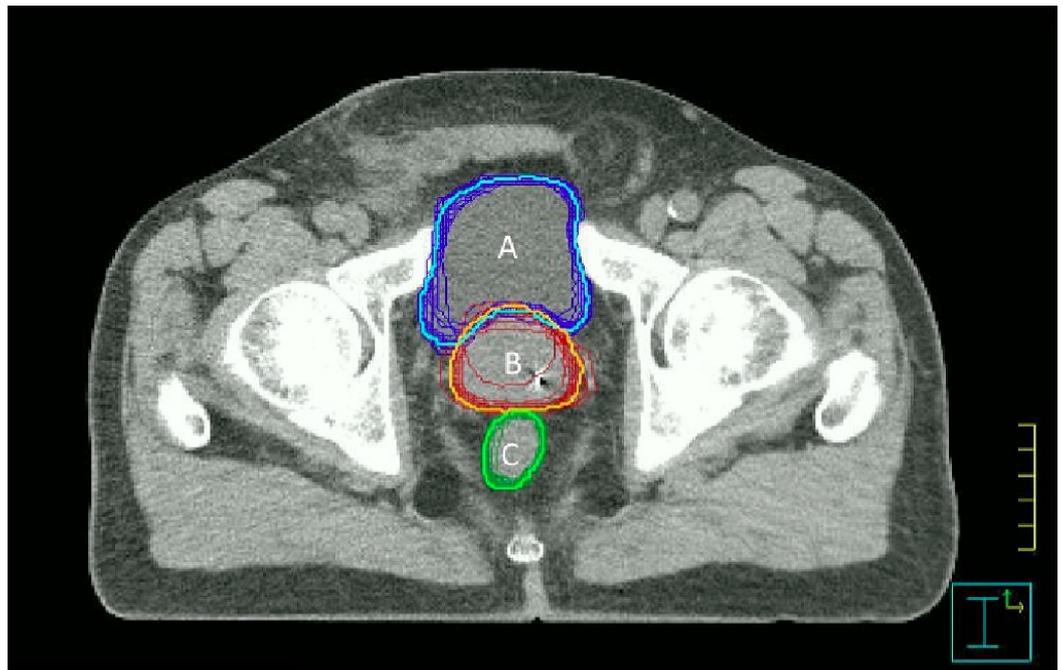
FIGURE 2: A representative central slice of the CT study with all observer contours superimposed
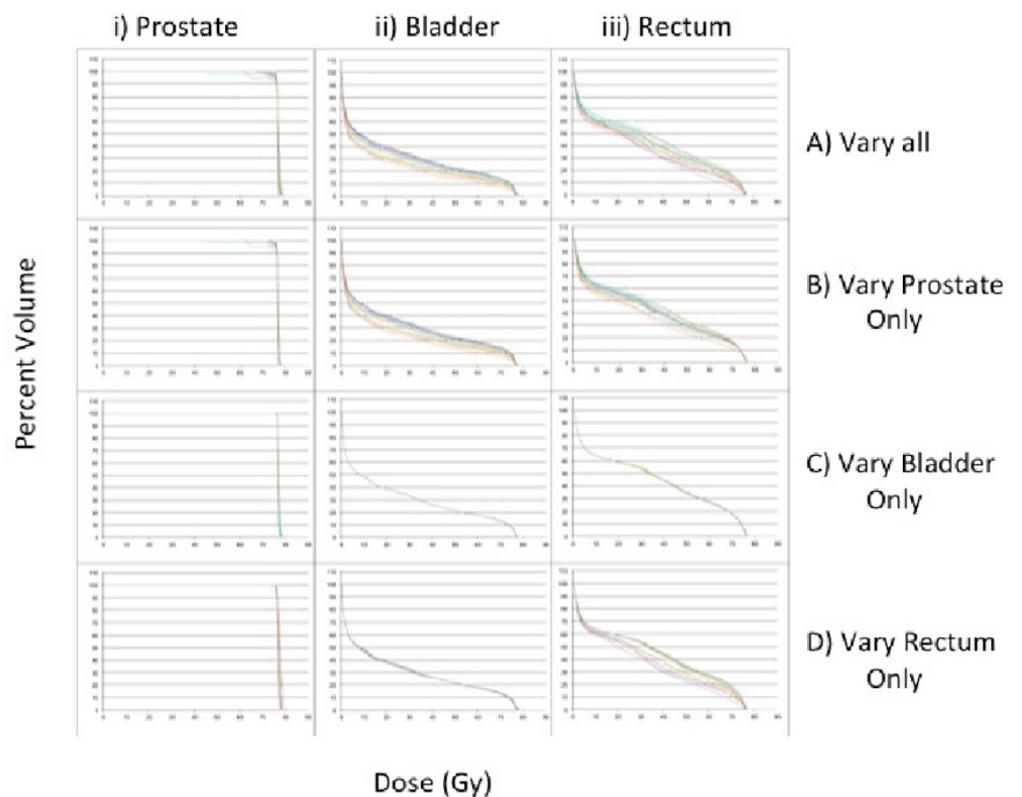
FIGURE 3: An illustration of the superimposed DVHs of treatment plans when varying all contours (A), as well as when

varying the contours of prostate, bladder, and rectum individually (B-D).

Structures Varied in Contour Series:	% Volume Variance at Selected Dose Ranges of Interest (mean ± SD):		
	65 ± 2.5 Gy	70 ± 2.5 Gy	75 ± 2.5 Gy
Vary all	5.00 ± 0.31	3.82 ± 0.29	2.13 ± 0.98
Vary prostate	4.92 ± 0.26	3.96 ± 0.30	2.17 ± 0.95
Vary bladder	0.01 ± 0.01	0.02 ± 0.01	0.07 ± 0.06
Vary rectum	0.03 ± 0.01	0.07 ± 0.01	0.23 ± 0.14
Contour Series Compared Using Paired T-Test:	Difference in Mean Dose (Gy) Mean (95% CI) p-Value		
	65 ± 2.5 Gy	70 ± 2.5 Gy	75 ± 2.5 Gy
All vs prostate	0.075 (0.052, 0.098) < 0.001	-0.145 (-0.184, -0.105) < 0.001	-0.037 (-0.083, 0.009) 0.114
All vs bladder	4.989 (4.858, 5.119) < 0.001	3.801 (3.683, 3.920) < 0.001	2.057 (1.636, 2.477) < 0.001
All vs rectum	4.966 (4.833, 5.098) < 0.001	3.752 (3.628, 3.877) < 0.001	1.899 (1.442, 2.357) < 0.001

TABLE 1: Summary of variance

Summary of variance in BLADDER volume for contour series with all contours varied (Vary All), Prostate varied only (Vary Prostate), Bladder varied only (Vary Bladder), and rectum varied only (Vary Rectum) at dose ranges of interest: V65 ± 2.5 Gy. V70 ± 2.5 Gy. V75 ± 2.5 Gy. Paired t-tests between "Vary All" series vs "Vary Prostate", "Vary Bladder", and "Vary Rectum".

Structures Varied in Contour Series:	% Volume Variance at Selected Dose Ranges of Interest (mean ± SD):		
	V65 ± 2.5 Gy	V70 ± 2.5 Gy	V75 ± 2.5 Gy
Vary all	11.25 ± 0.55	8.94 ± 0.94	3.17 ± 2.46
Vary prostate	5.17 ± 0.63	2.57 ± 0.87	0.47 ± 0.30
Vary bladder	0.05 ± 0.01	0.04 ± 0.00	0.16 ± 0.17
Vary rectum	11.09 ± 0.34	9.22 ± 1.05	2.86 ± 2.34
Contour Series Compared Using Paired T-Test:	Difference in Mean Dose (Gy) Mean (95% CI) p-Value		
	V65 ± 2.5 Gy	V70 ± 2.5 Gy	V75 ± 2.5 Gy
All vs prostate	6.077 (6.026, 6.128) < 0.001	6.370 (6.326, 6.414) < 0.001	2.704 (1.801, 3.607) < 0.001
All vs bladder	11.21 (10.98, 11.43) < 0.001	8.897 (8.510, 9.285) < 0.001	3.014 (1.967, 4.060) < 0.001
All vs rectum	0.158 (0.051, 0.266) 0.006	-0.277 (-0.339, -0.216) < 0.001	0.311 (0.198, 0.424) < 0.001

TABLE 2: Summary of variance

Summary of variance in RECTUM volume for contour series with all contours varied (Vary All), Prostate varied only (Vary Prostate), Bladder varied only (Vary Bladder), and rectum varied only (Vary Rectum) at dose ranges of interest: V65 ± 2.5 Gy. V70 ± 2.5 Gy. V75 ± 2.5 Gy. Paired t-tests between "Vary All" series vs "Vary Prostate", "Vary Bladder", and "Vary Rectum".

Structures Varied in Contour Series:	Dose Variance (Gy) at D95 ± 2.5 % Mean ± SD
Vary all	9.47 ± 12.43
Vary prostate	11.89 ± 17.59
Vary bladder	0.00 ± 0.00
Vary rectum	0.04 ± 0.10
Contour Series Compared Using Paired T-Test:	Difference in mean dose (Gy) Mean (95% CI) p-Value
All vs prostate	-2.421 (-9.111, 4.270) 0.462
All vs bladder	9.474 (4.227, 14.721) 0.001
All vs rectum	9.437 (4.195, 14.679) 0.001

TABLE 3: Summary of variance

Summary of variance in PROSTATE dose of contour series with all contours varied (Vary All), Prostate varied only (Vary Prostate), Bladder varied only (Vary Bladder), and rectum varied only (Vary Rectum) at volume range of interest: D95 ± 2.5 %. Paired t-tests between "Vary All" series vs "Vary Prostate", "Vary Bladder", and "Vary Rectum".

Collectively, both the qualitative and quantitative DVH analyses demonstrate that the variance in prostate DVH is primarily driven from differences in prostate contouring and that differences in rectum and bladder contouring have less impact on prostate metrics variations. Observed rectal DVH variation is primarily driven by differences in rectal contouring as well as prostate contouring, whereas bladder contouring variation did not have a similar impact. In terms of the bladder DVH variation, the differences in the contouring of the prostate (and not the bladder itself) were primarily responsible. The 95% confidence intervals mostly show statistically significant differences in DVH variance between the DVHs of the 'vary all' series compared to the 'vary bladder only', 'vary rectum only', and 'vary prostate only' series. Two comparisons did not show statistical significance: the bladder DVH variance and the prostate DVH variance comparisons of the "vary all" to the "vary prostate only" DVH series. This demonstrates that varying all contours together as a group did not significantly change the DVH variance of prostate or bladder DVH when compared to varying only the prostate contour.

Discussion

The DSC metric is commonly used in studies of contouring variability [10]. In this study, DSC demonstrated the highest compliance value for bladder (0.971), followed by prostate (0.838), and then rectum (0.771) There are several possible reasons for the high similarity among bladder contours compared to other contours. Firstly, the bladder is the largest structure of the three, meaning that variability itself must be larger to influence DSC significantly. Furthermore, the bladder has a relatively well-defined border observed in CT imaging, facilitating delineation. In comparison, the rectum is generally smaller with less well-defined borders, particularly at the prostate boundary and inferiorly towards the pelvic diaphragm and anal canal.

Remarkably, the dosimetric hierarchy observed from DVH results is *not* in agreement with the DSC results. It may have been suspected that the structures that produce the lowest DSC values

would cause the greatest variability in the dosimetric parameters. Although bladder had the highest DSC score as well as the least variable DVH results, the remaining two structures do not show correlation between DSC score and DVH impact. The rectum had the lowest DSC score, and yet the prostate contributed the most to DVH variability, despite having a higher DSC score than the rectum. Furthermore, varying only rectum contours caused notable "isolated" variability in DVH of only rectum, whereas varying only prostate contours caused notable generalized variability in DVH for prostate, rectum, and bladder. The true dosimetric impact of contouring variability in this case therefore could not have been predicted by assaying simple similarity measurements alone. This decoupling of dosimetric effect and contouring variability metrics suggests that future strategies of reducing contouring variability should seek to reduce dose variation impact, not just contour compliance. There is no one-to-one correspondence between contour and dose volumes. There is a non-linear interplay between dose optimization, dose deposition, and contour topology.

There are a number of possible reasons for the different contributions of each structure to DVH variability. Firstly, the optimization of treatment plans ranked the prostate voxels as dominant over voxels of either OAR through a weighting assignment. Thus, the prostate volume contributes more heavily to driving the dose optimization algorithm towards a large uniform dose to the prostate. Secondly, the prostate is geographically the central structure, since it is the target. It is in close proximity with two OARs, which may contribute to its high level of contribution to the treatment plan optimization.

Most previous studies of contouring variability have not "followed through" on dosimetric impact and have mostly only provided measurements of similarity, such as DSC. Studies that have demonstrated DVH results [5-7] have shown that contouring variability can lead to dosimetrically relevant treatment planning variability, although a previous prostate study [7] suggests that contouring variability in prostate cancer may not have large dosimetric impacts.

Some notable limitations exist in this study. Firstly, as with all studies involving a contouring challenge, the definition of a true gold standard is difficult. In this study, the gold standard was generated using the STAPLE algorithm to achieve a consensus representative contour. This approach reduces the subjectivity of a gold standard contour definition by a panel of experts. Another limitation was the five-field IMRT treatment planning process that was used. Although we used a consistent set of plan optimization parameters and a class solution that generally produces clinically acceptable plans, no attempt was made to re-optimize individual plans based on the test contours beyond the class solution. Other centres, or other operators, may use different selections and approaches. Thus, the results of this study must be interpreted within the context of a fixed treatment plan protocol.

Future work could include repeating similar experiments for other tumour sites, deriving hierarchies for organ sets in regions other than pelvis where only similarity measurements have been performed to date. Furthermore, the analytic approach used in this study could also be used in the validation of automated contouring techniques.

There is known inter- and intra-fraction variability in organ volume and position during radiotherapy. This contributes to dosimetric variability during actual treatment and is influenced by other observations, such as interpretation of in-room image guidance information used in patient setups or beam gating. Placing the contouring effects noted here into the context of overall uncertainties in prostate cancer planning and delivery will be important in order to identify weakest links in the planning-delivery chain where future effort should be concentrated. For example, would perfectly congruent OAR and PTV contouring make a significant dosimetric difference, given the other downstream sources of uncertainty that would still persist?

Conclusions

We examined the effect of combined contouring variability on DVH curves of OARs and PTV for a low risk prostate cancer case contoured by multiple observers. Variability in prostate contouring followed by rectal contouring caused the largest sources of dosimetric variation with less impact of variations due to bladder contouring. The dosimetric impact of contouring variations did not track well with DSC scores, suggesting that future studies of contouring quality must incorporate dosimetric endpoints as well as measures of contour congruence.

Additional Information

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** Glenn Bauman, Jerry Battista, and George Rodrigues have a non-financial software access agreement with Standard Imaging Inc for access to StructSure Software and with Philips Medical Systems for developments of Pinnacle software.

References

1. Vorweck H, Beckmann G, Bremer M, et al.: The delineation of target volumes for radiotherapy of lung cancer patients. Radiother Oncol. 2009, 91:455-460.
2. Petersen RP, Truong PT, Kader HA, et al.: Target volume delineation for partial breast radiotherapy planning: Clinical characteristics associated with low interobserver concordance. Int J Radiat Oncol Biol Phys. 2007, 69:41-48.
3. Yamamoto M, Nagata Y, Okajime K, et al.: Differences in target outline delineation from CT scans of brain tumours using different methods and different observers. Radiother Oncol. 1999, 50:151-160.
4. Rasch C, Barillot I, Remeijer P, Touw A, van Herk M, Lebesque JV: Definition of the prostate in CT and MRI: A multi-observer study. Int J Radiat Oncol Biol Phys. 1999, 43:57-66.
5. Li XA, Tai A, Arthur DW, et al.: Variability of target and normal structure delineation for breast-cancer radiotherapy: A RTOG multi-institutional and multi-observer study. Int J Radiat Oncol Biol Phys. 2009, 73:944-951.
6. Nelms BE, Tome WA, Robinson G, et al.: Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys. 2012, 82:368-378.
7. Livsey JE, Wylie JP, Swindell R, et al.: Do differences in target volume definition in prostate cancer lead to clinically relevant differences in normal tissue toxicity?. Int J Radiat Oncol Biol Phys. 2004, 60:1076-1081.
8. Warfield SK, Zou KH, Wells WM: Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging. 2004, 23:903-921.
9. Dice LR: Measures of the amount of ecologic association between species . Ecology. 1945, 26:297-302.
10. Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalf PE: A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol. 2010, 54:401-410.