Published in final edited form as:

Phys Med Biol. 2011 July 21; 56(14): 4557–4577. doi:10.1088/0031-9155/56/14/021.

Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study

M A Deeley^{1,‡}, A Chen², R Datteri², J Noble², A Cmelak¹, E Donnelly³, A Malcolm¹, L Moretti^{1,5}, J Jaboin^{1,§}, K Niermann¹, Eddy S Yang^{1,||}, David S Yu^{1,¶}, F Yei⁴, T Koyama⁴, G X Ding¹, and B M Dawant²

M A Deeley: matthew.deeley@vtmednet.org

¹Department of Radiation Oncology, Vanderbilt University, Nashville, TN

²Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN

³Department of Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN

⁴Department of Biostatistics, Vanderbilt University, Nashville, TN

⁵Department of Radiation Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

Abstract

The purpose of this work was to characterize expert variation in segmentation of intracranial structures pertinent to radiation therapy, and to assess a registration-driven atlas-based segmentation algorithm in that context. Eight experts were recruited to segment the brainstem, optic chiasm, optic nerves, and eyes, of 20 patients who underwent therapy for large space-occupying tumors. Performance variability was assessed through three geometric measures: volume, Dice similarity coefficient, and Euclidean distance. In addition, two simulated ground truth segmentations were calculated via the simultaneous truth and performance level estimation (STAPLE) algorithm and a novel application of probability maps. The experts and automatic system were found to generate structures of similar volume, though the experts exhibited higher variation with respect to tubular structures. No difference was found between the mean Dice coefficient (DSC) of the automatic and expert delineations as a group at a 5% significance level over all cases and organs. The larger structures of the brainstem and eyes exhibited mean DSC of approximately 0.8–0.9, whereas the tubular chiasm and nerves were lower, approximately 0.4–0.5. Similarly low DSC have been reported previously without the context of several experts and patient volumes. This study, however, provides evidence that experts are similarly challenged. The average maximum distances (maximum inside, maximum outside) from a simulated ground truth ranged from (−4.3, +5.4) mm for the automatic system to (−3.9, +7.5) mm for the experts considered as a group. Over all the structures in a rank of true positive rates at a 2 mm threshold from the simulated ground truth, the automatic system ranked second of the nine raters. This work underscores the need for large scale studies utilizing statistically robust numbers of patients and experts in evaluating quality of automatic algorithms.

[‡]Present address: Medical Physics Division, Fletcher Allen Health Care and Department of Radiology, University of Vermont, Burlington, VT.

[§]Present address: Department of Radiation Oncology, Washington University in St. Louis

^{||}Present address: Department of Radiation Oncology, University of Alabama-Birmingham

[¶]Present address: Department of Radiation Oncology, Emory University, Atlanta, GA.

1. Introduction

Three-dimensional imaging advances have revolutionized the treatment planning process in external beam radiation therapy. They provide physical information by which to calculate dose and specify external geometry, and as highly conformal treatments have become prevalent, they provide increasingly important information regarding patient anatomy both diseased and at risk. As a result image segmentation has become a central part and often rate limiting step in the planning process. Radiation oncologists must make judgments incorporating implicit and explicit anatomic, histologic and physiologic information in the presence of varying image quality to partition an image volume into normal and diseased tissue. This is a time consuming process that must occur before designing fields or calculating dose and thus can be a significant contributor to the overall efficiency of the process. The need for segmentation is only expected to increase in the future as additional conformal and adaptive techniques are implemented (Mell, Roeske & Mundt 2003, Mell, Mehrotra & Mundt 2005).

Until recently segmentation of all but the simplest structures was accomplished manually. Of late, however, a number of semi- and fully-automated methods have been developed to segment normal tissues in a radiotherapy clinical context (Gorthi, Duay, Houhou, Bach Cuadra, Schick, Becker, Allal & Thiran 2009, Malsch, Thieke, Huber & Bendl 2006, Lu, Chen, Olivera, Ruchala & Mackie 2004, Lu, Olivera, Chen, Ruchala, Haimerl, Meeks, Langen & Kupelian 2006, Xie, Chao & Xing 2008, Reed, Woodward, Zhang, Strom, Perkins, Tereffe, Oh, Yu, Bedrosian, Whitman, Bucholz & Dong 2008, Zhang, Chi, Meldolesi & Yan 2007, Pasquier, Lacornerie, Vermandel, Rosseeau, Lartigau & Betrouni 2007, Isambert, Dhermain, Bidault, Commowick, Bondiau, Malandain & Lefkopoulos 2008). Evaluation of these methods has been a persistent challenge as medical image segmentation unfortunately lacks a known ground truth, or gold standard, in its real world application. Phantoms provide an easily identifiable ground truth but are an unrealistic surrogate for patient imaging. The same can be said for synthetic images and cadaver sections. As noted by Warfield et al., the accuracy of a reference standard and the degree to which it reflects the clinical concerns are often inversely related. Accordingly, a single manual rater provides realistic data but can suffer from intra- and inter-rater variance. Recognizing the need for a useful reference standard, Warfield and colleagues introduced a method known as the simultaneous truth and performance level estimation (STAPLE) algorithm (Warfield, Zou & Wells 2004) to simulate a ground truth from a cohort of manual segmentations.

In addition to the absence of a known ground truth, evaluation methods have also lacked consensus as to comparison metrics. The choice of comparison metrics is quite important, as each yields different information and must be considered in the appropriate context. Generally, these measures fall into one of two categories: volume-based or distance-based. Measurement of nominal segmentation volume is a simple measure that does not require a reference standard for calculation, which makes it computationally inexpensive and allows for easy cross-study comparison with minimal background information. Spatial overlap measures such as the Dice similarity coefficient (DSC) (Dice 1945) and related Jaccard coefficient (1908) have been most broadly adopted in the literature in recent years. While these yield a good sense of volume overlap of two segmentations, they provide little in terms of the scale of mismatch (Crum, Camara & Hill 2006). Specificity and sensitivity are also commonly applied. Specificity, however, is plagued by its dependence on the number of true negatives; that is, the number of voxels in the image space not contained within the segmentation. This value may change quite considerably between studies simply as a function of image or region of interest size. A weakness of volume, DSC, and specificity and sensitivity, is that they are fairly insensitive to edge differences when those differences

have a small impact on overall volume. For example, two segmentations with large total volume may show a high degree of spatial overlap while exhibiting clinically relevant differences at their edges. Distance measures, however, such as the Hausdorff and Euclidean, or surface normal, distances offer yet another means of comparison by providing information regarding the differences in edges of two segmentations. The distance calculations generally result in a vector of distances that may be summarized as mean or median, or may be used in further statistical analyses. Thus, our experience has been that a combination of several volume and distance measures is required to gain a deep perspective of the dataset.

Our work is motivated by the observation that medical image segmentation is inherently a problem lacking a known ground truth. Accordingly, clinical evaluation studies should be behavioural in nature, employing a number of raters and patient volumes such as to provide good statistical power in the targeted clinical context. We designed a study to quantify variation amongst physicians in segmenting organs at risk in the brain and to assess our automated system in this context. Several other multiple observer studies have focused on evaluating automatic or semi-automatic systems within the brain (Bondiau, Malandain, Chanalet, Marcy, Habrand, Fauchon, Paquis, Courdi, Commowick, Rutten & Ayache 2005, Isambert et al. 2008, Babalola, Patenaude, Aljabar, Schnabel, Kennedy, Crum, Smith, Cootes, Jenkinson & Rueckert 2009) and head and neck (Chao, Xie & Xing 2008, Stapleford, Lawson, Perkins, Edelman, Davis, McDonald, Waller, Schreibmann & Fox 2010), but we know of no other study as comprehensive in terms of patient numbers, expert raters, and organs segmented. In addition, to be as clinically relevant as possible, we chose to conduct the study on volumes with large space-occupying lesions. We chose this anatomical site for the wealth of matched computed tomography (CT) and magnetic resonance (MR) imaging available, the clinical relevance to intensity-modulated radiation therapy (IMRT), as well the ubiquity in physician training in intracranial anatomy. We tested the hypothesis that the automatic system would produce segmentations that could serve as surrogates to the manual physician segmentations. An ancillary goal of this work was to collect a large and statistically robust dataset, which is useful for evaluating not only our algorithms but also those being developed by other groups. The recent release of several commercial radiotherapy segmentation systems underscores the need for a strong multi-rater data set for evaluation.

2. Methods

2.1. Study design

We selected 20 patients that had been previously treated in our department with IMRT for high grade gliomas. We chose difficult cases with large space-occupying tumors, often close to the critical structures, which would present a challenge for the non-rigid registration-based segmentation algorithm we use as well as yield pertinent dosimetry for the next phase of analysis. The mean gross tumor volume (GTV) and clinical tumor volumes (CTV) were 49 and 199 cm³, respectively. As a point of reference, these volumes roughly translate into a mean spherical equivalent of 4 and 7 cm in diameter. Each patient underwent stereotactic biopsy for which high resolution T1 MRs were acquired under 1.5T (N=10) or 3T (N=10) magnetic fields and reconstructed into image volumes of voxel size approximately 1×1×1.2 mm³. A helical CT of dimensions approximately 0.6–0.7 mm in the axial plane and either 2 mm (N=14) or 3 mm (N=6) in slice thickness was acquired for treatment planning.

Eight physicians were enlisted in this study as expert raters: 4 junior physicians (J₁–J₄) and 4 senior physicians (P₁–P₄). The senior physicians were comprised of 3 radiation oncologists and a diagnostic radiologist, while the junior physicians were PGY5 radiation oncology residents. Before initiating the study, we reviewed images and our atlas

delineations with them as a group to set general anatomical guidelines. One important guideline reiterated throughout the process was to set the inferior border of the brainstem at the foramen magnum, as the brainstem lacks a physical boundary with the the spinal cord. Another concern was where the brainstem meets the cerebellum in the lower pons. Here there is no significant contrasted boundary, so we developed an implicit rule whereby the experts should begin the contour anteriorly at the basilar sulcus of the pons, extend laterally to include the middle cerebellar peduncles, and continue posteriorly and medially toward the median sulcus of the fourth ventricle making an angle of approximately 45 degrees to the anterior-posterior axis.

The patient volumes were anonymized and loaded into a commercial treatment planning system (Eclipse version 8.5, Varian Medical Systems, Palo Alto, CA). This workstation was identical to the clinical systems in our department while reserved for research only. Computed tomography and MR images were registered within the planning system and fused. Each physician was given the opportunity to change window and level settings to his liking and received instructions to use all imaging information available to them to the point at which each felt confident delineating a critical structure. They were asked to delineate the brainstem, optic chiasm, optic nerves, and eyes. An in-house graphical user interface was constructed to inform the physicians where they stood in the task queue and to provide a mechanism to record time. The timing mechanism allowed the rater to pause momentarily or leave the system entirely and return later. Each rater was blinded to the work of the others. The delineations were collected over approximately one year.

Each physician was given all of the tools afforded by the clinical treatment planning system for contouring. A “paintbrush” tool produces an opaque segmentation as the expert traces out the structure. A “pencil” tool is similar without producing the opacity and can be used in a continuous or stepwise mode. There was also an “eraser” tool and the ability to stretch and deform contours after delineating. Three orthogonal views were present on screen at all times, though only axial were available for contouring. This is a limitation of the clinical software. We advised the experts to use the same tools they would use clinically and with which each was comfortable. We also advised them to inspect the final product of their work before completely the task. Lastly, above all we instructed the experts to perform these tasks in the context of real world clinical relevance.

The final result of each contouring session was a set of points in DICOMRT standard format that were stored at sub-voxel resolution.

2.2. Automatic segmentation

Two methods were utilized for the automatic segmentations in this study. The first method utilizes atlas-based registration (Crum, Hartkens & Hill 2004) to segment the eyes and the brainstem, while the second method utilizes a general technique we have developed for the segmentation of tubular organs, which we call the atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) (Noble, Warren, Labadie & Dawant 2008, Noble & Dawant 2009).

We first manually delineated the brainstem and the eyes in an atlas image. Then, a global affine registration was computed and used to register the atlas image (panel 1a, bottom row) onto the target image (panel 1a, top row) that we want to segment. A predefined bounding box around each organ is extracted from both the atlas and target image after the global affine registration (panel 1b). Another affine registration is performed locally between the extracted boxes of the atlas and target images, again resulting in a transformation that is used to project the atlas onto the target image. This second affine registration is performed to limit the registration on a local area within the image. The size of the boxes is determined

by the size and shape of the organ of interest within the atlas image, with an arbitrary amount of padding to aid in the local affine registration. This registration utilizes the normalized mutual information (NMI) (Studholme, Hill & Hawkes 1999) as the similarity measure. Lastly, local non-rigid registration is then performed between the results of the local affine registration and the atlas image. The manual contours drawn on the atlas are then projected onto the target image utilizing the deformation fields that were the result of the three registrations (panel 1c).

The non-rigid registration approach is an algorithm we termed the adaptive bases algorithm (ABA) (Rohde, Aldroubi & Dawant 2003). This algorithm uses normalized mutual information (Studholme et al. 1999) as the similarity measure and models the deformation field that registers the two images as a linear combination of radial basis functions (Wu 1995) with finite support.

Both the forward and the backward transformations are computed simultaneously, and the transformations are constrained to be inverses of each other using the method proposed by Burr (1981). Although this cannot be proven analytically, experience has shown that the inverse consistency error (Christensen & Johnson 2001) achieved with this approach is below the voxels' dimension. In our experience, enforcing inverse consistency improves the smoothness and regularity of the transformations.

In this work, we segment the optic nerves by applying the NOMAD algorithm. The NOMAD algorithm first computes the medial axis of the structure as the optimal path with respect to a cost function based on image and shape features. The medial axis is then expanded into the full structure using a level-set algorithm. Unlike other methods (Feng, Ip & Cheng 2004, Yim, Cebal, Mullick, Marcos & Choyke 2001), NOMAD uses a statistical model and image registration to provide the above segmentation framework with a priori, spatially varying intensity and shape information, thus accounting for unique local structure features. The statistical models were trained on volumes not included in this study.

In order to compensate for the lack and changing contrast of the structures, we take advantage of both the CT and MRI to build the models used by the algorithm. To ensure that the intensity information will consist of the best possible contrast, we rely solely on the CT in the region of the optic nerves, and solely on the MR in the region of the optic tracts and chiasm. The model consists of the set of points that compose the center line of the structure and their associated expected values for intensity and shape features extracted from the rigidly aligned MRs and CTs. Once the models are built, new sets of images can be segmented.

2.3. Calculation of simulated ground truths

We calculated two simulated ground truths for comparison to individual raters (P_1 - J_4) and our automatically generated segmentations, A_1 .

First, we used the STAPLE algorithm (Warfield et al. 2004) to calculate a consensus estimate from the physician segmentations. The STAPLE algorithm uses expectation-maximization to provide a probabilistic estimate of the underlying ground truth. It is designed to be robust to outliers within the input segmentation group. A second simulated ground truth was calculated through the creation of probability maps, termed p-maps (Meyer, Johnson, McLennan, Aberle, Kazerooni, MacMahon, Mullan, Yankelevitz, van Beek, Armato III, McNitt-Gray, Reeves, Gur & et al. 2006). A separate probability map was created for each rater across critical organ structures and patients to remove potential bias explicitly. The p-maps were created by summing the binary masks of each rater for a particular organ, omitting the rater for which the p-map will be used in comparison. For

example, the p-map for rater P1 would be formed by summing the 7 binary masks of raters P2-J4. The 3D array is then normalized to the number of raters included and smoothed using a 3×3 pixel Gaussian kernel applied in-plane with a standard deviation of 0.65 pixel width. The smoothing increases correlation between adjacent voxels, but it also improves the validity of later statistical tests that rely on assumptions of normality. We chose the filter parameters heuristically as a balance between reduction in gross quantization and an increase in spatial correlation between voxels. Additionally, we removed rater P2 from the p-maps, as an initial statistical analysis showed this rater produced several outliers within the complete dataset. The ground truth estimate was then created by thresholding the p-map at a desired level to form a binary mask. The choice of threshold level presents a challenge in using p-maps for ground truth estimation. A common interpretation is to choose a static, fixed value. For example, 0.5 would represent majority vote in which at least half of the raters agree. We chose to threshold at the mean value of the distributions, thus yielding a threshold specific to each p-map. That is, each voxel with a value greater than or equal to the mean of that p-map was included in the ground truth segmentation. While the mechanics of p-map creation and thresholding are identical for a static level, our method recognizes that rater consensus may vary considerably between structure and even between cases within structure. Another way to think about this is that the level of spatial independence within p-maps, an assumption violated for both STAPLE and p-map methods, varies over structures and cases. Choosing a static level such as simple majority vote, 0.5, assumes that value to be most representative of the group preferences over all structures and cases. However, in calculating a measure of central tendency we treat each scored voxel as a sampling distribution, and we take the mean of these sampling distributions as an appropriate level of consensus, thereby adjusting the level in response to the nature of the data.

We chose the STAPLE method as it is designed to produce a probabilistic ground truth estimate robust to deviations in rater performance. Use of STAPLE has become prevalent in segmentation evaluation work, and thus its inclusion herein should facilitate comparison with current and future work. While STAPLE was easily applicable to our imaging data, we also calculated the p-map-derived ground truth for its computational simplicity and the statistical value of the p-maps in future studies. We will refer to these simulated ground truths as STAPLE and PMAP_{mean}.

2.4. Comparison metrics

The data obtained in this study are most basically three-dimensional coordinate sets. To make judgments and draw conclusions about these data, we compare them using several metrics sensitive to different aspects of geometry. For this study we calculated two volumetric measures and one distance measure: volume, Dice similarity coefficient, and Euclidean distance from a simulated ground truth. The volume is calculated quite straightforwardly as the sum of the voxels contained within the binary mask of a segmentation multiplied by the voxel dimensions, which in our case were in CT space. The Dice similarity coefficient (DSC) has been used broadly in the field of segmentation as a measure of spatial overlap (Dice 1945, Jaccard 1908, Zijdenbos, Dawant, Margolin & Palmer 1994). The volumetric DSC is defined in equation 1 as the intersection of two masks normalized to their mean volume, where A and B are the masks and N is an operator yielding the number of voxels.

$$DSC = \frac{N(A \cap B)}{\frac{1}{2}(N(A) + N(B))} \quad (1)$$

Its range is [0,1] where zero indicates no overlap and 1 indicates exact overlap. Measures such as volume and DSC can be insensitive to differences in edges if these differences lead

to an overall small volumetric effect in relation to the total volume. The relative sensitivity of DSC to edge differences is a function of shape, or more explicitly the number of edge voxels in comparison to the number of inner voxels. For example, DSC will be more sensitive to edge variation in thin tubular structures such as the optic chiasm and nerves than in the brainstem and eyes, where the majority of voxels are not at the edges. Edge variation, however, could be quite important in a radiotherapy inverse planning context.

To gain information about differences at the edges of segmentations, we used the three-dimensional coordinates obtained from individual physician and automatic segmentations. We used these points to sample a distance map. The Euclidean distance map, or transform, is a pregenerated 3D array in which each voxel contains the value of the straight-line, or surface-normal, distance to the nearest non-zero voxel. We used $PMAP_{mean}$ as the source from which to calculate the distance maps. To determine the distribution of distances for an individual segmentation, the appropriate distance map was sampled at the contour points of the segmentation. This method yields a distance from each point drawn by a physician or the automatic system to the simulated ground truth. The distances were signed such that a rater's contour point lying inside the boundary of the ground truth was scored negative and outside scored positive. There are several ways to utilize distances. Often only the absolute distance from the ground truth is considered where direction is unimportant. In the context of radiotherapy we feel it is important to know whether a rater segments consistently small or consistently large as compared to the ground truth estimate. From this signed distribution of distances one can then do a number of things. We chose to generate boxplots of the distributions to get a sense of overall variability and understanding of whether there were instances of systematically positive or negative distances. We further used the absolute values of these distances to calculate true positive rates as a gauge of overall quality of segmentation.

It is important to recognize that this calculation provides information about where a rater made the decision to segment. It says nothing about where the rater decided to not segment. For example, we can imagine a simulated ground truth that extends for several axial slices of a CT image. A rater in question may draw an exact match to the simulated ground truth but on one slice only. The resulting distance distribution for this rater would be a vector of zeros, indicating that in every place the rater made a decision to draw a line, that decision was correct. The distance distribution says nothing regarding the failure of the rater to segment the other slices.

We chose two volume based measures coupled with the distance measure to provide more complete information about how segmentations differ. Alone each measure has a weakness. Volume and DSC tend to integrate edge differences that are small relative to the overall size of the segmentation. Meanwhile, the distance measure captures information only in the context of edges that were drawn, ignoring areas that a rater opted not to segment.

3. Results

Figures 2 and 3 present manual and automatic segmentations from a subject chosen randomly from the 20 patients used in this study. The eight physician-segmentations for the brainstem, chiasm, eyes, and optic nerves, can be seen in multiple colours, while the automatically generated segmentations are purple for each structure. The tumor volume is shown in red on the coronal slice. Figure 3 similarly presents axial contours of the brainstem and eyes, illustrating the variation that can be seen qualitatively between experts. We found this area of the brainstem at the cerebellar peduncles to be a consistent source of variation amongst the experts. The results we present here are an attempt to quantify the variation

geometrically such that we can make judgments about the expert and automatic segmentations, as well as the interaction of the two.

We calculated several quantitative measures to make comparisons between segmentations: volume, Dice similarity coefficient (DSC) and Euclidean distance from a simulated ground truth. Figures 4–7 use the boxplot to represent the results of these calculations. The boxplot presents the range of the distribution with a thin vertical line through the box. Dots above and below this line represent statistical outliers, or values outside 1.5 times the interquartile range. The thicker vertical line, the box, is bounded by the 25th and 75th percentiles of the distribution, and the median is shown via a short horizontal line. In these plots the automatic results are represented in the far left column labeled A_1 on the abscissa, followed by the senior physicians, P_1 – P_4 , and the junior physicians, J_1 – J_4 . In addition, when appropriate the two simulated ground truths are included and labeled S and P for STAPLE and $PMAP_{mean}$, respectively.

3.1. Volume

Figure 4 plots the volume distributions of the automatic, expert, and simulated ground truth segmentations for each of the six organs investigated. First, we note that physician P_2 segmented smaller structures than the others except in the case of the optic chiasm. The brainstem as segmented by P_2 was on average 40% smaller than the other physicians' segmentations with twice the coefficient of variation, 24%. The mean volumes [and 95% confidence intervals] across all physician segmentations were 25.88 [25.08, 26.70], 0.66 [0.60, 0.74], 8.5 [8.20, 8.73], 8.69 [8.40, 8.96], 0.88 [0.81, 0.94], and 0.87 [0.82, 0.92] cm^3 for the brainstem, optic chiasm, left and right eyes, and left and right optic nerves, while the automatic volumes were 23.99 [22.82, 24.87], 0.41 [0.39, 0.45], 9.00 [8.53, 9.42], 9.26 [8.65, 9.71], 0.64 [0.61, 0.68], and 0.63 [0.61, 0.67] cm^3 , respectively. The junior physicians as a group segmented larger structures than the senior physicians as a group. Although there were small differences in volume significant at the 5% level between the automatic structures and the physicians as a group, this difference disappears at an individual level. That is, the distribution of the automatic volumes falls within the variation of the individual physicians. It is clear, however, for the smaller tubular structures of the optic nerves and chiasm, the automatic structures were closer in volume to the smallest of physician segmentations. Additionally, the coefficient of variation of the automatic structures, 11–16%, was consistent across all organ structures. The individual physicians produced similar variation to the automatic system for the brainstem and eyes. For tubular structures, however, the physicians displayed more variation than the automatic segmentations, with coefficients of variation over the 20 patient cases ranging from 21–93% of mean structure volume.

Volumes for the two simulated ground truth segmentations were also calculated and can be seen in Figure 4. STAPLE consistently produced segmentations with larger volumes than the p-map derived method.

3.2. Dice similarity coefficient

The Dice similarity coefficient (DSC), a measure of volumetric overlap, was calculated and plotted in Figures 5 and 6. Each boxplot contains several columns representing distributions of non-redundant pairwise DSC comparisons for each of the raters. Figure 5 assesses inter-rater performance and variance. In the first column, A_1 , represents the distribution of DSC between automatic segmentations and individual physician segmentations. Columns P_1 – J_4 represent inter-physician comparisons: P_1 – P_4 senior and J_1 – J_4 junior physicians. Each distribution in these columns represents pair-wise comparisons of the expert in question to each of the other experts. The automatic segmentations are included only in the first column.

In this way we are able to gauge automatic performance in the context of all experts as well as inter-expert performance. Table 1 provides the mean DSC and 95% confidence intervals. The distributions of DSC are often skewed and depart from assumptions of normality required for statistical inference. To avoid making assumptions of the underlying population or transforming the data (Zou, Warfield, Bharatha, Tempany, Kaus, Haker, Wells III, Jolesz & Kikinis 2004), confidence intervals were calculated via bias corrected and accelerated bootstrap (Davison & Hinkley 1997) with 1000 replicates, about the mean DSC for each distribution plotted in figure 5. In individual comparisons only P_2 produced segmentations with mean DSC different from the other physicians and the automatic system. Additionally, we calculated the same statistic grouping the experts as a single group and as two groups representing senior and junior physicians. At the 5% significance level across all raters, cases, and organs, no difference exists between the mean DSC of the automatic segmentations and the physicians as a single group. The junior physicians and A_1 performed better than the senior at the 5% level, but the magnitude of the difference was small.

Figure 6 plots Dice coefficients against two ground truth estimations. The first two left most columns represent the distribution of Dice for the automatic segmentations compared to STAPLE and $PMAP_{mean}$, respectively. The same is plotted for the physician group in columns three and four. Lastly, the fifth column compares the two ground truth estimations. First, we note a high degree of overlap between STAPLE and $PMAP_{mean}$. Generally, the physician segmentations had a slightly higher spatial overlap with the ground truths than did the automatic system. However, the automatic system was more consistent, with smaller standard deviations and fewer outliers.

Figures 5 and 6 make essentially three types of comparisons: automatic-physician, physician-physician, and automatic- and physician-simulated ground truth. Another valuable comparison is that of individual groups to the simulated ground truths. For some structures, there was a small but significant ($p < 0.05$) difference between senior and junior physicians. This difference was almost entirely a result of P_2 as a member of the senior group. Looking across all the structures, the automatic segmentations produced a mean DSC against the simulated ground truths of 0.71 compared to 0.76 for the physicians. When decomposed into structures, again the biggest challenge was presented by the tubular chiasm and nerves, for both physicians and the automatic system. Whereas the mean for the brainstem and eyes was typically greater than 0.8, the chiasm and nerves were approximately 0.4 and 0.5, respectively. The tubular structures also had standard deviations on average over twice that of the brainstem and eyes.

3.3. Euclidean distance

Euclidean, or surface normal, distances were calculated in 3D between the segmentations and $PMAP_{mean}$. Signed distance maps for $PMAP_{mean}$ were pregenerated using an implementation of the algorithm proposed by Maurer et al. (2003) and then evaluated at the contour points of the automatic and physician segmentations. In Figure 7 the distances are signed to differentiate a contour point lying inside from a point outside $PMAP_{mean}$. Table 2 provides the minimum (furthest inside), mean, and maximum (furthest outside) distances for each structure averaged over the 20 patients. When the distance distribution is decomposed by structure, all raters had a mean distance between 0 and +2 mm except for P_2 's segmentation of the chiasm, which on average was 3 mm from the simulated ground truth. The average maximum distances (inside and outside) across the 20 cases ranged from -to +5.4 mm (inside and outside) for the automatic segmentations. The same for individual physicians ranged from -5.8 to +10.8, and when physicians are considered as a group, -3.9 to +7.5 mm.

Figure 8 plots the proportion of contour points that fall within 2 mm of the simulated ground truth as a function of rater and structure. This value can be thought of as the true positive rate, whereby any contour point drawn within a 2 mm shell of the simulated ground truth scores positive. The abscissa is partitioned by rater and structure, the ordinate is the 2 mm true positive rate, and the whiskers represent the 95% confidence interval on the proportion. This plot shows a broader variation amongst the physicians than within the automatic system. When we rank true positive rates, a senior physician, P₃, ranked the best overall and was the most consistent. The automatic system was second only to P₃ in terms of overall false positive rate and consistency.

3.4. Time

Segmentation time was recorded for each physician and is presented in Table 3. The average physician time-to-segment was 14.5 minutes with a standard deviation of 6.2 minutes. These times include only the task of segmenting the organs and explicitly exclude all time required to open the software or make adjustments before delineation began.

4. Discussion

In this work we desired to evaluate our automated segmentations in a real-world clinical study, to test the hypothesis that automatically segmented structures could serve as a surrogate to manual delineations. Accordingly, we designed a large study and chose a cohort of 20 challenging patient cases containing large space-occupying tumors, which are generally challenging for registration algorithms (Dawant, Hartmann, Pan & Gadumsetty 2002, Bach Cuadra, Pollo, Bardera, Cuisenaire, Villemure & Thiran 2004, Bach Caudra, De Craene, Duay, Macq, Pollo & Thiran 2006). To our knowledge no other clinically evaluative study of this scale has presented data on segmentation under these circumstances. In the absence of a well-defined or well-suited ground truth, Warfield (2004) and Meyer (2006) have presented alternatives. The Simultaneous truth and performance level estimation (STAPLE) algorithm produces a simulated ground truth from a cohort of expert delineations and can be compared directly with the automatic segmentations. STAPLE is a complex algorithm that has been shown to yield quality estimates of ground truth. However, early in our investigation we noted qualitatively that STAPLE could be influenced disproportionately by volumetrically larger segmentations within a group. Biancardi and colleagues (2010) have noted a similar phenomenon. To provide an additional basis of comparison, we simulated a second ground truth using the computationally simple concept of probability maps, which is analogous to the idea of voting rule. We chose to threshold the probability maps at a variable level, the non-zero mean of each probability distribution, to form the mask. Previously, Biancardi chose to threshold at fixed levels such as 0.5 or 0.75, which tended to produce consistently large or small estimates, respectively. Thresholding the p-maps at a static, predetermined level is problematic for two reasons. First, determination of a threshold level presents a challenge. A reasonable first choice is 50% as it is the threshold for majority vote. However, with a statistically small number of raters of unknown individual variance, 50% may not be reliable depending on whether false positives or false negatives are more important. This suggests that a threshold appropriate for one cohort of experts may not be appropriate for another cohort. Likewise, the same logic applies to different organ types. We believe our results show that consensus among experts is quite dependent on organ structure. In a large structure such as the brainstem we found significant areas over which 100% of the experts agreed, but in the optic chiasm and nerves such agreement was far rarer. Second, this method does not address the concern of spatial homogeneity. Both STAPLE and probability maps assume spatial independence of voxels. The STAPLE algorithm attempts to overcome deviations from this assumption through either incorporation of a priori information or using a Markov random field model. In our

method we recognize that adjacent voxels are correlated, and in fact we increase that correlation through Gaussian smoothing. The smoothing, however, helps achieve an approximately normal distribution of p-map values from which we calculate the mean probability as a threshold. Therefore, the appropriate threshold level will be unique to each cohort of expert segmentations and each structure. The end result shows that STAPLE and the probability map method produced ground truths of a high degree of spatial overlap (figure 6). However, while a full investigation of STAPLE was beyond the scope of this work, we did find that STAPLE produced volumetrically larger segmentations than the pmap method (figure 4).

In this work we used three principal metrics to characterize and compare segmentations: volume, the Dice similarity coefficient, and Euclidean distance calculated from a simulated ground truth. These measures offer several advantages. Volume is quite simple to calculate and stands alone, requiring no direct comparison to or use of a reference standard. The Dice similarity coefficient (DSC) is likely the most ubiquitous of metrics used in present literature. The Euclidean distances are particularly useful in the radiation therapy context, as their unit has implications to dose distributions and are well understood by the community. Volume, DSC, and Euclidean distance are invariant to image or mask size in terms of calculation, and thus do not suffer some of the pitfalls of specificity. A major goal of this work has been to provide a resource for others in algorithm assessment.

Each of the geometric measures showed the automatic segmentations to fall within the variation of the expert group, shown visually in boxplots (figures 4–7). Generally, there were few statistical differences between the automatic system and the ground truth estimations or the physicians as a group and the ground truth estimations, which were evaluated via bootstrapping 95% confidence intervals. The automatic system produced less variance than the physicians as a group over all the organs, and the magnitude of variance was more consistent across organs than within the physician group. This can be seen in figure 8, the 2 mm true positive rates.

Looking at individuals and groups within the larger physician group provides some trends. Junior physicians tend to segment volumetrically larger than their senior physician counterparts. We postulate this could result from a tendency to avoid risk of anatomically missing a portion of organ, while the more experienced physicians may be more confident in delineating a tighter border. We did not find, however, any evidence of reduced variance or higher spatial overlap in the senior physician group. In fact, one senior physician, P₂, was found to be different from the other physicians on all measures. A portion of this variance can be explained through the 2 mm true positive rates in figure 8. The 2 mm true positive rate for P₂ is low for most structures but ranks fifth of nine for the brainstem, which was grossly different as measured by volume and DSC. Upon closer inspection we found that for the brainstem this rater was inconsistent when marking inferior and superior extent of the organ, often not extending the slices as far in either direction as the rest of the group. This underscores the importance of choosing complementary metrics, as each examines a different aspect of geometry.

This work is not the first to evaluate automatic segmentation in the context radiation therapy organs at risk in the brain. Direct comparisons to other work are often compromised by the choice of metrics and differences in data acquisition. Bondiau and colleagues (2005) investigated atlas-based segmentation of the brainstem using MR images of 6 patients and 7 experts. Here we compare their observations to (our observations). Inter-expert volumes varied from 16.70 to 41.26 (8.82 to 35.89) cm³ across all cases. The mean expert delineations varied from 20.58 to 27.67 (19.66 to 29.15) cm³, and the automatic delineations varied from 17.75 to 24.54 (17.47 to 28.28) cm³ as a function of patient. Isambert (2008)

also segmented the brainstem, optic chiasm, optic nerves and eyes, for 11 patients against a single reference standard jointly delineated by a radiation oncologist and neurosurgeon. They concluded that automatic segmentation was well suited for organs greater than approximately 7 cm³, as they measured DSC above 0.8 for the eyes and chiasm, and concluded the small structures (DSC approximately 0.4) should be manually delineated by an expert. We noted a similarly low DSC for the chiasm in our study, though our optic nerves showed higher agreement of approximately 0.6 with respect to simulated ground truths. Though indeed spatial overlap is lower amongst the small tubular structures, we found that these structures are equally a challenge for the experts. In fact, in our study the automatically generated structures exceed the experts in some respects such as consistency, or robustness. This is seen in the variance of Dice index distribution (figure 6) of the automatic against the simulated ground truths, which is smaller than the physicians. The automatic system also scored near the top of the expert group with respect to the 2 mm true positive rates plotted in figure 8. Lastly, one must also consider the large variation in manual delineations for the optic nerves and chiasm reduces the accuracy of the ground truth produced with these contours.

4.1. Limitations and future work

We are undertaking a comprehensive clinical evaluation of our fully automatic segmentation system. Our experimental design is motivated by the following three observations. First, medical image segmentation is inherently a problem lacking a known ground truth. Accordingly, clinical evaluation studies should be behavioural in nature. Such a study requires a number of raters and patient volumes such as to provide good statistical power in the targeted clinical context. Second, realistically, the segmentation product of any automated system will require review and most likely modification by a qualified professional. Evaluation should characterize the impact of the modification process on efficiency, individual and group rater variance, and accuracy. Third, in the radiotherapy context organ segmentations are an important variable in a complex process culminating in the delivery of radiation dose to a patient. Traditional approaches to evaluation focus on the geometric properties of the resultant segmentations. While these are certainly the first and an important part of any evaluation, much value exists in understanding the impact of an automated system with respect to dosimetry.

There were several complementary goals in this work. The first was to evaluate our automatic segmentation methods in the brain on clinically relevant organs at risk. To our knowledge, this study is the largest and most robust that has been offered to date for such organs, specifically in the presence of large space-occupying brain lesions. Second, we hope that this work will provide a framework and a basis for comparison to others implementing similar algorithms. We emphasize the importance of using multiple complementary and easily reproducible metrics, as well as experimental designs that recognize the behavioural nature of human medical image segmentation.

Lastly, there are several limitations to the current study. First, we evaluate only our own algorithm for automatic segmentation. There are now scores of segmentation methods based on a seemingly equal number of algorithms and body sites. It is difficult to make comparisons to other algorithms without making those comparisons directly within the same dataset. Second, we implemented this investigation at only a single site with physicians who have often trained and work together, and accordingly, may be systematically biased in their understating of anatomy or manual delineation in general. This is in part a result of time and logistics as these studies are time intensive and costly. We spent over a year collecting the manual segmentations for this analysis. Third, we have made considerable effort to characterize inter-physician variance but have not evaluated intra-physician variance, which could be important in parsing variance into real differences and randomness. Lastly, we have

presented what we believe to be thorough though initial assessment of automatic segmentation within the context of radiation therapy. As these segmentations will undoubtedly be reviewed and modified by physicians in clinical practice, it is important to understand the impact of such a process on the workflow, consistency, and accuracy of segmentation as well as the final planned dose distribution.

Acknowledgments

This work was funded in part through a grant from the National Institute of Biomedical Imaging and Bioengineering (award R01EB006193). The authors would like to acknowledge Kenneth Lewis and Charles Coffey for their advice and suggestions through numerous discussions on methodology, as well as Dennis Duggan who was instrumental in initiating this study. In addition we acknowledge the Department of Radiation Oncology at Vanderbilt-Ingram Cancer Center for providing the images and clinical systems utilized.

References

- Babalola K, Patenaude B, Aljabar P, Schnabel J, Kennedy D, Crum W, Smith S, Cootes T, Jenkinson M, Rueckert D. An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *Neuroimage*. 2009; 4:1435–1447. [PubMed: 19463960]
- Bach Caudra M, De Craene M, Duay V, Macq B, Pollo C, Thiran J. Dense deformation field estimation for atlas-based segmentation of pathological MR brain images. *Compt. Methods Programs Biomed*. 2006; 84:66–75.
- Bach Cuadra M, Pollo C, Bardera A, Cuisenaire O, Villemure J, Thiran J. Atlas-based segmentation of pathological MR brain images using a model of lesion growth. *IEEE Trans. Med. Imag*. 2004; 23:1301–1314.
- Biancardi A, Jirapatnakul A, Reeves A. A comparison of ground truth estimations. *Int. J. Comput. Assist. Radiol. Surg*. 2010; 5:295–305. [PubMed: 20033494]
- Bondiau P, Malandain G, Chanalet S, Marcy P, Habrand J, Fauchon F, Paquis P, Courdi A, Commowick O, Rutten I, Ayache N. Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. *Int J Radiat Oncol Biol Phys*. 2005; 61(1):289–298. [PubMed: 15629622]
- Burr D. A dynamic model for image registration. *Comput. Graph. Image Process*. 1981; 15:102–112.
- Chao M, Xie Y, Xing L. Auto-propagation of contours for adaptive prostate radiation therapy. *Phys Med Biol*. 2008; 53(17):4533–4542. [PubMed: 18677041]
- Christensen G, Johnson H. Consistent image registration. *IEEE Trans. Med. Imag*. 2001; 20:568–582.
- Crum W, Camara O, Hill D. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging*. 2006; 25(11):1451–1461. [PubMed: 17117774]
- Crum W, Hartkens T, Hill D. Non-rigid image registration: theory and practice. *Br J Radiol*. 2004; 77(Spec 2):S140–S153. [PubMed: 15677356]
- Davison, A.; Hinkley, D. *Bootstrap methods and their applications*. Cambridge University Press; 1997.
- Dawant B, Hartmann S, Pan S, Gadamsetty S. Brain atlas deformation in the presence of small and large space-occupying tumors. *Comput. Aided Surg*. 2002; 7:1–10. [PubMed: 12173876]
- Dice L. Measures of the amount of ecologic association between species. *Ecology*. 1945; 26(3):297–302.
- Feng, J.; Ip, H.; Cheng, S. A 3d geometric deformable model for tubular structure segmentation. In: Chen, Y-PP., editor. *10th International Multimedia Modeling Conference*; IEEE Computer Society; 2004.
- Gorthi S, Duay V, Houhou N, Bach Cuadra M, Schick U, Becker M, Allal A, Thiran J. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. *IEEE Journal of Selected Topics in Signal Processing*. 2009; 3:135–147.
- Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau P, Malandain G, Lefkopoulos D. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiother Oncol*. 2008; 87(1):93–99. [PubMed: 18155791]

- Jaccard P. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*. 1908; 44:223–270.
- Lu W, Chen M, Olivera G, Ruchala K, Mackie T. Fast free-form deformable registration via calculus of variations. *Phys Med Biol*. 2004; 49(14):3067–3087. [PubMed: 15357182]
- Lu W, Olivera G, Chen Q, Ruchala K, Haimel J, Meeks S, Langen K, Kupelian P. Deformable registration of the planning image (kVCT) and the daily images (MVCT) for adaptive radiation therapy. *Phys Med Biol*. 2006; 51(17):4357–4374. [PubMed: 16912386]
- Malsch U, Thieke C, Huber P, Bendl R. An enhanced block matching algorithm for fast elastic registration in adaptive therapy. *Phys. Med. Biol*. 2006; 51:4789–4806. [PubMed: 16985271]
- Maurer C, Rensheng Q, Raghavan V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans. Pattern Anal. Mach. Intell*. 2003; 25(2):265–270.
- Mell L, Mehrotra A, Mundt A. Intensity-modulated radiation therapy use in the U.S. 2004. *Cancer*. 2005; 104:1296–1303. [PubMed: 16078260]
- Mell L, Roeske J, Mundt A. A survey of intensity-modulated radiation therapy use in the United States. *Cancer*. 2003; 98:204–211. [PubMed: 12833473]
- Meyer C, Johnson T, McLennan D, Aberle D, Kazerooni E, MacMahon H, Mullan B, Yankelevitz D, van Beek J, Armato S III, McNitt-Gray M, Reeves A, Gur D, et al. Evaluation of lung MDCT nodule annotation across radiologists and methods. *Acad Radiol*. 2006; 13(10):1254–1265. [PubMed: 16979075]
- Noble J, Dawant B. Automatic segmentation of the optic nerves and chiasm in ct and mr using the atlas-navigated optimal medial axis and deformable-model algorithm. *Proceedings of the SPIE: Medical Imaging*. 2009
- Noble J, Warren F, Labadie R, Dawant B. Automatic segmentation of the facial nerve and chorda tympani in ct images using spatially dependent feature values. *Med. Phys*. 2008; 35:5375–5384. [PubMed: 19175097]
- Pasquier D, Lacornerie T, Vermandel M, Rosseeau J, Lartigau E, Betrouni N. Automatic segmentation of pelvic structures from magnetic resonance images for prostate cancer radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys*. 2007; 68:592–600. [PubMed: 17498571]
- Reed V, Woodward W, Zhang L, Strom E, Perkins G, Tereffe W, Oh J, Yu T, Bedrosian I, Whitman G, Bucholz T, Dong L. Automatic segmentation of whole breast using atlas approach and deformable image registration. *Int. J. Radiat. Oncol. Biol. Phys*. 2008; 73:1493–1500. [PubMed: 18804333]
- Rohde G, Aldroubi A, Dawant B. The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Trans Med Imaging*. 2003; 22(11):1470–1479. [PubMed: 14606680]
- Stapleford L, Lawson J, Perkins C, Edelman S, Davis L, McDonald M, Waller A, Schreibmann E, Fox T. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *Int J Radiat Oncol Biol Phys*. 2010; 77:959–966. [PubMed: 20231069]
- Studholme C, Hill D, Hawkes D. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*. 1999; 32:71–86.
- Warfield S, Zou K, Wells W. Simultaneous Truth and Performance Level Estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004; 23(7):903–921. [PubMed: 15250643]
- Wu Z. Compactly supported positive definite radial functions. *Comput. Math*. 1995; 4:283–292.
- Xie Y, Chao M, Xing L. Feature-based rectal contour propagation from planning CT to cone beam CT. *Med Phys*. 2008; 35(10):4450–4459. [PubMed: 18975692]
- Yim P, Cebal J, Mullick R, Marcos H, Choyke P. Vessel surface reconstruction with a tubular deformable model. *IEEE Trans. Med. Imag*. 2001; 20:1411–1421.
- Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: toward on-line adaptive radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys*. 2007; 68:522–530. [PubMed: 17418960]
- Zijdenbos A, Dawant B, Margolin A, Palmer A. Morphometric analysis of white matter lesions in MR images: Method and validation. *IEEE Trans Med Imag*. 1994; 13(4):716–724.

Zou K, Warfield S, Bharatha A, Tempany C, Kaus M, Haker S, Wells W III, Jolesz F, Kikinis R.
Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol.*
2004; 11(2):178–189. [PubMed: 14974593]

Figure 1.

Atlas-based segmentation process for the brainstem and eyes. Panel (a): Orthogonal slices of a patient (top row) with a large right sided lesion and the atlas (bottom row) before registration. Panel (b): Volumes are then globally, affinely registered, and a bounded atlas region (white box) is projected onto the patient. Panel (c): Local affine and local non-rigid registration are performed on the bounded region where the top row represents the final product of the patient brainstem deformed to the atlas.

Figure 2.

A randomly chosen patient from the 20 cases used in this study. Eight physician raters segmented the brainstem, optic chiasm, eyes, and optic nerves using a fused CT/MR image set. The automatically generated segmentations are shown in purple. The large red contour in the right parietal is the gross tumor volume.

Figure 3.

Axial slice showing an area of high physician variability within the brainstem. In this area of the cerebellar peduncles there is little anatomical contrast, such that the physicians rely primarily on implicit knowledge. The automatic contour is represented in purple.

Figure 4.

Volume [cm^3] for the automatic (A_1), senior physician (P_1 – P_4), junior physician (J_1 – J_4), and simulated ground truth, STAPLE (S) and $\text{PMAP}_{\text{mean}}$ (P) segmentations. The horizontal line through each box indicates the median of the volume distribution while the rectangular box represents the interquartile range. Small dots are outliers for the distribution.

Figure 5.

Dice similarity coefficients across the 20 patients per structure to assess inter-rater performance and variance. Columns P₁–J₄ plot inter-physician comparisons: P₁–P₄ senior and J₁–J₄ junior physicians. Each distribution in these columns is comprised of pair-wise comparisons of the expert in question to each of the other experts. The automatic segmentations are included only in the first column.

Figure 6. Dice similarity coefficients for each rater group with respect to the simulated ground truths. The first two columns from the left compare A_1 to STAPLE (S) and $PMAP_{mean}$ (P), followed by comparison with the physician group, followed by comparison between S and P in the far right column.

Figure 7. Distance (mm) distributions from rater segmentations to $PMAP_{mean}$. Positive distances indicate a contour point lying outside the ground truth segmentation while negative distances indicate a contour point lying within the ground truth.

Figure 8.

True positive rate of contour points drawn within a 2 mm shell around the simulated ground truth. The abscissa is partitioned by rater and structure, the ordinate is the 2 mm true positive rate, and the whiskers represent the 95% confidence interval on the proportion.

Table 1

Mean Dice similarity coefficients (DSC) with 95% confidence interval and standard deviation. Row A₁ gives the mean, 95% CI, and standard deviation on the distribution of non-redundant pairwise DSC comparisons of A₁ versus each expert rater. Similarly, rows P₁-J₄ provide the statistics for physician-physician comparison (A₁ not included in these comparisons). The final three rows provide the same statistics grouping the physicians as senior, junior, or as one group including all experts.

Rater	All structures				Brainstem				Chiasm				Eyes				Optic nerves			
	Mean	Mean CI	std		Mean	Mean CI	std		Mean	Mean CI	std		Mean	Mean CI	std		Mean	Mean CI	std	
A ₁	0.656	0.642	0.671	0.223	0.830	0.818	0.839	0.064	0.374	0.345	0.403	0.179	0.843	0.831	0.853	0.070	0.523	0.501	0.544	0.138
P ₁	0.647	0.626	0.664	0.252	0.845	0.828	0.856	0.082	0.400	0.360	0.439	0.235	0.836	0.820	0.851	0.095	0.482	0.450	0.515	0.193
P ₂	0.543	0.528	0.560	0.248	0.691	0.670	0.706	0.105	0.253	0.218	0.293	0.225	0.754	0.734	0.770	0.102	0.403	0.377	0.429	0.162
P ₃	0.666	0.652	0.681	0.245	0.856	0.839	0.866	0.078	0.344	0.308	0.379	0.219	0.855	0.838	0.868	0.090	0.543	0.517	0.567	0.158
P ₄	0.686	0.669	0.700	0.223	0.855	0.840	0.867	0.084	0.437	0.407	0.470	0.194	0.851	0.835	0.862	0.083	0.560	0.525	0.587	0.178
J ₁	0.683	0.668	0.697	0.224	0.843	0.826	0.857	0.091	0.414	0.377	0.446	0.193	0.860	0.845	0.872	0.085	0.560	0.532	0.586	0.161
J ₂	0.667	0.651	0.683	0.223	0.851	0.837	0.862	0.078	0.416	0.381	0.455	0.239	0.818	0.800	0.834	0.099	0.549	0.523	0.572	0.149
J ₃	0.652	0.638	0.668	0.222	0.836	0.820	0.847	0.080	0.447	0.409	0.478	0.209	0.824	0.812	0.836	0.073	0.490	0.466	0.514	0.154
J ₄	0.626	0.609	0.643	0.255	0.822	0.806	0.834	0.078	0.423	0.389	0.461	0.231	0.849	0.834	0.861	0.077	0.407	0.381	0.431	0.150
All senior	0.619	0.606	0.633	0.256	0.798	0.781	0.813	0.122	0.322	0.298	0.353	0.224	0.810	0.799	0.819	0.110	0.486	0.468	0.503	0.195
All junior	0.669	0.659	0.680	0.219	0.859	0.855	0.863	0.034	0.476	0.450	0.499	0.192	0.842	0.834	0.848	0.076	0.497	0.484	0.511	0.152
All experts	0.646	0.641	0.652	0.241	0.825	0.819	0.83	0.099	0.392	0.378	0.405	0.226	0.831	0.827	0.835	0.094	0.499	0.493	0.507	0.174

Table 2

Distances [mm] to PMAP_{mean} simulated ground truth for each rater or rater group. The positive direction is outward looking from the ground truth while the negative direction is inward looking.

	Brainstem			Chiasm			Eyes			Nerves		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A ₁	-4.30	0.15	5.39	-2.39	0.04	2.95	-2.33	0.82	3.81	-2.65	-0.40	2.41
P ₁	-5.70	0.40	8.06	-2.26	-0.23	2.46	-3.35	0.14	2.90	-2.29	0.69	5.43
P ₂	-5.84	-0.36	5.40	-0.65	3.13	8.46	-5.43	-1.07	2.74	-2.64	-0.41	2.56
P ₃	-4.12	0.43	5.31	-2.39	-1.47	0.29	-1.72	1.04	3.19	-2.24	-0.09	5.17
P ₄	-2.77	0.88	6.64	-2.25	0.62	4.16	-1.71	1.18	3.33	-2.19	0.07	2.91
J ₁	-2.50	1.48	8.28	-2.03	1.89	8.86	-1.85	0.79	3.24	-2.10	0.25	2.83
J ₂	-3.70	0.61	7.43	-1.99	0.13	3.56	-2.25	1.41	4.81	-2.20	0.59	5.27
J ₃	-2.95	1.10	8.26	-2.15	0.50	4.70	-3.64	0.08	3.90	-2.36	0.76	4.52
J ₄	-3.39	1.60	10.78	-2.06	1.24	4.79	-2.49	0.51	3.28	-2.39	0.13	3.17
All senior	-4.69	0.34	6.45	-1.86	0.55	3.92	-3.04	0.33	3.07	-2.36	0.05	3.89
All junior	-3.14	1.20	8.69	-2.06	0.94	5.48	-2.56	0.70	3.81	-2.26	0.43	3.95
All experts	-3.87	0.77	7.52	-1.97	0.73	4.66	-2.80	0.51	3.42	-2.30	0.25	3.98

Table 3

Mean and standard deviation of segmentation times for the physician raters.

	Time [minutes]	
	Mean	std
P ₁	9.6	2.5
P ₂	14.1	4.5
P ₃	19.8	3.4
P ₄	21.1	6.2
J ₁	18.8	3.3
J ₂	14.8	4.4
J ₃	6.6	1.4
J ₄	11.1	3.0
All experts	14.5	6.2