Radiotherapy and Oncology 112 (2014) 321–325

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Randomised trial

Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer ☆

Gary V. Walker [a,b], Musaddiq Awan [a], Randa Tao [a], Eugene J. Koay [a], Nicholas S. Boehling [a], Jonathan D. Grant [a], Dean F. Sittig [b], Gary Brandon Gunn [a], Adam S. Garden [a], Jack Phan [a], William H. Morrison [a], David I. Rosenthal [a], Abdallah Sherif Radwan Mohamed [a], Clifton David Fuller [a,c,*]

[a] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center; [b] School of Biomedical Informatics, University of Texas Health Science Center; and [c] University of Texas Graduate School of Biomedical Sciences, Houston, USA

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form 7 August 2014
Accepted 12 August 2014
Available online 9 September 2014

Keywords:
Atlas-based autosegmentation
Normal tissue
Autocontouring
Head and neck
Automatic segmentation
Organs-at-risk

ABSTRACT

Background and purpose: Target volumes and organs-at-risk (OARs) for radiotherapy (RT) planning are manually defined, which is a tedious and inaccurate process. We sought to assess the feasibility, time reduction, and acceptability of an atlas-based autosegmentation (AS) compared to manual segmentation (MS) of OARs.

Materials and methods: A commercial platform generated 16 OARs. Resident physicians were randomly assigned to modify AS OAR (AS + R) or to draw MS OAR followed by attending physician correction. Dice similarity coefficient (DSC) was used to measure overlap between groups compared with attending approved OARs (DSC = 1 means perfect overlap). 40 cases were segmented.

Results: Mean ± SD segmentation time in the AS + R group was 19.7 ± 8.0 min, compared to 28.5 ± 8.0 min in the MS cohort, amounting to a 30.9% time reduction (Wilcoxon $p < 0.01$). For each OAR, AS DSC was statistically different from both AS + R and MS ROIs (all Steel–Dwass $p < 0.01$) except the spinal cord and the mandible, suggesting oversight of AS/MS processes is required; AS + R and MS DSCs were non-different. AS compared to attending approved OAR DSCs varied considerably, with a chiasm mean ± SD DSC of 0.37 ± 0.32 and brainstem of 0.97 ± 0.03.

Conclusions: Autosegmentation provides a time savings in head and neck regions of interest generation. However, attending physician approval remains vital.

The advent of conformal radiotherapy (RT), as well as the developing paradigm of image-guided RT, affords delivery of tumoricidal radiation doses to user-defined target volumes while minimizing dose to organs-at-risk (OARs). However, since the optimization software responsible for dose-reduction to OARs requires accurate region of interest (ROI) segmentation for all subsequent dose calculations, accurate ROI definition is recognized as of paramount importance for RT planning, representing perhaps contributing to the most uncertainty in RT planning.

Target volumes and OARs for RT planning are typically manually defined by human users, which is a tedious and time consuming process [1]. Manual segmentation (MS), which uses a digital paintbrush controlled by the user to outline the ROI, remains highly variable, particularly for novices [2]. Inexperienced trainees or those who treat particular oncologic sites infrequently may under or over-contour particular OAR regions-of-interest (ROIs), especially those that are difficult to visualize on computed tomography (CT). Atlas-based autosegmentation (AS) algorithms have been shown to accurately delineate OARs ROIs for a variety of disease sites and early results show a time saving advantage [3–10].

Head and neck cancers present a unique set of challenges in terms of target delineation for conformal RT [11,12]. With data from Fuller et al. suggesting that head and neck cancer target volumes [13] are comparatively difficult to contour, time-consuming, and have higher inter-observer variability than other anatomic sites, we sought to ascertain whether time-savings might be achieved using OAR AS as part of our clinical workflow. To this end we constructed a randomized, blinded prospective *in silico* study with the following specific aims:

☆ This work was presented in abstract form at the 55th annual meeting of the American Society for Radiation Oncology on September 23, 2013 in Atlanta, GA.

* Corresponding author at: Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Unit 97, Houston, TX 77030, USA.

E-mail address: cdfuller@mdanderson.org (C.D. Fuller).

http://dx.doi.org/10.1016/j.radonc.2014.08.028
0167-8140/

(1) Determine the real-time workflow feasibility and capacity for delineation time-reduction using AS software for head and neck OARs in a representative clinical population.
(2) Evaluate individual OAR acceptability of AS/AS-assisted definition of head and neck ROIs using resident and expert physician comparators.

Materials and methods

This prospective, randomized, double-blind study was approved by the Institutional Review Board. Patients being simulated for definitive RT ± chemo for head and neck malignancies were included. Patients were excluded for the following reasons: (1) under the age of 18 years, (2) prior history of RT or surgery to the head and neck, and (3) cutaneous malignancies. These patients were excluded so as to reduce bias as these patients could have anatomical variations. Patients were simulated supine with chins extended utilizing a custom thermoplastic mask for immobilization. Non-contrast axial CT slices were obtained using 2.5–3.0 mm slices. All patients were treated using intensity modulated RT (IMRT) with either the whole-field IMRT technique or the half-beam technique with the upper IMRT field matched to a static low neck field as previously described [14].

Study design

Following CT acquisition, all DICOM files were processed using a commercial, proprietary AS platform (Pinnacle 9.4, SPICE AS algorithm, Philips Healthcare, Andover, MA, USA). This software package performs an initial registration, dense deformable registration and then probabilistic refinement using an automated platform in the background while the user performs other tasks. Residents were randomly assigned to manually modify AS OARs ROIs (AS + R) or to draw MS OARs ROIs *de novo* using a pairwise randomization technique so that each resident did equal number of AS + R and MS cases. These 16 OARs included the spinal cord, brainstem, optic chiasm, mandible, oral cavity, soft palate, larynx, pharyngeal constrictors as well as bilateral optic nerves, parotid glands, submandibular glands and cochlea. For the purposes of this study we did not evaluate gross or clinical tumor volumes as these display significant variation between patients based on clinical presentation necessitating MS followed by multi-physician examination and real-time ROI quality assurance, using a methodology described previously [12,15]. Total resident segmentation/correction time was recorded. Residents rotate through the head and neck service during the first, second and in either their third or fourth year of training. Attending physicians subsequently reviewed all OARs and manually corrected them as necessary, blinded to AS or MS ROIs priors. All OARs underwent established processes including approval by the attending physician and the head and neck service quality assurance team consisting of multiple attending physicians to minimize inter-observer variation in OAR delineation [12,15]. ROIs were never used for treatment planning without attending approval.

Endpoints and analysis

Specified outcome variables included overall segmentation time (including AS processing time) and segmentation accuracy (using attending physician approved contours as a reference). For this study, nonparametric statistical comparison was implemented using the paired Wilcoxon rank sum test for between-group comparison of ordinal and scalar variables (Supplement 1). Based on these calculations, a maximal enrollment of 106 patients (53 per arm) was approved by the institutional review board, with an interim analysis conducted when enrollment met 40 patients. As the Wilcoxon rank-sum test reached statistical significance, enrollment was stopped early as per protocol specifications.

For ROI accuracy assessment, Dice similarity coefficient (DSC) was used to measure overlap between the AS, AS with resident correction (AS + R), and MS compared with attending approved ROIs, where DSC = 1 means perfect ROI overlap, and DSC = 0 is totally discordant. An informal *a priori* threshold DSC of $\geqslant 0.85$ was utilized as an acceptability threshold (e.g., ROIs with 85% DSC agreement were considered to require clinically acceptable modification, whereas >15% disagreement was considered clinically unacceptable within our current workflow). A paired Kruskal–Wallis test was performed using the Steel–Dwass methodology (the non-parametric analog comparable to Tukey's range test) was used to compare the total process time between the groups stratified by OAR ROI. Univariate descriptive statistics among the groups were compared using a Pearson χ^2 test. A non-Bonferroni-corrected alpha of 0.05 was specified for statistical significance for all analyses. Data analysis was performed using Stata/SE 12.0 (Stata Corporation, College Station, TX, USA) and JMP v10/SAS v9.2 (both by SAS Institute, Cary, NC, USA) statistical software.

Results

Baseline demographic and tumor characteristics were similar between the 40 cases undergoing MS and AS (Supplemental Table 1). The median age was 59.8 with a mean body mass index (BMI) of 29.2. The oropharynx was the most common site of disease (65.0%), followed by larynx (12.5%) with a range of TNM stages (Table 1).

All 40 cases were segmented by one of 8 residents and approved by one of 7 head and neck section expert attending physicians. Mean ± standard deviation (SD) resident segmentation time in the AS + R group was 19.7 ± 8.0 min, compared to 28.5 ± 8.0 min in the MS cohort, amounting to a 30.9% time reduction (Wilcoxon $p = 0.002$) (Table 1). There was no difference in the number of cases done by residents in each year of training (p = n.s.). Average time savings per resident ranged from 2.3–15.7 min.

Fig. 1 displays an example of the AS and attending physician approved OAR ROIs. The mean ± SD DSC for the uncorrected AS, AS + R, and MS ROIs compared to final attending approved ROIs was 0.74 ± 0.29, 0.96 ± 0.10, and 0.95 ± 0.16, respectively (Table 2, Fig. 2). For each OAR, AS DSC was statistically different from both AS + R and MS ROIs (all Steel–Dwass $p < 0.01$) except the spinal cord ($p = 0.4$) and the mandible (0.4); AS + R and MS DSCs were non-different (all p = n.s.) suggesting significant oversight of AS/MS processes is still required.

AS accuracy compared to attending approved OAR DSCs varied considerably with the following structures showing the most overlap (i.e., most accurate): mandible (DSC 0.98 ± 0.2), brainstem (0.97 ± 0.03), pharyngeal constrictors (0.93 ± 0.09), spinal cord (0.90 ± 0.14), parotid glands (0.89 ± 0.11), and soft palate (0.86 ± 0.15). However, specific structures showed reduced DSC values, representative of poor agreement with expert ROIs. Notably, the optic chiasm (0.37 ± 0.32), oral cavity (0.53 ± 0.24), cochlea (0.56 ± 0.38), larynx (0.65 ± 0.26), optic nerves (0.71 ± 0.26), and submandibular glands (0.73 ± 0.25) were notable for comparatively suboptimal segmentation, with a statistically different performance across OAR ROIs (Steel–Dwass $p < 0.01$).

Discussion

These data represent, to our knowledge, the largest prospective randomized controlled study evaluating the ROI accuracy and time-efficiency gains of OAR AS software in head and neck RT.