BJR

Received: 17 April 2014 | Revised: 4 August 2014 | Accepted: 27 August 2014

doi: 10.1259/bjr.20140299

Cite this article as:
Langmack KA, Perry C, Sinstead C, Mills J, Saunders D. The utility of atlas-assisted segmentation in the male pelvis is dependent on the interobserver agreement of the structures segmented. Br J Radiol 2014;87:20140299.

FULL PAPER

The utility of atlas-assisted segmentation in the male pelvis is dependent on the interobserver agreement of the structures segmented

[1]K A LANGMACK, DPhil, FIPEM, [2]C PERRY, MB ChB, FRCR, [3]C SINSTEAD, BSc, [2]J MILLS, MRCP, FRCR and [2]D SAUNDERS, MRCP, FRCR

[1]Radiotherapy Physics Department, City Hospital Campus, Nottingham University Hospitals NHS Trust, Nottingham, UK
[2]Oncology Department, Nottingham University Hospitals NHS Trust, Nottingham, UK
[3]Radiotherapy Department, Nottingham University Hospitals NHS Trust, Nottingham, UK

Address correspondence to: Dr Keith A. Langmack
E-mail: *keith.langmack@nuh.nhs.uk*

Objective: To investigate the relationship between the ability of atlas-based autosegmentation to reduce outlining time in the male pelvis (body, bladder, rectum, femoral heads, prostate and seminal vesicles) and the interobserver agreement in the delineation of these structures. To examine any increase of the interobserver agreement with the use of an autosegmentation tool.
Methods: We created atlases in the ABAS™ system v. 2.0 (Elekta, Crawley, UK) and recorded the time to delineate the above structures on eight patients with and without its aid. We also measured the interobserver agreement in the structure definitions using several metrics [Dice's similarity coefficient (DSC), mean distance to conformity, percentage volume difference] with and without the aid of ABAS.
Results: There is a high degree of correlation between the time saving with the use of ABAS and the degree of interobserver agreement ($r = 0.90$ for DSC). This indicates that for structures where the interobserver agreement is low (DSC < 0.65), the ABAS does not reduce outlining time. We found that the interobserver agreement is increased with ABAS only for the prostate.
Conclusion: Outlining time saved in the male pelvis is highly correlated with the interobserver agreement of the structures. Only for the prostate does the use of ABAS significantly reduce the amount of interobserver variation in contouring.
Advances in knowledge: The use of autosegmentation software increases the outlining time for structures where the interobserver agreement is low. Any increase in the interobserver agreement in contouring with the aid of such software may be limited to those structures where there is currently mid-range agreement between observers.

The use of automated segmentation tools has been extensively studied. One advantage of the use of such software is to reduce the time spent contouring. Studies have shown that the time saving varies from around 25–50% of the manual outlining time.[1–5] Only one of these studies documents the time saving per structure, and this indicated that the time saved depends on the structure being outlined.[4] Gambacorta et al[5] postulate that for structures where their autocontouring system was unreliable, the correction time becomes important.

It has also been proposed that the use of autocontouring software improves contouring agreement between observers.[1–3] However, the evidence for this is limited. Young et al[1] state that although they found autosegmentation software useful for contouring the postoperative endometrial nodal volumes, it was not useful for the vaginal cuff owing to its positional variability. This study found only a statistically significant increase in contouring overlap from 0.77 to 0.79 for one out of three pairs of oncologists. Teguh et al[2] found that an expert panel scored 88% of manually edited autocontours as "good" *vs* 83% of clinical contours in the head and neck region. No statistical analysis of this result was given. Chao et al[3] demonstrated an increase in the interobserver contour agreement using their own software. This computer-assisted target volume delineation system (CAT) presented exemplar template contours alongside the computer-delineated contours requiring editing in a slice-by-slice manner, so acted as a combined atlas/autocontouring system. This is unusual, as many autocontouring systems present the user with computer-delineated structures alone, and the manual editing is performed unaided.

The purpose of this study was to measure the time saved by the use of an atlas segmentation software for delineating individual organs in the male pelvis and to investigate if there was a correlation between the time saving and the interobserver outlining agreement. It also aimed to examine the hypothesis that the use of this software significantly increased the interobserver outlining agreement compared with the agreement observed without the aid of the software.

METHODS AND MATERIALS

We used the ABAS™ system v. 2.0 (Elekta, Crawley, UK) to autosegment the male pelvis for prostate cancer planning (body, femurs, bladder, rectum, prostate and seminal vesicles), as it has been found to be an effective tool for autosegmentation.[2,6,7] ABAS performs a segmentation using a previously outlined example (an atlas) and uses both rigid and deformable registration to warp the atlas contours onto the new CT data set.[8] For both the male pelvis and the head and neck region, ABAS in addition uses a modelling approach to refine the contours created by the registration process.[8] Finally, ABAS can combine the results from several atlases to provide a single structure set by the use of the simultaneous truth and evaluation level (STAPLE) algorithm.[8,9]

All the cases used in the atlas preparation and the subsequent study were low-to-intermediate-risk prostate cancer patients (prostate size range, 24.7–94.2 cm^3; median, 42.3 cm^3) with no known nodal involvement.

Imaging, structure definition and atlas preparation

Atlas and study subjects were imaged using our standard prostate outlining protocol that uses both CT and MRI scans. Approximately 20 min prior to CT scanning, patients drank 150 ml of water. They were scanned supine and immobilized using knee rests and ankle stocks (ProStep™ immobilization system; Oncology Imaging Systems, Lewes, UK), from mid-sacroiliac joints to below the lesser trochanters on a Toshiba Aquilion™ LB CT (Toshiba Medical Systems, Crawley, UK). Images of 3-mm thickness were reconstructed at 3-mm intervals and transferred to ProSoma (Oncology Systems Limited, Shrewsbury, UK).

MRI images were acquired using a GE Signa® Excite 1.5-T scanner (GE Healthcare, Crawley, UK) with a Medibord flat couch top (Medibord Ltd, Nottingham, UK) and the ProStep immobilization system in place.[10] Axial T_2 weighted fast-spin echo images of 3-mm thickness were sent to ProSoma. CT and MRI images were fused in ProSoma using the mutual information algorithm and checked for accuracy by an experienced dosimetrist.

As per our clinical practice, a dosimetrist outlined the body, femoral heads (to the level of the lesser trochanter) and bladder. A clinical oncologist delineated the rectum, prostate and seminal vesicles. The rectum was outlined from the anal canal to the rectosigmoid junction. This was performed using the CT/MRI-fused images in ProSoma.

The atlases in the ABAS system were constructed using anonymized CT and structure sets from 11, randomly chosen, previously manually outlined studies. Our commissioning study of ABAS indicated that this was an optimal number of atlases, and we use this number clinically (data not shown). These outlines were produced using the CT/MRI fusion protocol described above; however, ABAS does not use the MRI images for segmentation. Prior to inclusion in the atlas set, all the structures were verified by two consultant clinical oncologists with extensive experience in prostate radiotherapy (JM, DS) to ensure that the atlases were of a high standard and that structures had been correctly delineated.

Outlining studies

For an overview of the study design, see Figure 1. We randomly selected eight previously outlined prostate cancer patients who had not been used for the atlas creation to act as "gold standard" outlines. The outlining was manual, apart from the body contour; this had been automatically outlined in ProSoma using its body autodelineation tool. These cases were then outlined *de novo* by a clinical oncologist (CP) and a dosimetrist (CS) as per the above protocol. Each operator recorded the time taken to outline individual structures. This gave us a reoutlined structure set for comparison with the gold standard set. The comparison of these two structure sets acted as a local baseline for interobserver variability.

The CT series for these subjects were also sent to ABAS to create ABAS outlined structure sets. All 11 atlases were used in the creation of these structures, using the "prostate-specific" algorithm available within the ABAS. The STAPLE algorithm was used to combine the results of using each atlas into a single structure set for each subject. To minimize memory bias a minimum of 2 weeks later, these structure sets were edited in ProSoma by CS (body, femoral heads and bladder) and CP (prostate, rectum and seminal vesicles) to be clinically acceptable. This resulted in edited ABAS structure sets to be compared with the gold standard sets. The editing time per structure was noted to compare with the manual delineation time.

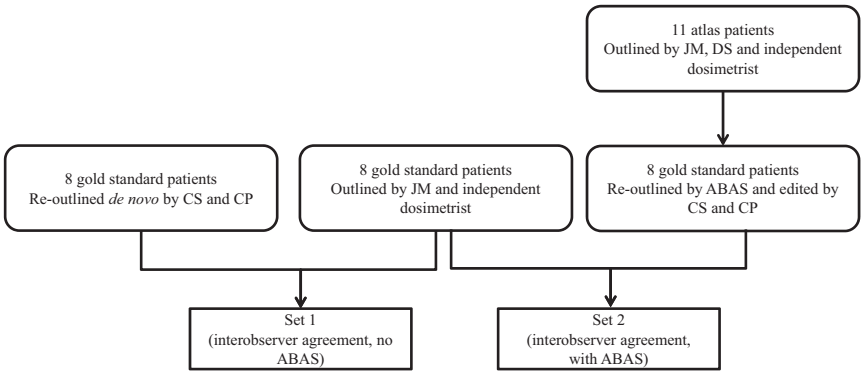
Recent reviews have concluded that there is no one metric type that can be used for totally assessing the agreement between sets of outlines.[11–13] As recommended by these reviews, we used three different types of metrics to evaluate the outlining agreement achieved. The simplest was using a comparison of volume [in terms of percentage volume difference (PVD)]. Secondly, we used a measure of volume overlap, Dice's similarity coefficient (DSC; Appendix A for definition). The final metric was the mean distance to conformity (MDC). This is a measure of the mean distance between points on one contour to the contour being compared.[14] To generate these metrics, the CTs and structure sets were transferred to IMSimQA™ (Oncology Systems Limited).

Data analysis

Timing study

The data were analysed using Excel® (Microsoft 2010, Redmond, WA). Student's *t*-tests were used to compare the mean times to produce a contour *de novo* and by editing the ABAS-created contours. We plotted the ratio of the time taken to edit the ABAS contours to the time taken to outline *de novo vs* logit (DSC) (Appendix A) for the DSC between the gold standard contours and the reoutlined contours to determine any correlation between

Figure 1. Overview of study design. ABAS™ system v. 2.0 (Elekta, Crawley, UK).



contouring agreement and the time saved. We repeated this for DSC, PVD and MDC.

Interobserver variation

We examined any change in the agreement between the gold standard contours and those contours outlined *de novo*, or those segmented with the aid of ABAS, using several measures. For these, two sets of agreement metrics were calculated. Set 1 was the agreement between the gold standard contours and the *de novo* contours. These were our local baseline agreements. Set 2 measured the agreement between the gold standard outlines and the edited ABAS-generated outlines (Figure 1). Student's *t*-test was used to test for differences between Sets 1 and 2.

Also, for DSC and MDC, these metrics were split into two groups. For DSC, two cut-off levels were used: 0.7 and 0.8 (Appendix A). For MDC, a 2-mm cut-off was taken as being "an acceptable threshold for variation between expert observers".[14,15] We tested for differences between this now categorical data representation of Sets 1 and 2 by the use of the χ^2 test.

RESULTS

Timing studies

The mean time for ABAS to outline a subject, using all 11 atlases and STAPLE, was 15.75 ± 0.40 min. However, as this was performed offline, it has not been included in our timing data. The mean time to outline each subject *de novo* was 26.43 ± 2.80 min. The mean time to edit the ABAS-outlined subjects to produce clinically acceptable structures was 15.70 ± 2.25 min. This decrease in mean outlining time by 10.73 min per subject was highly statically significant ($p = 0.0001$, Student's *t*-test). For each structure, a statistically significant reduction in time to edit the ABAS-outlined structures was seen in comparison with outlining *de novo* except for the body, rectum and seminal vesicles (Table 1).

Figure 2 shows the correlation between the time saved by the use of ABAS (expressed as the ratio of the times taken to edit ABAS-generated structures and the time taken to contour a structure *de novo*) *vs* the interobserver agreement as measured by logit (DSC) between the gold standard contours and the *de novo* contours. The linear fit shows that the ratio is <1 (*i.e.* ABAS is time saving) if DSC is ≥0.65. Similarly, the linear fit to the plot using DSC as the metric gives DSC = 0.63 as the breakeven point (data not shown). The correlation coefficients for all metrics are given in Table 2. This demonstrates that the use of the atlas software is not time saving for structures where the interobserver agreement is low (*e.g.* the seminal vesicles).

Contour comparison studies

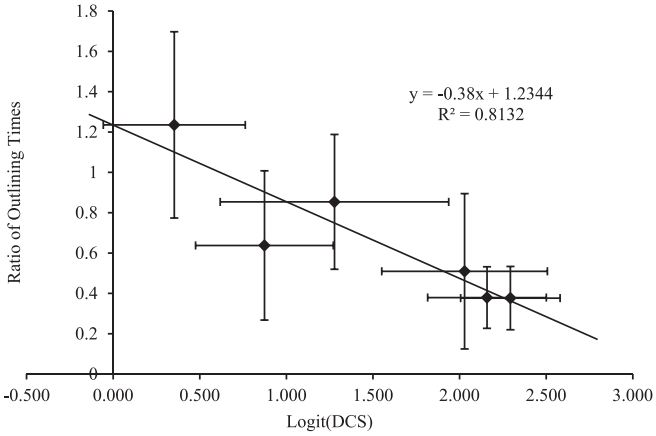
Table 3 shows the agreement metrics between the gold standard contours and the *de novo* contours (Set 1). This table is a measure of the interobserver agreement in outlining without the aid of ABAS and is out local baseline for contouring agreement. The

Table 1. Timing results showing the mean times to outline *de novo* or edit ABAS™-created specific structures

Structure	Time to outline *de novo* (min)	Time to edit ABAS outlines (min)	Ratio (*de novo*/edited)	*p*-value
Body	0.28 (0.06)	0.13 (0.35)	0.44	0.2726
Left femoral head	5.60 (0.78)	2.13 (0.80)	0.38	0.0002
Right femoral head	5.70 (1.27)	2.15 (0.75)	0.38	0.0008
Bladder	3.80 (2.26)	1.94 (0.90)	0.51	0.0312
Rectum	4.20 (0.78)	3.59 (1.24)	0.85	0.3640
Prostate	4.46 (0.66)	2.84 (1.59)	0.64	0.0234
Seminal vesicles	2.38 (0.71)	2.94 (0.66)	1.24	0.2317

Values are mean (standard deviation). *p*-values are calculated using paired Student's *t*-test.
ABAS system v. 2.0 (Elekta, Crawley, UK).

Figure 2. The variation of the time saved by the use of ABAS™ *vs* the interoperator variability as measured by logit [Dice's similarity coefficient (DSC)]. ABAS system v. 2.0 (Elekta, Crawley, UK).



metrics measuring the agreement between the gold standard contours and those obtained by editing ABAS-generated contours is also shown in Table 3 (Set 2). This is a measure of the interobserver agreement with the aid of ABAS. If these measures are improved over the baseline values, then the use of ABAS has increased the interobserver agreement in outlining. The differences reached statistical significance for only the prostate (DSC, $p = 0.014$; logit DSC, $p = 0.014$; and MDC, $p = 0.008$) using the Student's *t*-test. Similarly, the χ^2 test indicated an improvement for only prostate outlines using DSC ≥ 0.8 (1/8 *vs* 4/8; $p = 0.001$) and MDC ≤ 2 mm (0/8 *vs* 5/8; $p = 0.0003$). No difference in agreement metrics reached statistical significance for any other structure. So, we found that the use of ABAS improves the interobserver agreement in contouring for only the prostate.

DISCUSSION

We have found that editing structures created by autosegmentation software can significantly save time over a totally manual process in agreement with other studies.[1,4,5,16] Two of these[4,16] address the male pelvis. Lin et al[4] gives the timings for individual structures (femurs, prostate, bladder and rectum) and found broadly similar time savings—the rectum was the structure that showed the least time saving in their study. We found statistically significant reductions in delineation times for all but three structures—the body, rectum and seminal vesicles. For the body, the standard outlining procedure already involved using the Hounsfield number-based autosegmentation algorithm in ProSoma, so the *de novo* time was in fact the time to edit the ProSoma generated contours. With ABAS, seven of the eight examples used did not require their body contours editing. However, the body contour produced by ABAS for the eighth subject required extensive editing in places, so increasing the mean editing time for ABAS-assisted contouring for this structure. For the rectum, it was found that the ABAS-based contours did not always agree well at the rectosigmoid junction or the rectum–anal canal junction. This may be because these limits are not defined by easily discernible tissue boundaries. These differences are reflected by the time saving in delineating the rectum with the aid of ABAS not reaching statistical significance. The agreement for the seminal vesicles for all metrics is especially poor. This is likely to be a reflection of both the difficulty in visualizing them and determining which part of them to outline.[17,18] This resulted in the editing times for the automatically delineated seminal vesicles being greater than the *de novo* outlining times. Additional work is required to improve our outlining protocol for this organ.

Table 2 indicates a strong correlation between the interobserver agreement and the time saving for individual structures for both logit (DSC) and DSC. The other metrics did not show as strong a correlation with time saved. We are not aware of any other studies that have investigated the relationship between interobserver agreement and time saving. This study provides the first quantitative evidence for the hypothesis that more accurate autocontouring would increase the time saving and, for structures where the interobserver agreement is low, it is of no use.[1,5] Our data indicate that if the average DSC is <0.65 between observers for a structure, then the time to edit the autocontours is likely to exceed the time to delineate them manually. We found this for the seminal vesicles, which are accepted to have a relatively low interobserver agreement.[17,18] We postulate that DSC may be used predictively to determine if the use of an autocontouring software will decrease outlining time. However, this result should be validated for other body sites and other software tools.

The second aim of this study was to investigate if the use of an atlas-outlining tool increased the interobserver agreement in outlining. Table 3 gives several measures of the interobserver agreement with and without the aid of autocontouring. These data are in broad agreement with other studies in this area.[5,16,17] The statistical analysis of the various agreement metrics used in this study does not indicate any significant change in the interobserver agreement with and without the use of ABAS apart from the prostate. For this structure, interobserver agreement was

Table 2. Correlation between the time saved in contouring with the use of ABAS™ and the various metrics of agreement

Metric	Correlation coefficient
Logit (DSC)	0.90
DSC	0.89
Percentage volume difference	0.79
Mean difference to conformity (mm)	0.36

DSC, Dice's similarity coefficient.
ABAS system v. 2.0 (Elekta, Crawley, UK).

Table 3. Interobserver agreement showing the comparison between the original gold standard contours *vs* those created *de novo* (Set 1) and the comparison between the original contours *vs* those created by editing ABAS™-created contours (Set 2)

Structure	Set	DSC	Logit DSC	Mean difference to conformity (mm)	Percentage volume difference
Bladder	1	0.88 (0.05)	2.029 (0.478)	2.2 (0.3)	9.2 (6.7)
	2	0.92 (0.06)	2.724 (1.032)	2.1 (0.6)	3.6 (3.7)
Body	1	0.99 (0.01)	4.499 (0.769)	2.6 (2.8)	1.9 (1.1)
	2	0.99 (0.01)	5.239 (0.907)	1.4 (0.4)	0.9 (1.2)
Left femoral head	1	0.89 (0.03)	2.158 (0.343)	3.0 (1.0)	7.4 (6.9)
	2	0.88 (0.05)	2.104 (0.471)	4.9 (3.4)	10.0 (3.6)
Right femoral head	1	0.91 (0.02)	2.293 (0.287)	2.6 (0.8)	5.9 (4.3)
	2	0.90 (0.07)	2.407 (0.878)	4.5 (3.3)	7.2 (5.4)
Prostate	1	0.70 (0.08)	0.873 (0.398)	3.1 (0.7)	28.3 (27.1)
	2	0.83 (0.09)	1.689 (0.639)	2.1 (0.5)	20.9 (14.0)
Rectum	1	0.76 (0.11)	1.278 (0.660)	4.8 (2.4)	18.3 (18.6)
	2	0.79 (0.10)	1.403 (0.597)	5.2 (2.0)	14.1 (11.7)
Seminal vesicles	1	0.58 (0.10)	0.353 (0.410)	3.4 (0.9)	27.7 (16.1)
	2	0.57 (0.13)	0.310 (0.533)	3.5 (0.8)	23.7 (15.8)

DSC, Dice's similarity coefficient.
Values are mean (standard deviation).
ABAS system v. 2.0 (Elekta, Crawley, UK).

significantly improved on three metrics (DSC, logit DSC and MDC) with the use of ABAS using paired Student's *t*-test. The χ^2 test indicates that the ABAS-aided contouring of the prostate shows greater agreement than the manual contours on two measures, DSC is $\geq$0.8 and MDC is $\leq$2 mm. Examining the outlines of the prostate on a slice-by-slice basis showed that the improved consistency of the outlining was throughout the extent of the organ.

These observations lead us to propose that the degree of interobserver agreement and the ability of the use of auto-segmentation to increase interobserver agreement are linked. For organs with already high levels of agreement (*e.g.* DSC is >0.8), a statistically significant increase in agreement may not be achievable or will require a much larger sample size to detect. For structures with low agreement levels between observers, the manual editing of the autocontours could tend to diminish any increase in outlining consistency resulting from the use of the autocontouring software. Again a much larger study will be required to find any difference. However, the use of a tool like CAT,[3] where exemplar contours are presented alongside the contours requiring editing, is likely to mitigate this effect. For tools similar to ABAS, where manual editing is required without the aid of exemplar contours, increasing the interobserver agreement may only be for structure with mid-range interobserver agreement. This hypothesis requires further studies to confirm it.

CONCLUSIONS

Our timing studies indicate that the use of ABAS can significantly reduce the time taken to outline structures within the male pelvis. The time saved is highly correlated with some measures of interobserver agreement in organ delineation [logit (DSC) and DSC]. If DSC is <0.65 between observers for a structure, then the use of an autocontouring software may not significantly decrease the delineation time for that structure. Only for the prostate did the use of ABAS significantly increase the agreement of the contours with the gold standard in comparison with those produced without its use. We have found indications that the use of autocontouring software may only increase interobserver agreement for those structures where the level of agreement is mid-range before its introduction.

FUNDING

The work was carried out within the National Health Service as a service development.

REFERENCES

1. Young AV, Wortham A, Wernick I, Evans A, Ennis RD. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. *Int J Radiat Oncol Biol Phys* 2011; **79**: 943–7. doi: 10.1016/j.ijrobp.2010.04.063
2. Teguh DN, Levendag PC, Voet PW, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue

(swallowing/mastication) structures in the head and neck. *Int J Radiat Oncol Biol Phys* 2011; **81**: 950–7. doi: 10.1016/j.ijrobp.2010.07.009
3. Chao KS, Bhide S, Chen H, Asper J, Bush S, Franklin G, et al. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. *Int J Radiat Oncol Biol Phys* 2007; **68**: 1512–21.
4. Lin A, Kubicek G, Piper JW, Nelson AS, Dicker AP, Valicenti RK. Atlas-based segmentation in prostate IMRT: timesavings in the clinical workflow. *Int J Radiat Oncol Biol Phys* 2008; **72**: S328–9.
5. Gambacorta MA, Valentini C, Dinapoli N, Boldrini L, Caria N, Barba MC, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. *Acta Oncol* 2013; **52**: 1676–81. doi: 10.3109/0284186X.2012.754989
6. Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. *Radiother Oncol* 2012; **102**: 68–73. doi: 10.1016/j.radonc.2011.08.043
7. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol* 2012; **7**: 160. doi: 10.1186/1748-717X-7-160
8. *Atlas-based auto-segmentation technical reference*. Document ID: LRMAAS0001. Crawly, UK: Elekta; 2010.
9. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–21.
10. Langmack KA. The use of an advanced composite material as an alternative to carbon fibre in radiotherapy. *Radiography* 2012; **18**: 74–7.
11. 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* 2010; **54**: 401–10. doi: 10.1111/j.1754-9485.2010.02192.x
12. Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. *Clin Oncol (R Coll Radiol)* 2010; **22**: 515–25. doi: 10.1016/j.clon.2010.05.006
13. 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. doi: 10.1007/s00066-011-0027-6
14. Jena R, Kirby NF, Burton KE, Hoole AC, Tan LT, Burnet NG. A novel algorithm for the morphometric assessment of radiotherapy treatment planning volumes. *Br J Radiol* 2010; **83**: 44–51. doi: 10.1259/bjr/27674581
15. Burton K, Jefferies S, Jena R, Estall V, Burnet N. Inter and intra observer variation in the gross tumour volume (GTV) delineation for glioblastoma (GBM). *Radiother Oncol* 2008; **88**: S27.
16. Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, et al. Technology assessment of automated atlas based segmentation in prostate bed contouring. *Radiat Oncol* 2011; **6**: 110. doi: 10.1186/1748-717X-6-110
17. Jeanneret-Sozzi W, Moeckli R, Valley JF, Zouhair A, Ozsahin EM, Mirimanoff RO; on behalf of SASRO. The reasons for discrepancies in target volume delineation: a SASRO study on head-and-neck and prostate cancers. *Strahlenther Onkol* 2006; **182**: 450–7.
18. Valicenti RK, Sweet JW, Hauck WW, Hudes RS, Lee T, Dicker AP, et al. Variation of clinical target volume definition in three dimensional conformal radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 1999; **44**: 931–5.
19. Zou KK, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. *Acad Radiol* 2004; **11**: 178–89.
20. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging* 1994; **13**: 716–24.
21. Fleiss JL. *Statistical methods for rates and proportions*. 2nd edn. New York, NY: John Wiley; 1981.
22. Mattiucci GC, Boldrini L, Chiloiro G, D'Agostino GR, Chiesa S, De Rose F, et al. Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? *Acta Oncol* 2013; **52**: 1417–22. doi: 10.3109/0284186X.2013.813069

APPENDIX A. DICE'S SIMILARITY COEFFICIENT

The IMSimQA™ (Oncology Systems Limited, Shrewsbury, UK) tool calculates conformity index (CI)[14] (also known as the Jaccard index[13]) rather than Dice's similarity coefficient (DSC). We computed the DSC using equation (5) of Fotina et al[13] [Equation (A1) below], as this metric has been extensively reported in other studies, and is therefore a useful comparator.

$$\mathrm{DSC}=\frac{2\mathrm{CI}}{1+\mathrm{CI}} \qquad \text{(A1)}$$

It is an open question, requiring further research, as to the value of DSC that is indicative of good interobserver agreement. Many authors have used a value of 0.7 as a threshold.[5,6,19,20] Zijdenbos et al[20] have shown that DSC is a special instance of the κ statistic, and if DSC is >0.70, then κ is >0.75—a value where κ is accepted to indicate an excellent agreement between observers.[21] However, Fotina et al[13] consider DSC "can provide a false impression of high agreement", and Mattiucci et al[22] suggest DSC >0.8 to be acceptable. We examined two levels of DSC (0.7 and 0.8) as a cut-off in our increase in interobserver agreement study to see if it changed our analysis. It had a limited effect with the stricter limit showing statistical significance in the χ^2 test.

The very limited range of DSC means the assumption that this variable is normally distributed is less likely. For this study, we also included the logit (DSC), as this is more likely to be normally distributed than DSC itself.[20] The advantage of the logit transform is that it expands the range of DSC from (0, 1) to $(-\infty, \infty)$. The formula for this transformation is given by:

$$\mathrm{logit}(X)=\ln\left(\frac{x}{1-x}\right) \qquad \text{(A2)}$$