# STREAMLINE.
# OPTIMIZE.
# TRUST.

ADAPTIVO | LINACVIEW | DOSEVIEW | PIPSPRO | QA PILOT | IMSURE | STRUCTSURE AI QA

## COMPLETE INTEGRATED QA

STANDARD IMAGING®

CLICK HERE TO LEARN MORE

# Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015

Patrik F. Raudaschl[a)]
*Department of Biomedical Computer Science and Mechatronics, Institute for Biomedical Image Analysis, UMIT, Hall, Tyrol 6060, Austria*

Paolo Zaffino
*Department of Experimental and Clinical Medicine, Magna Graecia University of Catanzaro, Catanzaro 88100, Italy*

Gregory C. Sharp
*Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA*

Maria Francesca Spadea
*Department of Experimental and Clinical Medicine, Magna Graecia University of Catanzaro, Catanzaro 88100, Italy*

Antong Chen
*Merck and Co., Inc., West Point, PA 19422, USA*

Benoit M. Dawant
*Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA*

Thomas Albrecht and Tobias Gass
*Varian Medical Systems, Baden 5404, Switzerland*

Christoph Langguth and Marcel Lüthi
*University of Basel, Basel 4001, Switzerland*

Florian Jung, Oliver Knapp, and Stefan Wesarg
*Fraunhofer, 64283 Darmstadt, Germany*

Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, and Graham Vincent
*Imorphics Ltd., Kilburn House, Manchester Science Park, Manchester, M15 6SE, UK*

Mauricio Orbes-Arteaga, David Cárdenas-Peña, and German Castellanos-Dominguez
*Signal Processing and Recognition Group, Universidad Nacional de Colombia, Colombia*

Nava Aghdasi, Yangming Li, Angelique Berens, Kris Moe, and Blake Hannaford
*University of Washington, Seattle, WA 98105, USA*

Rainer Schubert and Karl D. Fritscher
*Department of Biomedical Computer Science and Mechatronics, Institute for Biomedical Image Analysis, UMIT, Hall, Tyrol 6060, Austria*

(Received 19 May 2016; revised 13 October 2016; accepted for publication 22 February 2017; published 21 April 2017)

**Purpose:** Automated delineation of structures and organs is a key step in medical imaging. However, due to the large number and diversity of structures and the large variety of segmentation algorithms, a consensus is lacking as to which automated segmentation method works best for certain applications. Segmentation challenges are a good approach for unbiased evaluation and comparison of segmentation algorithms.
**Methods:** In this work, we describe and present the results of the Head and Neck Auto-Segmentation Challenge 2015, a satellite event at the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2015 conference. Six teams participated in a challenge to segment nine structures in the head and neck region of CT images: brainstem, mandible, chiasm, bilateral optic nerves, bilateral parotid glands, and bilateral submandibular glands.
**Results:** This paper presents the quantitative results of this challenge using multiple established error metrics and a well-defined ranking system. The strengths and weaknesses of the different auto-segmentation approaches are analyzed and discussed.
**Conclusions:** The Head and Neck Auto-Segmentation Challenge 2015 was a good opportunity to assess the current state-of-the-art in segmentation of organs at risk for radiotherapy treatment. Participating teams had the possibility to compare their approaches to other methods under unbiased and standardized circumstances. The results demonstrate a clear tendency toward more general purpose and fewer structure-specific segmentation algorithms.  [https://doi.org/10.1002/mp.12197]

Key words: atlas-based segmentation, automated segmentation, model-based segmentation, segmentation challenge

## 1. INTRODUCTION

Advances in medical imaging have greatly improved radiation oncology. Treatment planning uses computer tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) imaging, while treatment delivery achieves high accuracy through image-guided (adaptive) radiotherapy, i.e., IG(A)RT, advanced treatment planning, and sophisticated delivery techniques. Precise control of the planning target volume and avoidance areas allows high radiation dosage to be targeted to smaller volumes, increasing the sparing of healthy tissues[1] and reducing the risk for radiation-induced secondary malignancies.[2,3]

Delineation of target structures (e.g., tumor) and organs at risk (OARs) is one key step during the treatment planning process. Because manual segmentation of these structures is challenging and time-consuming, developing accurate automated segmentation methods is crucial to aid pretreatment radiotherapy planning and IGART. In recent years, a variety of automated segmentation approaches have been introduced. However, a consensus is lacking as to which segmentation approach is best. This may be due to the large number and variety of anatomic structures, each presenting specific challenges. Indeed, some automated segmentation approaches are designed for a specific region or modality, and may be more accurate in one domain and less accurate in others.

The application of automated segmentation in clinical practice can benefit from an evaluation and comparison of different segmentation approaches. When researchers test a segmentation method using a proprietary dataset, comparison with other methods is not easy, even if the same structures and the same modalities are used. Several projects have the aim to overcome this problem. Previous studies[4,5] evaluated multiple segmentation approaches on the same dataset in order to compare them objectively. However, a fair evaluation of each approach is still difficult, as most segmentation algorithms need a considerable amount of expert knowledge to achieve optimal performance.

Another approach for the unbiased evaluation and comparison of segmentation algorithms is to conduct a "challenge". In such an event, all participating teams use the same training and testing datasets to evaluate their algorithms. Aside from increasing the validity of the evaluation, such an event also allows an efficient evaluation and comparison of different segmentation approaches, performed by an impartial third party. The nontrivial problem of optimal parameterization is solved because the algorithm developers are also responsible for parameter selection and parameter optimization.

Previous challenges have demonstrated the potential to offer a great and powerful opportunity for the evaluation of registration and segmentation algorithms. The competitive character of this approach attracts participants of the best research groups and companies. In addition, these events stimulate new algorithm development and promote scientific discussions among participants. In the course of the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2007 conference, a liver segmentation challenge was conducted.[6] A similar challenge was performed in MICCAI 2009 for prostate segmentation.[7] In MICCAI 2009 and 2010, Head and Neck Auto-segmentation Challenges were performed for the evaluation of segmentation algorithms.[8,9] In the 2010 Head and Neck Auto-segmentation challenge organized by Pekar et al.,[9] 25 anonymized datasets were provided to evaluate the performance of automatic segmentation approaches of six participating teams for the left and right parotid glands. At the segmentation challenge of 2009, five participating teams used the same data for segmenting mandible and brainstem.[8]

This paper presents the results from the Head and Neck Auto-Segmentation Challenge 2015 held as a satellite event of MICCAI 2015 in Munich, Germany. The aim of the challenge was the evaluation of state-of-the-art automatic segmentation approaches for the head–neck area under standardized conditions in CT images. No manual interventions were allowed, except of the specification of seed points. Compared to previous segmentation challenges, a higher number of OARs (brainstem, mandible, chiasm, and bilateral optic nerves, parotid glands and submandibular glands) in the head and neck region was selected as target for segmentation. As in previous challenges, treatment target delineation was not covered in this study. The performance of automatic segmentation approaches for all structures was determined by using a common set of well-established metrics. The approaches presented in this work summarize the results of all submissions to the challenge. The most common approaches for automated segmentation in current practice,[10] model- and atlas-based segmentation (ABS) approaches, were used by multiple teams. In addition, algorithmic pipelines which combine statistical shape/appearance models (SSM/SAM) with ABS were used in this challenge. The most commonly used commercial treatment planning systems are also using model- and/or atlas-based approaches for segmentation (Table I).[10] Hence, this challenge can be seen as a snapshot of the state-of-the-art. The presented segmentation algorithms are not actually the same algorithms as used in the treatment planning systems shown in Table I, but they are of the same type.

The paper is organized as follows: In Section 2, the characteristics of the datasets for the challenge are described in detail. Section 3 elucidates the organization of the challenge, while Section 4 outlines the evaluation process. Section 5 introduces the participants and briefly describes their specific segmentation algorithms. Quantitative segmentation results are presented in Section 6. Furthermore, in this section, specific segmentation results are visualized and discussed in

TABLE I. Commercial software tools for automated image segmentation in the head and neck region.[10]

| Vendor | Product Name | Segmentation Approach | Reference |
|---|---|---|---|
| Varian | Eclipse (smart segmentation) | Atlas-based | [11] |
| MIM software | MIM Maestro 6+ | Atlas-based | [12] |
| Velocity | VelocityAI 3.0.1 | Atlas-based | [13] |
| BrainLab | iPlan | Atlas-based | [14] |
| Dosisoft | IMAgo | Atlas-based | [15] |
| Mirada | RTx 1.4, workflow box | Atlas-based | [16] |
| OSL | OnQ RTS | Atlas-based | [17] |
| Elekta | ABAS 2.01 | Atlas- and model-based | [18] |
| Philips | SPICE 9.8 | Atlas- and model-based | [19] |
| RaySearch | RayStation 4.0 | Atlas- and model-based | [20] |

order to emphasize interesting findings. Based on the results of this challenge, in Section 7 fundamental pros and cons of different segmentation approaches are discussed. Finally, the paper ends with a conclusion followed by a short outlook in Section 8.

## 2. DATA

The imaging data for this segmentation challenge are publically available via the Cancer Imaging Archive (TCIA).[21] Originally the data come from the RTOG 0522 clinical trial by Ang et al.[22] with treatment planning CT scans of 111 patients available. For this challenge, a subset of 40 images was used: 25 images were used as training data, 10 images were used for off-site testing, and 5 images were used for on-site testing. The subset was chosen to ensure that all structures were completely included within the CT images, image quality was adequate, and that structures minimally overlapped tumor volumes. No restriction with respect to age or gender was made.

### 2.A. Characteristics of image data

CT images and manual contouring data were provided. For all data, the reconstruction matrix was 512 × 512 pixels. The in-plane pixel spacing was isotropic, and varied between 0.76 mm × 0.76 mm and 1.27 mm × 1.27 mm. The number of slices was in the range of 110–190 slices. The spacing in z-direction was between 1.25 mm and 3 mm.

### 2.B. Manual delineation of target structures

Nine anatomical structures in the head and neck region were used as target for segmentation: brainstem, optic chiasm (OC), mandible, bilateral optic nerves (ONs), bilateral parotid glands (PGs), and bilateral submandibular glands (SGs). All structures are highly relevant OARs for radiation therapy treatment in the head and neck.[23] Although some of these OARs were delineated on some of the CT images for the clinical trial, all structures used for the challenge were resegmented by experts to provide uniform quality and consistency. Segmentation guidelines were developed by performing an extensive literature research. Manual delineations were provided as binary labels (value of "1" for inside and "0" for outside). A summary of these guidelines follows, and more detailed description can be found in Ref (24).

#### 2.B.1. Brainstem

The segmentation protocol for the brainstem follows specific recommendations of RTOG protocols 0920 and 1216.[25,26] In radiation oncology, it is common to truncate the inferior and superior brainstem borders at a discrete axial slice, instead of tilting the boundary across several axial slices. The inferior border of the brainstem was located at the top of C1 vertebra and the superior was located at the top slice containing the posterior clinoid.

#### 2.B.2. Optic chiasm and optic nerve

The ON was contoured from the posterior of the retina, through the optic canal up to the OC. Because there are no anatomical boundaries for the anterior and posterior part of the OC, artificial boundaries were defined. The boundary between the ON and OC was defined by a virtual line between the anterior clinoid process and the tuberculum sellae. A short portion of the optic tract posterior to the chiasm is included in the contour, truncated to a length of 8 mm beginning at an imaginary line connecting the lateral boundary of the ipsilateral ON with the contralateral optic tract. More details can be found in Refs (27–29)

#### 2.B.3. Mandible

The mandible is the largest bone in the human head. It forms the lower jaw and locates the lower teeth. It was contoured starting from the bottom (chin area) and finishing at the mandible conoid processes and condyles. Particular attention was set on the discrimination of the boundary between bone and teeth.

#### 2.B.4. Parotid glands

The paired PGs are the major salivary glands located below the ears. Contouring of the PG follows the guidelines of van de Water et al.[30] Several nerves and blood vessels pass through the PG, including branches of the facial nerve, external artery, and retromandibular vein. These vessels are included in the contour when they are contained within the enclosing envelope of the PG.

#### 2.B.5. Submandibular glands

SGs are also paired salivary glands located beneath the floor of the mouth. They were delineated according to guidelines defined by van de Water et al.[30]

### 2.C. Quality assurance of manual delineations

Three different medical imaging experts performed the segmentation of the datasets, with each structure segmented by the same observer for all 40 datasets. To ensure that the manually delineated structures were correctly and consistently segmented, a quality assessment was performed. For this purpose, a medical doctor checked all segmentations of each structure and recommended modifications until all structures adhered to the segmentation guidelines.

## 3. CHALLENGE ORGANIZATION

Being a satellite event of the MICCAI conference, the Head and Neck Auto-Segmentation Challenge 2015 was launched in June 2015 in the form of announcements via several mailing lists and a dedicated website.[31] The data as described in Section 2 were available for download via the website. In addition to the images and labels, detailed information about the segmentation guidelines used for manual delineation was also provided.[24] The segmentation challenge was divided into two phases:

### 3.A. Phase 1 ("off-site phase")

The participants downloaded a training dataset (25 labeled images) and a testing dataset (10 unlabeled images). The usage of the training datasets was not mandatory. The segmentation results for the test dataset had to be submitted by September 11, 2015. In addition, all participating teams had to submit a (short) manuscript, which described the approach that was used for the challenge. Submissions ranging 2–8 pages were accepted.

### 3.B. Phase 2 ("on-site phase")

The second phase of the challenge was organized as a satellite event of the MICCAI conference in Munich on October 9, 2015. During this event, all participating teams had to segment five new test images within 2 hours. The datasets were provided via thumb drive. For participating teams which were not able to attend the challenge on-site, or required off-site computational resources, the dataset was also distributed via a download link on the challenge website. Teams which were not able to finish the segmentation of the five additional test datasets on time were allowed to submit their results within 3 days after the challenge. Results submitted after the challenge are denoted so in the results section (Section 6). In addition, each participating team gave a 12-min presentation as part of the MICCAI satellite event.

## 4. EVALUATION METHODS

Evaluation of the segmentation approaches was performed separately for off-site and on-site segmentation results, and segmentation accuracy was assessed independently for all structures. An overall performance for all teams was computed by summation of the individual ranks for each structure (see section 4.B).

### 4.A. Evaluation Metrics

For the evaluation of the segmentation performance, four different metrics were used: Dice similarity coefficient, 95% Hausdorff distance (HD), maximum HD, and contour mean distance. These are the most common used metrics for evaluating 3D medical image segmentations and include volume- and overlap-based metric types.[32] Multiple metrics are used because different metrics reflect different types of errors.[33] For example, when segmentations are small, distance-based metrics such as HD are recommended over overlap-based metrics such as Dice coefficient. Overlap-based metrics are recommended if volume-based statistics are important.[32] In the following, the metrics used are described in more detail:

- The Dice coefficient measures the volumetric overlap between the automatic and manual segmentation. It is defined as:[34,35]

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

A and B are the labeled regions that are compared and |.| is the volume of a region. The Dice coefficient can have values between 0 (no overlap) and 1 (complete overlap).

- The maximum HD measures the maximum distance of a point in a set A to the nearest point in a second set B. Commonly it is defined as:[36]

$$\begin{aligned} HD(A,B) &= \max\Big(h(A,B), h(B,A)\Big) \\ h(A,B) &= \max_{a\in A} \min_{b\in B} \|a-b\| \end{aligned} \tag{2}$$

||.|| is the Euclidean distance, a and b are points on the boundary of A and B, and h (A, B) is often called the directed HD. It should be mentioned that maximum HD is sensitive to outliers but appropriate for nonsolid segmentations.[32]

- The 95% HD is similar to maximum HD. However, in contrast to maximum HD, 95% HD is based on the calculation of the 95th percentile of the distances between boundary points in A and B. The purpose for using this metric is to eliminate the impact of a very small subset of inaccurate segmentations on the evaluation of the overall segmentation quality.

- For the contour mean distance, the distance between the boundaries of non-zero regions of two images is computed. It is defined as:[37]

$$\begin{aligned} CM(A,B) &= \max\Big(k(A,B), k(B,A)\Big) \\ k(A,B) &= \underset{a\in A}{mean} \min_{b\in B} \|a-b\| \end{aligned} \tag{3}$$

The mean contour distance quantifies the average mismatch between the boundary of A and B.

The previously mentioned metrics were computed for all structures. For this purpose, the software tool Plastimatch[38] and the Insight Segmentation and Registration Toolkit (ITK)[39] were used. For paired organs (PGs, SGs, and ONs), the metric values for both lateralities were computed and averaged. All metrics were computed from the voxelized representations of the segmentations. The overall challenge rankings were generated based on the computed Dice coefficient and on the 95% HD. Maximum HD and contour mean distance were computed after the challenge to provide additional information.

### 4.B. Evaluation and rankings

Participant' segmentation results were ranked according to their average Dice values and the average 95% HDs on all images separately for each structure, with paired substructures treated as one structure. Ranking was performed independently for off-site and on-site datasets. Both metrics contributed equally to the ranking. For all teams who submitted segmentation labels for all structures and both datasets (off-site and on-site), the individual rankings were summed independently for both datasets to find an "overall ranking". The overall ranks for both datasets were summed to get an overall result for each team who submitted all structures.

## 5. CHALLENGE SUBMISSIONS

In phase 1, six different teams participated in the challenge (see Table II). In the following paragraphs, a brief description of the participants' segmentation approaches will be given. Subsequently, the main features of the submitted segmentation approaches are compared and summarized in Table III.

### 5.A. Team FH[40]

The approach is based on an articulated atlas trained using the labeled training data.[41] A coupled shape model called "CoSMo" consisting of rigid and deformable model items was used based on a previous work of team FH.[42] Bones are represented as rigid objects, whereas the remaining structures were modeled as deformable items. For rigid items, training labels were used to calculate a probability image and an average intensity image. In addition, a relative rigid transform with respect to the image center of each training image was computed.[41] Structures showing higher shape variability were represented using SSMs/SAMs.[43] In addition to the relative transform to the center of the articulated atlas, additional shape-specific parameters were stored. For all deformable model items, a SSM was created, and the respective model parameters and rigid transforms were stored. Following, the model adaption process was performed in multiple stages. During the segmentation process, model adaptation begins by segmenting bony structures. Next, deformable items (PGs, SGs, ONs, and OC) are segmented using the SSM, and are fit to the image using a gradient-based approach. In the final stage of the adaption process, the remaining model items (brainstem) were adapted. The final shape and position were obtained using the trained "CoSMo" model without performing a specific segmentation.

TABLE II. Affiliation of participating teams and the respective abbreviations used in this paper.

| Team affiliation | Abbreviation |
|---|---|
| Institut für Graphische Datenverarbeitung, Fraunhofer IGD, Germany | FH |
| IMorphics Ltd, Manchester UK | IM |
| Graphics and vision group, university of basel, Switzerland, and varian medical systems imaging laboratory, Baden, Switzerland | UB |
| Signal processing and recognition research group -Universidad Nacional De Colombia, Colombia | UC |
| BioRobotics Lab, University of Washington, USA | UW |
| Electrical Engineering and Computer Science Department, Vanderbilt University, USA | VU |

### 5.B. Team IM[44]

A segmentation approach based on an active appearance model (AAM) was applied.[45] For this purpose, a variant of the minimum description length (MDL)[46] approach was used to perform groupwise registration of the signed distance images of each structure in all training images. By this means, a mean shape for each structure as well as a set of deformations which map the mean image to each example image can be obtained. In addition to the provided training structures, a mean shape and a set of deformations were created for the orbit (which was not part of the challenge). This orbit model was used to initialize the ON segmentation. From the mean shapes and deformation for each structure, an appearance model (AM) containing shape and texture information was created. For the actual segmentation, the AM was matched to the image using an AAM approach.[45] An AAM can match its AM to an image from a rough initial estimate by optimizing the model parameters. Using AAM, the best-fitting model instance was identified for each structure starting with the mandible. From the result of the mandible, the remaining structures were segmented in the same manner using a search region relative to the mandible.

### 5.C. Team UB[47]

A multiatlas-based segmentation (MABS) approach was used in combination with a subsequent refinement based on active shape models (ASMs).[48] The MABS registration step was performed using an initial rigid alignment, which was performed by automatically detecting a

TABLE III. Comparison of the main features of the participants' segmentation approaches.

| Team | Segmentation approach | Nonrigid registration | Initialization | Similarity measure | Remarks |
|---|---|---|---|---|---|
| FH | Model-based (SSM) | Clamped plate spline warp[58] | Atlas-based | Mutual information | Multilevel segmentation: (1) mandible; (2) PGs, SGs, ONs, OC; (3) brainstem |
| IM | Model-based (AAM) | Groupwise image registration[46] | Alignment of center of gravity/scale | Minimum description length | Based on mandible segmentation, remaining structures were searched/ segmented |
| UB | Atlas- and model-based (ASM) | DEEDs algorithm[49] | Atlas-based | Self similarity context | MABS: brainstem, PGs, OC, mandible MABS+ASM: ON, SG |
| UC | Atlas-based | Elastic transformation (ELAST)[52] | Atlas-based | Mutual information | Multiresolution registration was used; |
| UW | Basic image processing | - | Landmark detection | Sum of squared distances (SSD) | Nasal tip detection forms the basis of the algorithm; Not atlas- or model-based |
| VU | Atlas-based | Adaptive bases algorithm (ABA)[52,56] | Atlas-based | Normalized mutual information | Multiresolution registration; |

set of landmark points. Subsequently, a nonrigid registration was performed using the DEEDS algorithm.[49] After the registration of all atlas images, label voting was executed by applying a variant of the majority voting approach.[50] In contrast to "classical" majority voting where a pixel is considered as part of a structure if more than 50% of the votes classify it as foreground, a voting bias was introduced. Using this bias, a voxel was considered as part of a structure if more than one-third of all votes were classifying it as "foreground". For selected organs (ONs and SGs), an additional model-based segmentation step was performed. SSMs were created for these structures and an ASM-based segmentation approach was applied for the actual segmentation. The result of the MABS step was used to initialize the ASM. Hence, a set of landmarks was transformed from each atlas image to the target image for each structure. Using this set of landmarks, the initial rigid alignment of the SSM was computed. ASM fitting was performed using image intensity profiles along surface normal as introduced by Cootes et al.[48] Hence, boundary points of the shape model were placed at image points that have similar intensity profiles.

### 5.D. Team UC[51]

Team UC also used a MABS approach. The registration part of MABS was performed by using an elastic transformation model[52] with Gaussian regularization and mutual information as the similarity metric. A three-step multiresolution registration approach was applied using a multiresolution pyramid with isotropic downsampling factors of 8, 2, and 1. The label voting step was performed using a generative probabilistic approach based on image patches. For this purpose, the probability of a target patch to be given by the model of an atlas patch was computed. Underlying model parameters were spatially constrained using a Gibbs distribution. By this means, the label for an image patch could be estimated from the class conditional probability for each voxel. Using overlapping voxel neighborhoods, the final segmentation were obtained by combining multiple estimations using a 3D sliding Gaussian window. Only the brainstem, PGs, and mandible were segmented by this team.

### 5.E. Team UW[53]

Team UW presented a semiautomatic method, and only submitted results for mandible and ONs. For the segmentation of the mandible, a point distribution model of the mandible was created. The image to be segmented was first cropped at the center of mass in order to extract a region within the head and neck. In the second step, the nasal tip position was identified in a thresholded version of the image volume. For subsequent steps, voxels superior to the nasal tip were ignored. To fit the mandible model to the image, a thresholded version of the image was created to obtain bone surface points. Subsequently, the iterative closest point algorithm (ICP)[54] was applied in order to align the atlas to the image of surface points. The transformation resulting from ICP was applied to a binarized version of the mandible model. Finally, the segmentation was refined by applying heuristics based on the specific shape of the organ in a slice-by-slice manner. The segmentation of the ONs was performed equivalently, using structure-specific heuristics and landmarks.

### 5.F. Team VU[55]

A MABS approach was used by team VU. In the initial step, one image of the training set was selected as template on which all other images were registered using an affine transform. An average atlas was created for all structures using the adaptive bases algorithm (ABA)[56] in combination with mutual information metric and a gradient descent optimization scheme. Registration was performed by applying a multiresolution scheme using isotropic downsampling of 4,

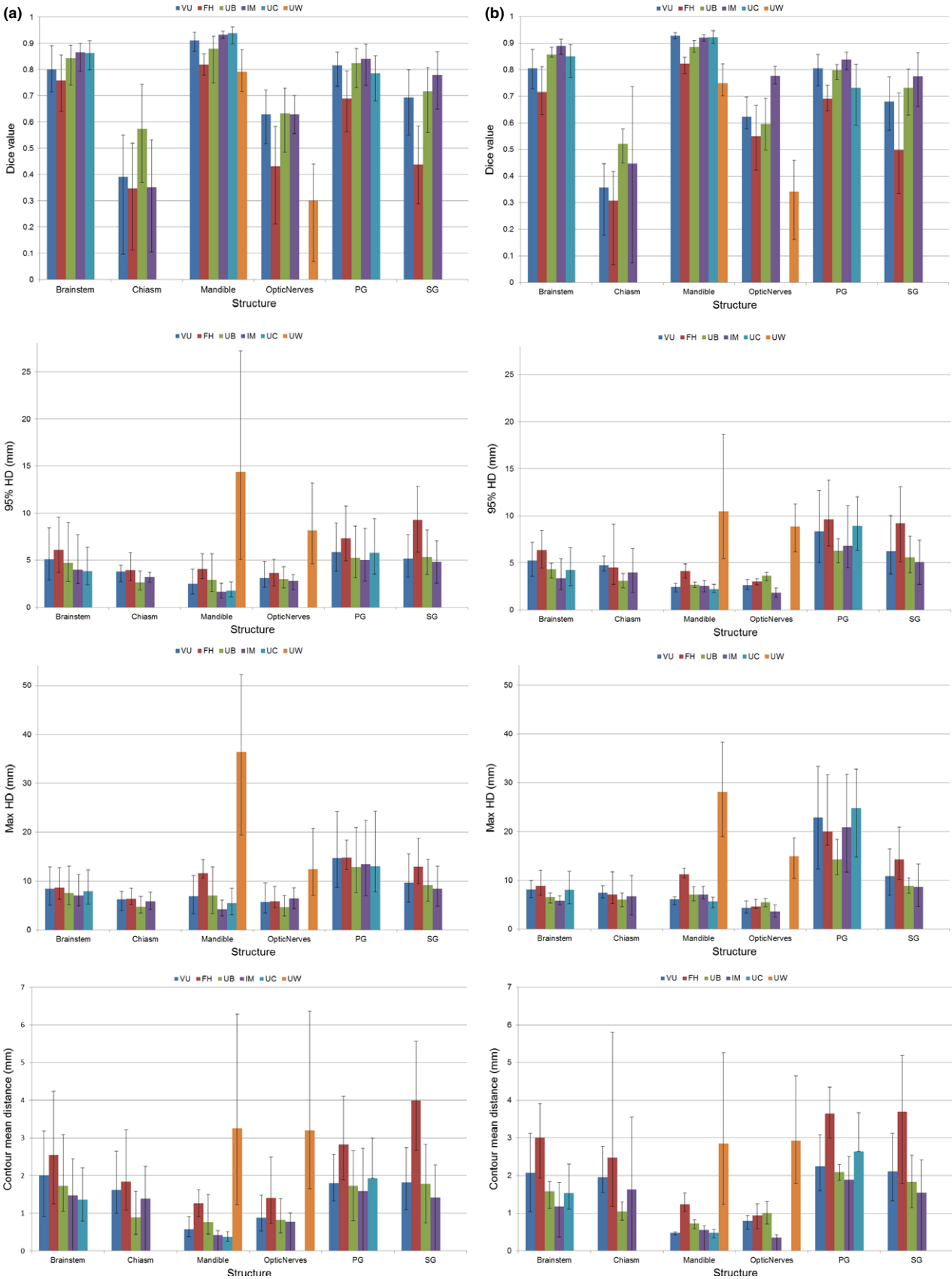


FIG. 1. Average metric values of segmentation results of all participating teams; (a) off-site (b) on-site; Colored bars represent the average metric value of segmentation results of each structure per team (first row: average dice value; second row: average 95% HD [mm]; third row: maximum HD [mm]; and fourth row: contour mean distance [mm]). Whiskers indicate 0.95 and 0.05 percentiles. (PG – Parotid glands; SG – Submandibular glands). [Color figure can be viewed at wileyonlinelibrary.com]

2, and 1. The same registration scheme (affine + nonrigid registration) was used to register the average atlas with unseen images. Bounding boxes around each structure of interest were defined and mapped onto the new image after the global alignment step. Using the subregions defined by these boxes, a local nonrigid registration based on all atlas images was performed. For this registration step, ABA was used in combination with the symmetric normalization algorithm (SyN).[52,57] By applying the resulting deformation to the training labels for the respective structures, segmentation is obtained for each training subject. The final segmentation result was achieved by generating a weighted sum of all segmentations based on the correlation coefficient between the new image and the deformed atlases.

### 5.G. Comparison of submitted segmentation approaches

In Table III, main features of the participants' segmentation approaches are compared and summarized.

## 6. RESULTS

Four teams submitted results for all structures, and two teams submitted results for a subset of the structures. Team UC provided labels for brainstem, mandible, and PGs, and team UW provided results for mandible and ONs. All six teams participated in phase 2 of the challenge. Three teams, UB, FH and IM, completed the segmentation of five test images within the 2-hour time limit. The remaining three teams submitted their final results within 72 h after the challenge.

### 6.A. Evaluation metrics

In Fig. 1, the average error metric values of all participating teams for all six structures are illustrated. The diagrams on the left side (Fig. 1(a)) show off-site segmentation results, whereas the diagrams on the right side (Fig. 1(b)) show the results of the on-site tests. Colored bars represent the respective average metric value of each structure over all test subjects. Team affiliation is color coded. Whiskers indicate 0.95 and 0.05 percentiles. Mandible was segmented by all six participating teams. Brainstem, ONs, and PGs were segmented by five teams and SGs and chiasm by four teams.

### 6.B. Structure-specific rankings

In Table IV, structure-specific rankings of off-site and on-site segmentation results can be seen. In this table, for rank determination, all metric types were used.

### 6.C. Overall rankings

As described previously, the official overall ranking is based on the summation of rank results for all structures, using only the Dice score and the 95% HD. This is shown in Table V. Only participants who segmented all types of structures, both off-site and on-site, were considered for this final overall ranking. Participants who submitted their results after the challenge are highlighted with parentheses.

### 6.D. Comparison and discussion of specific segmentation results

In the following, the segmentation results for each structure are discussed by comparing the results of different approaches. Exemplary cases were selected in order to show strengths and weaknesses of different segmentation approaches based on similarity metric values. It was not the primary aim to only show best and worst cases. In addition, the specific challenges concerning the segmentation of each structure are highlighted.

#### *6.D.1. Mandible*

The mandible was the only bony structure to be segmented. It shows a very high contrast to neighboring tissue, and the degree of shape variation is rather low compared to most soft-tissue organs. One major challenge concerning the segmentation of the mandible is the correct exclusion of the teeth (having similar gray values as the bone) from the mandible segmentation. In addition, the mandibular region is partly affected by image noise caused by dental implants (six off-site test datasets and three on-site test datasets had dental implants). An example CT-slice including the manual segmentation result in a typical dataset is shown in Fig. 2. Looking at the quantitative segmentation results, it can be

TABLE IV. Off-site/On-site segmentation ranking of the participating teams for each structure.

| Team | Brainstem | Chiasm | Mandible | ONs | PGs | SGs |
|---|---|---|---|---|---|---|
| FH | 5/5 | 4/3 | 5/5 | 4/3 | 5/5 | 4/4 |
| IM | 1/1 | 2/2 | 1/3 | 2/1 | 1/1 | 1/1 |
| UB | 3/2 | 1/1 | 5/4 | 1/4 | 2/2 | 2/2 |
| UC | 2/2 | -/- | 1/1 | -/- | 3/4 | -/- |
| UW | -/- | -/- | 6/6 | 5/5 | -/- | -/- |
| VU | 4/4 | 3/3 | 3/2 | 3/2 | 4/3 | 3/3 |

TABLE V. Official overall rank sum and final overall ranking of the challenge (Challenge metrics and all metrics).

| Team | Rank sum challenge | Overall rank challenge | Rank sum all metrics | Overall rank all metrics |
|---|---|---|---|---|
| FH | 52 | 4 | 52 | 4 |
| IM | 16 | 1 | 17 | 1 |
| UB | 27 | 2 | 29 | 2 |
| UC | - | - | - | - |
| UW | - | - | - | - |
| VU | (33) | (3) | (37) | (3) |

observed that highest Dice scores and lowest error rates among all structures were achieved for the mandible. Five of six teams achieved Dice scores > 0.8 and an average 95% HD < 5 mm. This is probably due to the high tissue contrast and low shape variability of bony structures already mentioned above.

When comparing the segmented label of the teams (Fig. 3), it can be seen that the model-based approach by team IM provided very accurate segmentations for the mandible resulting in anatomically plausible shapes. Using this model-based approach, the teeth are very well excluded from the segmentation (see also Fig. 3, left).

However, teeth exclusion is not a "self-evident" property of model-based segmentation approaches. This can be seen, when looking at the result of team FH shown in Fig. 3 (center): Although this group is also applying a model-based approach, teeth are partly included in the final segmentation labels. This is probably caused by the fact that in this approach, the mandible is modeled as a rigid structure and final contours are determined by using adaptive thresholding. Adaptive thresholding could also be the reason for partial leaking of the contour (visible in Fig. 3, center). Fig. 3 (right) shows a segmentation result with a comparably large error, in which a portion of the upper part of the mandible is not included in the segmentation. Due to the fact that the upper part of mandible is rather thin, however, the difference of the Dice score to qualitatively better segmentation results is rather small. The large error becomes evident when looking at the Hausdorff and contour mean distances (see also Table VI) and demonstrates the importance of using multiple metrics for the evaluation of segmentation results.

### 6.D.2. Brainstem

The brainstem is the part of the brain adjoining the spinal cord. It has somewhat low contrast to the surrounding brain tissue (see Fig. 4). In contrast to most other structures of the challenge (with the exception of OC), it does not have completely clear anatomical boundaries in some parts. On the other hand, the brainstem segmentation is generally not affected by image noise, except for occasional noise caused by dental implants, and compared to other soft-tissue organs, shows less variability in shape and appearance. Perhaps for these reasons, the brainstem had the second best segmentation results among all organs in the challenge. Four of five teams provided results with an average Dice score ≥ 0.8 and average 95% HD ≤ 5 mm. Looking at the 0.95 and 0.05 percentiles (Fig. 1), the consistency of the results within each team is also rather high for brainstem segmentation.

Team UC and IM produced the best segmentation results for the brainstem. Although other groups used similar approaches, their results were not quite as good for this

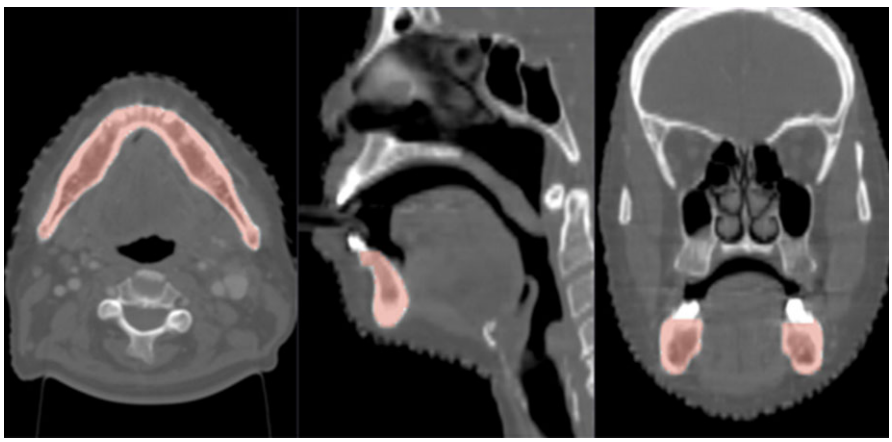

FIG. 2. Example CT-slice showing the mandible and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

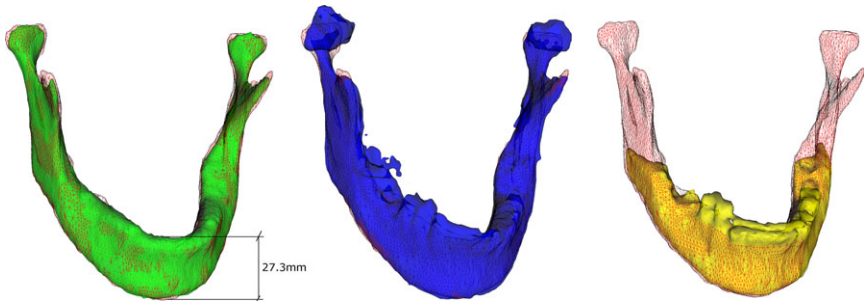


FIG. 3. Example segmentation result for the mandible by team IM (left), team FH (center), and team UW (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

structure. The difference becomes obvious when looking at the example segmentations for the brainstem in Fig. 5. These results were achieved using an atlas-based segmentation approach (yellow for team VU and blue for team UC) and a model-based algorithm (green for team IM). Although teams UC and VU both use atlas-based approaches, team UC successfully recovered the overall shape of the structure more correctly. This is reflected in the Dice and mean distance values for the different teams for this example dataset in Table VII.

The superior boundary is captured better by team VU in this example, although it is showing a rather low amount of regularization. This leads to a higher maximum HD for team VU for this dataset. Qualitatively comparing both atlas-based approaches, it can also be observed that the result of team VU is smoother and more regularized than the result provided by team UC. This is probably caused by the fact that team VU is using a weighted sum of the label resulting from multiple deformable registrations in order to generate a final label. In contrast to this, team UC uses a 3D sliding Gaussian window, which rather provides locally smooth surfaces and a lower overall regularization. In the case of the brainstem, this leads to better results by preventing over-regularization. The model-based approach of team IM provides very smooth and anatomically plausible results for brainstem segmentation. In contrast to the result provided by team VU, however, the higher amount of regularization does not seem to have a negative influence on brainstem segmentation. This also becomes evident when comparing mean distances of team UC and team IM, which are very similar for this dataset (Table VII).

TABLE VI. Metric values of the visualized example datasets of the mandible in Fig. 3.

| Team | Dice | 95% HD [mm] | Max HD [mm] | Contour Mean [mm] |
|---|---|---|---|---|
| UW | 0.728 | 29.459 | 56.128 | 6.433 |
| FH | 0.785 | 5.919 | 15.391 | 1.683 |
| IM | 0.930 | 2.041 | 4.749 | 0.459 |

TABLE VII. Metric values for the exemplary segmentation of the brainstem shown in Fig. 5.

| Team | Dice | 95% HD [mm] | Max HD [mm] | Contour Mean [mm] |
|---|---|---|---|---|
| VU | 0.717 | 8.859 | 13.909 | 3.248 |
| UC | 0.869 | 3.466 | 8.557 | 1.308 |
| IM | 0.865 | 4.299 | 8.864 | 1.510 |

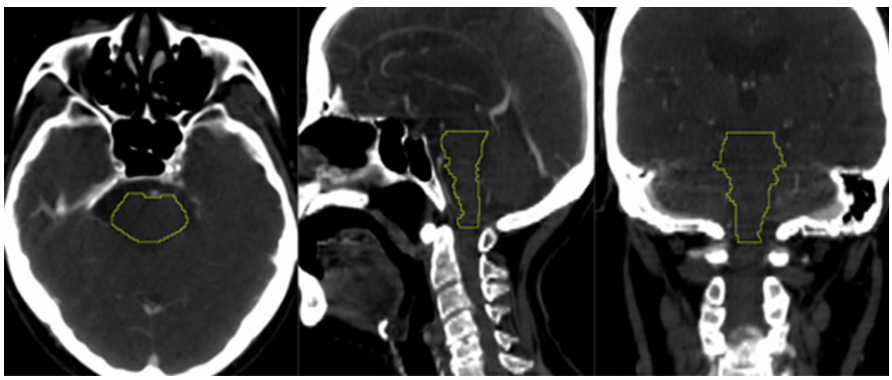

FIG. 4. Example showing the brainstem and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

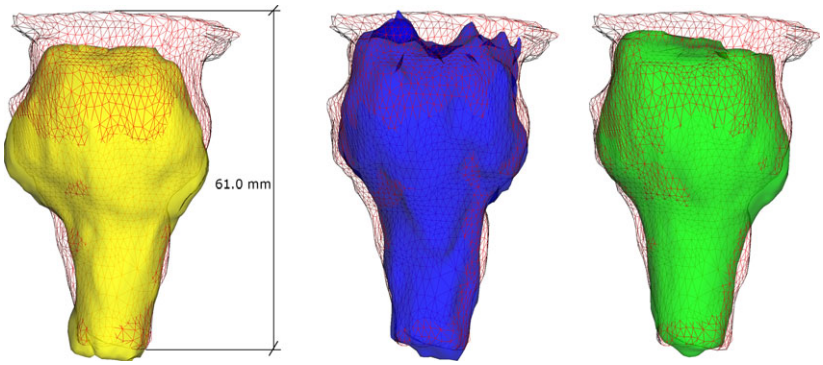


FIG. 5. Example segmentation result for the brainstem provided by team VU (left), team UC (center), and team IM (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

### 6.D.3. Parotid glands

The PGs are the major salivary glands and crucial for mastication and swallowing. Located directly behind the mandibular ramus, the PGs show comparably large shape variations and partly poor contrast to the surrounding tissue, especially along the medial border. Another challenge for automated segmentation approaches is that the interior part of the PG is rather heterogeneous including vessels and ducts (see also Fig. 6 and Section 2.B).

In terms of segmentation quality, three of five teams could obtain an average Dice score $\geq 0.8$ for PG segmentation, which is similar to the results for the brainstem. However, PG results show the largest values for maximum HD among all structures, consistently for all teams. This is probably caused by the fact that the PG sometimes shows elongated shapes and partly even dislocated parts (also referred to as accessory glands) in the superior–anterior part of the organ. This specific PG shape represents ectopic salivary tissue, which can be observed in ~20% of the general population.[59] For automatic segmentation approaches, these accessory glands are very hard to deal with. As their volume is comparably small, their influence on Dice score is limited. However, mis-segmentation of accessory glands can cause large HDs. This can also be seen in Fig. 7 which is showing two exemplary results for erroneous segmentation of the superior part of a left PG.

It can be observed that team UB (second highest Dice scores for PG) is showing better maximum HD values for both test datasets. This tendency can also be observed in Fig. 8, illustrating another exemplary segmentation of teams IM and UB of the same dataset. Although showing similar Dice scores, the maximum HD of the result obtained by team UB is distinctly lower (see also Table VIII), because of the more accurate segmentation of the superior–anterior part of the PG by team UB (see Fig. 8, left).

### 6.D.4. Submandibular glands

Considerably smaller than the PGs, the SGs are located beneath the lower jaws (see also Fig. 9) and produce about 70% of the saliva in the oral cavity.

The anterio-medial part of the SGs has poor contrast with the surrounding tissue. In addition, the region around the SGs is partly subject to intense artifacts caused by dental implants, and the shape variation of the structure is also rather high especially in the inferior part. All these facts make SG segmentation very challenging. This can also be seen when looking at the results for SG segmentation obtained by the participating teams: only two teams (IM and UB) could achieve Dice scores > 0.7, which is clearly lower than the result for PG. However, contour distance values (contour mean distance and HD distance) are better than the results that were obtained for PGs. This is most possibly caused by the fact that the SG is smaller, and does not show the same degree of shape variations as the superior part of the PG. For organs like the SG, which can be affected by heavy image

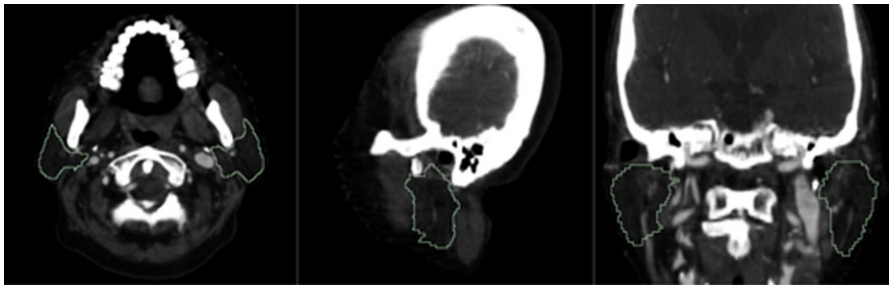

FIG. 6. Example showing the PGs and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

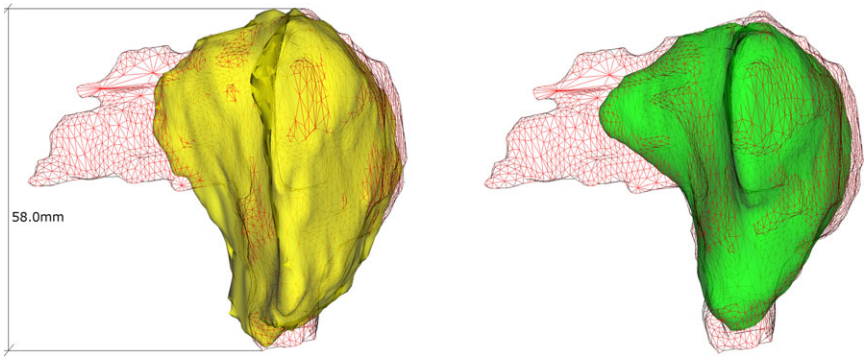


FIG. 7. Example segmentation result for a left PG provided by team UB (left) and team IM (right) vs. manual segmentation (wireframe). Missing parts can be seen in the superior–anterior part of the structure (see text for additional explanation). [Color figure can be viewed at wileyonlinelibrary.com]

noise, the stronger shape regularization performed by model-based approaches can be beneficial. This is also the reason why SG was one of only two structures (aside from ONs) for which team UB created SSMs in order to refine the results of their MABS approach. However, as already mentioned above, model-based approaches are also very sensitive to initialization. This sensitivity can result in smooth and anatomically plausible segmentation results, which however do not correctly reflect the true shape of the organ. An example for this can be seen in Fig. 10 where one exemplary SG segmentation result of team UB and team IM is compared. Although both results look anatomically plausible (based on using SSM), only team IM could successfully create a segmentation result, which accurately reflects the correct shape of the SG for this dataset (Table IX).

### 6.D.5. Optic nerve

The ON is one of 12 paired cranial nerves. It leaves the orbit via the optic canal, running postero-medially toward the OC (see Fig. 11).

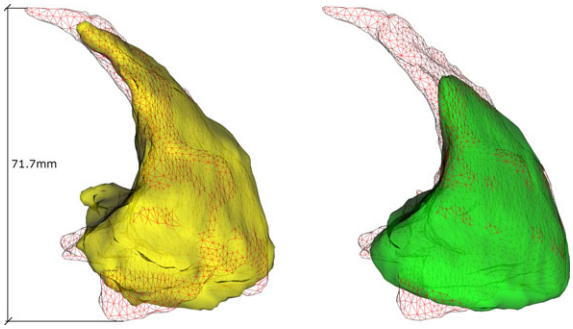


FIG. 8. Example segmentation result for a left PG provided by team UB (left) and team IM (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE VIII. Metric values for the exemplary segmentation of the left PG illustrated in Fig. 8.

| Team | Dice | Max HD [mm] |
|---|---|---|
| UB | 0.814 | 12.000 |
| IM | 0.826 | 30.490 |

Compared to the other structures of the challenge, its shape is quite different, as it is the only thin and tubular structure included in the challenge. Due to the slice thickness of 2–3 mm for the CT data that was used for the challenge and also caused by its rather horizontal (anterior–posterior) path, the ON is commonly only visible in 3–4 CT-slices. This difficulty makes ON segmentation very challenging. As a result, the Dice scores for ON were generally worse than the results for the larger organs consistently for all teams.[60] Out of five teams who provided a segmentation for ON, only three of five teams for the off-site tests (two of five for on-site tests) could provide results with a Dice score > 0.6. On the other hand, segmentation results with a contour mean distance for ON below 1 mm were provided by four of five teams for the off-site tests (three of five for on-site tests). In addition, three of five teams for off-site tests (two of five for on-site tests) could obtain an average 95% HD below 3.1 mm. These are the lowest 95% HDs for all structures in the challenge. These small contour distance errors are most probably caused by the comparably low amount of shape variation of the ON and the rather high contrast to the surrounding tissue for most parts of the nerve.

In Fig. 12, an example segmentation result of the left ON can be seen. It is obvious that in the segmentation result of team VU (yellow) the left ON is disconnected (Fig. 12 left; Table X). This is primarily caused by the thin tubular part of the structure. For the same dataset, the segmentation result of

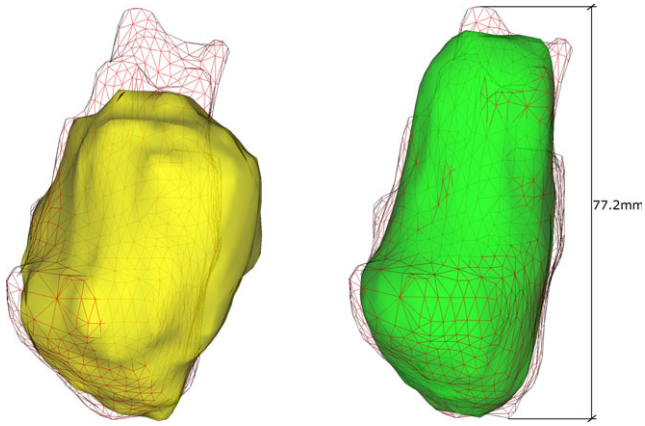


FIG. 10. Example segmentation result for a left SG provided by team UB (left) and team IM (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

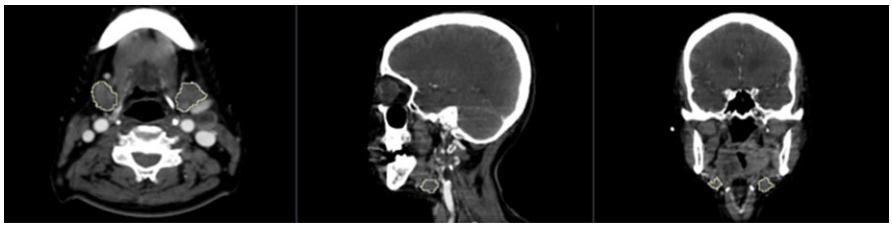

FIG. 9. Example CT-slice showing the SGs and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

team IM (green) provides a connected ON (Fig. 12 right; Table X). The AAM-based approach was able to ensure a valid, connected shape of the ON.

### 6.D.6. Optic chiasm

The OC is located below the hypothalamus and is the part of the brain where the paired ONs partially cross (see also Fig. 13). Similar to the brainstem, the OC does not show anatomical boundaries for all parts of the structure. As a consequence, boundaries had to be defined in order to separate the OC from the ONs (anterior) and the optic tract (posterior). See Section 2.B for more detailed segmentation guidelines.

The OC has very low contrast with surrounding brain tissue and is only visibly in 1–3 CT-slices, using a typical slice thickness of 2–3 mm. These facts make the OC probably the most challenging structure to segment among all structures included in the challenge. The difficulty becomes obvious when looking at the quantitative results of the teams. On one hand, OC has the lowest average Dice scores of all structures and at the same time the largest variation concerning the Dice values for all teams, where average Dice scores range between 0.3 and 0.57. This large variation is also demonstrated in Fig. 14, showing two very different segmentation results for the same dataset. Quantitative results can be found in Table XI.

In contrast to the ON, for most teams the contour mean distance for OC is rather large and very similar to the results for SG. The SG, however, is considerably larger and showing a higher amount of shape variability than the OC. Although the low Dice scores already mentioned above are most probably caused by the small size of the OC compared with the large slice thickness, the reason for this large contour mean distances and mediocre results for HD are most likely caused by the poor contrast to neighboring tissue. As a result, even for teams that were showing consistently good results for the other structures using very robust segmentation approaches, the segmentation of the OC was extremely challenging and led to very large errors.

## 7. DISCUSSION

It is not a primary goal of the Head and Neck Challenge to determine the best segmentation approach for head and neck organs. Rather, quantitative evaluation and comparison of cutting-edge approaches for automatic segmentation provides the possibility for researchers to compare the performance of their methods with other approaches in an unbiased and standardized manner. Furthermore, the results have shown that a

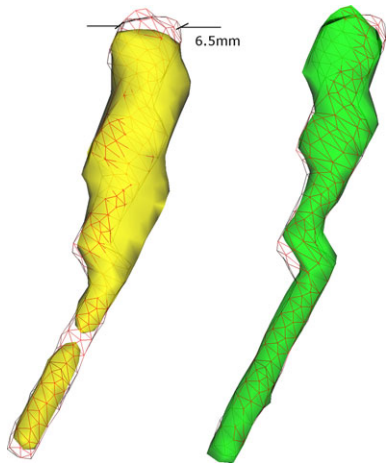


FIG. 12. Example segmentation result for a left ON provided by team VU (left) and team IM (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE IX. Metric values for the exemplary segmentation of the left SG illustrated in Fig. 10.

| Team | Dice | Max HD [mm] |
|---|---|---|
| UB | 0.760 | 6.546 |
| IM | 0.895 | 1.997 |

TABLE X. Metric values for the exemplary segmentation of the left ON illustrated in Fig. 12.

| Team | Dice | Max HD [mm] |
|---|---|---|
| VU | 0.709 | 2.085 |
| IM | 0.828 | 1.170 |

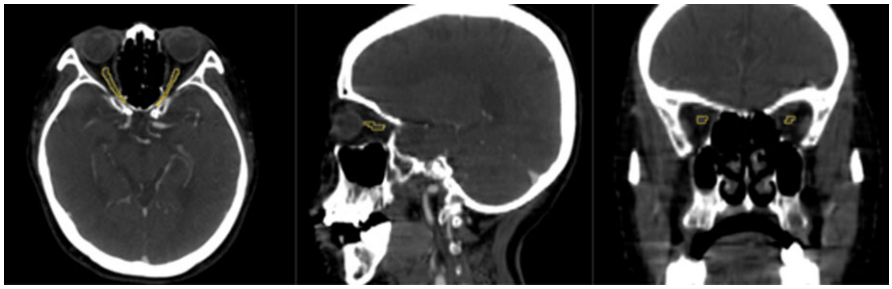

FIG. 11. Example CT-slice showing the paired ON and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

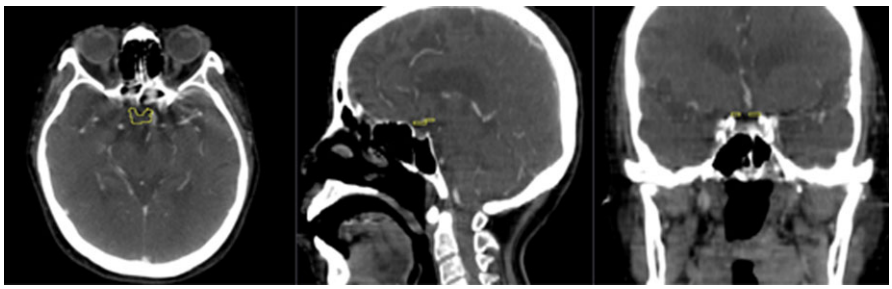

FIG. 13. Example CT-slice showing the OC and the result of the manual segmentation for one dataset in three orthogonal views. [Color figure can be viewed at wileyonlinelibrary.com]

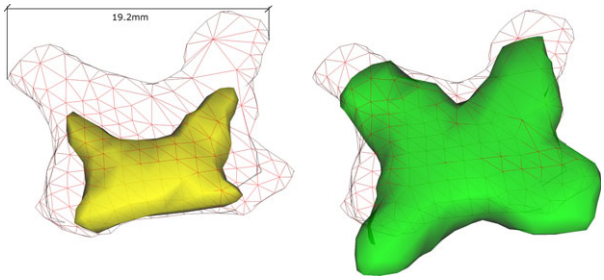


FIG. 14. Example segmentation result for an OC provided by team IM (left) and team UB (right) vs. manual segmentation (wireframe). [Color figure can be viewed at wileyonlinelibrary.com]

TABLE XI. Metric values for the exemplary segmentation of the OC illustrated in Fig. 14.

| Team | Dice | Max HD [mm] |
|---|---|---|
| UB | 0.795 | 1.997 |
| IM | 0.318 | 3.567 |

combination of volume- and distance-based metrics is meaningful as they describe different characteristics and therefore complement each other.

In the remainder of this section, the main results and findings obtained by different segmentation approaches will be discussed. By this means, the current state-of-the-art concerning the automated segmentation of OARs in the head and neck will be summarized and strengths and weaknesses of certain approaches will be highlighted.

## 7.A. Comparison of different segmentation approaches

Participating teams had the possibility to compare their approaches to other methods under unbiased and standardized conditions. The obtained quantitative results are not only interesting for the participating teams but also highly relevant for the medical imaging community in general. Although the challenge focused on a fixed set of OARs, insights into the strengths and weaknesses of different segmentation approaches can also be transferred to the segmentation of other organs.

When comparing the two main groups of algorithms which are most frequently used for head and neck segmentation (model-based vs. atlas-based segmentation), a clear winner cannot be found. This is because the results of different teams using rather similar approaches were very different, so that no clear tendency toward one algorithm type could be observed. This is an indication that considerable effort must be dedicated to multiple, sometimes inter-related steps of an algorithmic pipeline to make the difference between good and outstanding results. An atlas-based initialization of the segmentation has shown to be a good approach. Four of six teams have chosen this attempt. Furthermore, the results have shown that ABS delivered very good segmentations of the mandible and brainstem. This is probably due to the fact that these structures have a relative small amount of variability compared to other structures. Moreover from the results it is obvious that the label fusion technique has an important influence on the final segmentation result for specific structures (e.g., over-regularization as described in Section 6.D.2). For smaller structures (i.e., ON and OC in this study), ABS approaches performed worse compared to model-based approaches. The small structure size combined with the relatively high slice distance (2–3 mm) is one reason therefore.

Based on the results of this challenge, it can be said that model-based approaches have shown to appropriate for nearly all structures. Compared to ABS approaches, model-based attempts have advantages for structures with high shape variability (PGs) and smaller structures (ON and OC). However, in order to model variability correctly, the training dataset must also contain this variability appropriately.

It can be concluded that the results also show that a combination of atlas-based and model-based segmentation are a very promising approach. An accurate registration with an atlas provides a close initialization for the models which can then effectively exploit local image information.

CT is the predominant imaging technique in the course of treatment of head and neck cancers. Hence, for this challenge CT data were used. However, for some structures, MRI is additionally used as further information is available. For example, bad contrast conditions of the brainstem in CT scans are challenging for automated segmentation approaches. Within MRI images, the brainstem is clearly

TABLE XII. Comparison of the best segmentation results between this and previous challenges with respect of dice overlap. Furthermore, results of recent methods in literature are also included. (- structure was not analyzed in this challenge/work).

| Structure (# of datasets) | 2015 (15) | 2009[8] (7) | 2010[9] (8) | Fritscher[61] (16) | Fortunati[62] (18) | Thomson[63] (20) | Deeley[64] (20) | Harrigan[65] (30) |
|---|---|---|---|---|---|---|---|---|
| Brainstem | 0.88 | 0.88 | - | 0.86 | 0.78 | - | 0.82 | - |
| Mandible | 0.93 | 0.93 | - | - | - | - | - | - |
| PG | 0.84 | - | 0.85 | 0.83 | - | 0.78 | - | - |
| ON | 0.62 | - | - | - | 0.62 | - | 0.52 | 0.39–0.79 |
| SG | 0.78 | - | - | - | - | 0.70 | - | - |
| OC | 0.55 | - | - | - | - | - | 0.37 | - |

better visible. Atlas- and model-based segmentation approaches can also be used for MRI images.

### 7.B. Comparison with previous segmentation challenges and other studies

In Table XII, the best segmentation results of previous challenges are compared with the best results of the Head and Neck Auto-segmentation Challenge 2015 for the Dice overlap score. Furthermore, Table XII also shows segmentation results of other recent studies. It is seen that the resulting Dice scores of this challenge are comparable with the results of the two previous auto-segmentation challenges in 2009 and 2010 for brainstem, mandible, and PGs.[8,9] As these structures were often segmentation targets in past studies, automated segmentation approaches already reached a very good level. Consequently, further improvements will be difficult because segmentation accuracy is converging to inter-rater variability.[64,65] For other structures, no comparison can be made because these structures were not target of previous challenges. When looking at the segmentation results of approaches in other recent work, it can be seen that the Dice scores of the results of this challenge are within the same range or even slightly better for nearly all structures.

## 8. CONCLUSION

Segmentation of OARs is a key step in the radiotherapy planning process. Events like the Head and Neck Auto-Segmentation Challenge 2015 give the opportunity to see an overview of state-of-the-art automatic segmentation approaches. In addition, such events provide a possibility to evaluate the performance of independent algorithms under standardized circumstances, which are comparable to clinical practice. Comparisons with previous challenges and recent works have shown that the results of this challenge are state-of-the-art (Table XII).

A positive observation is that teams which submitted results for all organs of the challenge ranked quite similarly for all structures. It was also seen that there is a clear tendency toward more general purpose, and fewer structure-specific algorithms. This is useful not only for end-users but also for the scientific community and medical imaging companies. Efforts devoted to creating general approaches have a higher impact than development of structure-specific algorithms.

In order to give the participating teams as well as the scientific community in general the possibility to objectively evaluate their segmentation approaches, all training and test datasets are available for download from the Internet.[31] In addition, the challenge hosts aim at organizing similar events at least biannually in order to regularly enforce not only objective comparison but also fruitful discussions between participating teams and scientists who visit the challenge. Although the number of datasets and especially the number of structures has already been considerably increased compared to similar events in the past,[8,9] additional improvements concerning the provided data are planned. Aside from further increasing the number of test and training datasets, multiple manual segmentations per structure coming from different medical experts should be provided in future challenges. By this means, inter-rater variability can be assessed. This is another important factor for judging the quality of autosegmentation approaches for different structures. Finally, structures like larynx, spinal cord, or pharyngeal muscles are excellent candidates to be included in future challenges, in order to further increase the benefit for the scientific community and the improvement of radiotherapy treatment of the head–neck area in general.

## ACKNOWLEDGMENTS

Antong Cheng works as senior scientist at Merck. Thomas Albrecht and Tobias Gass are scientific software engineers at Varian Medical Systems. Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, and Graham Vincent are scientists at Imorphics Ltd.

## CONFLICT OF INTEREST

Patrik F. Raudaschl, Paolo Zaffino, Gregory C. Sharp, Maria Francesca Spadea, Rainer Schubert, Karl D. Fritscher, Benoit M. Dawant, Christoph Langguth, Marcel Lüthi, Florian Jung, Oliver Knapp, Stefan Wesarg, Mauricio Orbes-Arteaga, David Cárdenas-Peña, German Castellanos-Dominguez, Nava Aghdasi, Yangming Li, Angelique Berens, Kris Moe and Blake Hannaford have no relevant conflicts of interest to disclose.

[a]Author to whom correspondence should be addressed. Electronic mail: patrik.raudaschl@umit.at.

## REFERENCES


1. Dawson LA, Sharpe MB. Image-guided radiotherapy: rationale, benefits, and limitations. *Lancet Oncol*. 2006;7:848–858.
2. Hall EJ, Wuu C. Radiation-induced second cancers: the impact of 3D-CRT and IMRT. *Int J Radiation Oncology Biol Phys*. 2003;56:83–88.
3. Tubiana M. Can we reduce the incidence of second primary malignancies occuring after radiothearpy? A critical review. *Radiother Oncol*. 2009;91:4–15.
4. Lalaoui L, Mohamadi T. A comparative study of Image Region-Based Segmentation Algorithms. *IJACSA*. 2013;4:198–206.
5. Feng Y, Kawrakow I, Olsen J, et al. A comparative study of automatic image segmentation algorithms for target tracking in MR-IGRT. *J Appl Clin Med Phys*. 2016;17:441–460.
6. Heimann T, Styner M, vanGinneken B. *Segmentation of the liver 2007*. Available: http://sliver07.org; 2016, Jan. 14.
7. Hata N, Fichtinger G, Oguro S, Elhawary H, vanWalsum T. *2009 Prostate segmentation challenge MICCAI*. Available: http://wiki.na-mic.org/Wiki/index.php/2009_prostate_segmentation_challenge_MICCAI; 2016, Jan. 14.
8. Pekar V, Allaire S, Qazi AA, Kim JJ, Jaffray DA. *Head and neck auto-segmentation challenge*. Available: http://www.midasjournal.org/browse/publication/703; 2016, Jan. 14.
9. Pekar V, Allaire S, Qazi AA, Kim JJ, Jaffray DA. Head and neck auto-segmentation challenge: segmentation of the parotid glands. In: *MICCAI 2010, Grand Challenges in Medical Image Analysis: Head & Neck Auto-segmentation Challenge*, 2010. http://www.researchgate.net/publication/264893852_Head_and_Neck_Auto-segmentation_Challenge_Segmentation_of_the_Parotid_Glands.
10. Sharp G, Fritscher K, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41:1–13.
11. Haas B, Coradi T, Scholz M, et al. Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies. *Phys Med Biol*. 2008;53:1751–1771.
12. Hu K, Lin A, Young A, et al. Time savings for contour generation in head and neck IMRT: multi-institutional experience with an atlas-based segmentation method. *Int J Radiat Oncol Biol Phys*. 2008;72:391.
13. Stapleford LJ, Lawson JD, Perkins C, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *Int J Radiat Oncol Biol Phys*. 2010;77:959–966.
14. Grosu A-L, Lachner R, Wiedenmann N, et al. Validation of a method for automatic image fusion (BrainLAB System) of CT data and 11C-methionine-PET data for stereotactic radiotherapy using a LINAC: first clinical experience. *Int J Radiat Oncol Biol Phys*. 2003;56:1450–1463.
15. Commowick O, Gregoire V, Malandain G. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiother Oncol*. 2008;87:281–289.
16. Gooding MJ, Chu K, Conibear J, et al. Multicenter clinical assessment of DIR atlas-based autocontouring. *Int J Radiat Oncol Biol Phys*. 2013;87:714–715.
17. Oncology Systems Limited (OSL); OnQ rts. see http://www.osl.uk.com; 2015.
18. Han X, Hoogeman M, Levendag P, et al. Atlas-based auto-segmentation of head and neck CT images. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008*. New York, NY: Springer; 2008: 434–441.
19. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. *Med Phys*. 2011;38:6160–6170.
20. Stewart J, Lim K, Kelly V, et al. Automated weekly replanning for intensity-modulated radiotherapy of cervix cancer. *Int J Radiat Oncol Biol Phys*. 2010;78:350–358.
21. University of Arkansas for Medical Sciences. *Cancer treatment and diagnosis, national cancer institute, cancer imaging archive*. Available: http://www.cancerimagingarchive.net; 2015.
22. Ang KK, Zhang Q, Rosenthal DI, et al. Randomized Phase III Trial of Concurrent Accelerated Radiation Plus Cisplatin With or Without Cetuximab for Stage III to IV Head and Neck Carcinoma: RTOG 0522. *J Clin Oncol*. 2014;32:2940–2950.
23. Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117:83–90.
24. Auto-Segmentation Challenge. *Segmentation Guidelines*. www.imagenglab.com/data/pddca/pddca.odt; 2015.
25. Machtay M, Siu L, Thorstad W, et al. *A phase III study of postoperative radiation therapy (IMRT) +/- Cetuximab for locally-advanced resected head and neck cancer*. Available: http://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=0920; 2015, Dec. 02.
26. Harari PM, Rosenthal DI, et al. *Randomized Phase II/III Trial of surgery and postoperative radiation delivered with concurrent Cisplatin versus docetaxel versus docetaxel and cetuximab for high-risk squamous cell cancer of the head and neck*. Available: https://www.rtog.org/ClinicalTrials/ProtocolTable/StudyDetails.aspx?study=1216; 2015, Dec. 02.
27. Sadun AA, Glaser JS, Bose S. Chapter 34: anatomy of the visual sensory system. In: *Duane's Ophthalmology*. Lippincott Williams & Wilkins; 2005.
28. Bergman RA, Afifi AK, Miyauchi R. *Illustrated encyclopedia of human anatomic variation*. Available: www.anatomyatlases.org; 2015.
29. Karesh JW, Yassur I, Hirschbein MJ. Chapter 34: advanced neuroimaging techniques for the demonstration of normal orbital, periorbital, and intracranial Anatomy. In: *Duane's Ophthalmology*. Lippincott Williams & Wilkins; 2005.
30. van de Water TA, Bijl HP, Westerlaan HE, Langendijk JA. Delineation guidelines for organs at risk involved in radiation-induced salivary dysfunction and xerostomia. *Radiother Oncol*. 2009;93:545–552.
31. Auto-Segmentation Challenge. *Head-Neck Auto-Segmentation Challenge*. http://www.imagenglab.com/wiki/mediawiki/index.php?title=2015_MICCAI_Challenge; 2015.
32. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
33. Taha AA, Hanbury A, Jimenez del Toro O. A formal method for selecting evaluation metrics for image segmentation. Sydney: IEEE International Conference on Image Processing (ICIP); 2014: 932–936.
34. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302.
35. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*. 2009;46:726–738.
36. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing Images Using the Hausdorff Distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions*. 1993;15:850–863.
37. Kitware *Contour mean distance image filter*. Available: http://www.itk.org/Doxygen/html/classitk_1_1ContourMeanDistanceImageFilter.html; 2015.
38. Plastimatch; Image Computation. Available: http://www.plastimatch.org/; 2015, Dec. 03.
39. Insight Segmentation and Registration Toolkit (ITK). ITKv4. Available: http://www.itk.org; 2015, Dec. 03.
40. Jung F, Knapp O, Wesarg S. CoSMo - coupled shape model segmentation. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/970
41. Steger S, Kirschner M, Wesarg S. Articulated atlas for segmentation of the skeleton from head and neck CT datasets. Barcelona: 9th IEEE International Symposium on Biomedical Imaging (ISBI); 2012: 1256–1259.
42. Jung F, Steger S, Knapp O, Noll M, Wesarg S. COSMO - coupled shape model for radiation therapy planning of head and neck cancer. In: *Clinical Image-Based Procedures. Translational Research in Medical Imaging, LNCS*, vol. 8680. Cham: Springer; 2014: 25–32.
43. Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: a review. *Med. Image Analy*. 2009;13:543–563.
44. Mannion-Haworth R, Bowes M, Ashman A, Guillard G, Brett A, Vincent G. Fully Automatic Segmentation of Head and Neck Organs using Active Appearance Models. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/967
45. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. *IEEE Trans Pattern Anal Mach Intell*. 2001;6:681–685.

46. Cootes TF, Twining CJ, Petrovic V, Schestowitz R, Taylor C. Groupwise construction of appearance models using piece-wise affine deformations. In: *16th British Machine Vision Conf.*, Vol. 2. Oxford: British Machine Vision Association, BMVA; 2005: 879–888.
47. Albrecht T, Gass T, Langguth C, Lüthi M. Multi atlas segmentation with active shape model refinement for multi-organ segmentation in head and neck cancer radiotherapy planning. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/968
48. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active Shape Models-Their Training and Application. *Comput Vis Image Underst*. 1995;61:38–59.
49. Heinrich MP, Jenkinson M, Brady M, Schnabel JA. MRF-based deformable registration and ventilation estimation of lung CT. *IEEE Trans Med Imaging*. 2013;32:1239–1248.
50. Heckemann RA, Hajna JV, Aljabar P, Rueckert D, Hammers A. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*. 2006;33:115–126.
51. Orbes-Arteaga M, Cárdenas-Peña D, Castellanos-Dominguez G. Head and neck auto segmentation challenge based on non-local generative models. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/965
52. ANTS - Advanced Normalization Tools; ANTs. Available: http://stnava.github.io/ANTs/; 2016, Oct. 14.
53. Aghdasi N, Li Y, Berens A, Moe K, Hannaford B. Head and neck segmentation based on anatomical knowledge. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/971
54. Besl PJ, McKay ND. A method for registration of 3-D shapes. *IEEE Trans Pattern Anal Mach Intell*. 1992;14:239–256.
55. Chen A, Dawant BM. A multi-atlas approach for the automatic segmentation of multiple structures in head and neck CT images. In: *Presented in Head and Neck Auto-Segmentation Challenge 2015*. Munich: MICCAI; 2015 [Online] Available: http://midasjournal.org/browse/publication/964
56. Rohde GK, Aldroubi A. B. M. Dawant", The adaptive bases algorithm for intensity-based nonrigid image registration". *IEEE Trans Med Imaging*. 2003;22:1470–1479.
57. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal*. 2008;12:26–41.
58. Marsland S, Twining CJ. Clamped-plate splines and the optimal flow of bounded diffeomorphisms. In: *Proceedings of Leeds Annual Statistical Research Workshop: statistics of Large Datasets*. Manchester: Division of Imaging Science and Biomedical Engineering Manchester, University of Manchester; 2002: 91–95.
59. Frommer J. The human accessory parotid gland: its incidence, nature, and significance. *Oral Surgery, Oral Medicine, Oral Pathology*. 1977;43:671–676.
60. Isambert A, Dhermain F, Bidault F, et al. 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;87:93–99.
61. Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Med Phys*. 2014;41.
62. Fortunati V, Verhaart RF, dervan Lijn F, et al. Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. *Med Phys*. 2013;40:1–14.
63. Thomson D, Boylan C, Liptrot T, et al. Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. *Radiat Oncol*. 2014;2:1–12.
64. Deeley MA, Chen A, Datteri R, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *Phys Med Biol*. 2011;21:4557–4577.
65. Harrigan RL, Panda S, Asman AJ, et al. Robust optic nerve segmentation on clinically acquired computed tomography. *J Med Imaging*. 2014;1:1–10.