Published in final edited form as:
Acta Oncol. 2015 April ; 54(4): 557–560. doi:10.3109/0284186X.2014.953638.

Auto-segmentation of the brachial plexus assessed with TaCTICS - A software platform for rapid multiple-metric quantitative evaluation of contours

Musaddiq Awan[1], **Brandon Alan Dyer**[2], **Jayashree Kalpathy-Cramer**[2,3], **Eva Bongers**[4], **Max Dahele**[4], **Jinzhong Yang**[1], **Gary V. Walker**[1], **Nikhil G. Thaker**[1], **Emma Holliday**[1], **Andrew J. Bishop**[1], **Charles R. Thomas Jr.**[2], **David I. Rosenthal**[1], and **Clifton David Fuller**[1,2]

[1]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[2]Department of Radiation Medicine, Oregon Health & Science University, Portland, OR, USA

[3]Athinoula A. Martinos Center for Biomedical Imaging; Department of Radiology, Harvard Medical School; Department of Radiology and Neuroscience, Massachusetts General Hospital, Charlestown, MA, USA

[4]Department of Radiation Oncology, VU University Medical Center, Amsterdam, NL

Introduction

Segmentation of organs-at-risk (OARs) remains a highly variable yet critical operator-dependent step in radiation planning [1]. With the increased conformality of intensity-modulated radiotherapy (IMRT) delivery, the ability to spare OARs is markedly increased, enabling more targeted treatment with sparing of specific tissues. However, manual segmentation of target volumes and OARs remains highly variable. For this reason, auto-segmentation approaches are attractive mechanisms to potentially reduce inter-observer region of interest (ROI) variation [2,3], allow assessment of OARs that might otherwise be subject to beam path toxicities [3,4] and improve workflow-time parameters [4-6].

Corresponding Author: Dr. Clifton David Fuller, MD, PhD, Head and Neck Section, Division of Radiation Oncology, Dept. of Radiation Oncology, University of Texas Health MD Anderson Cancer Center, 1515 Holcombe Blvd, Unit 97, Houston, TX 77030, cdfuller@mdanderson.org, (P) 713-563-2334, (F) 713-746-8273.

Disclosures: Department of Radiation Oncology VU University Medical Center has research agreements with Varian Medical Systems and Brainlab. MD has received travel support honoraria from Varian and Brainlab. CDF received/receives funding support from the SWOG-Hope Foundation Dr. Charles A. Coltman, Jr. Fellowship in Clinical Trials, the National Institutes of Health Paul Calabresi Clinical Trial Program (K12 CA088084-14) and Clinician Scientist Loan Repayment Program (L30 CA136381-02), an MD Anderson/Elekta AB MR-LinAc Seed Grant, the MD Anderson Center for Advanced Biomedical Imaging/ General Electric Healthcare In-Kind support, the Center for Radiation Oncology Research at MD Anderson Cancer Center Seed Grant, and the MD Anderson Institutional Research Grant Program. JKC received funding support from a National Institute of Health/National Library of Medicine grant (4R00LM009889). TaCTICS software was developed by JKC and CDF with funding from the Society of Imaging Informatics in Medicine (SIIM) Product Development Grant. The remaining authors have no conflicts of interest to declare. These funders played no role in the study design, collection, analysis, interpretation of data, manuscript writing, or decision to submit the report for publication. Acknowledgements: The authors wish to gratefully thank Suresh Senan, PhD, Niels Haasbeek, M.D. and Joost Bot, MD for their efforts and participation in the aforementioned study. Portions of this data were selected for a presentation at the 14th World Conference on Lung Cancer (WCLC), July 2011 in Amsterdam, The Netherlands.

Auto-segmentation techniques have been developed that implement *a priori* atlas libraries of normal tissue ROIs, with deformable image registration to transfer these ROIs from the reference library to a patient DICOM file [7]. While several commercial and in-house auto-segmentation approaches have been presented and show promise, rigorous quality assessment should be performed before clinical implementation [1,6] given the clinical implications of over-or under-contouring [8].

However, individual institutions may have significant difficulty systematically evaluating competing auto-segmentation platforms, as evaluation of registration and segmentation typically requires substantial effort for multi-ROI segmentation assessment [9,10]. Consequently, we surmised that there exists an unmet need for an open-source, web-based software solution for comparison of auto-segmented ROIs with reference manually segmented ROIs. We have previously reported the development of an open-source web-based software called TaCTICS (*Ta*rget *C*ontour *T*esting/*I*nstructional *C*omputer *S*oftware, https://github.com/kalpathy/tacticsRT) that provides quantitative and qualitative comparison of submitted and reference manually segmented ROIs in order to provide feedback to users about their performance on contouring target volumes and OARs [11,12]. For this reason we sought to investigate the feasibility and utility of TaCTICS in evaluating the quality of auto-segmentation algorithms by comparing their results to composite expert contours using two brachial plexus ROIs as index OARs. The specific aims of the current study were to assess the feasibility of utilizing TaCTICS to report multi-metric analysis of an auto-segmentation algorithm of the brachial plexus relative to a TaCTICS-generated probabilistic multi-expert manual segmentation, define a performance benchmark comparison of an auto-segmentation algorithm of the brachial plexus to that of a set of reference resident contours and finally, to establish a quality-assessment workflow for the future evaluation of commercial/in-house auto-segmentation algorithm performance.

Materials and Methods

Institutional Review Board approval was obtained, allowing collection of anonymized DICOM files. Clinical datasets were anonymized and stripped of all identifiers, and fictionalized case histories were constructed for all resultant efforts detailed herein.

Five radiation oncology trainees (each with less than 2 years of residency training) and four expert radiation oncology attending physicians were asked to contour right-sided brachial plexuses on a *head and neck* case (patient simulated arms down) and on a *chest* case (patient simulated arms up) with the ability to reference an existing contouring atlas [13]. DICOM files were then auto-segmented using a previously described in-house intensity-based accelerated “DEMONS” deformable registration/auto-segmentation algorithm [14] to derive brachial plexus contours both of *head and neck* and *chest* case ROIs.

The RT Structure sets for both cases for all five trainees, four experts and the auto-segmentation mechanism were imported into TaCTICS. Using the TaCTICS software a composite Warfield’s Simultaneous Truth And Performance Level Estimation (STAPLE) of the four expert contours was developed and was used as a “gold-standard” for comparison [15,16]. A number of existing literature-derived [17,18] metrics comparing the residents/

auto-segmented contours to the reference composite STAPLE were calculated using the TaCTICS software. A brief description and list of these metrics is found in Supplemental Materials, Table 1.

After tabulation, each metric was calculated for all residents for each case and compared to the calculated metrics for the auto-segmented contours using the non-parametric one-tailed Wilcoxon Signed-Rank test, with $p = 0.05$ considered statistically significant. Non-parametric analysis was selected owing to the obviously limited sample size.

Results and Discussion

In both the *head and neck case* and the *chest* case the auto-segmentation algorithm was found to have both lower False Negative Dice (0.34 and 0.31, respectively) and higher target overlap (0.49 and 0.49, respectively), implying it missed fewer gold-standard voxels than the average trainee (0.47 and 0.61, respectively; 0.36 and 0.30, respectively). However, the auto-segmentation algorithm had a higher overall volumetric difference for both the *chest* case and the *head and neck* case (1.03 and 1.31, respectively versus 0.72 and 0.38), implying that for both cases the auto-segmented contours volumes were significantly more disparate from the STAPLE than the trainee contours. Interestingly, neither the 95% Hausdorff distance nor the False Positive Dice were significantly different from the trainee contours. This implies that though there was a volumetric difference between resident and auto-segmented ROIs, the auto-segmentation algorithm did not seem to significantly over-contour (FPD), nor were contoured ROI surfaces on average farther away from the expert composite ROI surface than ROIs of trainees (HD, Table 2). Importantly, both the Dice and Jaccard coefficients in both cases were not significantly different from the trainee contours (Table 2). This combined analysis seems to imply that the auto-segmentation algorithm as implemented at our institution performs at least comparably if not superior to that of junior radiation oncology trainees. However, the discordance between resident trainees, the tested algorithm, and attending physicians was striking, with both autosegmentation and resident ROIs far inferior to pre-determined thresholds of acceptability.

Admittedly, there was also large variability between experts within our study and thus raising the important question of what can be used as a “gold-standard truth.” In particular, the Dice coefficients for the “experts” for both cases against the multi-expert STAPLE were between 0.23-0.27 for the chest case and between 0.25-0.52 for the head and neck case. This points to a larger issue: whether OAR delineation remains only an issue for novice trainees. Our data, and that of cooperative group analyses [19], suggest otherwise. It is critical that target and OAR delineation is not seen as solely an issue for the inexperienced clinician. Creating standardized agreements between “experts” is essential for the next era of radiation treatment planning quality improvement efforts, particularly if auto-segmentation algorithms are to be assessed for efficacy [20]. Already auto-segmentation or semi-automated segmentation assessment solutions are likely to become a part of radiotherapy clinical trial efforts sooner rather than later [21]. A flexible, robust software solution, capable of both manual and auto-segmentation assessment might also have applicability for both “fixed-location” [22] and “remote” [11,12] web-based training solutions that are likely to become increasingly important as the availability of new technologies increases. Having a no-cost open-source solution, as presented herein, also opens the possibility of end-users adding desired metrics [9,17,18] on a clinical trial or training needs-based situation.

Integrating auto-segmentation algorithms of OARs into a stable clinical workflow is often hindered by the uncertainty of the efficacy of such algorithms relative to institutional expert manual segmentation [3,6]. We have presented and demonstrated the feasibility of utilizing TaCTICS, an open-source web-based system, for the utility of such analyses. By uploading DICOM RT structures into the TaCTICS system, one can readily obtain the aforementioned metrics within a matter of minutes. Performing a similar analysis as described illuminates whether such an algorithm meets the end user's standards for integration into an existing workflow. Unfortunately, standards that universally define adequacy of contours are of crucial importance. Of the seven metrics examined, we specifically highlight the utility of the False Negative Dice coefficient in this particular scenario, as it places a particularly high cost on missing gold-standard voxels, spotlighting inadequate auto-segmentation of organs-at-risk.

It is also important that auto-segmentation algorithms be tested in multiple clinical scenarios (e.g. distinct treatment positions as we have presented) to establish the efficacy of such algorithms across multiple workflows. Ideally, use of such a quality assessment process can be combined with rigorous assessment of other treatment planning quality assurance practices (e.g. rigorous deformable image registration benchmarking [23]) to provide a quantifiable assessment of the potential gains from implementation. In the absence of readily available and user-friendly platforms, only large academic centers are likely to have the necessary physics and computer science infrastructure to perform independent analysis of commercial or open-source auto-segmentation solutions.

In our estimation, the presented data suggest that the tested auto-segmentation algorithm performs at a level comparable to a resident trainee brachial plexus segmentation. At our institution, this would be an acceptable standard in scenarios where brachial plexus doses are far below thresholds associated with toxicity (e.g. if the low neck is treated to <60 Gy). However, if brachial plexus doses approach meaningful dose constraints, we do not advocate use of unevaluated auto-segmented structures. As the Dice coefficients for both, tested residents and the auto-segmentation platform, fell below what we consider acceptable Dice and False Negative Dice thresholds, we continue to recommend attending approval of resident and DEMONS-derived ROIs. However, auto-segmentation could be routinely used to "pre-contour" brachial plexus volumes for subsequent modification, especially in scenarios where a resident is not present; based on our data the utilized algorithm would be potentially useful in such a time saving application.

Our hope is that, as individual institutions/users see other unmet needs in the TaCTICS software, user-developed software updates or modifications may be readily incorporated (e.g. MAP STAPLE [24,25]). Future efforts will focus on expansion of evaluated auto-segmentation solutions as our process has demonstrated feasibility within an established workflow.

In conclusion, our data suggest that TaCTICS is a feasible platform for auto-segmentation assessment, and further, that the tested DEMONS algorithm can segment brachial plexus ROIs to a degree better or comparable to resident trainees. However, based on low concordance compared to, and between, reference attendings we strongly recommend individual expert physician confirmation of segmentation for both resident trainees and autosegmentation algorithms when dose constraint to the brachial plexus is of clinical importance. Additionally, we recommend that, before implementation, site-specific OAR auto-segmentation quality assurance be performed against institutional expert ROI benchmarks with a method such as TaCTICS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

[1]. Mukesh M, Benson R, Jena R, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? Br J Radiol. Aug; 2012 85(1016):e530–536. [PubMed: 22815423]

[2]. Hardcastle N, Tome WA, Cannon DM, et al. A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy. Radiat Oncol. 2012; 7:90. [PubMed: 22704464]

[3]. Mattiucci GC, Boldrini L, Giuditta C, D'Agostino GR, Chiesa S, et al. Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark? Acta Oncol. 2013; 52(7):1417–1422. [PubMed: 23957565]

[4]. Rosenthal DI, Chambers MS, Fuller CD, et al. Beam path toxicities to non-target structures during intensity-modulated radiation therapy for head and neck cancer. International journal of radiation oncology, biology, physics. Nov 1; 2008 72(3):747–755.

[5]. La Macchia M, Fellin F, Amichetti M, 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. Sep 18.2012 7(1):160. [PubMed: 22989046]

[6]. Gambacorta MA, Valentini C, Dinapoli N, Boldrini L, Caria N, 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(8):1676–1681. [PubMed: 23336255]

[7]. Chaney EL, Pizer SM. Autosegmentation of images in radiation oncology. J Am Coll Radiol. Jun; 2009 6(6):455–458. [PubMed: 19467494]

[8]. Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. Mar; 2011 98(3):373–377. [PubMed: 21269714]

[9]. Yang J, Wei C, Zhang L, Zhang Y, Blum RS, Dong L. A statistical modeling approach for evaluating auto-segmentation methods for image-guided radiotherapy. Comput Med Imaging Graph. Sep; 2012 36(6):492–500. [PubMed: 22673541]

[10]. Chen A, Niermann KJ, Deeley MA, Dawant BM. Evaluation of multiple-atlasbased strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. Physics in medicine and biology. Jan 7; 2012 57(1):93–111. [PubMed: 22126838]

[11]. Kalpathy-Cramer J, Bedrick SD, Boccia K, Fuller CD. A pilot prospective feasibility study of organ-at-risk definition using Target Contour Testing/Instructional Computer Software (TaCTICS), a training and evaluation platform for radiotherapy target delineation. AMIA Annu Symp Proc. 2011; 2011:654–663. [PubMed: 22195121]

[12]. Kalpathy-Cramer J, Fuller CD. Target Contour Testing/Instructional Computer Software (TaCTICS): A Novel Training and Evaluation Platform for Radiotherapy Target Delineation. AMIA Annu Symp Proc. 2010; 2010:361–365. [PubMed: 21347001]

[13]. Kong FM, Ritter T, Quint DJ, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. International journal of radiation oncology, biology, physics. Dec 1; 2011 81(5):1442–1457.

[14]. Wang H, Dong L, Lii MF, et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. International journal of radiation oncology, biology, physics. Mar 1; 2005 61(3):725–735.

[15]. Commowick O, Warfield SK, Malandain G. Using Frankenstein's creature paradigm to build a patient specific atlas. Med Image Comput Comput Assist Interv. 2009; 12:993–1000. Pt 2. [PubMed: 20426208]

[16]. Warfield SK, Zou KH, Wells WM. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. IEEE transactions on medical imaging. Jul; 2004 23(7):903–921. [PubMed: 15250643]

[17]. 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). Sep; 2010 22(7):515–525. [PubMed: 20554168]

[18]. Fotina I, Lutgendorf-Caucig C, Stock M, Potter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol. Feb; 2012 188(2):160–167. [PubMed: 22281878]

[19]. Peters LJ, O'Sullivan B, Giralt J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. J Clin Oncol. Jun 20; 2010 28(18):2996–3001. [PubMed: 20479390]

[20]. Martin S, Rodrigues G, Patil N, et al. A Multiphase Validation of Atlas-Based Automatic and Semiautomatic Segmentation Strategies for Prostate MRI. International journal of radiation oncology, biology, physics. May 8.2012

[21]. Gwynne S, Spezi E, Wills L, et al. Toward Semi-automated Assessment of Target Volume Delineation in Radiotherapy Trials: The SCOPE 1 Pretrial Test Case. International journal of radiation oncology, biology, physics. Aug 6.2012

[22]. Nijkamp J, de Haas-Kock DF, Beukema JC, et al. Target volume delineation variation in radiotherapy for early stage rectal cancer in the Netherlands. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology. Jan; 2012 102(1):14–21. [PubMed: 21903287]

[23]. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Physics in medicine and biology. Apr 7; 2009 54(7):1849–1870. [PubMed: 19265208]

[24]. Commowick O, Akhondi-Asl A, Warfield SK. Estimating A Reference Standard Segmentation With Spatially Varying Performance Parameters: Local MAP STAPLE. IEEE transactions on medical imaging. Aug; 2012 31(8):1593–1606. [PubMed: 22562727]

[25]. Commowick O, Warfield SK. Incorporating priors on expert performance parameters for segmentation validation and label fusion: a maximum a posteriori STAPLE. Med Image Comput Comput Assist Interv. 2010; 13:25–32. Pt 3. [PubMed: 20879379]

Table 1

Calculated Metrics Derived From TaCTICS

	Symbolic Expression	Represents	Perfect Segmentation	Complete Discordance
Target Overlap	$TO = \frac{\mid A \cap G \mid}{\mid G \mid}$	The portion of the gold-standard which is overlapped by the segmentation	1	0
Union Overlap (Jaccard)	$J = 2\frac{\mid A \cap G \mid}{\mid A \cap G \mid}$	The portion of overlap between the gold-standard and segmentation relative to the size of the union of the gold-standard and segmentation	1	0
Mean Overlap (Dice)	$D = \frac{2 \mid A \cap G \mid}{\mid A \mid + \mid G \mid}$	The portion of overlap between the gold-standard and segmentation relative to the size of the gold-standard plus the size of the segmentation	1	0
Volumetric Difference	$VD = \frac{V_a \cap V_g}{V_g}$	The difference in volume between the segmentation and the gold-standard as a portion of the volume of the gold standard	0	Varies (> 0)
False Negative Dice	$FND = \frac{2 \mid \bar{A} \cap G \mid}{\mid A \mid + \mid G \mid}$	The volume that the segmentation missed of the gold-standard relative to the size of the gold-standard plus the size of the segmentation	0	Varies (> 0)
False Positive Dice	$FPD = \frac{2 \mid A \cap \bar{G} \mid}{\mid A \mid + \mid G \mid}$	The volume of the segmentation not found within the gold-standard relative to the size of the gold-standard plus the size of the segmentation	0	Varies (> 0)
95% Hausdorff Distance (HD)	$H(A, G) = \max(h(A, G), h(G, A)) \cap$ $h(A, G) = \max_{a \in A}(d(A, G)) \& d(A, G) = \min_{g \in g} \|\ a \cap g \|$	The maximum distance between a point in the segmentation and that of the gold-standard if 5% of the outliers are thrown out	0	Large

Table 2

Results of TaCTICS Analysis of Auto-segmentation and Resident Contours

	User	TO	|VD|	D	J	FND	FPD	HD (mm)
Head and	Trainee 1	0.34	0.81	0.24	0.14	0.47	1.04	22.97
	Trainee 2	0.44	0.75	0.32	0.19	0.41	0.95	13.92
	Trainee 3	0.25	0.74	0.19	0.10	0.54	1.08	31.25
	Trainee 4	0.31	0.85	0.22	0.12	0.48	1.08	27.50
	Trainee 5	0.44	0.52	0.35	0.21	0.44	0.86	10.34
	Trainee Average	0.36	0.73	0.26	0.15	0.47	1.00	22.20
	Autosegmentation	0.49	1.03	0.32	0.19	0.34	1.02	15.44
	p-value by Wilcoxon Signed-Rank	*0.031	*0.031	0.094	0.094	*0.031	0.500	0.156
Chest Case	Trainee 1	0.25	0.08	0.26	0.15	0.78	0.70	30.27
	Trainee 2	0.38	1.03	0.25	0.14	0.41	1.09	25.33
	Trainee 3	0.32	0.09	0.31	0.18	0.65	0.74	7.96
	Trainee 4	0.30	0.08	0.29	0.17	0.67	0.75	15.83
	Trainee 5	0.24	0.80	0.17	0.09	0.54	1.12	15.23
	Trainee Average	0.30	0.38	0.25	0.15	0.61	0.88	18.92
	Autosegmentation	0.49	1.31	0.29	0.17	0.31	1.10	22.49
	p-value by Wilcoxon Signed-Rank	*0.031	*0.031	0.094	0.094	*0.031	0.094	0.312

Note that the auto-segmentation was significantly different from the residents in total overlap (TO), volumetric difference (VD) and false negative Dice coefficients (FND).