2015

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

Elise M. Pogson
University of Wollongong, elisep@uow.edu.au

Jarrad Begg
Liverpool Hospital

Michael Jameson
University of Wollongong, mgj77@uowmail.edu.au

Claire Dempsey
University of Newcastle

Drew Latty
Crown Princess Mary Cancer Centre

See next page for additional authors

Publication Details

Pogson, E. M., Begg, J., Jameson, M. G., Dempsey, C., Latty, D., Batumalai, V., Lim, A., Kandasamy, K., Metcalfe, P. E. & Holloway, L. C. (2015). A phantom assessment of achievable contouring concordance across multiple treatment planning systems. Radiotherapy and Oncology, 117 (3), 438-441.

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

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. DICEBreast_Shape = 0.99 ± 0.01, were achieved. Decreasing image resolution decreased contouring conformity.

Keywords
assessment, systems, achievable, contouring, concordance, across, planning, multiple, phantom, treatment

Disciplines
Engineering | Science and Technology Studies

Publication Details
Pogson, E. M., Begg, J., Jameson, M. G., Dempsey, C., Latty, D., Batumalai, V., Lim, A., Kandasamy, K., Metcalfe, P. E. & Holloway, L. C. (2015). A phantom assessment of achievable contouring concordance across multiple treatment planning systems. Radiotherapy and Oncology, 117 (3), 438-441.

Authors
Elise M. Pogson, Jarrad Begg, Michael Jameson, Claire Dempsey, Drew Latty, Vikneswary Batumalai, Andrew Lim, Kankean Kandasamy, Peter E. Metcalfe, and Lois C. Holloway

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

Elise M. Pogson[1,2], Jarrad Begg[2], Michael G. Jameson[1,2], Claire Dempsey[3,9], Drew Latty[4], Vikneswary Batumalai[2,8], Andrew Lim[5], Kankean Kandasamy[6], Peter E. Metcalfe[1,2], Lois C. Holloway[1,2,7,8].

1. Centre for Medical Radiation Physics, University of Wollongong, Wollongong, NSW, Australia.
2. Liverpool and Macarthur Cancer Therapy Centres and the Ingham Institute, Liverpool Hospital, Liverpool, NSW, Australia.
3. Department of Radiation Oncology, Calvary Mater Newcastle Hospital, Newcastle, NSW, Australia.
4. The Crown Princess Mary Cancer Centre, Westmead, NSW, Australia.
5. Peter MacCallum Cancer Centre, Melbourne, VIC, Australia.
6. Prince of Wales Hospital, Randwick, NSW, Australia
7. Institute of Medical Physics, University of Sydney, NSW, Australia
8. South Western Clinical School, University of New South Wales, NSW, Australia
9. School of Health Sciences, University of Newcastle, NSW, Australia

Corresponding authors email: *elisep@uow.edu.au*
Corresponding authors phone: *+612 42214054*
Corresponding authors address: *School of Physics, Engineering and Information Sciences, University of Wollongong, Wollongong, NSW, 2522, Australia.*

Total number of pages: 4
Total number of tables: 0
Total number of figures: 2
Total number of supplementary material: 5
Running head: A multi-observer concordance baseline
Keywords: Radiotherapy, Contouring, Delineation, Inter-observer, DICE, Hausdorff

Disclaimer: The views expressed in this article are my own and not an official position of the institution or funding support.
Sources of Support: Cancer Australia and National Breast Cancer Foundation grant project number 1033237.
Total Word Count: 1930
Abstract Word Count: 50
Conflict of Interest Declaration: The authors report a grant from Cancer Australia and National Breast Cancer Foundation, during the conduct of the study.

Abstract

In this paper, metrics including the Dice similarity coefficient (DICE) and Hausdorff Distance determine the highest level of inter- and intra-observer conformity achievable with different treatment planning systems (TPSs), contouring tools, shapes, and sites. High conformity values, e.g. $DICE_{Breast\_Shape}=0.99\pm0.01$, are achieved with differing TPSs. Decreasing image resolution decreased contouring conformity.

INTRODUCTION

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 CI(%) of 0.66-0.90% has been demonstrated to result in variation in Tumour Control Probability (TCP) of 0.19–0.68% [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 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) including; Jaccard Index (JI also known as conformity index or concordance index (CI) [6, 12]), CI_{pairs} the average of all possible pairs of the JI (equates to CI_{gen} when mutual variability between all observers is the same [13]), Dice Coefficient (DICE or DSC), Volume Overlap Index (VOI), the generalised kappa statistic and Hausdorff Distance (HD), 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

Image Datasets

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 resolution (512×512) and field of view of 350 mm. This phantom had threeinserts 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 a Computational Environment for Radiotherapy Research CERR [14, 15] and MATLAB. High intensities were utilised to obtain optimal image contrast. The Quasar phantom with inserted shapes is displayed, with inter-observer contours, in Supplementary Fig. 1.

Inter-Observer Contouring Protocol

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): 2 sites, Oncentra (Elekta, Stockholm Sweden): 2 sites, Pinnacle^3 9.0 (Philips, Netherlands): 2 sites, and FocalSim 4.80.01 (Elekta, Stockholm Sweden): 2 sites. These contours were then exported and collated in CERR.

Intra-Observer Contouring

The same original 512×512 data-set was contoured five times by four observers, with a minimal 24 hour time lapse between contouring. Pairwise analysis CI_{pairs}, VOI and HD's were calculated for each observer and averaged. This was performed for all manually contoured structures.

Inter observer contouring at lowering image resolutions

Different studies have different image resolutions. As such the Quasar phantom was resampled and contoured by 5 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).

Analysis Metrics

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 utilised 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 Hausdorff Distance, 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 (κ) 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). There were no observable trends between automatically or manually delineated contours. 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% confidence limit) for all shapes.

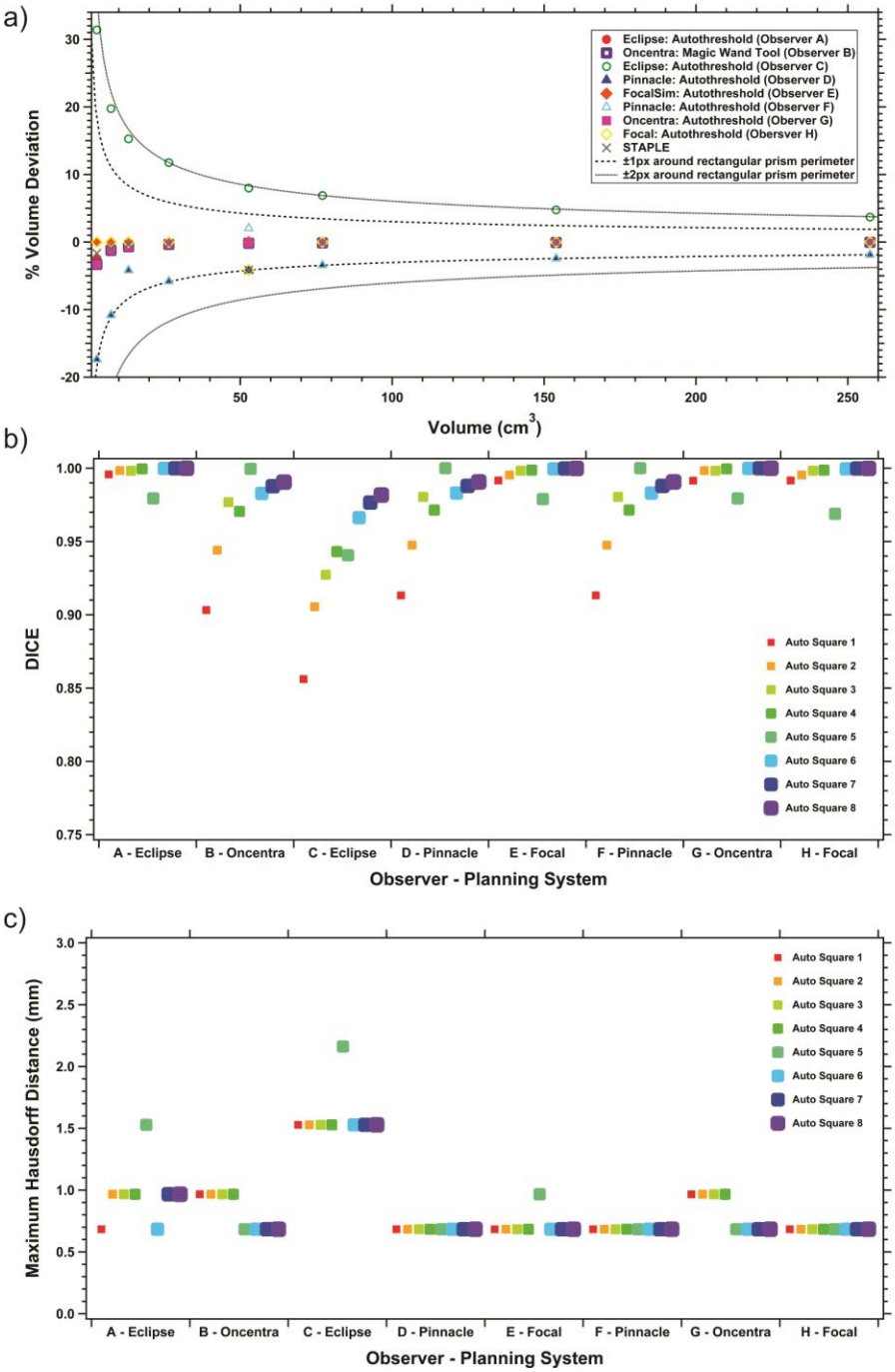


Fig. 1. Auto-contoured squares; a) Percentage deviation of volume from the true volume. Majority of contours are within 1 px^2 and the rest within 2 px^2, b) DICE c) maximum HD from the STAPLE volume. Observer C display's the largest deviation from the STAPLE.

The JI, average DICE and kappa for the manually delineated shapes are summarized in Supplementary Table 2.

Inter-observer generalized kappa statistics for differing shapes is shown in Fig. 2(a). Decreasing image resolution reduces concordance, especially for smaller structure volumes e.g. triangular prism (47 cm^3). This is evident in the average DICE compared to the STAPLE volume in each image (Fig. 2(b)) and

the average maximal HDs (Fig. 2(c)). The HDs are increasing due to lengthening pixel sizes. This was similar to results shown in another study [27]. The breast contour and some rectangular prisms with an image resolution of 0.250 px/mm and 0.125 px/mm were excluded as the outline was not visible at recommended window levels due to resampling.

As resolution decreases below 0.250 px/mm, the relative inter-observer DICE also decreases for manual contours, despite Fig. 2(b) showing good concordance compared to the STAPLE generated on each individual resolution dataset. Supplementary Fig. 3, displays the relative DICE of contours with lowering resolution compared to the highest resolution image (1.49).

Varying the slice thickness from 1 mm to 2 mm, 4 mm and 8 mm had no significant effect on inter-observer conformity.

DISCUSSION

Inter-observer variation is shown to increase with lower resolution. Intra-observer variation is either in agreement or smaller than inter-observer variation similarly to previously reported clinical findings [5]. Disagreement between the same TPS is evident for contours generated using auto-threshold tools in the same TPS by different observers, (Fig. 1(c)). Hounsfield Units (HUs) used for Auto-thresholding were requested, and showed significantly different HUs had been used. This ambiguity is likely due to conversion between TPSs. We recommend that the conversion between multiple TPSs for inter-observer studies be performed and sent out with the study dataset in future studies. The highest achievable values are dependent upon image resolution, contour volume, number of observers, image contrast, window level and adherence to the protocol.

Previously reported values in breast radiotherapy CTV inter-observer studies include a JI of; 0.81 for radiation oncologist breast contouring [9], 0.84 for radiation therapist breast contouring [9], 0.87 for glandular breast volumes [12], 0.56 for partial breast volumes [12] and 0.82 for glioblastoma GTV's (Gross Tumour Volumes) [28]. An inter-observer breast contour generalized kappa of 0.97 ($p<0.05$), maximal HD of 3.42 mm, average JI of 0.98±0.01 and average DICE of 0.99±0.01 was found in this study. This demonstrates the highest achievable values for future expert clinician contours compared to a STAPLE volume, for an acceptable number of observers (five or more, with a recommendation to have as large a number of expert observers as possible for small volumes [27]) and a standard CT image resolution (512×512). The gold standard STAPLE volume has been generated by the contours assessed here, whilst this has minimal effect, in an ideal study the aim would be to have a separate group of contours to generate a gold

standard STAPLE and compare to this. To avoid this metrics such as CI_{pairs} or VOI may be utilised instead.

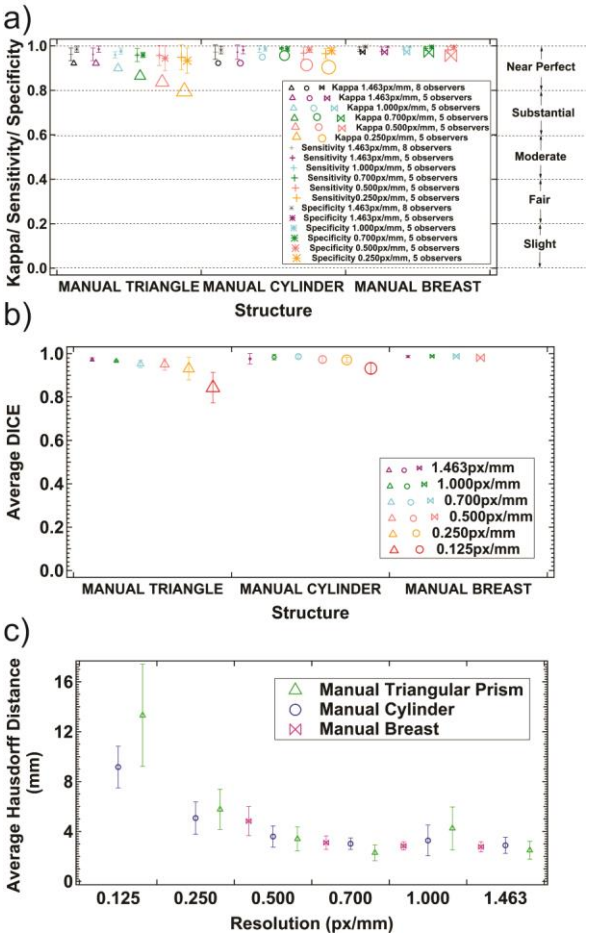


Fig. 2. Manually delineated Inter-observer a) STAPLE parameters with differing image resolution; Kappa, Specificity, Sensitivity and Volume, b) 5 observer average DICE and c) 5 observer average Hausdorff Distances. Error bars represent 1SD. The STAPLE in the resampled images have lower specificity and sensitivity with lowering resolution. The 95% confidence intervals also become larger, for small volumes, with worsening resolution (as the amount of data is reduced).

Complexity of shape showed no observable effect in conformity, as the complicated breast contour achieved a higher average DICE, average JI and Kappa than the cylinder and rectangular prism, of similar volumes. However an assessment of more complicated irregular shapes than rounded breast contours still needs to be undertaken.

Multi-observer results from multiple TPSs, differing TPS tools, image resolution, image slice thickness, contour shapes and volumes has been established for average DICE, average JI, CI_{pairs}, VOI, kappa, average HD and maximum HD. Values obtained in this phantom study suggest that multiple sites and systems do not have significant impact on concordance metrics for these particular volumes. Values presented here may provide an upper bound as to what is achievable in future studies. Alternatively if images are of significantly different image resolution, extremely small volumes (such as a head and neck study), of more irregular shape, or with less observers, future studies might consider including another object/dataset to determine their highest achievable kappa, average DICE or average JI under these circumstances. This could be undertaken on a study by study basis.

ACKNOWLEDGEMENTS

This work was supported by a Cancer Australia and National Breast Cancer Foundation grant project number 1033237.

REFERENCES

[1] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol.112:317-20.
[2] Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother Oncol. 2011;98:373-7.
[3] Jameson MG, Kumar S, Vinod SK, Metcalfe PE, Holloway LC. Correlation of contouring variation with modeled outcome for conformal non-small cell lung cancer radiotherapy. Radiother Oncol. 2014;112:332-6.
[4] Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, Ang K, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract Radiat Oncol. 2014;4:e31-7.
[5] Weiss E, Hess CF. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy. Strahlenther Onkol. 2003;179:21-30.
[6] 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. Strahlenther Onkol. 2012;188:160-7.
[7] Zhou W, Xie Y. Interactive contour delineation and refinement in treatment planning of image-guided radiation therapy. J Appl Clin Med Phys. 2014;15:4499-522.
[8] Simmat I, Georg P, Georg D, Birkfellner W, Goldner G, Stock M. Assessment of accuracy and efficiency of atlas-based autosegmentation for prostate radiotherapy in a variety of clinical conditions. Strahlenther Onkol. 2012;188:807-15.
[9] Holloway LC, Jameson MG, Batumalai V, Koh E, Papadatos G, Lonergan D, et al. Estimating a Delineation Uncertainty Margin to Account for Inter-observer Variability in Breast Cancer Radiotherapy. Int J Radiat Oncol Biol Phys. 2010;78:S741.
[10] Yamazaki H, Shiomi H, Tsubokura T, Kodani N, Nishimura T, Aibe N, et al. Quantitative assessment of inter-observer variability in target volume delineation on stereotactic radiotherapy treatment for pituitary adenoma and meningioma near optic tract. Radiat Oncol. 2011;6.
[11] Lütgendorf-Caucig C, Fotina I, Stock M, Pötter R, Goldner G, Georg D. Feasibility of CBCT-based target and normal structure delineation in prostate cancer radiotherapy: multi-observer and image multi-modality study. Radiother Oncol. 2011;98:154-61.
[12] Struikmans H, Wárlám-Rodenhuis C, Stam T, Stapper G, Tersteeg RJHA, Bol GH, et al. Interobserver variability of clinical target volume delineation of glandular breast tissue and of boost volume in tangential breast irradiation. Radiother Oncol. 2005;76:293-9.
[13] 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.
[14] Apte A, Khullar D, Alaly J, Deasy J, O. The Computational Environment for Radiotherapy Research (CERR). 2010.
[15] Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003;30:979-85.
[16] Dowling JA. Opportunities for image analysis in radiation oncology. Australas Phys Eng Sci Med. 2014;37:275-7.
[17] Dowling JA, Fripp J, Chandra S, Pluim JPW, Lambert J, Parker J, et al. Fast automatic multi-atlas segmentation of the prostate from 3D MR images. Prostate Cancer Imaging Image Analysis and Image-Guided Interventions: Springer; 2011. p. 10-21.

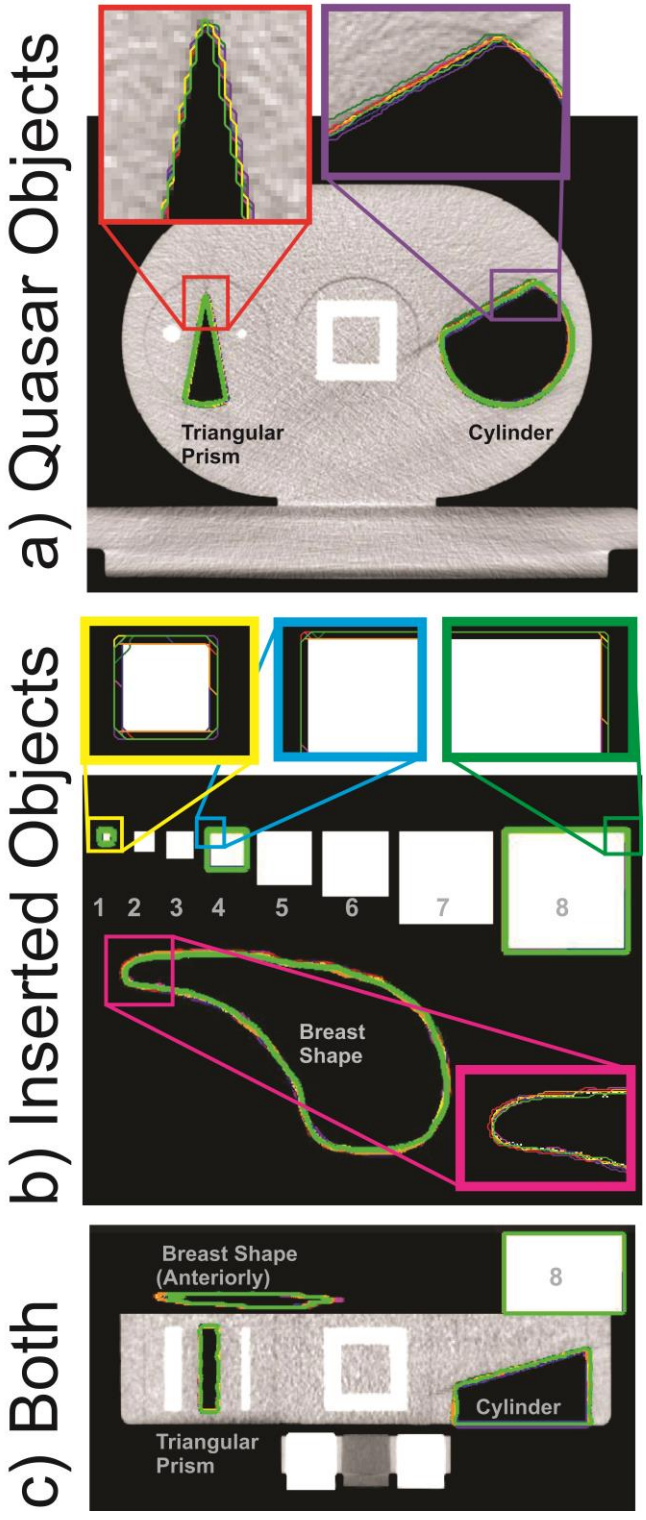
[18] Petersen RP, Truong PT, Kader HA, Berthelet E, Lee JC, Hilts ML, et al. Target Volume Delineation for Partial Breast Radiotherapy Planning: Clinical Characteristics Associated with Low Interobserver Concordance. Int J Radiat Oncol Biol Phys. 2007;69:41-8.
[19] Fein DA, McGee KP, Schultheiss TE, Fowble BL, Hanks GE. Intra- and interfractional reproducibility of tangential breast fields: A prospective on-line portal imaging study. Int J Radiat Oncol Biol Phys. 1996;34:733-40.
[20] Feuvret L, Noël G, Mazeron J-J, Bey P. Conformity index: A review. Int J Radiat Oncol Biol Phys. 2006;64:333-42.
[21] Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. J Med Imaging Radiat Oncol.54:401-10.
[22] Ebert M, McDermott L, Haworth A, van der Wath E, Hooton B. Tools to analyse and display variations in anatomical delineation. Australas Phys Eng Sci Med. 2012;35:159-64.
[23] Lim K, Small W, Jr., Portelance L, Creutzberg C, Jurgenliemk-Schulz IM, Mundt A, et al. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. Int J Radiat Oncol Biol Phys. 2011;79:348-55.
[24] Fedorov A, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323-41.
[25] Fedorov A, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, et al. 3D Slicer. 4.3 ed.
[26] Pinter C, Lasso A, Wang A, Jaffray D, Fichtinger G. SlicerRT: radiation therapy research toolkit for 3D Slicer. Med Phys. 2012;39:6332-8.
[27] Commowick O, Warfield SK. Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE. IEEE Trans Med Imaging. 2010;29:771-80.
[28] Ryuji M, Toshinori H, Ryo T, Hideo N, Yasuyuki Y. Double reading for gross tumor volume assessment in radiotherapy planning. J Solid Tumors. 2012;2:38.

Supplementary material

Supplementary . Table 1. Concordance measures and tools.

Metric	Equation/Outline	Description	Metric Advantages/Disadvantages
Jaccard Index (JI)	$JI = \frac{A \cap B}{A \cup B}$	Relative Overlap method between two volumes. In this case the JI between each observers contour (A) is taken with the STAPLE contour (B) and an average calculated.	As an overlap metric, is not sensitive enough to large deviations of small volume that may significantly alter beam coverage if the structure was a target volume. Provides no quantitative information on contour variation in terms of size, shape or location.
Dice Coefficient (DICE)	$DICE = \frac{2(A \cap B)}{(A + B)}$	Overlap method, similar to JI. An average is taken of every observers contour (A) with the STAPLE volume (B).	An overlap metric with same issues as JI. This metric places double value to overlap area and may give false interpretations of high agreement.
CI_{pairs} (pairwise analysis)	$CI_{pairs} = \frac{2}{k(k-1)} \sum_{pairs\ ij} \frac{\lvert A_i \cap B_j \rvert}{\lvert A_i \cup B_j \rvert}$	Conformity Index (CI) pairs is an overlap calculated by taking the JI over all possible observers pairs ($A_{1\text{-}i}$) and ($B_{1\text{-}j}$), where k is the number of delineations.	An overlap metric with same issues as JI. This metric does not require a gold standard reference volume to compare to and is performed over all possible contour pairs.
VOI (pairwise analysis)	$VOI = \sum_{pairs\ ij} \frac{2\lvert A_i \cap B_j \rvert}{\lvert A_i \cup B_j \rvert}$	Volume Overlap Index (VOI) is an overlap metric calculated by taking the DICE over all possible observers pairs (A_i) and (B_j).	An overlap metric with same issues as JI. This metric does not require a gold standard reference volume to compare to and is performed over all possible contour pairs. This metric places double value to overlap area and may give false interpretations of high agreement, As such CI_{pairs} is preferred.
Kappa	$Kappa = \frac{(Apparent_{agreement} - Chance_{agreement})}{(1 - Chance_{agreement})}$	In the range of 0.81-1 for almost perfect agreement, 0.61-0.8 substantial agreement, 0.41-0.60 moderate agreement, 0.21-0.4 fair agreement, 0.01-0.20 slight agreement, and 0 is poor agreement.	Is clearly defined what any output means. Will tend to overestimate agreement due to the difference in actual measured data compared to intended use (categorical data). The probability of agreement between observers will be low, thus making this metric high. This metric is also sensitive to the number of observers.
Hausdorff Distance (HD)	$H(A,B) = max(h(A,B), h(B,A))$ where, $h(A,B) = max_{a \in A} min_{b \in B} \lVert a - b \rVert$	Measure of the resemblance of two contours (A and B) to each other. Where A is an observers contour and B the STAPLE contour.	Gives a measure of any large deviations in the structure, which complements overlap metrics. However, this metric does not describe where this deviation is, and is limited to one single value. Average HDs are less sensitive to outliers than maximum HDs.
STAPLE	STAPLE is an expectation-maximization algorithm that computes a probabilistic estimate of the true segmentation and a measure of the performance level represented by each segmentation.	The source of each segmentation is an expert's contour.	Provides a good gold standard contour, but varies in use across system, number of observers, and observer weighting.

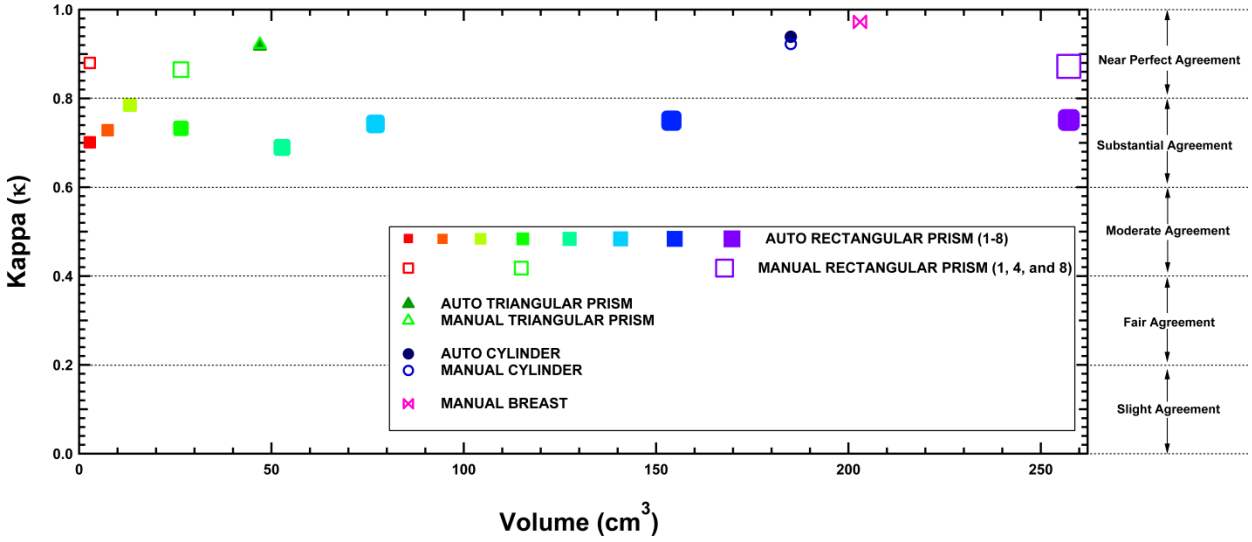
Supplementary Fig.2 1. The eight manually drawn inter-observer contours are displayed for a) the transverse quasar phantoms triangular prism and cylinder, b) the transverse inserted breast contour and squares 1,4 and 8, and c) inserted breast contour, square 8, triangular prism and cylinder on coronal slice.



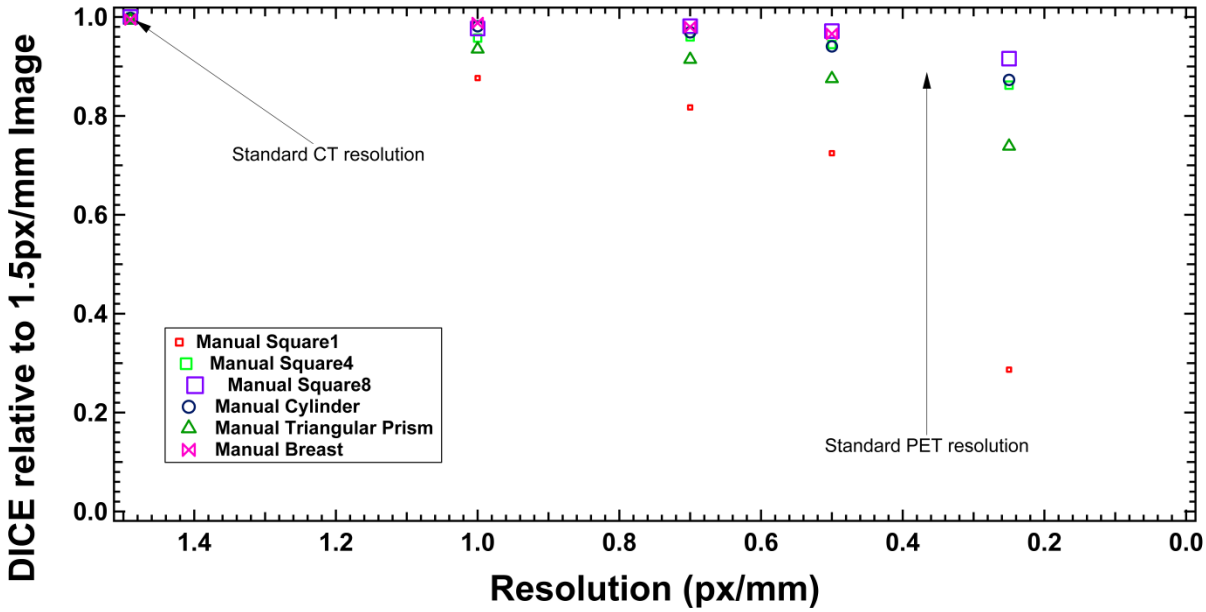
Supplementary Table 2. Manually delineated Inter –observer indices for all 8 observers (±1SD), including average JI, CI_{pairs}, average DICE, VOI, kappa statistics, maximum HD's and average HD's. The intra-observer indices for 5 observers 5 times each in shown in italics for CI_{pairs} and HDs.

Manual Contour	**Breast**	**Triangular prism**	**Cylinder**	**Square 1**	**Square 4**	**Square 8**
Volume (cm^3)	203.3±3.5	46.6±2.0	185.0±6.5	2.8±0.1	26.8±0.6	258.1±1.1
JI (Mean±1SD)	0.975±0.009	0.944±0.019	0.948±0.040	0.973±0.038	0.990±0.021	0.998±0.007
CI_{pairs} (Inter-)	0.961±0.009	0.904±0.026	0.914±0.039	0.901±0.081	0.970±0.019	0.988±0.006
CI_{pairs} (Intra-)	*0.976±0.007*	*0.946±0.021*	*0.965±0.016*	*0.977±0.043*	0.962±0.031	0.993±0.006
DICE (Mean±1SD)	0.987±0.005	0.971±0.010	0.973±0.021	0.986±0.020	0.995±0.011	0.998±0.003
VOI	0.980±0.005	0.950±0.015	0.955±0.021	0.946±0.045	0.985±0.010	0.994±0.003
Kappa (p<0.05)	0.972	0.921	0.923	0.880	0.865	0.872
Sensitivity (Mean±1SD)	0.985±0.010	0.962±0.028	0.972±0.032	0.998 ±0.006	1.000±0.000	0.999±0.000
Specificity (Mean±1SD)	0.996±0.004	0.985±0.014	0.981±0.017	0.910±0.141	0.879±0.240	0.898±0.139
Maximum HD (mm)	3.42 (Inter-)	3.52	4.19	1.37	0.97	0.97
	3.49 (Intra-)	*3.42*	*2.46*	*1.21*	*1.39*	*1.53*
Average HD (mm)	2.77±0.41 (Inter-)	2.49±0.72	2.89±0.65	0.81±0.26	0.74±0.19	0.80±0.14
	2.06±0.38(Intra-)	*1.60±0.45*	*1.74±0.44*	*0.72±0.12*	*0.92±0.22*	*0.83±0.16*

Supplementary Fig. 2. Kappa statistic for all shapes, calculated over all 8 inter-observers.



Supplementary Fig. 3. DICE comparing normal resolution (resolution=1.49 pixels/mm) STAPLE contours to those of lowering resolution.


DICE relative to 1.5px/mm Image
1.0
0.8
0.6
0.4
0.2
0.0
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Resolution (px/mm)
Standard CT resolution
Standard PET resolution
Manual Square1
Manual Square4
Manual Square8
Manual Cylinder
Manual Triangular Prism
Manual Breast