Medical Physics
The International Journal of Medical Physics Research and Practice

Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer

Albert K. Hoang Duc, Gemma Eminowicz, Ruheena Mendes, Swee-Ling Wong, Jamie McClelland, Marc Modat, M. Jorge Cardoso, Alex F. Mendelson, Catarina Veiga, Timor Kadir, Derek D'Souza, and Sebastien Ourselin

Citation: Medical Physics **42**, 5027 (2015); doi: 10.1118/1.4927567
View online: http://dx.doi.org/10.1118/1.4927567
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/42/9?ver=pdfcov
Published by the American Association of Physicists in Medicine

Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer

Albert K. Hoang Duc[a)]
Center for Medical Image Computing, University College London, London WC1E 6BT, United Kingdom

Gemma Eminowicz, Ruheena Mendes, and Swee-Ling Wong
Radiotherapy Department, University College London Hospitals, 235 Euston Road, London NW1 2BU, United Kingdom

Jamie McClelland, Marc Modat, M. Jorge Cardoso, and Alex F. Mendelson
Center for Medical Image Computing, University College London, London WC1E 6BT, United Kingdom

Catarina Veiga
Department of Medical Physics and Bioengineering, University College London, London WC1E 6BT, United Kingdom

Timor Kadir
Mirada Medical UK, Oxford Center for Innovation, New Road, Oxford OX1 1BY, United Kingdom

Derek D'Souza
Radiotherapy Department, University College London Hospitals, 235 Euston Road, London NW1 2BU, United Kingdom

Sebastien Ourselin
Centre for Medical Image Computing, University College London, London WC1E 6BT, United Kingdom

(Received 30 September 2014; revised 30 June 2015; accepted for publication 14 July 2015; published 5 August 2015)

Purpose: The aim of this study was to assess whether clinically acceptable segmentations of organs at risk (OARs) in head and neck cancer can be obtained automatically and efficiently using the novel "similarity and truth estimation for propagated segmentations" (STEPS) compared to the traditional "simultaneous truth and performance level estimation" (STAPLE) algorithm.
Methods: First, 6 OARs were contoured by 2 radiation oncologists in a dataset of 100 patients with head and neck cancer on planning computed tomography images. Each image in the dataset was then automatically segmented with STAPLE and STEPS using those manual contours. Dice similarity coefficient (DSC) was then used to compare the accuracy of these automatic methods. Second, in a blind experiment, three separate and distinct trained physicians graded manual and automatic segmentations into one of the following three grades: clinically acceptable as determined by universal delineation guidelines (grade A), reasonably acceptable for clinical practice upon manual editing (grade B), and not acceptable (grade C). Finally, STEPS segmentations graded B were selected and one of the physicians manually edited them to grade A. Editing time was recorded.
Results: Significant improvements in DSC can be seen when using the STEPS algorithm on large structures such as the brainstem, spinal canal, and left/right parotid compared to the STAPLE algorithm (all $p < 0.001$). In addition, across all three trained physicians, manual and STEPS segmentation grades were not significantly different for the brainstem, spinal canal, parotid (right/left), and optic chiasm (all $p > 0.100$). In contrast, STEPS segmentation grades were lower for the eyes ($p < 0.001$). Across all OARs and all physicians, STEPS produced segmentations graded as well as manual contouring at a rate of 83%, giving a lower bound on this rate of 80% with 95% confidence. Reduction in manual interaction time was on average 61% and 93% when automatic segmentations did and did not, respectively, require manual editing.
Conclusions: The STEPS algorithm showed better performance than the STAPLE algorithm in segmenting OARs for radiotherapy of the head and neck. It can automatically produce clinically acceptable segmentation of OARs, with results as relevant as manual contouring for the brainstem, spinal canal, the parotids (left/right), and optic chiasm. A substantial reduction in manual labor was achieved when using STEPS even when manual editing was necessary. [http://dx.doi.org/10.1118/1.4927567]

Key words: radiotherapy, OAR, head and neck, atlas segmentation

1. INTRODUCTION

Intensity-modulated radiotherapy (IMRT) enables normal tissue sparing by allowing better conformal dose distribution in head and neck cancer tissue. This technology requires the accurate delineation of several target volumes (TVs) and surrounding organs at risk (OARs). This delineation is typically performed manually by trained experts on computed tomography (CT) or magnetic resonance (MR) images and sometimes complemented with functional imaging techniques such as positron emission tomography (PET).[1,2] This process may need to be repeated multiple times during radiotherapy treatment to accommodate tumor response and physiological changes in the patient.

In practice, manual contouring is time-consuming and labor intensive, especially for large TVs and irregular OARs. It is also subject to large inter-rater variability,[3,4] despite universally accepted delineation guidelines.[5–7] Mean volume variations of up to 50% were reported in parotid delineation across three radiation oncologists on CT images.[8] Further investigations showed that the effects of inter-rater variability in delineating OARs have a significant dosimetric impact.[9] In addition, the range of inter-rater variability has been found to be greater in some cases than errors due to positioning and organ motion.[10] Consequently, the development of accurate and reproducible automatic segmentation method is crucial to allow clinicians to focus on other aspects of patients' treatment.

Recently, automatic atlas-based segmentation methods have shown promising results in segmenting head and neck CT images.[11,12] Different methods have been developed based on either a single-patient atlas,[13] a population-based average atlas,[14] or multiple atlases.[15] Multiatlas methods have been shown to yield better results than single atlas methods.[7,15] For the fusion of multiple atlases, the "simultaneous truth and performance level estimation" (STAPLE) algorithm[16] has been used in several studies to generate contours in the head and neck region.[12,15,17] Since the introduction of the original STAPLE algorithm, other segmentation methods that build upon it have been proposed to take into account the similarity between the atlases and the image to segment. In particular, Jorge Cardoso *et al.*[18] developed the "similarity and truth estimation for propagated segmentations" (STEPS) algorithm. In STEPS, atlases are locally ranked based on their similarity with the image to segment using the locally normalized cross-correlation. For a local region to segment, only the top ranked atlases for that region are used during the fusion process. In contrast, all atlases carry the same global weight in STAPLE. STEPS has previously been validated on brain structure segmentation[19,20] and has been shown to perform better than STAPLE. This is in line with the fact that local fusion strategies outperform global methods.[21]

A standard evaluation of accuracy has been the direct comparison of manual and automatic segmentations using overlap measures such as the Dice similarity coefficient (DSC).[22] However, the accuracy of automatic methods as measured this way is limited by the degree of inter-rater variability in manual contouring. In the presence of such variability, even an algorithm that performs as well as an expert cannot be expected to achieve total agreement with manual segmentations. Furthermore, it is possible that an automatic segmentation does not resemble the gold standard but is still acceptable for use in radiotherapy planning. This judgment cannot reliably be made based on overlap measures, and an expert rater decision is required.

Automated methods can reduce physician contouring time by up to 30%–40% as seen in studies of head and neck cancer[7] and also reduce the inherent inter-rater variability in volume delineation.[12] The improvement in time and consistency is valuable only if segmentation accuracy is not undermined. Assessing the accuracy of automatic segmentation is a challenging task and manual editing is usually required to achieve clinically acceptable results.[11,12] Nevertheless, the workload of manual editing can be significantly shorter than manual contouring.[7]

In this study, we compare STAPLE against STEPS in producing accurate segmentations for radiotherapy planning. Both algorithms are used to segment the following OARs in head and neck cancer: the brainstem, the spinal canal, the left and right parotids, the optic chiasm, and the eyes. The accuracy of both algorithms was measured using the DSC.[22] In addition to accuracy, we wanted to measure the clinical acceptability of each automatic method. To account for the variability in overlap measures, manual contours and automatic segmentations produced by STAPLE and STEPS were graded on a three-point scale for clinical acceptability in a blind experiment by three distinct trained physicians. The comparison through blindly obtained grades of manual and automatic segmentations represents a novel approach for their evaluation. Traditional evaluation has been to directly compare manual and automatic segmentations using the DSC. Although a high DSC should guarantee clinical acceptability, a lower DSC does not necessarily mean that an automatic segmentation is not clinically useful. To our knowledge, methods classifying segmentations for clinical acceptability on a point scale by expert raters have not been published before. Time gain by using automatic segmentation was also assessed.

2. MATERIALS AND METHODS

2.A. Overview

First, 6 OARs were delineated by two radiation oncologists in a dataset of 100 patients with head and neck cancer on CT images. Each patient in the dataset was automatically segmented with both the STAPLE and STEPS algorithms using those manual contours. DSC was then used to measure the accuracy of the automatic segmentations. Second, three separate and distinct trained physicians graded the manual and automatic segmentations generated by both methods into one of the following three grades in a blind experiment: clinically acceptable without modification, fulfilling universal delineation guidelines[23] for radiotherapy planning (grade A), reasonably acceptable for clinical practice upon manual editing (grade B), and not acceptable (grade C). DSC for the STEPS algorithm and for each grade was then calculated.

Finally, STEPS segmentations graded B were selected and given to one of the three physicians who manually edited them to grade A. Editing time was recorded.

2.B. Atlas dataset

The atlas dataset consisted of $N = 100$ planning CT images of patients with different diagnoses of head and neck cancer. These were cases treated with IMRT at the radiotherapy department for any head and neck cancer diagnosis (squamous cell cancer and adenocarcinoma), including postoperative and primary radiotherapy with diagnoses including pharyngeal, laryngeal, oral cavity, unknown primary, and maxillary sinus cancer. Staging ranged from T2N0M0 to T4N3M0. Each CT image was acquired using a General Electric RT CT scanner and was composed of 100–205 slices (2.5 mm thick) containing 512×512 pixels each. All patients were scanned head-first supine with their head blocked by an anatomical cushion and an individual thermoplastic mask. Our study involved 100 patients: a first radiation oncologist contoured 43 patients and a second distinct radiation oncologist contoured the remaining 57 patients. For each patient, six OARs in the head and neck region were manually contoured for radiotherapy purposes. This included the brainstem, the spinal cord, the parotids (left/right), the optic chiasm, and the eyes. The eyes volume comprises the left and right sides of the orbits, lenses, and optic nerves. This grouping was deliberate. Since those structures are small, spreading only a couple of axial slices, and are generally delineated successively one side after the other, it was coherent to group them under a single label. Also, this was done to align the time scoring of the eyes with the time scoring of the other OARs [i.e., brainstem, the spinal cord, the parotids (left/right), and the optic chiasm].

Some traditional OARs (i.e., lymph nodes and mandible) used in head and neck planning were not investigated. Indeed, not all traditional OAR segmentations were available for all patients. In a large amount of cases, the lymph nodes (either left or right), the mandible, or the vocal cord was not available to us for this study. As a result, we only considered the OARs that were available for every patient, which were the brainstem, the spinal canal, the left and right parotids, the optic chiasm, and the eyes.

2.C. Atlas-based segmentation

A registration algorithm is used to create automatic segmentations of regions of interest for a new image by transforming existing segmentations of the corresponding structures in existing images. Those automatic segmentations are then combined into a single consensus using a fusion algorithm.

2.C.1. Registration algorithm

A leave-one-out experiment was used in which each patient (referred to as a target) in the dataset was automatically segmented using the remaining atlases. A registration algorithm[24] was used to deform the atlases onto the target image space. The target image space is defined as the space of the patient to segment. As a result, each target image is in a different individual space rather than in a same common space. The manual contours were then mapped onto the target using the resulting transformation from registration and fused with either the STAPLE or STEPS algorithm to yield estimated segmentations. The registration first determined an affine registration using translation, rotation, and scaling. The affine registration used a symmetric approach of the block-matching algorithm developed by Ourselin *et al.*[25] A multi-level nonrigid registration step using free-form deformations with a B-spline control point parameterization[26] was subsequently applied. The locally normalized cross-correlation was used as a similarity measure. The control point spacing was 5 voxels in all directions and a bending energy penalty term was used to regularize the deformation. The time to perform affine and nonrigid atlas registration onto a patient target image is about 45 min using a regular CPU.

2.C.2. Fusion using the STAPLE and STEPS algorithms

The STAPLE and STEPS algorithms are both based on an expectation–maximization (EM) framework. The framework starts with computing an estimate of the ground truth using a simple segmentation method. Based on this initial guess, it is possible to calculate the performance of each individual label. In the expectation step (E-step), labels are combined to estimate the true segmentation depending on their performance. In the maximization step (M-step), given an estimate of the true segmentation, the performance values of each labels are reassessed and are maximized. In general, the performance is dependent on certain parameters and the M-step is used to find the parameters which maximize the performance of each label, while in the E-step, the estimate of the true segmentation is improved based on these parameters. In STAPLE, each segmentation is weighted globally depending upon their estimated performance level in the E-step, and the sensitivity and specificity of each label is calculated in the M-step. In STEPS, the sensitivity and specificity is only calculated in areas where each classifier is considered an expert by the LNCC ranking strategy. This results in a two-step performance estimation that decouples the two sources of error: one based on the LNCC image similarity metric observation characterizing the nonuniform registration accuracy and shape differences, and the other step characterizing the specificity and sensitivity of each classifier when compared with the consensus classification. Due to the local nature and smoothness of the metric, the similarity between the images is described on a smooth voxel by voxel basis, enabling a voxel by voxel ranking with reduced discontinuity effect. The raw HU units were used to compute the LNCC metric.

When a dataset of atlases is available, it is best to select the most similar atlases to the target when using STAPLE rather than using the whole dataset.[27,28] To apply STAPLE in this study, we followed the method in Ref. 29 based on manifold learning for atlas selection as the method showed consistently good results in selecting atlases. In Ref. 29, three dimensionality reduction techniques (Isomap, locally linear embedding,

and Laplacian eigenmaps) were compared for the selection of atlases to use in multiatlas segmentation. This study also investigated the optimal number of atlases to fuse for each technique. Optimal results were obtained by choosing the best seven atlases using locally linear embedding. Therefore, for each target, the best seven atlases were selected using the locally linear embedding method.[30] In contrast, STEPS does not require an explicit atlas selection as the algorithm already integrates a local ranking scheme. In this study, the whole dataset was registered to the target. Once all registrations are done, the top seven ranked registered atlases for each local region (i.e., a patch of 5×5 voxels) to segment were used in the fusion process. As a result, STEPS does not require an atlas selection strategy but more registrations need to be performed than in STAPLE. Indeed, STEPS requires as many registrations as the size of the atlas dataset. The time to perform atlas fusion is about 5 min using a regular CPU. So total time to obtain an automatic segmentation (registration and fusion) is about 50 min.

2.D. Evaluation

The first objective was to compare the STAPLE against the STEPS algorithm in producing accurate segmentations. DSC and the Hausdorff distance between manual contouring and the two automatic segmentation methods were reported. The DSC is defined as $D(U,V) = 2|U \cap V|/(|U| + |V|)$, where $|U|$ (respectively, $|V|$) is the number of voxels in the automated (respectively, manual) region. Its value ranges from 0 to 1, where 0 means no overlap and 1 signifies a perfect match. The Hausdorff distance is defined as the maximum of the minimum distances for each point between the automated and manual regions.

2.E. Segmentation grading

The second objective was to assess whether the STAPLE and STEPS algorithms could produce segmentations as clinically relevant as manual contouring. All segmentations were imported into a treatment planning system (Varian Eclipse version 11) and graded by a trained physician. Three distinct physicians, with the same level of expertise as the two radiation oncologists, graded in a blind experiment manual and automatic segmentations using one of the following three grades:

- Grade A: the segmentation is clinically acceptable and satisfies universal OAR delineation guidelines[23] and can be used as created for radiotherapy planning.
- Grade B: the segmentation is reasonably acceptable but needs some manual editing. Some contour lines need to be corrected to meet universal guidelines.
- Grade C: the segmentation does not meet universal guidelines. Some slices show gross misdelineation that cannot be attributed to segmentation variability.

On this scale, grade A is considered higher than grade B and grade B higher than grade C. The three distinct physicians graded manual and automatic segmentations in a random order. To reduce bias from assessing the same structure multiple times, associated automatic and manual segmentations were graded at least 1 week apart. The first physician graded the 6 OARs of 100 patients. Due to time constraint, the second and third physicians could only grade the 6 OARs of 50 and 30 patients, respectively. Comparison between grades of manual and automatic segmentations by the three trained physicians is used as an indicator of clinical acceptability. Although radiation oncologists contours were graded by three distinct trained physicians, this does not imply that one expert rater was better than another. A total of 1200 automatic and 600 manual segmentations were graded ($1200 = 6$ OARs $\times$ 100 patients $\times 2$ and $600 = 6$ OARs $\times$ 100 patients).

2.F. Manual editing time

The third objective was to quantify manual contouring time saved by using the STEPS algorithm. When patients were originally contoured for radiotherapy treatment, contouring time was not recorded. In order to estimate this contouring time and to keep manual contouring to an acceptable level, one of the three trained physicians recontoured the OARs of five patients and the time was recorded. Those five patients were chosen to be representative of the whole dataset by an external researcher. Time reported for the eyes volume was the aggregated time to contour the component parts. For each OAR, the physician was given 15 randomly selected STEPS segmentations graded B and edited them to grade A. Editing time was recorded. A brush to push in/out the contour lines, freehand, and eraser tools were used for contouring and editing.

3. RESULTS

3.A. STAPLE vs STEPS

The DSC and Hausdorff distance are reported in Fig. 1. Significant improvements can be seen when using the STEPS algorithm on large structures such as the brainstem, spinal canal, and left/right parotid compared to the STAPLE algorithm. Using a Wilcoxon rank-sum test, STEPS segmentations yielded significantly higher DSC than STAPLE segmentations (all $p < 0.001$) for those structures. For smaller structures, such as optic chiasm and the eyes, the difference is not significantly different ($p > 0.300$ and $p > 0.170$). The DSC for those structures is significantly lower compared to larger ones. This can be explained by their size, where even small voxel misclassification in the automatic segmentation will result in large DSC discrepancy. Figure 2 shows some examples of manual, STEPS, and STAPLE segmentations of the brainstem, the spinal canal, and the parotids (left/right). The optic chiasm and eyes are not shown as they are small structures and hard to depict in a single view. The clinical acceptability of our method could not have been reliably determined with the DSC, and verification by means of separate trained physicians was required.

3.B. Grading

Results of grading by the three trained physicians are shown in Fig. 3. A surprising number of manual contours for the

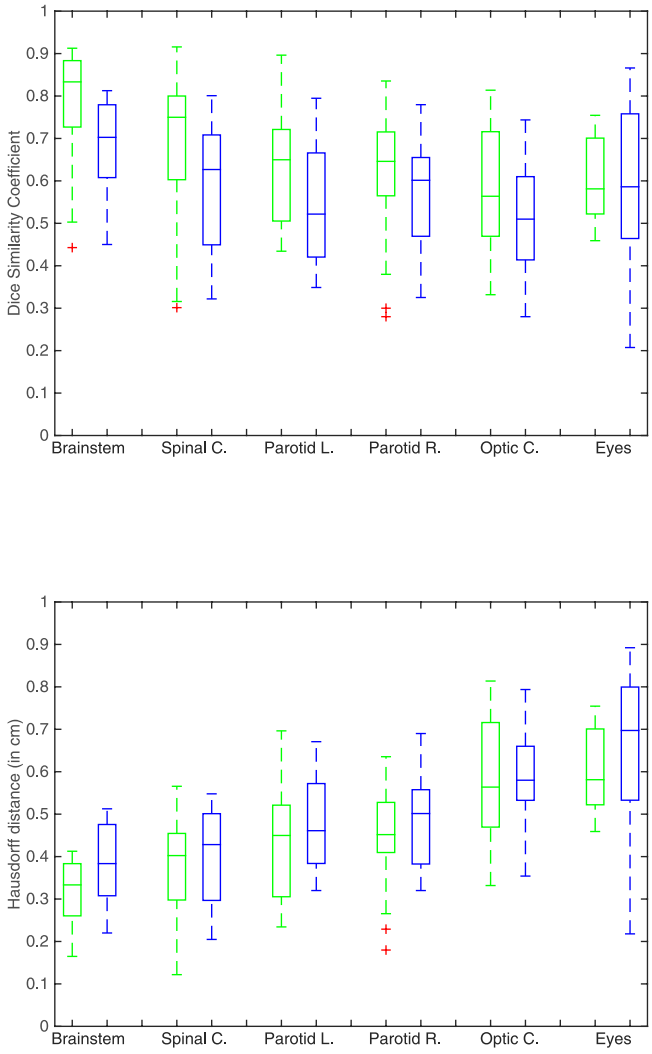

FIG. 1. Dice similarity coefficient (top) and Hausdorff distance (bottom) of the STEPS (left) and STAPLE (right) algorithms against manual contouring.

eyes and optic chiasm were graded B and C, corresponding to high inter-rater variability. This is consistent across the three trained physicians. This may be due to the poor contrast of those areas in CT images. Manual and STEPS segmentations of the parotids (left/right) and the optic chiasm were given

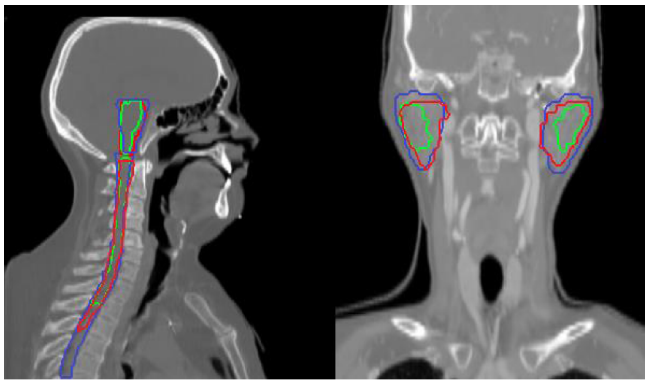

FIG. 2. Examples of manual (blue), STEPS (red), and STAPLE (green) segmentations of the brainstem, spinal canal, and parotids (left/right).

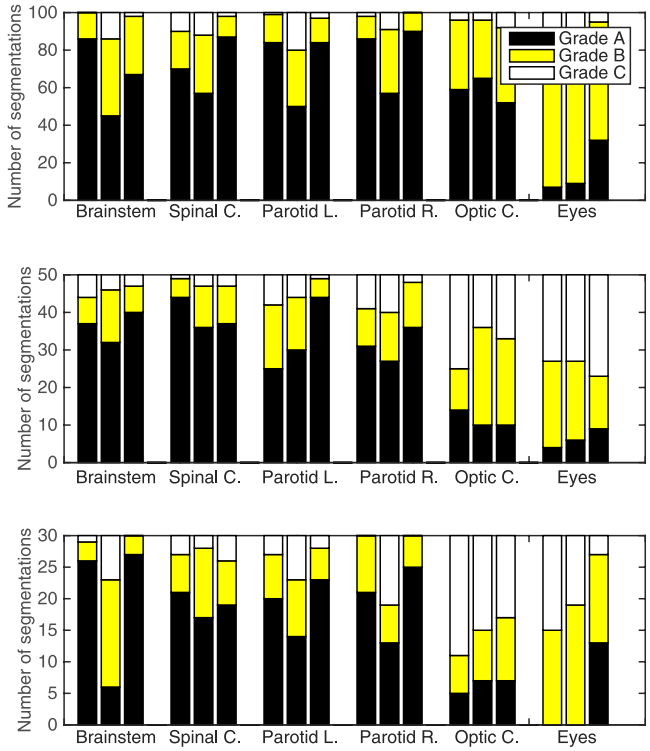

FIG. 3. Grading of manual and automatic segmentations by three distinct trained physicians. Each graph represents grading done by a physician. For each OAR: STEPS = left bar, STAPLE = middle bar, manual = right bar. Grade A: clinically acceptable, no editing required. Grade B: reasonably acceptable, some editing required. Grade C: not acceptable.

similar grades by two trained physicians. The third physician, except for the left parotid, drew a similar conclusion. When similar grades were given, a Wilcoxon signed-rank test did not show any significant difference for those OARs (all $p > 0.100$). For the brainstem and the spinal canal, STEPS segmentations were overall graded similarly as well. In some cases, STEPS segmentations of those OARs were graded higher than manual segmentation and those differences were statistically significant ($p < 0.010$). In contrast, with STEPS segmentations the eyes were graded significantly lower ($p < 0.005$).

Overall, STAPLE segmentations were graded significantly lower than both manual and STEPS segmentations (all $p < 0.01$), except for the optic chiasm and the eyes ($p > 0.273$ and $p > 0.382$).

Figure 4 shows the grade distribution of STEPS, which gave the best results out of the two automatic methods, and manual segmentations. Only distribution from the trained physician who graded all 100 patients is shown. We note that a substantial number of STEPS segmentations of the spinal canal (27 cases) and the eyes (30 cases) were graded lower than their associated manual contours and offer some explanation. The well-defined boundaries of the spinal canal make it one of the easier OARs to segment for an expert rater, but atlas-based methods were seen to suffer from two key problems there. High neck flexion confounded registration in ten cases, and discrepancies in the length of the lower part segmented in the atlas set (vertebrae below C1) caused failure in 17 more. No atlas-based method can overcome such discrepancies, and they

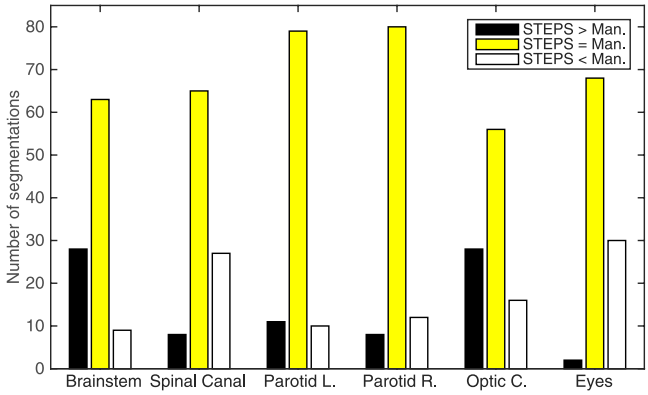

FIG. 4. Grade distribution of automatic and associated manual segmentations. STEPS > Man.: STEPS segmentation has a higher grade than its associated manual contour. STEPS = Man.: STEPS and manual segmentations have the same grade. STEPS < Man.: STEPS segmentation has a lower grade than its associated manual contour.

must be fixed by standards in the templates used. For the eyes, since the structures involved are small, a slight deviation in the automatic segmentation will inevitably result in some manual editing being required.

Across all OARs, STEPS was observed to outperform STAPLE and produce segmentations graded as well as or better than manual contours with a rate of 83%. A one sided confidence interval based on the t-statistic places the true rate above 80% with 95% confidence.

3.C. Dice similarity coefficient and clinical acceptability

To examine the relationship between acquired grades and DSC, we calculated the DSC between clinically acceptable (grade A) manual contours only and the STEPS segmentations graded A–C. As the results across the 3 trained physicians are similar, only the segmentations from the physician who graded all 100 patients are examined. Results are presented in Fig. 5. Using a Wilcoxon rank-sum test, STEPS segmentations graded A did not yield significantly higher DSC than STEPS segmentations graded B. The median DSC was also seen to vary significantly between OARs, for instance, the median DSC of the left/right parotids was significantly different from all other regions (all $p < 0.020$). Therefore, it may not be meaningful to compare segmentation quality between different regions using this measure. For all OARs, DSC of STEPS segmentations graded C was significantly lower (all $p < 0.005$) compared to segmentations graded A and B. Since STEPS segmentations graded A and B yielded similar DSC, the clinical acceptability of our method could not have been reliably determined with DSC, and verification by means of a separate trained physician was required.

3.D. Time scoring

Figure 6 shows the time taken to obtain a grade A result using the STEPS algorithm with manual editing, without it, and using fully manual contouring. Using the Wilcoxon rank-sum

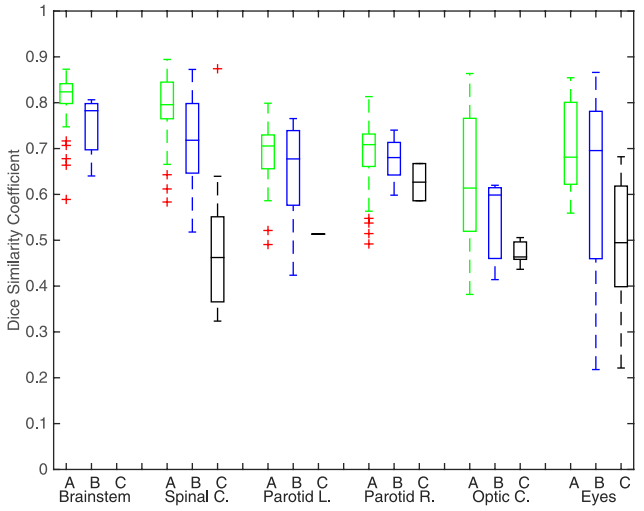

FIG. 5. Dice similarity coefficient of STEPS segmentations graded A (left), graded B (middle), and grade C (right) versus manual contours graded A. Only the segmentations from the physician who graded all 100 patients are shown.

test, these results demonstrate that STEPS yielded significant time saving, even when automatic segmentation needed editing. Time saved is relatively lower for the eyes; these being a grouping of six different structures, the trained physician spent a significant amount of time switching between editing tools, which added to the effective editing time. Time gained and p-values are reported in Table I. Time gained is calculated using the following ratio: (grading time + editing time)/(manual contouring time) if the automatic segmentation needed editing and (grading time)/(manual contouring time) if the automatic segmentation did not need editing.

4. DISCUSSION

In this study, the STAPLE and STEPS algorithms used multiple manual contours to generate the most likely segmentation

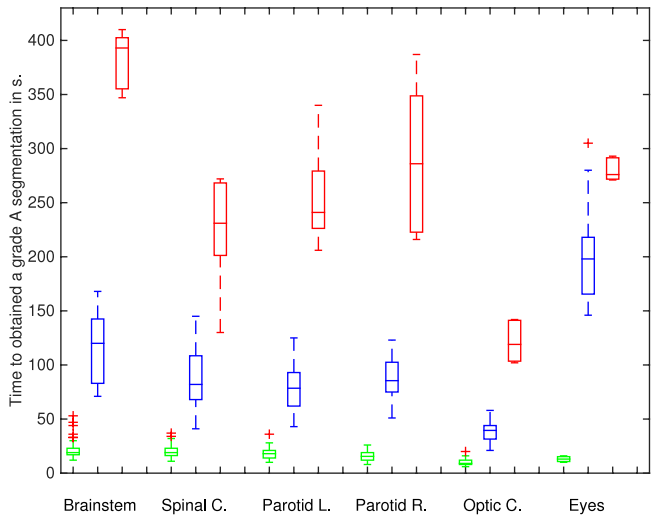

FIG. 6. Time in seconds to obtain a grade A segmentation using STEPS algorithm without (left) or with (middle) manual editing and with fully manual contouring (right).

TABLE I. Relative gain (%) in segmentation time. *P*-values are the results of the Wilcoxon rank-sum test.

	Auto. without editing vs man. scratch	Auto. with editing vs man. scratch
Brainstem	95.16%, $p < 10^{-3}$	69.46%, $p < 10^{-3}$
Spinal canal	91.77%, $p < 10^{-3}$	64.50%, $p < 0.005$
Parotid left	92.53%, $p < 10^{-3}$	67.42%, $p < 10^{-3}$
Parotid right	94.58%, $p < 10^{-3}$	70.10%, $p < 10^{-3}$
Optic chiasm	92.43%, $p < 10^{-3}$	66.80%, $p < 10^{-3}$
Eyes	95.28%, $p < 0.01$	28.26%, $p < 0.005$

using information from the radiation oncologists. Inter-rater variability is one of the most challenging issues in IMRT and is a motivation for the development of methods that improve consistency. The results showed the advantages of STEPS over STAPLE in segmenting OARs in head and neck cancer. In summary, DSC from STEPS was higher compared to DSC from STAPLE for the brainstem, spinal canal, and left/right parotid. This showed that the local combination strategy introduced in STEPS outperforms the global fusion method in STAPLE. In addition, STEPS produced segmentations that were as clinically acceptable as manual contouring for structures such as the brainstem, spinal canal, parotids (left/right), and optic chiasm. In contrast, STEPS segmentation grades of the eyes were lower than grades from manual contouring. DSC reported in this study compares well with DSC reported in the literature (0.78 and 0.79 for the brainstem and parotids gland in Ref. 15 and 0.75 and 0.72 in Ref. 31). Across all OARs, we found a reduction in time of 61% and 93% on average when STEPS segmentation did and did not, respectively, require manual editing. This time gain was superior to numbers previously reported in the literature (40% in Ref. 31, 26% in Ref. 12, and 47% in Ref. 32).

The better results generated by STEPS over STAPLE are in line with findings in the literature. In Ref. 18, the robustness and accuracy of STEPS were evaluated on a database of cross-sectional and longitudinal brain MRI scans. In that study, STEPS performed better than STAPLE. STEPS has also been successfully used in other papers[19,20] to segment MR images. However, only our studies and the one from Ref. 18 directly compared the performance of STEPS and STAPLE and further investigation will need to be done across various ranges of image modalities to check if this statement holds.

A standard evaluation approach in radiotherapy has been to directly compare manual and automatic segmentations using the DSC. However, this study demonstrated that the DSC does not reliably reflect clinical acceptability of an automatic segmentation. Although a high DSC should guarantee clinical acceptability, a lower DSC does not necessarily mean that an automatic segmentation is not clinically useful. It may then be counterproductive to use a particular minimum DSC as a threshold for clinical acceptance of an automatic method, even if this is calibrated for a particular OAR.

Atlas-based segmentation is highly dependent on the similarity between the underlying atlas and the patient.[33] In our study, the failure in delineating the spinal canal in some cases could be due to multiple factors: (a) bad performance of the registration algorithm around that area, (b) lack of images in the atlas dataset with the same overall spinal morphology, (c) labeling discrepancies in the manual segmentation of the spinal canal [i.e., discrepancies in the length of the lower part segmented in the atlas set (vertebrae below C1)], and (d) patient head and neck position in the scanner when images are acquired. Different segmentation strategies based on either a single-patient atlas, a population-based average atlas, or multiple atlases have intrinsic limitations due to large deformations of normal anatomy that cannot be corrected with registration algorithms. Importantly, when thinking about applying automated segmentation, clinical concern arises due to abnormal anatomy in patients developing head and neck cancer. Our dataset included a variety of cases including some with bulky tumors, and results with our method were still comparable to manual contouring for the brainstem, spinal canal, left/right parotid, and optic chiasm across the cohort. In any case, automatic segmentations should always be checked and corrected if necessary by an expert before planning.

Starting contouring from an existing template (either automatic or manual) may have influenced the trained physicians' perception of gold standard. In general, relatively minor editing to the segmentations was performed and the lack of modifications may be attributed to the fact that the segmentations closely resembled physicians' definition of gold standard. However, this scenario represents the common clinical situation of verifying contours from less experienced clinicians, where relatively minor modifications are usually made overall.

Finally, there are some limitations to this study. Limitations include the small number of OARs edited and manually contoured to measure time cost and the lack of assessment of intrarater variability. However, these limitations should not affect the conclusion drawn as the significant *p*-values are all below 0.01 despite a wide confidence interval. In addition, this study did not include TVs. Multimodality imaging is often used to improve the visibility of TVs by coregistering CT with MR or PET images. Unfortunately, we did not have access to any imaging modalities other than CT. We note that atlas-based methods perform well when the shape of the target is well represented in the dataset of atlases, which is rarely the case in radiotherapy as tumors have no predefined shape.

5. CONCLUSIONS

The STEPS algorithm shows better performance than the STAPLE algorithm in segmenting OARs for radiotherapy of the head and neck. It is clinically useful and can considerably save time for clinicians in contouring OARs for radiotherapy planning. Even though automatically generated segmentations should always be checked and approved by an expert before radiotherapy planning, the STEPS segmentation method was found to be comparable to manual contouring for the brainstem, spinal canal, and left/right parotid.

ACKNOWLEDGMENTS

Sebastien Ourselin receives funding from the EPSRC (Nos. EP/H046410/1, EP/J020990/1, and EP/K005278), the MRC

(No. MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (No. FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL, and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative). Jamie McClelland acknowledges the support of the Intelligent Imaging EPSRC Program grant (No. EP/H046410/1). Marc Modat is supported by the UCL Leonard Wolfson Experimental Neurology Centre. M. Jorge Cardoso receives funding from EPSRC (No. EP/H046410/1). Alexander F. Mendelson is funded by EPSRC doctoral training Grant No. EP/J500331/1 and receives support from a CASE studentship with 355 the EPSRC and GE Healthcare. Catarina Veiga is funded by Funda¸cao para a Ciˆencia e a Tecnologia (FCT) Grant No. SFRH/BD/76169/2011, cofinanced by ESF, POPH/QREN, and EU. The authors would like to acknowledge Dr. Ruheena Mendes for data collection.

[a]Author to whom correspondence should be addressed. Electronic mail: albert.hoangduc.ucl@gmail.com

[1]T. J. Kruser *et al.*, "The impact of hybrid PET-CT scan on overall oncologic management, with a focus on radiotherapy planning: A prospective, blinded study," Technol. Cancer Res. Treat. **8**(2), 149–158 (2009).

[2]K. Newbold, M. Partridge, G. Cook, S. A. Sohaib, E. Charles-Edwards, P. Rhys-Evans, and C. Nutting, "Advanced imaging applied to radiotherapy planning in head and neck cancer: A clinical review," Br. J. Radiol. **79**(943), 554–561 (2006).

[3]T. S. Hong, W. A. Tome, R. J. Chappell, and P. M. Harari, "Variations in target delineation for head and neck IMRT: An international multi-institutional study," Int. J. Radiat. Oncol., Biol., Phys. **60**(1), S157–S158 (2004).

[4]W. Jeanneret-Sozzi, R. Moeckli, J. F. Valley, A. Zouhair, E. M. Ozsahin, and R. O. Mirimanoff, "The reasons for discrepancies in target volume delineation," Strahlenther. Onkol. **182**(8), 450–457 (2006).

[5]V. Gregoire, P. Levendag, K. K. Ang, J. Bernier, M. Braaksma, V. Budach, and H. Reychler, "CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines," Radiother. Oncol. **69**(3), 227–236 (2003).

[6]V. Gregoire, A. Eisbruch, M. Hamoir, and P. Levendag, "Proposal for the delineation of the nodal CTV in the node-positive and the post-operative neck," Radiother. Oncol. **79**(1), 15–20 (2006).

[7]C. Sjoberg, M. Lundmark, C. Granberg, S. Johansson, A. Ahnesj, and A. Montelius, "Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients," Radiat. Oncol. **8**, 229–235 (2013).

[8]X. Geets, J. F. Daisne, S. Arcangeli, E. Coche, M. D. Poel, T. Duprez, and V. Gregoire, "Inter-observer variability in the delineation of pharyngo-laryngeal tumor, parotid glands and cervical spinal cord: Comparison between CT-scan and MRI," Radiother. Oncol. **77**(1), 25–31 (2005).

[9]B. E. Nelms, W. A. Tom, G. Robinson, and J. Wheeler, "Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer," Int. J. Radiat. Oncol., Biol., Phys. **82**(1), 368–378 (2012).

[10]E. Weiss and C. F. Hess, "The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy," Strahlenther. Onkol. **179**(1), 21–30 (2003).

[11]A. V. Young, A. Wortham, I. Wernick, A. Evans, and R. D. Ennis, "Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes," Int. J. Radiat. Oncol., Biol., Phys. **79**(3), 943–947 (2011).

[12]L. J. Stapleford, J. D. Lawson, C. Perkins, S. Edelman, L. Davis, M. W. McDonald, and T. Fox, "Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer," Int. J. Radiat. Oncol., Biol., Phys. **77**(3), 959–966 (2010).

[13]O. Commowick and G. Malandain, "Efficient selection of the most similar image in a database for critical structures segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2007* (Springer, Berlin, Heidelberg, 2007), pp. 203–210.

[14]O. Commowick, V. Gregoire, and G. Malandain, "Atlas-based delineation of lymph node levels in head and neck computed tomography images," Radiother. Oncol. **87**(2), 281–289 (2008).

[15]D. N. Teguh, P. C. Levendag, P. W. Voet, A. Al-Mamgani, X. Han, T. K. Wolf, and M. S. Hoogeman, "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. **81**(4), 950–957 (2011).

[16]S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," IEEE Trans. Med. Imaging **23**(7), 903–921 (2004).

[17]X. Han, M. S. Hoogeman, P. C. Levendag, L. S. Hibbard, D. N. Teguh, P. Voet, and T. K. Wolf, "Atlas-based auto-segmentation of head and neck CT images," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2008* (Springer, Berlin, Heidelberg, 2008), pp. 434–441.

[18]M. Jorge Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, and S. Ourselin, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," Med. Image Anal. **17**(6), 671–684 (2013).

[19]S. R. Irani, C. J. Stagg, J. M. Schott, C. R. Rosenthal, S. A. Schneider, P. Pettingill, and M. R. Johnson, "Faciobrachial dystonic seizures: The influence of immunotherapy on seizure control and prevention of cognitive impairment in a broadening phenotype," Brain **136**(10), 3151–3162 (2013).

[20]D. Ma, M. J. Cardoso, M. Modat, N. Powell, J. Wells, H. Holmes, and S. Ourselin, "Automatic structural parcellation of mouse brain MRI using multi-atlas label fusion," PloS One **9**(1), e86576 (2014).

[21]X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solrzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," IEEE Trans. Med. Imaging **28**(8), 1266–1277 (2009).

[22]L. R. Dice, "Measures of the amount of ecologic association between species," Ecology **26**(3), 297–302 (1945).

[23]D. Genovesi, C. A. Perez, and A. Vinciguerra, *A Guide for Delineation of Lymph Nodal Clinical Target Volume in Radiation Therapy* (Springer, New York, NY, 2008), p. 173.

[24]M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, and S. Ourselin, "Fast free-form deformation using graphics processing units," Comput. Methods Programs Biomed. **98**(3), 278–284 (2010).

[25]S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache, "Reconstructing a 3D structure from serial histological sections," Image Vision Comput. **19**(1), 25–31 (2001).

[26]D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," IEEE Trans. Med. Imaging **18**(8), 712–721 (1999).

[27]P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," Neuroimage **46**(3), 726–738 (2009).

[28]K. K. Leung, J. Barnes, G. R. Ridgway, J. W. Bartlett, M. J. Clarkson, K. Macdonald, and S. Ourselin, "Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease," Neuroimage **51**(4), 1345–1359 (2010).

[29]A. K. Hoang Duc, M. Modat, K. K. Leung, M. J. Cardoso, J. Barnes, and T. Kadir, "Using manifold learning for atlas selection in multi-atlas segmentation," PloS One **8**(8), e70059 (2013).

[30]S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science **290**(5500), 2323–2326 (2000).

[31]J. F. Daisne and A. Blumhofer, "Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation," Radiat. Oncol. **8**(1), 154–165 (2013).

[32]K. S. Chao, S. Bhide, H. Chen, J. Asper, S. Bush, G. Franklin, and L. Dong, "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. **68**(5), 1512–1521 (2007).

[33]T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, Jr., "Quo vadis, atlas-based segmentation?," in *Handbook of Biomedical Image Analysis* (Springer, 2005), pp. 435–486.