Radiotherapy and Oncology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

# A phantom assessment of achievable contouring concordance across multiple treatment planning systems

Elise M. Pogson [a,b,*], Jarrad Begg [b], Michael G. Jameson [a,b], Claire Dempsey [c,d], Drew Latty [e], Vikneswary Batumalai [b,i], Andrew Lim [f], Kankean Kandasamy [g], Peter E. Metcalfe [a,b], Lois C. Holloway [a,b,h,i]

[a] Centre for Medical Radiation Physics, University of Wollongong; [b] Liverpool and Macarthur Cancer Therapy Centres and the Ingham Institute, Liverpool Hospital; [c] Department of Radiation Oncology, Calvary Mater Newcastle Hospital; [d] School of Health Sciences, University of Newcastle; [e] The Crown Princess Mary Cancer Centre, Westmead; [f] Peter MacCallum Cancer Centre, Melbourne; [g] Prince of Wales Hospital, Randwick; [h] Institute of Medical Physics, University of Sydney; and [i] South Western Clinical School, University of New South Wales, Australia

ARTICLE INFO

*Article history:*
Received 17 December 2014
Received in revised form 1 September 2015
Accepted 18 September 2015
Available online xxxx

*Keywords:*
Radiotherapy
Contouring
Delineation
Inter-observer
DICE
Hausdorff

ABSTRACT

In this paper, the highest level of inter- and intra-observer conformity achievable with different treatment planning systems (TPSs), contouring tools, shapes, and sites have been established for metrics including the Dice similarity coefficient (DICE) and Hausdorff Distance. High conformity values, e.g. $DICE_{Breast\_Shape} = 0.99 \pm 0.01$, were achieved. Decreasing image resolution decreased contouring conformity.



Delineation of radiotherapy structures has direct clinical consequences. Contouring of nodal CTV sub-volumes in particular, is critical [1]. Even moderate geometrical differences in small neck Planning Target Volumes (PTVs) can impact on the target dose (up to 11 Gy reductions in D99 for DICE above 0.8) [2]. For non-small lung cancer variation a concordance index (CI) has been demonstrated to result in variation in Tumour Control Probability (TCP) [3], highlighting the correlation between contour variation and TCP. However, there are no reported contour variation metric baseline values considering uncertainties in the process such as different treatment planning systems (TPSs), importing and exporting processes, contour shapes, volumes and image resolution. Knowledge of these baseline values is important for clinical trials which commonly occur across multiple centres and TPSs. Current literature does not give clear guidelines for reporting contouring variability in inter-observer studies [4] with variation in methodology and metrics only enabling comparison between inter-observer studies in a limited fashion [5]. As such, calculating multiple metrics including a combination of descriptive statistics, overlap measures and statistical measures of agreement is recommended for multiple observer studies [6].

The number of studies reporting on auto-segmentation [7,8], and the inter- [9,10] and intra- [11] observer conformity of volumes is growing. Inadequate definition of the Gross Tumour Volume (GTV) or Clinical Target Volume (CTV) leads to systematic uncertainty which may result in geometric miss of the tumour throughout the course of patient radiation therapy [5]. As such there has been an increasing trend to assess, and reduce, the variability of these target volumes. This study determined the highest concordance metrics achievable, and how these metrics (details given in Supplementary Table 1) may vary in a best case phantom scenario considering: multiple sites, variation between TPSs, shapes, volume, tools utilized and adherence to auto-threshold settings within the protocol.

## Methods

A Quasar Body phantom (Modus Medical Devices Incorporated, Ontario Canada) was used to provide an initial CT dataset. The Quasar phantom was scanned on a Brilliance Big Bore CT (Phillips Healthcare, The Netherlands) using a helical abdomen scanning sequence: 1 mm slice spacing, 2 mm slice thickness, standard

* Corresponding author at: School of Physics, Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2522, Australia.
*E-mail address:* elisep@uow.edu.au (E.M. Pogson).

http://dx.doi.org/10.1016/j.radonc.2015.09.022

resolution (512 × 512) and field of view of 350 mm. This phantom had three inserts containing structures providing a range of surface contours and edges. In this study the 20-degree air wedge contained in the first insert (referred to as the triangular prism) and the entire empty third insert (an 8 cm diameter cylinder with semi-conic top) were used for contouring.

The Quasar phantom CT dataset was imported into MATLAB R2012a (Mathworks Incorporated, Natick USA). Uniform rectangular prisms and a patient breast volume (203 $cm^3$) were inserted into the CT dataset using 'Computational Environment for Radiotherapy Research' (CERR) [12,13] and MATLAB. High intensities were utilized to obtain optimal image contrast. The Quasar phantom with inserted shapes is displayed, with inter-observer contours, in Supplementary Fig. 1.

A contouring protocol set image window levels to Window/Level = 400/800 HU and described allowable techniques/tools. All eight rectangular prisms were auto-contoured using auto-threshold at recommended threshold values or other automated tools (e.g. Oncentra's magic-wand tool). Rectangular prisms 1, 4 and 8 (Supplementary Fig. 1) were manually contoured. Bounding boxes in auto-contouring and zoom functions were allowed. The breast contour was manually delineated; allowing interpolation between slices and/or copy to next slice. The triangular prism and cylinder were both delineated using automated tools (such as auto-threshold) and manually. All eight observers were blind to others contours. The TPSs used for contouring were; Eclipse Planning System 11.0.64 (Varian Medical Systems, Palo Alto Canada): two sites, Oncentra (Elekta, Stockholm Sweden): two sites, Pinnacle$^3$ 9.0 (Philips, Netherlands): two sites, and FocalSim 4.80.01 (Elekta, Stockholm Sweden): two sites. These contours were then exported and collated in CERR.

The same original 512 × 512 data-set was contoured five times by four observers, with a minimal 24 h time lapse between contouring. Pairwise analysis of the Jaccard Index (JI) also known as conformity index or concordance index (CI) [6,14] ($CI_{pairs}$ the average of all possible pairs of the JI which equates to $CI_{gen}$ when mutual variability between all observers is the same [15]), Volume Overlap Index (VOI) and Hausdorff Distances (HDs) were calculated for each observer and averaged. This was performed for all manually contoured structures.

Different studies have different image resolutions. As such the Quasar phantom was resampled and contoured by five different observers, to show the expected inter-observer effects for differing sample/dataset pixel size and slice thickness. The resampling was performed in MATLAB with the overall volume maintained. Slice thickness was also set to the spacing of 2 mm, 4 mm and 8 mm keeping the resolution at 512 × 512 px (1.463 px/mm) and saved as DICOM. The resampled DICOM data were of the following resolutions; 512 × 512 $px^2$ (1.463 px/mm – a typical high resolution CT), 350 × 350 $px^2$ (1.000 px/mm), 245 × 245 $px^2$ (0.700 px/mm), 175 × 175 $px^2$ (0.500 px/mm), 88 × 88 $px^2$ (0.250 px/mm), and 44 × 44 $px^2$ (0.125 px/mm).

To allow comparison between observers, simultaneous truth and performance level estimation (STAPLE) volumes were generated as consensus gold standard reference volumes in CERR, using a 90% confidence interval with observers weighted equally. CERR was utilized to calculate the generalized kappa statistic as well as the DICE, and JI in three dimensions for all observers comparing to the gold standard STAPLE volume (Supplementary Table 1). The maximal HD, average Hausdorff Distance, $CI_{pairs}$ and VOI was calculated in a pairwise analysis over all volumes in MilxView (Australian e-Health Research Centre (AEHRC), Australia) [16,17] (Supplementary Table 2).

The JI [18–20], DICE [4], Hausdorff distance [21] and Kappa ($\kappa$) statistic [22,23] outlined in Supplementary Table 1, are metrics commonly used to establish inter-observer variation [6]. JI and DICE values from CERR were verified in 3D Slicer [24–26] and MILXview and were consistent to within 2 significant figures.

## Results

Eight auto-contoured, inter-observer rectangular prism contours from different TPSs were all within two pixels of the true volume on every slice, for every point within the contour (Fig. 1(a)). The maximum HD of these contours compared to the STAPLE ranged from 1 pixel width/height (0.68 mm) or 2 pixels added in quadrature (0.97 mm), with a maximum of 3 pixels (2.04 mm) for the auto-contoured rectangular prisms (Fig. 1(c)). As the STAPLE for square 5 is different to the true volume there are larger HDs and discrepancies for this volume. A pairwise HD measure, rather than to the STAPLE, is less sensitive to such errors and is used in all following analysis. Fig. 1(b) displays each inter-observer's DICE compared to the STAPLE. Inter- and intra- observer contour variation as measured by maximum HD relative to the STAPLE volumes was less than 7 mm for all volumes at normal resolution (1.463 px/mm). Kappa statistics comparing multiple shapes from the Quasar phantom show near perfect agreement for most shapes despite asymmetry from the breast contour (Supplementary Fig. 2).

Auto-contoured rectangular prisms were less conformal (kappa in the range of 0.61–0.80) than manually delineated shapes (kappa in the range of 0.81–1), (Supplementary Fig. 2), with other shapes having no difference. The contouring tool used did not show any observable effect in contour conformity. Average manual and auto-threshold DICE were in agreement (within the 95%

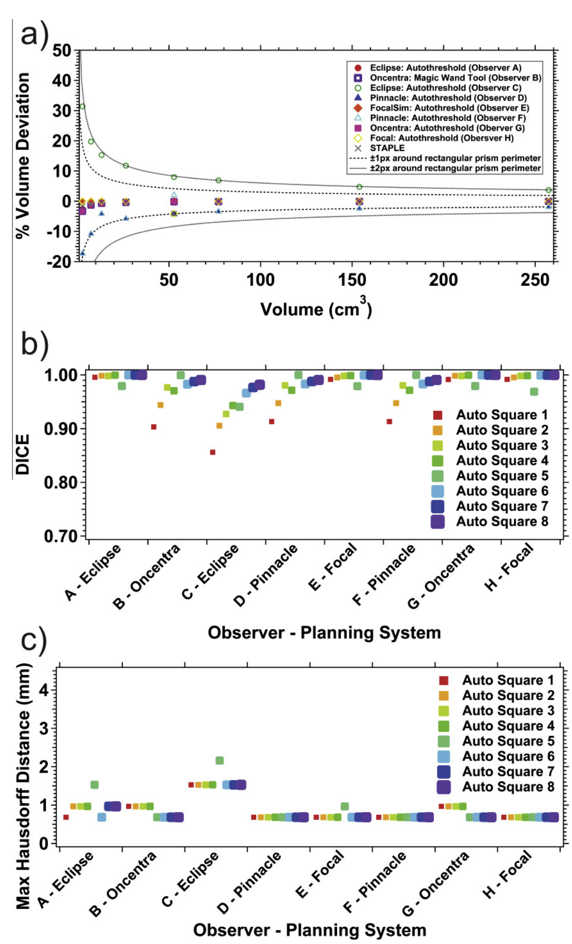


**Fig. 1.** Auto-contoured squares; a) Percentage deviation of volume from the true volume. Majority of contours are within 1 px and the rest within 2 px, b) DICE c) maximum HD from the STAPLE volume.

Please cite this article in press as: Pogson EM et al. A phantom assessment of achievable contouring concordance across multiple treatment planning systems. Radiother Oncol (2015), http://dx.doi.org/10.1016/j.radonc.2015.09.022