Technical Innovations & Patient Support in Radiation Oncology 3–4 (2017) 41–47

Contents lists available at ScienceDirect

Technical Innovations & Patient Support in Radiation Oncology

journal homepage: www.elsevier.com/locate/tipsro

Research article

A dose based approach for evaluation of inter-observer variations in target delineation

Ingrid Kristensen [a,b,*], Kristina Nilsson [c], Måns Agrup [d], Karin Belfrage [b], Anna Embring [e], Hedda Haugen [f], Anna-Maja Svärd [g], Tommy Knöös [b,h], Per Nilsson [b,h]

[a] Department of Oncology, Clinical Sciences, Lund University, Lund, Sweden
[b] Department of Haematology, Oncology and Radiation Physics, Skåne University Hospital, Lund, Sweden
[c] Department of Immunology, Genetics and Pathology, Experimental and Clinical Oncology, Clinical Oncology, Uppsala University Hospital, Uppsala, Sweden
[d] Department of Oncology, Linköping University Hospital, Linköping, Sweden
[e] Department of Oncology, Karolinska University Hospital, Stockholm, Sweden
[f] Department of Oncology, Sahlgrenska University Hospital, Gothenburg, Sweden
[g] Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden
[h] Department of Medical Radiation Physics, Clinical Sciences, Lund University, Lund, Sweden

ARTICLE INFO

Article history:
Received 24 August 2017
Received in revised form 6 October 2017
Accepted 9 October 2017
Available online 4 November 2017

Keywords:
Inter-observer variation
Pediatric radiotherapy
Target delineation
Treatment planning

ABSTRACT

Background and purpose: Substantial inter-observer variations in target delineation have been presented previously. Target delineation for paediatric cases is difficult due to the small number of children, the variation in paediatric targets, the number of study protocols, and the individual patient's specific needs and demands. Uncertainties in target delineation might lead to under-dosage or over-dosage. The aim of this work is to apply the concept of a consensus volume and good quality treatment plans to visualise and quantify inter-observer target delineation variations in dosimetric terms in addition to conventional geometrically based volume concordance indices.

Material and methods: Two paediatric cases were used to demonstrate the potential of adding dose metrics when evaluating target delineation diversity; Hodgkin's disease (case 1) and rhabdomyosarcoma of the parotid gland (case 2). The variability in target delineation (PTV delineations) between six centres was quantified using the generalised conformity index, *CIgen*, generated for volume overlap. The STAPLE algorithm, as implemented in CERR, was used for both cases to derive a consensus volumes. STAPLE is a probabilistic estimate of the true volume generated from all observers. Dose distributions created by each centre for the original target volumes were then applied to this consensus volume.

Results: A considerable variation in target segmentation was seen in both cases. For case 1 the variation was 374–960 cm^3 (average 669 cm^3) and for case 2; 65–126 cm^3 (average 109 cm^3). *CIgen* were 0.53 and 0.70, respectively. The DVHs in absolute volume displayed for the delineated target volume as well as for the consensus volume adds information on both "compliant" target volumes as well as outliers which are hidden with just the use of concordance indices.

Conclusions: The DVHs in absolute volume add valuable and easily understood information to various indices for evaluating uniformity in target delineation.

Introduction

Substantial inter-observer variations in target delineation have been presented in a number of previous studies [1–9]. The variations can be due to differences in interpretation of the diagnostic material, ambiguities in treatment protocols, lack of guidelines and/or inadequate, differences in local policies, the availability and use of multi-modality imaging, the subjective assessment of disease dissemination and/or the individual training and experience of the radiation oncologists. In a recent review, Vinod et al. [10] concluded that guidelines and atlases or atlas-based delineation tools would improve delineation [11,12], as well as training and the use of multi-modality imaging. Studies have also shown that delineation workshops [13] and peer reviews [14] can

* Corresponding author at: Department of Haematology, Oncology and Radiation Physics, Skåne University Hospital, Lund, Sweden.
E-mail address: ingrid.kristensen@skane.se (I. Kristensen).

https://doi.org/10.1016/j.tipsro.2017.10.002
2405-6324/

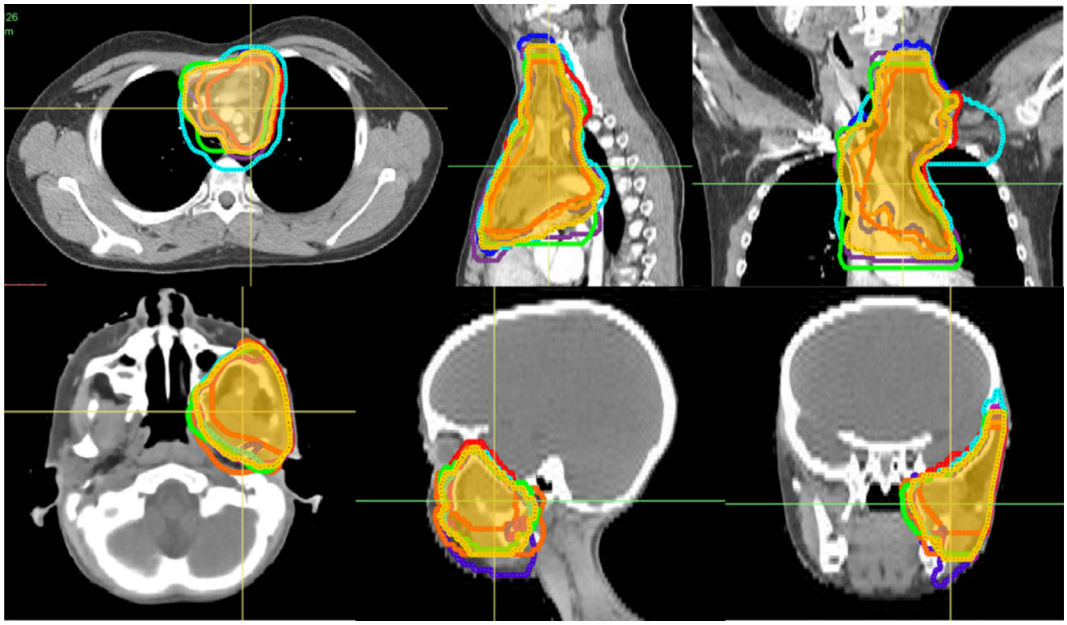

Fig. 1. Volume delineations fromall six centres for case 1 (top panel) and case 2 (bottom panel) as well as the consensus volume in transparent yellow.

Table 1
Volume related metrics for delineated target volumes.

	Case 1	Case 2
Volume (cm^3) average (range)	669 (374–960)	109 (65–126)
Intersection volume (cm^3)	293	53
Union volume (cm^3)	1189	131
CI_{gen}	0.53	0.70

Table 2
Volume related metrics for the STAPLE derived consensus volume.

	Case 1	Case 2
Volume (cm^3)	706	92
Agreement sensitivity (mean ± SD)	0.78 ± 0.20	0.88 ± 0.13
Agreement specificity (mean ± SD)	0.96 ± 0.03	0.98 ± 0.01
K	0.63	0.78

Table 3
Dose-volume metrics for each centre's target volume.

	Case 1 Average dose, (range)	Case 2 Average dose, (range)
$V_{95\%}$ (%)	91% (76–98)	95% (87–99)
$D_{98\%}$ (Gy)	18.3 (16.7–19.1)	38.9 (37.5–39.9)
$D_{50\%}$ (Gy)	19.9 (19.7–20.3)	41.5 (41.5–41.7)
$D_{2\%}$ (Gy)	20.6 (20.3–21.1)	42.8 (42.7–43.1)
HI	0.12 (0.09–0.19)	0.09 (0.07–0.13)
RCI	1.00 (0.90–1.16)	0.90 (0.68–1.09)

improve target delineation concordance and reduce inter-observer variability. Target delineation for paediatric cases is even more difficult due to the small number of children at most centres, the large variation in paediatric targets, the large number of study protocols, and the individual patient's specific needs and demands [15–17]. Uncertainties in target delineation might lead to under-dosage or over-dosage, causing a decrease in tumour control probability (TCP) or an increase in normal tissue complications (NTCP).

The evaluation of differences in segmented volumes in inter-observer studies can be done in several ways [18]. There is, however, no consensus among researchers on the methodologies to be applied and which metrics to report; e.g. differences in volume sizes, centre of mass variations, concordance indices, etc., making comparison between studies difficult to interpret. Valentini et al. describes a methodology for auto-segmentation which also could be used for studies on inter-observer variations [19]. Applying concordance indices is the most common method. It converts the variation in positions and sizes of delineated structures in relation to each other into a numerical value. The numerical value of different concordance indices are, however, dependent on the size of the structure studied and it is hence difficult to judge the resulting index value or when e.g. an improvement has occurred and to which degree. There is also an uncertainty in target delineation studies regarding which volume should be considered the "golden standard" or reference volume [20]. This volume is chosen in dissimilar ways in different studies. It could be segmented by an "expert" or a group of "experts". Another more objective method is to derive a "consensus volume" by applying an algorithm that computes a probabilistic estimate of the "true" segmentation based on the delineated volumes, e.g. STAPLE (Simultaneous Truth And Performance Level Estimation) [21]. This method has previously been introduced for radiotherapy [22] and used in target delineations studies [23–34].

Dose metrics are, however, not routinely reported in delineation studies, even though it might be helpful making the consequences of target delineation variations easier to interpret [20]. If treatment plans are created as a part of the target delineation process and these plans are clinically acceptable it would be an attractive complement to evaluate the quality of the resulting dose distribution on a consensus volume rather than only the volume metrics *per se*.

At an internal target delineation workshop, performed by The Swedish Workgroup for Paediatric Radiotherapy, these concepts were discussed. The group has previously performed and reported on an inter-observer study evaluated with conventional volume metrics [35].

The aim of this paper is to apply the concept of a consensus volume and good quality treatment plans for two paediatric cases to visualise target delineation variation in dosimetric terms in addition to conventional geometrically based volume concordance indices.

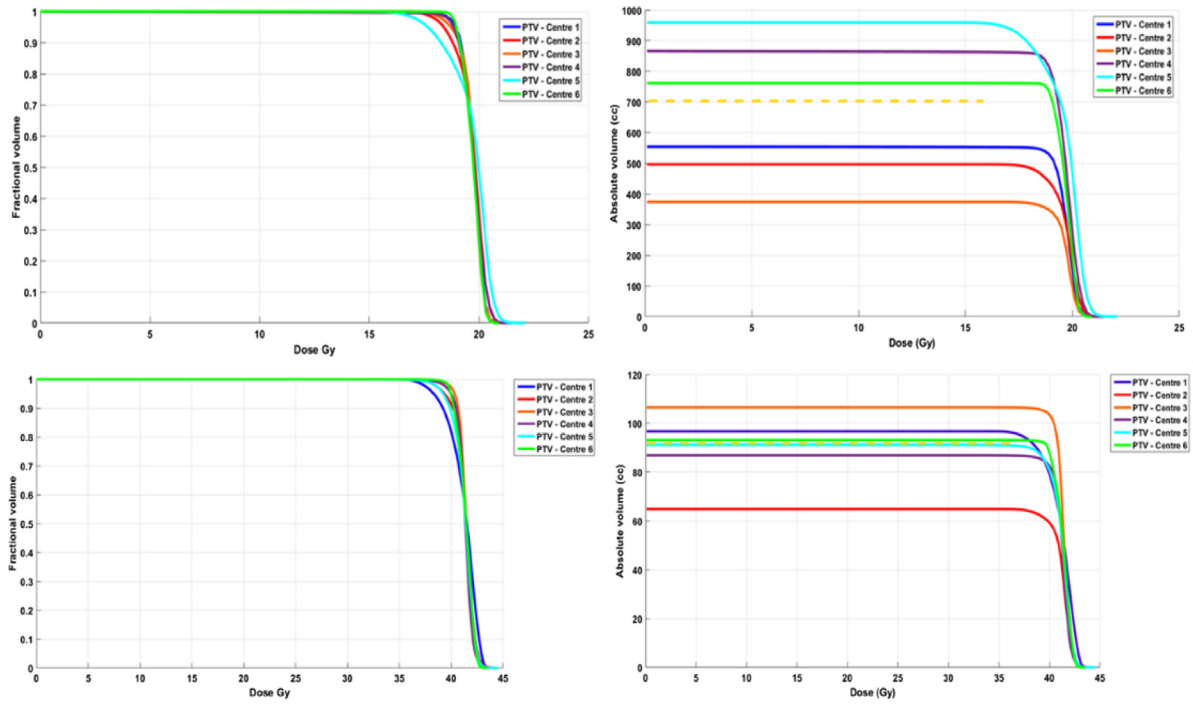

Fig. 2. DVHs for all target volumes for case 1 (top two DVHs) and case 2 (bottom two DVHs). To the left DVHs in relative volume and absolute dose and to the right absolute volume and dose. All targets with own treatment plan, yellow dashed line to the right represents the size of the consensus volume.

Material and methods

Six Swedish centres treating paediatric patients with cancer participated in this study. Two cases were used to demonstrate the potential of adding dose metrics when evaluating target delineation diversity; Hodgkin's disease (case 1) and rhabdomyosarcoma of the parotid gland (case 2). All necessary data were anonymised and sent to the participating centres. The package included the full set of the planning computed tomography (CT), diagnostic imaging; positron emission tomography (PET-CT) for the Hodgkin's case, magnetic resonance imaging (MRI) for the sarcoma case and medical records including histology reports. The package also included relevant study protocols. The planning CT was complemented with a structure set with pre-defined organs at risk (OARs) (all data supplied in Dicom-RT format).

The centres introduced the planning CT into their local treatment planning system (TPS) and continued the work as they would with their own patients. One paediatric radiation oncologist at each centre were asked to delineate GTV, CTV and PTV [36]. They were also asked to create treatment plans for the two cases.

TPSs used were Varian Eclipse, versions 11 and 13 (Varian Medical Systems, Inc. Palo Alto, CA, USA) at four sites and Oncentra Master-Plan, version 4.5 (Elekta, Stockholm, Sweden) at two sites. Dose calculations were performed with the anisotropic analytical algorithm (AAA) in Eclipse and with either the collapsed cone algorithm (case 1 in one centre) or pencil beam algorithm (case 1 at one centre and case 2 at both centres) in Oncentra Master-Plan.

Patient cases

Case 1. Hodgkin's disease (HD) – A 16-year old boy who presented with weight loss and an enlarged neck node. Examination revealed Hodgkin's disease, nodular sclerosis, with involvement of left lower part of neck and mediastinum corresponding to stage IIB, therapy group 2. Treatment was planned and delivered according to the EuroNet-PHL-C2 protocol. He received two cycles of OEPA,[1] with good PET response but a small positive residue was present, thus he also received two cycles of COPDAC.[2] He was then planned for radiotherapy, 19.8 Gy in 11 fractions, to left lower neck, left supraclavicular fossa and left lung hilus including mammary internal nodes in upper and middle mediastinum.

Case 2. Rhabdomyosarcoma (RMS) – A four year old girl who presented with an increasingly swollen cheek. MR showed involvement of the temporalis muscle, the masseter muscle and the attachment at processus coronoideus. Biopsy showed embryonal rhabdomyosarcoma of the left parotid gland. Treatment was planned and delivered according to CWS guidance SR group C. After three chemotherapy cycles, MR showed a 50% tumour regress. Macroscopic radical surgery was performed, however the surgery was not microscopically radical, but a small residue was present. Radiotherapy to 41.4 Gy in 23 fractions was planned with concomitant chemotherapy.

Data analysis

For comparison and analysis of both volumetric and dosimetric data from the different TPSs, the DICOM files containing CT, structure set, treatment plan and dose distribution were imported to and analysed with the CERR software package [37]. In this work we have chosen to use the PTV to explore the additional dose metrics for target delineation variability.

Target volumes

A total of six sets of target volumes, one per participating centre, were prepared for each case. Different strategies for delineation near the skin for case 2 were used at the participating centres. To be able to make fair comparisons, all target volumes were cropped to 4 mm from the skin surface.

[1] OEPA – Vincristine, Etopside, Prednisone, Doxorubicin
[2] COPDAC – Cyclophosfamide, Vincristine, Prednisone, Dacarbazine

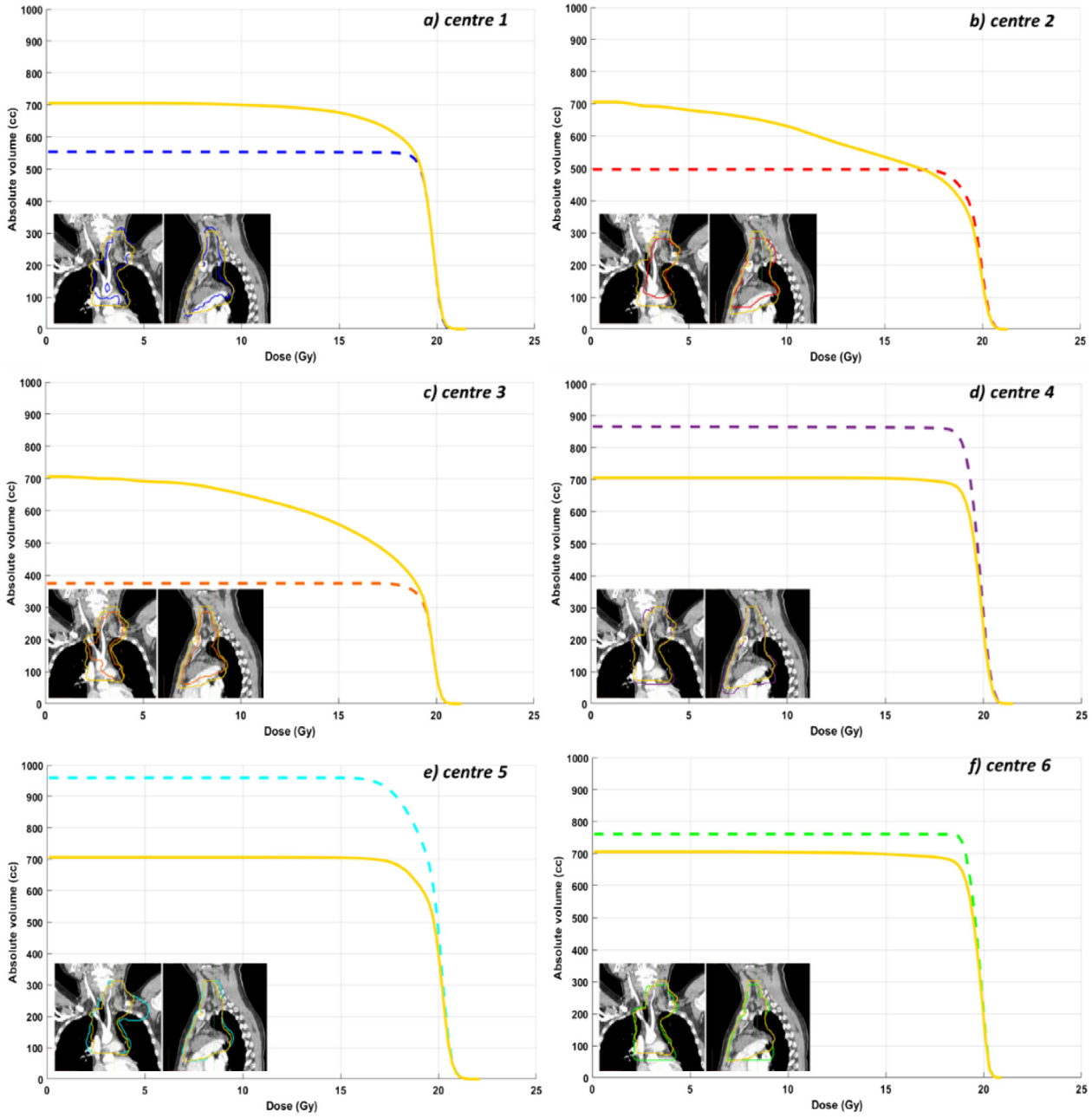

Fig. 3. Individual DVHs for all target volumes (dashed lines, centre 1–6) for case 1 compared to the consensus volume (yellow).

The variability in target delineation between the six centres were quantified using the generalised conformity index, CI_{gen} [38], generated for volume overlap (observer's agreement). The Centre of Mass (CoM) for the delineated volumes was also calculated.

The STAPLE algorithm [21], implemented in CERR, was used for both cases to derive the consensus volumes. STAPLE is a probabilistic estimate of the *true* volume generated from all observers (in this case all the individual delineations). The CERR consensus tool also reports mean sensitivity and mean specificity values as well as Kappa-statistics (K) [39], corrected for chance.

Dose distributions

A number of dose volume descriptors were analysed for each original dose distribution in order to verify the clinical plan quality. The dose-volume descriptors $V_{95\%}$, $D_{98\%}$ (near-minimum dose), $D_{50\%}$ (median dose) and $D_{2\%}$ (near-maximum dose) [40] were analysed for each centre specific PTV. In addition, the homogeneity index [$HI = (D_{2\%} - D_{98\%})/D_{50\%}$], and the radiation conformity index (RCI) [41], based on $V_{95\%}$ for the body, were calculated for each centre's treatment plan.

Dose distributions (Dicom RT dose) from each centre (i) were then applied on the STAPLE determined consensus volume. DVHs were derived and analysed pairwise for each dose distribution, applied to the original target ($DVH_{PTV,i}$) and consensus volume ($DVH_{con,i}$).

The study was approved by the Ethics board of Umeå, Sweden (Dnr 2012-465-31M) and Ethics Board of Lund, Sweden (EPN Lund, Dnr 2013/742).

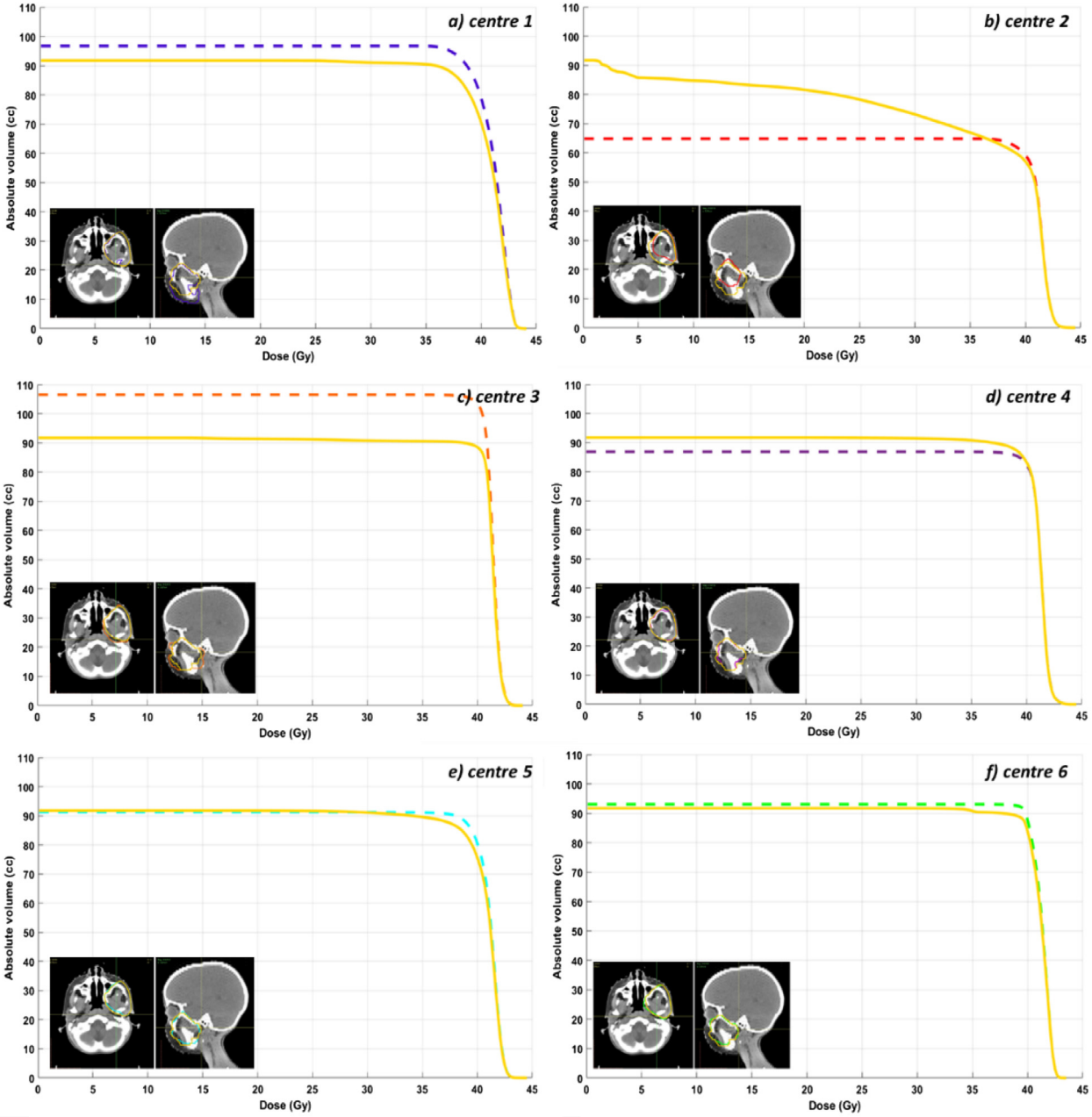

Fig. 4. Individual DVHs for all target volumes (dashed lines, centre 1–6) for case 2 compared to the consensus volume (solid yellow line).

Results and discussion

Target volume evaluation

Inter-physician variability in target delineation has been analysed among a small group of paediatric radiation oncologists. Two paediatric cases were used and both conventional volume comparisons and dose comparisons metrics were evaluated. A considerable variation in target segmentation was seen in both cases (Fig. 1). The largest variation between delineations in case 1 can be seen in the left supraclavicular fossa, towards the right lung and caudally towards the heart. Delineation for case 2 was more unified between the six centres, but there were substantial variations both in the anterior-posterior and in the cranio-caudal direction.

Volume related measures and indices are presented in Table 1.

The CI_{gen} were chosen for comparison of concordance. This parameter has been shown by e.g. Kuwenhoven et al., [38] and confirmed by Fotina et al. [18] to be applicable for any number of pairwise delineations. The CI_{gen} index (c.f. Table 1) is useful for simultaneous comparison of any number of delineations. CI_{gen} = 1 indicates a total overlap, while CI_{gen} = 0 indicates totally separated volumes. The CI_{gen} is, however, a value that can be difficult to interpret. For case 1, the largest variation in target delineation is shown. Compared to a previous study [35], it is in same areas the variation is observed; the width of supraclavicular fossa, and both the width and the caudal extension of the mediastinal part of the target volume. Lütgendorf et al. [17] did a target delineation exercise with experienced radiation oncologists testing two delineation concepts for delineating paediatric Hodgkin's lymphoma. For both concepts the CI_{gen} were less than 0.4. In our studies we had a CI_{gen} for PTV of 0.59 in the first study [35] and 0.53 in the current one.

The CoM standard deviations were 4 mm (*X*), 2 mm (*Y*), 7 mm (*Z*) and 2 mm (*X*), 1 mm (*Y*), 4 mm (*Z*) for case 1 and case 2, respectively. These figures indicate that there is a geometric agreement of the central part of the target volume. This is visualised in Fig. 1 as well.

We decided to use a "golden standard" volume for the analysis. The method used, giving a rather objective result, is the STAPLE algorithm implemented in CERR, which also reports mean sensitivity and mean specificity. STAPLE has been used by other authors [23–34] to create a common consensus volume. However it has to our knowledge not been used in combination with dose metrics to assess how well an external beam treatment plan covers the consensus volume.

The sensitivity for the volume is the relative frequency of an observer including a voxel within the consensus volume and the specificity is the frequency of an observer not including a voxel when it is outside the consensus volume. According to Landis and Kochs "strength of agreement" the *K* for case 1 (Table 2) shows "moderate to good agreement" (moderate; $K = 0.41–0.60$, good; $K = 0.61–0.8$) [42] while *K* for case 2 shows "good agreement".

Dose distributions

The resulting treatment plans from the participating centres were all clinically acceptable. Six volumetric modulated arc therapy (VMAT) plans were created for each of the two cases. Dose-volume metrics for each target volume are presented in Table 3.

The calculated DVHs with absolute volume for the target structures clearly show that there are variations in delineated volume. In Fig. 2, DVHs for all treatment plans applied to their own PTV volume are shown. The two DVHs to the left represent the plan quality, i.e. how well each treatment plan covers its own target. The two DVHs to the right are plotted with absolute volume and the intersection with the volume axis indicates the variation in delineated volumes.

Each dose distribution were applied to both its own target, $DVH_{PTV,i}$, as well as to the consensus volume $DVH_{con,i}$. These DVHs are shown as pairs in Figs. 3 and 4 for case 1 and 2, respectively. Provided that the consensus volume really represents what the observers would/should delineate, it is easy to observe which target volumes that fail regarding over-dosage of healthy tissues or under-dosage of the consensus volume.

There is a substantial variation in volume delineation for case 1. Which target volumes whose corresponding dose distribution will under-dose or over dose are, however, difficult to pinpoint with just a concordance index. CI_{gen} for all PTV volumes for case 1 are 0.53, but when removing the two centres with the largest *RCI* (smallest target volumes), clearly observed in the DVH, the CI_{gen} is 0.60. This was made to test the CI_{gen} index. However, the improvement in concordance isn't very large, making the "improvement" in target delineation difficult to assess. $DVH_{PTV,i}$ and $DVH_{con,i}$ for case 1 presented in Fig. 3a–c indicates that the consensus volume is under-dosed to varying degree. For b and c large parts of the target does not even receive 15 Gy. For the centres in Fig. 3d and e, substantial volumes outside the target receive large doses.

For case 2 the variation in delineated target volume is considerably smaller. The CI_{gen} for this case is 0.70. The $DVH_{PTV,i}$ and $DVH_{con,i}$ for case 2 presented in Fig. 4e and f indicates that the delineated target volumes are almost of identical size as the consensus volume. For centres in Fig. 4a and d the variation is rather small, one (a) over-dosing while the other (d) is under-dosing. However, for centre in Fig. 4c a substantial volume is over-dosed, while a substantial is under-dosed for centre in Fig. 4b. For case 2, with one centre removed, as done for case 1 (smallest target volume), the corresponding CI_{gen} are 0.70 and 0.77 respectively.

Volumetric concordance indices might be difficult to interpret while judging DVHs, as we do in clinical practice, is easier to interpret. Creating treatment plans and applying individual dose distributions to a consensus volume will quickly add more information on the impact and importance of target delineation variations. In this work we have chosen a computer generated consensus volume but the concept could equally well have been applied to e.g. a segmentation performed by an "expert" or a group of "experts".

Conclusions

By applying the treatment plans with its dose distributions for the original target volumes and overlaying them on the consensus volume, we conclude that the DVHs in absolute volume adds information that is more understandable and interpretable compared to various indices for evaluating uniformity in target delineation The DVHs displayed for the consensus volume adds information on both "compliant" target volumes as well as outliers which are hidden with just the use of concordance indices. This information should be reported together with descriptive statistics, concordance indices and statistical measures of agreement [18,20] to get a complete evaluation of delineation studies. More effort is needed to homogenize the segmentation among different centres to be able to truly compare clinical results. There is also a need to develop quality assurance processes in connection with target delineation.

Acknowledgements

We would like to express our sincere appreciation to those dosimetrists and physicists involved in the treatment planning of the cases for this study.

References

[1] Holliday E, Fuller CD, Kalpathy-Cramer J, Gomez D, Rimner A, Ying L, et al. Quantitative assessment of target delineation variability for thymic cancers: agreement evaluation of a prospective segmentation challenge. J Radiat Oncol 2016;5:55–61.
[2] Genovesi D, Cèfaro GA, Vinciguerra A, Augurio A, DiTommaso M, Marchese R, et al. Interobserver variability of clinical target volume delineation in supradiaphragmatic Hodgkin's disease. Strahlenther Onkol 2011;187:357–66.
[3] Louie AV, Rodrigues G, Olsthoorn J, Palma D, Yu D, Yaremo B, et al. Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. Radiother Oncol 2010;95:166–71.
[4] Steenbackers RJHM, Duppen JC, Fitton I, Deurloo KEI, Zijp L, Uitterhoeve ALJ, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: A 'Big Brother' evaluation. Radiother Oncol 2005;77:182–90.
[5] Li AX, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG multi-institutional and multiobserver study. Int J Radiat Oncol Biol Phys 2009;73:944–51.
[6] Nielsen M, Berg M, Pedersen AN, Andersen K, Glavicic V, Jakobsen EH, et al. Delineation of target volumes and organs at risk in adjuvant radiotherapy of early breast cancer: National guidelines and contouring atlas by the Danish Breast Cancer Cooperative Group. Acta Oncol 2013;52:703–10.
[7] Fokas E, Spezi E, Patel N, Hurt C, Nixon L, Chu K-Y, et al. Comparison of investigator-delineated gross tumour volumes and quality assurance in pancreatic cancer: Analysis of the on-trial cases for the SCALOP trial. Radiother Oncol 2016;120:212–6.
[8] Jeanneret-Sozzi W, Moeckli R, Valley J-F, Zouhair A, Ozsahin EM, Mirimanoff R-O. The reasons for discrepancies in target volume delineation. A SASRO study on head and neck and prostate cancers. Strahlenther Onkol 2006;182:450–7.
[9] Nakamura K, Shioyama Y, Tokomaru S, Hayashi N, Oya N, Hiraki Y, et al. Variation of clinical target volume definition among Japanese radiation oncologists in external beam radiotherapy for prostate cancer. Jpn J Clin Oncol 2008;38:275.
[10] Vinod A, Min M, Jameson M, Holloway L. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. J Med Imag Radiat Oncol 2016;60:393–6.
[11] Gambacorta MA, Boldrini L, Valentini C, Dinapoli N, Mattiucci GC, et al. Automatic segmentation software in locally advanced rectal cancer: READY (REsearch program in Auto Delineation sYstem)-RECTAL 02: prospective study. Oncotarget 2016;7:42579–84. https://doi.org/10.18632/oncotarget.9938.

[12] Mattiucci GC, Boldrini L, Chiloiro G, et al. Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? Acta Oncol 2013 Oct;52:1417–22.
[13] Grau Eriksen J, Salembier C, Rivera S, De Bari B, Berger D. Four years with FALCON – An ESTRO educational project. Achieve Perspect Radiother Oncol 2014;112:145–9.
[14] Marks LB, Adams RD, Pawlicki T, Blumberg AL, Hoopes D. Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: Executive summary. Practical Rad Onc 2013;3. 149-6.
[15] Coles CE, Hoole ACF, Harden SV, Burnet N, Twyman N, Taylor E, et al. Quantitative assessment of inter-clinician variability of target volume delineation for medulloblastoma: Quality assurance for the SIOP PNET 4 trial protocol. Radiother Oncol 2003;69:189–94.
[16] Padovani L, Huchet A, Claude L, Bernier V, Quetin P, Mahe M, et al. Inter-clinician variability in making decisions in pediatric treatment: A balance between efficacy and late effects. Radiother Oncol 2009;93:372–6.
[17] Lütgendorf-Caucig C, Fotina I, Gallop-Evans E, Claude L, Lindh J, Knäusl, et al. Multicenter evaluation of different target volume delineation concepts in pediatric Hodgkin's lymphoma. Strahlenter Onkol 2012;188:1025–30.
[18] Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenter Onkol 2012;188:160–7.
[19] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol 2014 Sep;112:317–20.
[20] 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;121:169–79.
[21] Warfield SK, Zou KH, Wells M. Simultaneous Truth and Performance Level Estimation (STAPLE), an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004; (23)7: p. 903–21.
[22] Allozi R, Li AX, White J, Apte A, Tai A, Michalski JA, et al. Tools for consensus analysis of experts' contours for radiotherapy structure definitions. Radiother Oncol 2010;97:572–8.
[23] Lawton CAF, Michalski J, El-Naqua I, Kuban D, Lee WR, Rosenthal SA, et al. Variation in the definition of clinical target volumes for pelvic nodal conformal radiation therapy for prostate cancer. Int J Radiaton Oncol Biol Phys 2009;74:377–82.
[24] Yang J, Woodward WA, Reed VK, Strom EA, Perkins GH, Tereffe VK, et al. Statistical modeling approach to quantitative analysis of interobserver variability in breast contouring. Int J Radiat Oncol Biol Phys 2014;89:214–21.
[25] Gillespie EF, Panjwani N, Golden DW, Gunther J, Chapman TR, Brower JV, et al. Multi-institutional randomized trial testing the utility of an interactive 3-dimensional contouring atlas among radiation oncology residents. Int J Radiat Oncol Biol Phys 2016. https://doi.org/10.1016/j.ijrobp.2016.11.050.
[26] Suk Kim Y, Won Lim J, Sup Yoon W, Kyu Kang M, Jae Lee I, Hyun Kim T, et al. Interobserver variability in gross tumor volume delineation for hepatocellular carcinoma. Strahlenter Onkol 2016;192:714–21.
[27] Jensen NKG, Mulder D, Lock M, Fisher B, Zener R, Beech B, et al. Dynamic contrast enhanced CT aiding gross tumor volume delineation of liver tumors: An interobserver variability study. Radiother Oncol 2014;111:153–7.
[28] Ost P, De Meerleer G, Vercauteren T, De Gersem W, Veldeman L, Vandecasteele K, et al. Delineation of the postprostatectomy prostate bed using computed tomography: interobserver variability following the EORTC delineation guidelines. Int J Radiat Oncol Biol Phys 2011;81:e143–9.
[29] Petriĉ P, Hudej R, Rogelj P, Blas M, Tanderup K, Fidarova E, et al. Uncertainties of target volume delineation in MRI guided adaptive brachytherapy of cervix cancer: A multi-institutional study. Radiother Oncol 2013;107:6–12.
[30] Viswanathan AN, Erickson B, Gaffney D, Beriwal S, Bhatia, S, Burnett III OL, et al. Comparison and consensus guidelines for delineation of clinical target volume for CT- and MR-based brachytherapy in locally advanced cervical cancer. Int J Radiaton Oncol Biol Phys 2014; 90: p. 320–8.
[31] Kosztyla R, Chan EK, Hsu F, Wilson D, Ma R, Cheung A, et al. High-grade glioma radiation therapy target volumes and patterns of failure obtained from magnetic resonance imaging and ^{18}F-FDOPA position emission tomography delineations from multiple observers. Int J Radiat Oncol Biol Phys 2013;87:1100–6.
[32] Awan M, Kalpathy-Cramer J, Gunn BG, Beadle BM, Garden AS, Phan J, et al. Prospective assessment of an atlas-based intervention combined with real-time software feedback in contouring lymph node levels and organs-at-risk in the head and neck: Quantitaive assessment of conformance to expert delineation. Practical Rad Onc 2013;3. 186-3.
[33] Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiaton Oncol Biol Phys 2010;77. 959-6.
[34] Hellebust TP, Tanderup K, Lervåg C, Fidarova E, Berger D, Malinen E, et al. Dosimetric impact of interobserver variability in MRI-based delineation for cervical cancer brachytherapy. Radiother Oncol 2013;107:13–9.
[35] Kristensen I, Agrup M, Bergström P, Engellau J, Haugen H, Martinsson U, et al. Assessment of volume segmentation in radiotherapy of adolescents; a treatment planning study by the Swedish Workgroup for Paediatric Radiotherapy. Acta Oncol 2014;53:126–30.
[36] Prescribing, recording, and reporting photon-beam intensity modulated radiation therapy (IMRT), ICRU report 50, Bethesda, USA; 1993.
[37] Deasy J, Blanco A, Clark V. CERR: A computational environment for radiotherapy research. Med Phys 2003;30:979–85.
[38] Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol 2009;54:2863–73.
[39] Altman DG. Practical statistics for medical research. New York: Chapman & Hall/CRC; 1997.
[40] Prescribing, recording, and reporting intensity modulated radiation therapy (IMRT). ICRU report 83, Bethesda, USA; 2010.
[41] Knöös T, Kristensen I, Nilsson P. Volumetric and dosimetric evaluation of radiation treatment plans: Radiation conformity index. Int J Radiat Oncol Biol Phys 1998;42:1169–76.
[42] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.