Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods

Tomaž Vrtovec[a)] and Domen Močnik
Faculty Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana SI-1000, Slovenia

Primož Strojan
Institute of Oncology Ljubljana, Zaloška cesta 2, Ljubljana SI-1000, Slovenia

Franjo Pernuš
Faculty Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana SI-1000, Slovenia

Bulat Ibragimov
Faculty Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana SI-1000, Slovenia
Department of Computer Science, University of Copenhagen, Universitetsparken 1, Copenhagen D-2100, Denmark

(Received 26 October 2019; revised 27 May 2020; accepted for publication 29 May 2020; published 28 July 2020)

Radiotherapy (RT) is one of the basic treatment modalities for cancer of the head and neck (H&N), which requires a precise spatial description of the target volumes and organs at risk (OARs) to deliver a highly conformal radiation dose to the tumor cells while sparing the healthy tissues. For this purpose, target volumes and OARs have to be delineated and segmented from medical images. As manual delineation is a tedious and time-consuming task subjected to intra/interobserver variability, computerized auto-segmentation has been developed as an alternative. The field of medical imaging and RT planning has experienced an increased interest in the past decade, with new emerging trends that shifted the field of H&N OAR auto-segmentation from atlas-based to deep learning-based approaches. In this review, we systematically analyzed 78 relevant publications on auto-segmentation of OARs in the H&N region from 2008 to date, and provided critical discussions and recommendations from various perspectives: *image modality* — both computed tomography and magnetic resonance image modalities are being exploited, but the potential of the latter should be explored more in the future; *OAR* — the spinal cord, brainstem, and major salivary glands are the most studied OARs, but additional experiments should be conducted for several less studied soft tissue structures; *image database* — several image databases with the corresponding ground truth are currently available for methodology evaluation, but should be augmented with data from multiple observers and multiple institutions; *methodology* — current methods have shifted from atlas-based to deep learning auto-segmentation, which is expected to become even more sophisticated; *ground truth* — delineation guidelines should be followed and participation of multiple experts from multiple institutions is recommended; *performance metrics* — the Dice coefficient as the standard volumetric overlap metrics should be accompanied with at least one distance metrics, and combined with clinical acceptability scores and risk assessments; *segmentation performance* — the best performing methods achieve clinically acceptable auto-segmentation for several OARs, however, the dosimetric impact should be also studied to provide clinically relevant endpoints for RT planning. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14320]

Key words: auto-segmentation, deep learning, head and neck, organs at risk, radiotherapy planning

1. INTRODUCTION

Cancer in the region of the head and neck (H&N), comprising malignancies of the lips, oral cavity, pharynx, larynx, nasal cavity and paranasal sinuses, salivary glands, and thyroid has a yearly incidence of approximately 1.5 million worldwide,[1] making it one of the most prominent cancers. In addition to surgery and chemotherapy, radiotherapy (RT) is an important treatment modality for the H&N cancer, with an optimal utilization rate in patients presented with this malignancy of around 80%.[2] The aim of RT is to deliver a high radiation dose to the targeted cancerous cells to ensure clinically required tumor control probability and, at the same time, spare the nearby healthy tissues to prevent acute radiation toxicity and serious late complications for the treated patient. The optimal radiation dose distribution is calculated in an optimization process using the inverse planning approach, which requires a precise spatial description of the target volumes as well as of the organs at risk (OARs). This knowledge is commonly obtained by trained radiation oncologists and, in some instances, also other experts from the field performing manual delineation, or *segmentation*, of the target volumes and OARs from the acquired three-dimensional (3D) images of the patient.

Medical image segmentation as the process of partitioning an image into multiple anatomical structures is, in general, a

challenging task that is hampered by the high variability of medical images. The source of variability is commonly represented by different imaging modalities revealing different characteristics of the human anatomy, for example, conventional radiographic (x rays), computed tomography (CT), and magnetic resonance (MR) imaging, various imaging artifacts causing weak or missing boundaries, for example, noise, intensity inhomogeneity, partial volume effect and motion, and variable image appearance of anatomical structures under segmentation, for example, due to pathological changes or the natural biological variability of the human anatomy. Nevertheless, image segmentation is important from the perspective of analyzing the properties of the obtained structures, and while manual delineation may still be the approach of choice, it is a time-consuming and tedious task subjected to intra/interobserver variability.[3] Alternatively, computerized techniques based on medical image processing and analysis have been developed that replace manual with automated segmentation, or *auto-segmentation*,[4,5] which eliminates the subjective bias of the observer, accelerates the whole process and, as a result, reduces the total workload in terms of human resources.

In the past decade, the field of computerized medical imaging has experienced an increased interest, with new emerging trends that are largely focused on deep learning (DL)[6] as a subset of machine learning that mimics the data processing of the human brain for the purpose of decision-making. In comparison to traditional approaches based on conventional atlases, shape models and feature classification, DL has shown superior image segmentation performance that was conveyed by several milestone auto-segmentation frameworks,[7] for example, the U-Net,[8] 3D U-Net,[9] V-Net,[10] SegNet,[11] DeepMedic,[12] DeepLab,[13] VoxResNet[14] and Mask R-CNN.[15] Several ideas have been adopted for RT,[16,17] including for image segmentation and detection, image phenotyping, radiomic signature discovery, clinical outcome prediction, image dose quantification, dose-response modeling, radiation adaptation, and image generation,[18] and therefore also impacted the area of auto-segmentation of OARs in the H&N region[19–21] so as to provide a qualitative support for guiding critical treatment planning and delivery decisions. In this review, we provide a detailed overview of the existing studies for auto-segmentation of OARs in the H&N region by systematically outlining, analyzing, and categorizing the relevant publications in the field from 2008 to date.

2. METHODOLOGY

In May 2020, a search was conducted on the *Web of Science* (https://apps.webofknowledge.com) and *PubMed* (https://www.ncbi.nlm.nih.gov/pubmed/) on-line citation indexing services, with the topic keyword (auto OR automatic) AND (segmentation OR contouring OR delineation) AND (head AND neck) with a time span from 2008 to date. Studies not concerned with OAR auto-segmentation in the H&N region, as well as longitudinal studies and dosimetric studies without geometric validations were excluded. The obtained relevant publications were further supplemented with selected publications found in their list of references. A detailed analysis of the resulting publications was then conducted from the perspective of *image modality*, *OAR*, *image database*, *methodology*, *ground truth*, *performance metrics*, and *segmentation performance*.

3. RESULTS

In the field of OAR auto-segmentation for RT planning in the H&N region, the search on the *Web of Science* and *PubMed* yielded, respectively, 281 and 257 results. After reviewing their abstracts, 49 were considered to be relevant and were further supplemented with selected publications from their list of references. In total, we collected 75 publications[22–96] focused on RT planning and three studies focused on hyperthermia therapy planning[97–99] from 2008 to date (Fig. 1), along with three review papers related to auto-segmentation in the H&N region.[19–21] The results of analyzing these publications from different perspectives are presented in the following subsections.

3.A. Image modality

The RT planning is primarily performed using *CT imaging* information because the data on electron density, required for the calculation of the radiation beam energy absorption and dose distribution, is derived directly from the CT image intensities[100,101]. As a result, segmentation of the target volumes and OARs has to be generated from the planning CT images, therefore making CT the prevailing image modality also for auto-segmentation approaches (Table I). While CT images provide a good visibility of the bony anatomy, the contrast differences between various soft tissues are relatively low, and can be to a certain degree improved by using an intravenous contrast enhancement agent.[68,84,95,98,99]

On the other hand, *MR imaging* gained a broad adoption because of its superior soft tissue contrast resolution

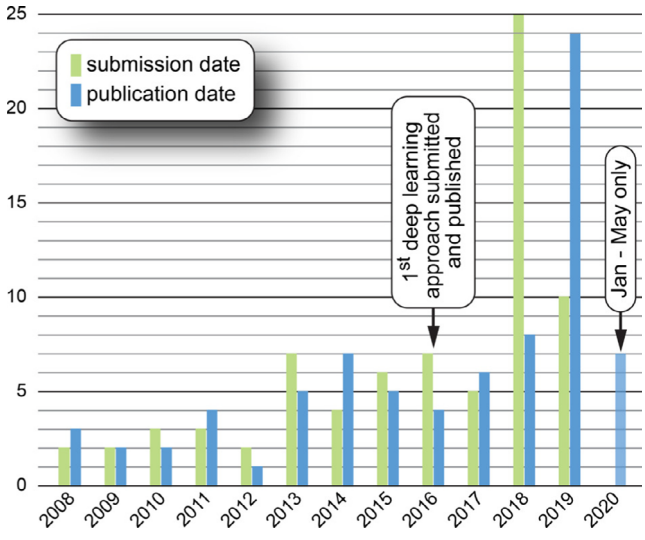

FIG. 1. The chronological distribution of 78 reviewed publications in the field of organ at risk auto-segmentation in the head and neck region.

TABLE I. Image modalities used for auto-segmentation of organs at risk in the head and neck region for the purpose of radiotherapy planning, and the corresponding references.

Image modality
Computed tomography (CT)
Conventional CT[22–24,26–28,30–42,44–57,59–62,64–70,72,73,76–82,84–93,95,96,98,99]
Dual-energy CT (DECT)[29]
Magnetic resonance (MR)
T1-weighted MR[38,40,43,57–59,63,68,74,88,89,94]
T2-weighted MR[38,63,75,83,97]
Ultrasound (US)[25]

compared to CT images and various imaging setups. In the recent consensus for CT-based manual delineation guidelines for OARs in the H&N region,[102] it is strongly recommended to use, besides CT, also MR images to facilitate the delineation of several soft tissue OARs. Auto-segmentation of OARs from MR images can be also performed independently,[58,63,68,74,94,97] and the resulting segmentation masks are then propagated to the planning CT images by applying the geometric transformations of the corresponding MR-to-CT image registration. Alternatively, image registration can be performed first, and auto-segmentation is then performed simultaneously on both image modalities.[57,88,89] While the obtained results combine the information of the CT and MR image modality, both approaches rely on an accurate intrapatient multimodal image registration.[103–105]

Similar challenges are present in the case of adaptive RT, when cone beam CT (CBCT) images are often obtained between sessions for verifying the patient setup or adjusting the treatment plan to anatomical changes, as they can be acquired faster and at lower radiation doses in comparison to classical CT images. As a pretreatment planning CT image is always acquired and segmented to plan the dose distribution, auto-segmentation of CBCT images can be obtained by CBCT-to-CT registration followed by propagation of presegmented OARs back to CBCT images.[106,107]

Other image modalities can be optionally provided to obtain complementary information, for example, positron emission tomography (PET) images can be acquired simultaneously with CT or MR images, however, they are not used for OAR but rather for target volume auto-segmentation.[68] On the other hand, specific OARs (e.g., the carotid artery) can be successfully auto-segmented only from ultrasound (US) images,[25] while the feasibility of using dual-energy CT (DECT) has been recently explored from the perspective of selecting the optimal energy level for generating the virtual monoenergetic image,[108] in which different H&N OARs can be segmented.[29]

3.B. Organ at risk

Auto-segmentation is commonly performed for OARs whose RT-induced damage proved to be linked to late complications that may endanger the life of the patient or considerably reduce its quality (Table II).[109–111] Major salivary glands, that is, the *parotid* and *submandibular glands*, are among the most frequently delineated OARs because of their importance for a sufficient secretion and proper composition of saliva, and therefore for the prevention of xerostomia, and associated problems with swallowing, speech, and oral health. The *eyeballs*, *vitreous humor*, *optic chiasm*, *optic nerves*, *lens*, *sclera*, *cornea*, and *lacrimal glands* have to be spared to prevent optic neuropathy leading to an impaired vision or even blindness, while the commonly delineated nervous tissues are the *spinal cord* and *brain*, including the *brainstem*, *cerebrum*, *cerebellum*, and *pituitary gland*. In particular, segmentation of the former is of critical importance due to potentially devastating consequences (i.e., tetraplegia) of its over-irradiation. The *pharyngeal constrictor muscles* and *cervical esophagus* with the *cricopharyngeal inlet* have to be spared to prevent the swallowing dysfunction.

Other relevant OARs include the *thyroid*, *larynx*, *trachea*, *cochlea*, *chewing muscles*, *oral cavity*, *mastoids*, *temporomandibular joints*, *mandible*, and *brachial plexus*, as their malfunction is connected with a variety of problems (e.g., hypothyroidism, swallowing problems, including aspiration with resulted pulmonary morbidity, hearing decrease, osteoradionecrosis, brachial plexopathy). Although the *lips* and *carotid arteries* are commonly delineated for the purpose of RT planning, reports on auto-segmentation of these OARs are very limited.[25]

3.C. Image database

Auto-segmentation methods are validated on a wide range of image databases (Table III). Several methods utilize a subset of all available samples as an atlas or as a training set, while the remaining samples then constitute the test set, which serves to evaluate the auto-segmentation performance and accuracy. When the set of all available samples is relatively small, *cross-validation* (*k*-fold or, when *k* equals the number of samples, leave-one-out) is commonly employed to enable all available samples to be used for testing.

Among the reviewed publications, one database[36] stands out as it was devised from CT images of 3495 patients resulting in 825−1702 training set samples for each studied OAR. On the other hand, there are several databases of H&N images that are publicly available. The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net/), an open-access resource platform of medical images for cancer research,[112,113] currently contains 12 databases of the H&N region, for example, the *Head-Neck Cetuximab* (https://doi.org/10.7937/K9/TCIA.2015.7AKGJUPZ),[22,30,46,60,66] *Head-Neck-PET-CT* (https://doi.org/10.7937/K9/TCIA.2017.8oje5q00),[22,30,46,114] *TCGA-HNSC* (https://doi.org/10.7937/K9/TCIA.2016.LXKQ47MS)[22,60] and *Data from Head and Neck Cancer CT Atlas* (https://doi.org/10.7937/K9/TCIA.2017.umz8dv6s)[22,115] CT image databases, the *RT-MAC* (https://doi.org/10.7937/tcia.2019.bcfjqfqb)[116] MR image database, or

TABLE II. Organs at risk in the head and neck region involved in auto-segmentation for the purpose of radiotherapy planning, and the corresponding references.

Organ at risk
Parotid glands[22–24,26–32,34–37,40,42,45–58,60,63–66,68–70,72,73,75–78,80,82–84,86,87,90,91,93,95]
Submandibular glands[22–24,26,30–32,34,35,40,42,46,50,51,53,55,60,65,66,69,70,77,78,80,82,86,87,95]
Brainstem[22–24,26,27,29–32,35,36,38,40,42,43,46–50,52–56,59,60,66,68,69,73,76,80,82,84,86,87,89,90,92–95,97–99]
Brain, cerebrum and cerebellum[23,36,60,82,94,97–99]
Temporal lobes[27,30]
Hippocampus[38]
Pituitary gland[30,33,94]
Spinal cord and spinal canal[22,23,26–28,30,32,34–36,42,47,48,51–53,58,60,63,65,68,73,80,82,87,90,95,97–99]
Cerebrospinal fluid[97]
Eyeballs and vitreous humor[22,29,30,33,36,38,43,47,48,59,60,62,65,68,73,79,82,89,94,96–99]
Optic chiasm[22,24,27,30,31,36,38,40,43,46,49,54,55,59,65,66,70,73,80,88,89,94]
Optic nerves[22,24,27,29–31,33,34,36–38,40,43,46,47,49,54,55,59,60,62,65,66,69,74,79,80,88,89,94,96,98,99]
Lens[29,30,33,36,47,59,60,96–99]
Sclera[97–99]
Cornea[99]
Lacrimal glands[60]
Extraocular muscle[62]
Mandible[23,24,26,28,30–32,34–36,39–42,44,46–49,51–56,58,60,65,66,69,78,80,82,86,90,92,93,95]
Oral cavity[23,26,28,30,32,35,42,47,50,52,53,80]
Temporo-mandibular joints[30,42,47]
Mastoids[47]
Chewing muscles[87,95]
Pharyngeal constrictor muscles[23,26,28,32,40,50–53,65,77,80,87]
Cervical esophagus and cricopharyngeal inlet[23,26,28,32,36,42,50–53,61]
Thyroid[23,30,37,44,85,98,99]
Larynx[26,28,30,32,35,40,42,47,50–53,65,77,80]
Trachea[30,52,63]
Cochlea[26,32,36,53,60,77,80]
Brachial plexus[30,67,71,81]
Carotid artery[23,25]

the *QIN-HEADNECK* (https://doi.org/10.7937/K9/TCIA.2015.K0F5CGLI)[117,118] PET-CT image database.

Although many TCIA databases include reference H&N OAR delineations, they are associated with considerable variability because of the lack of a standardized delineation protocol. As a result, some of them were augmented and/or combined into new publicly available databases, for example, the manual delineations of 28 OARs in 140 CT images from the *Head-Neck Cetuximab* and *Head-Neck-PET-CT* databases as well as in 175 CT images from an in-house database (https://github.com/uci-cbcl/UaNet#Data),[30] the manual delineations of 21 OARs in 31 CT images from the *Head-Neck Cetuximab* and *TCGA-HNSC* databases forming the *TCIA test & validation radiotherapy CT planning scan dataset* (TCIA-RT) (https://github.com/deepmind/tcia-ct-scan-dataset) database,[60] or the manual delineations of nine OARs in 48 CT images from the *Head-Neck Cetuximab* database forming the *Public domain database for computational anatomy* (PDDCA) (http://www.imagenglab.com/newsite/pddca/) database.[66]

Examples of publicly available databases that do not originate from TCIA include the *StructSeg* (https://structseg2019.grand-challenge.org/Dataset/) database consisting of 50 CT images with 22 manually delineated OARs, and the *MRI-RT* (https://figshare.com/s/a5e09113f5c07b3047df) database[105] consisting of 15 CT and 15 MR images of the same patients with 23 manually delineated OARs from the H&N region.

3.D. Methodology

The most common approach for segmenting OARs from H&N images is *atlas-based auto-segmentation* (ABAS), which has been frequently implemented in commercial tools.[5,66,119] In ABAS, the image undergoing segmentation is first registered to images with known reference segmentation masks that form the atlas, and then these reference masks are, according to the geometrical transformations obtained from the registration, propagated back and fused into the final segmentation. To improve the results of ABAS, contour and level set refinement methods were applied to enhance the boundaries of the segmented OARs. Also, models of intensity or models of shape and appearance were generated to restrain the registration, and machine learning techniques were used to improve feature classification (Table IV).

Recently, DL techniques have been applied to various steps of the RT workflow, including auto-segmentation,[17,18,120] resulting in a superior performance in comparison to other classification and regression methods. The most popular architecture for *DL-based auto-segmentation* of medical images is the U-Net,[9] which originates from the fully convolutional neural networks (CNNs) and consists of a contracting path and an expansive path in the shape of the letter U. Through convolution, activation, and pooling, the contracting path reduces spatial while increasing feature information, and the expansive path performs up-convolutions of the feature and spatial information with lateral concatenations of low- and high-level feature maps. The architecture was released as open-source (https://lmb.informatik.uni-freiburg.de/resources/opensource/unet/) and was, with additional augmentations, extended to the 3D U-Net,[10] V-Net[11] and AnatomyNet.[46] On the other hand, the DeepMedic[13] framework is based on 3D CNNs and consists of two parallel convolutional paths for processing the input at multiple scales to achieve a large receptive field for classification while using small convolutional kernels that are associated with relatively low computational costs. Although it was originally developed for segmenting brain lesions, it was also released as open-source (https://biomedia.doc.ic.ac.uk/software/deepmedic/) and consequently applied in many different fields, including H&N OAR auto-segmentation, as well as augmented into new architectures, such as the DeepVoxNet.[15]

TABLE III. Number of samples included in image databases used for auto-segmentation of organs at risk in the head and neck region for the purpose of radiotherapy planning, and the corresponding references.

	Image database (number of samples)
5−10	5:L,[83] 7,[98] 7,[38] 10,[77] 10:L,[95] 10:L,[78] 10:L [87]
11−18	11:L,[97] 12:L,[58] 12,[67] 13,[93] 14:L,[68] 14:L,[29] 5|10,[74] 15:5F,[28] 8|8,16:L,[72] 16,[90] 18:L,[76] 18:L,[64] 18:L[99]
20−25	20:L,[85]y 20,[89] 20,[84] 20,[59] 20,[27] 21:L,[61] 14|10,[88] 25:L▲,[69] 15|10,[86] 15|10,[40] 10|15[91]
30−33	30,[62] 15|15,[63] 30:L,[79] 20|10▲,[70] 32,[82] 22|10▲,[40,55] 33:L▲,[41] 33:2F▲ [56]
39−50	25|14▲,[49] 25|15▲,[66] 25|15▲,[39] 30|10,[52] 40,[80] 41,[96] 41,[25] 42,[75] 44:5F,[57] 45:L,32▲,[35] 33|15▲,[31,54] 32+6|10▲,[24] 50:5F,[65] 50:5F,[41] 40|10 [33]
70−95	70,[38] 74,[24] 48+12|20,[43] 70|17,[26] 10|80,[81] 70|20,[53] 70+10|15,[32] 100:L,[73] 100:5F[48]
>100	52+8|49,[39] 100|10,[44] 100+20|20,[37] 142|15,[50] 185:4F,[47] 160+20|20,[42] 246*,[51] 234|20,15▲,[45] 261|10▲,[46] 215|100,[30] 328|20,[22] 389+51|46,+6|24•,15▲,[60] 475+5|20 [34]
>500	549+40|104 [23]
>1000	(660+165−1362+340)|(48−168),24• [36]

Legend: *n* – number of cases with a model or without a training set; *m*\|*n* – *m* cases for training, *n* cases for testing; *m*+*k*\|*n* – *m* cases for training (if omitted, models are used), *k* cases for model selection, *n* cases for testing; *n*:*k*F – *n* cases with the *k*-fold cross-validation; *n*:L – *n* cases with the leave-one-out validation; * – for 30 patients, 2 or more images available, together 36\|262; ▲ – evaluated on the PDDCA database;[66] • – evaluated on the TCIA-RT database.[60]

Other DL architectures adopt specific mechanisms to improve the auto-segmentation of OARs in the H&N region. For example, the self-channel-and-spatial-attention neural network (SCSA-Net)[24] is equipped with attention learning, a technique for strengthening the discriminative ability of the segmentation network with minimal or no additional layers, the DenseNet[40] employs adversarial learning, a technique where two CNNs compete in generating more accurate predictions, while the regional CNN (R-CNN)[28] can be used for rapidly detecting the location of OARs before actual segmentation.

TABLE IV. Methodology applied for auto-segmentation of organs at risk in the head and neck region for the purpose of radiotherapy planning, and the corresponding references.

Methodology
Atlas[27,29,34,44,52,58,59,61,68,69,71,73,78–81,84,85,87,89–91,93,94,99]
with shape/appearance models[38,66,76,77,82,86,92,95]
with intensity models[97–99]
with feature classification[35,63,72,75,83,86]
with contour refinement[72,76,92]
with level set refinement[91]
Feature classification[64,74]
Localization model and feature classification[51,56]
Level-set statistical model[88,89]
Shape models[25,62,96]
Deep learning[23,24,37,40,47,49,54,57,65,70]
with U-Net and its versions[22,28–31,33,36,39,41–43,45,46,50,55,60]
with DeepMedic and its versions[26,32,53]

3.E. Ground truth

The quality of the resulting auto-segmentation is evaluated by the comparison against the corresponding reference segmentation, often referred to as the *ground truth*. Manual delineation (contouring) of OARs in images performed by human experts (e.g., radiation oncologists, diagnostic radiologists) is the main approach for generating the ground truth. However, it is a time-consuming (e.g., 3–6 hours per image for up to 20 OARs[19,87,98]), tedious, and costly task that is limited by the subjective human interpretation of organ boundaries, which is manifested through the intra- and interobserver variability in the delineation (Table V). Most studies therefore rely on a single set of ground truth per image, nevertheless, studies report also two,[32,60,63,79,88,93] three,[22,25,41,58,75,97,99] four,[71,98] five[77], or even eight[89] independently obtained sets of ground truth per image. An anatomically validated ground truth was introduced for a single OAR, that is, the brachial plexus,[6,121] so that its manual delineations obtained from high-resolution MR images of up to 12 cadavers were first validated by dissection and then registered to corresponding CT images to obtain the ground truth for the purpose of RT planning.

In some cases, multiple ground truth sets were combined into a consensus by generating probability maps,[89] (weighted) majority voting,[44,69] performing intensity-based patch-based label fusion (Patch),[67] applying the simultaneous truth and performance level estimation (STAPLE) expectation maximization algorithm[67,77,81,89] that estimates the correct segmentation by weighting each input by its estimated performance level, or applying the similarity and truth estimation for propagated segmentations (STEPS) algorithm[58] that introduces a spatially variant image similarity term into STAPLE. Alternatively, a less labor intensive but relatively biased approach for generating the ground truth is to manually correct the auto-segmentation boundaries[73,77,80,84,85,87,93] or to merge different auto-segmentation results with, for example, the STAPLE algorithm.[96]

To mitigate the intra- and interobserver delineation variability, well-defined guidelines have been proposed[102,121–128] that help ensuring the consistency and accuracy of manual delineation. The most established consensus[102] encompasses a complete set of OARs in the H&N region, with the expert recommendation to always include the parotid glands, submandibular glands, spinal cord, and pharyngeal constrictor muscles in the RT plan. Other guidelines are focused on OARs involved in the nasopharyngeal carcinoma (i.e., the temporal lobe, parotid glands, spinal cord, and inner and middle ear),[122] swallowing (i.e., the pharyngeal constrictor muscles, cricopharyngeal muscle, esophagus inlet muscles, cervical esophagus, base of tongue, and larynx),[124] salivary functioning (i.e., the parotid glands, submandibular glands, sublingual gland, and minor salivary glands in the soft palate, lips, and cheeks),[125] hearing and balance (i.e., the inner and middle ear),[126] brachial plexopathy (i.e., the brachial plexus and

TABLE V. Observer variability of manual delineations of organs of risk in the head and neck region, and the corresponding references (cf. Table VI for the list of metrics).

Observer variability	
Parotid glands	
DC (%)	91[m,f] (o = 5,p = 10,S),[77] 91 (o = 2,p = 32),[60] 89 ± 3,[32] 87 ± 3 (o = 2,p = 24,•),[60] 84 ± 4 (o = 3,p = 12),[58] 91[m,f],[22] 83 ± 2 (o = 8,p = 16),[145] 81 (o = 2,p = 13),[63] 77 ± 8 (o = 32,p = 1)[143]
SC (%)	sDC: 94.4 ± 2.8 (τ = 2.85mm,o = 2,p = 24,•)[60]
HD (mm)	HD91[m,f]: 10.7 ± 4.4 (o = 3,p = 12)[58]; DTA91[m,f]: 91[m,f] (o = 5,p = 10,S)[77]; HD91[m,f]: 91[m,f],[22] 5.0 ± 1.7 (o = 3,p = 12)[58]
ASD (mm)	ASSD: 1.8 ± 0.2[32]; ASD91[m,f]: 1.4 ± 0.5 (o = 3,p = 12)[58]; DTA91[m,f]: 91[m,f] (o = 5,p = 10,S)[77]
Submandibular glands	
DC (%)	91(o = 2,p = 64)[60] , 91[m,f] (o = 5,p = 10,S),[77] 87 ± 5,[32] 91[m,f],[22] 83 ± 20 (o = 2,p = 24,•),[60] 77 ± 5 (o = 8,p = 16)[145]
SC (%)	sDC: 89 ± 21.2 (τ = 2.02mm,o = 2,p = 24,•)[60]
HD (mm)	DTA91[m,f]: 91[m,f] (o = 5,p = 10,S)[77]; HD91[m,f]: 91[m,f22]
ASD (mm)	ASSD: 1.5 ± 0.2[32]; DTA91[m,f]: 91[m,f] (o = 5,p = 10,S)[77]
Brainstem	
DC (%)	91[m,f] (o = 3,p = 11),[97] 92 (o = 2,p = 45),[60] 90 ± 2 (o = 2,p = 24,•),[60] 91[m,f],[22] 84(82,85) (intra,o = 4,p = 7),[98] 83 ± 3 (o = 8,p = 16),[145] 83 ± 10 (o = 8,p = 20),[89] 91[m,f] (o = 3,p = 13),[99] 78(73,85) (o = 4,p = 7),[98] 68 ± 12,[32] 66 ± 17 (o = 31, p = 1)[143]
SC (%)	sDC: 96.7 ± 2.5 (τ = 2.5mm,o = 2,p = 24,•)[60]; sPPV: 91[m,f] (τ = 2mm,o = 8,p = 20)[89]
HD (mm)	HD91[m,f]: 91[m,f] (o = 3,p = 13)[99]; HD91[m,f]: 91[m,f] (o = 3,p = 11),[97] 91[m,f22]
ASD (mm)	ASSD: 2.2 ± 0.5[32]; ASD91[m,f]: 1.1(0.9,1.2) (intra,o = 4,p = 7)[98], 91[m,f] (o = 3,p = 13)[99], 1.7(1.1,2.4) (o = 4,p = 7)[98]
SSD (mm)	SDTA91[m,f]: 0.8 (o = 8,p = 20,p)[89]; SDTA91[m,f]: −3.9 (o = 8,p = 20,p)[89]; SDTA91[m,f]: 7.5 (o = 8,p = 20,p)[89]
Brain, cerebrum (CBR) and cerebellum (CBE)	
DC (%)	99 ± 0.3 (o = 2,p = 24,•),[60] 99 (o = 2,p = 75),[60] 99 (CBR,intra,o = 4,p = 7),[98] 98 ± 1 (o = 10,p = 1),[143] 91[m,f] (CBR,o = 3, p = 13),[99] 91[m,f] (CBR,o = 3,p = 11),[97] 94(93,95) (CBR,o = 4,p = 7),[98] 91[m,f] (CBE,o = 3,p = 11),[97] 94(91,95) (CBE,intra, o = 4,p = 7),[98] 91[m,f] (CBE,o = 3,p = 13),[99] 86(84,88) (CBE,o = 4,p = 7)[98]
SC (%)	sDC: 96.2 ± 1.1 (τ = 1.01mm,o = 2,p = 24,•)[60]
HD (mm)	HD91[m,f]: 91[m,f] (CBE,o = 3,p = 13),[99] 91[m,f] (CBR,o = 3,p = 13)[99]; HD91[m,f]: 91[m,f] (CBR,o = 3,p = 11),[97] 91[m,f] (CBE,o = 3, p = 11)[97]
ASD (mm)	ASD91[m,f]: 0.4 (CBR,intra,o = 4,p = 7),[98] 0.9(0.6,1.2) (CBE,intra,o = 4,p = 7),[98] 91[m,f] (CBR,o = 3,p = 13),[99] 91[m,f] (CBE, o = 3,p = 13),[99] 2.2(1.8,2.5) (CBE,o = 4,p = 7),[98] 2.4(2.0,2.9) (CBR,o = 4,p = 7)[98]
Temporal lobes	
DC (%)	82 ± 2 (o = 8,p = 16)[145]
Pituitary gland	
DC (%)	65 ± 8 (o = 8,p = 16)[145]
Spinal cord and spinal canal	
DC (%)	95 (canal,o = 2,p = 23),[60] 94 ± 2 (canal,o = 2,p = 24,•),[60] 91[m,f] (o = 2,p = 15),[63] 91[m,f] (o = 3,p = 11),[97] 88 (o = 2, p = 24),[60] 85(84,87) (intra,o = 4,p = 7),[98] 84 ± 5 (o = 2,p = 24,•),[60] 91[m,f],[22] 80 ± 7 (o = 29,p = 1),[143] 79 ± 7 (o = 3, p = 12),[58] 79(73,84) (o = 4,p = 7),[98] 91[m,f] (o = 3,p = 13),[99] 77 ± 4 (o = 8,p = 16),[145] 71 ± 7[32]
SC (%)	sDC: 99.8 ± 0.4 (τ = 2.93mm,o = 2,p = 24,•),[60] 95 ± 2 (canal,τ = 1.17mm,o = 2,p = 24,•)[60]
HD (mm)	HD91[m,f]: 91[m,f] (o = 3,p = 13),[99] 7.1 ± 5.2 (o = 3,p = 12)[58]; HD91[m,f]: 91[m,f] (o = 3,p = 11),[97] 91[m,f],[22] 4.6 ± 3.1 (o = 3, p = 12)[58]
ASD (mm)	ASSD: 4.4 ± 1.9[32]; ASD91[m,f]: 0.6 (intra,o = 4,p = 7),[98] 91[m,f] (o = 3,p = 13),[99] 1(0.81,1.3) (o = 4,p = 7)[98]; ASD91[m,f]: 1.6 ± 0.8 (o = 3,p = 12)[58]
Cerebrospinal fluid	
DC (%)	91[m,f] (o = 3,p = 11)[97]
HD (mm)	HD91[m,f]: 91[m,f] (o = 3,p = 11)[97]
Eyeballs and vitreous humor (VH)	
DC (%)	91[m,f] (VH,o = 3,p = 11),[97] 95 (o = 2,p = 19),[60] 93 ± 2 (o = 2,p = 24,•),[60] 91(90,92) (VH,intra,o = 4,p = 7),[98] 91[m,f],[22] 89 ± 1 (o = 8,p = 16),[145] 91[m,f] (VH,o = 3,p = 13),[99] 86(82,89) (VH,o = 4,p = 7),[98] 85 ± 3 (+ eye muscles,o = 2,p = 15),[79] 83 ± 9 (o = 8,p = 20)[89]
SC (%)	sDC: 96 ± 3 (τ = 1.65mm,o = 2,p = 24,•)[60]; sPPV: 91[m,f] (τ = 2mm,o = 8,p = 20)[89]
HD (mm)	HD91[m,f]: 91[m,f] (VH,o = 3,p = 13),[99] 4.9 ± 0.6 (+ eye muscles,o = 2,p = 15)[79]; HD91[m,f]: 91[m,f] (VH,o = 3,p = 11),[97] 91[m,f22]
ASD (mm)	ASD91[m,f]: 0.4 (VH,intra,o = 4,p = 7),[98] 91[m,f] (VH,o = 3,p = 13),[99] 0.7(0.5,1.1) (VH,o = 4,p = 7)[98]; ASD91[m,f]: 0.5 ± 0.2 (+ eye muscles,o = 2,p = 15)[79]
SSD (mm)	SDTA91[m,f]: 0.5 (o = 8,p = 20,p)[89]; SDTA91[m,f]: −2.8 (o = 8,p = 20,p)[89]; SDTA91[m,f]: 3.4 (o = 8,p = 20,p)[89]
Optic chiasm	
DC (%)	91[m,f] (o = 2,p = 10),[88] 91[m,f],[22] 39 ± 23 (o = 8,p = 20),[89] 38 ± 8 (o = 8,p = 16)[145]

TABLE V. Continued.

Observer variability	
SC (%)	sPPV: $91^{m,f}$ (τ = 2mm,o = 8,p = 20)[89]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 2,p = 10)[88]; $HD91^{m,f}$: $91^{m,f}$[22]
ASD (mm)	$ASD91^{m,f}$: $91^{m,f}$ (o = 2,p = 10)[88]
SSD (mm)	$SDTA91^{m,f}$: 0.7 (o = 8,p = 20,p)[89]; $SDTA91^{m,f}$: −2.0 (o = 8,p = 20,p)[89]; $SDTA91^{m,f}$: 4.7 (o = 8,p = 20,p)[89]
Optic nerves	
DC (%)	$91^{m,f}$ (o = 2,p = 10),[88] 79 ± 5 (o = 2,p = 24,•),[60] 77 ± 6 (o = 2,p = 17),[60] 73 ± 4 (o = 2,p = 15),[79] 70(65,76) (intra,o = 4, p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 60(50,66) (o = 4,p = 7),[98] $91^{m,f}$,[22] 57 ± 9 (o = 8,p = 16),[145] 50 ± 17 (o = 8,p = 20)[89]
SC (%)	sDC: 97 ± 3 (τ = 2.5mm,o = 2,p = 24,•)[60]; sPPV: $91^{m,f}$ (τ = 2mm,o = 8,p = 20)[89]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 2,p = 10),[88] 2.9 ± 0.5 (o = 2,p = 15),[79] $91^{m,f}$ (o = 3,p = 13)[99]; $HD91^{m,f}$: $91^{m,f}$[22]
ASD (mm)	$ASD91^{m,f}$: 0.6(0.4,0.7) (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 0.9(0.6,1.7) (o = 4,p = 7)[98]; $ASD91^{m,f}$: $91^{m,f}$ (o = 2, p = 10),[88] 0.5 ± 0.1 (o = 2,p = 15)[79]
SSD (mm)	$SDTA91^{m,f}$: 0.3 (o = 8,p = 20,p)[89]; $SDTA91^{m,f}$: −2.3 (o = 8,p = 20,p)[89]; $SDTA91^{m,f}$: 4.0 (o = 8,p = 20,p)[89]
Lens	
DC (%)	$91^{m,f}$ (o = 3,p = 11),[97] 88 ± 10 (o = 2,p = 73),[60] 87 ± 8(o = 2,p = 24,•),[60] 80(75,85) (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3, p = 13),[99] 70 ± 5 (o = 8,p = 16),[145] 68(55,76) (o = 4,p = 7)[98]
SC (%)	sDC: 98 ± 3 (τ = 0.98mm,o = 2,p = 24,•)[60]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 13)[99]; $HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 11)[97]
ASD (mm)	$ASD91^{m,f}$: 0.3(0.2,0.4) (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 0.7(0.4,1.2) (o = 4,p = 7)[98]
Sclera	
DC (%)	$91^{m,f}$ (o = 3,p = 11),[97] 63(62,67) (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 48(30,56) (o = 4,p = 7)[98]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 13)[99]; $HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 11)[97]
ASD (mm)	$ASD91^{m,f}$: 0.5 (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 0.9(0.6,1.8) (o = 4,p = 7)[98]
Cornea	
DC (%)	$91^{m,f}$ (o = 3,p = 13)[99]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 13)[99]
ASD (mm)	$ASD91^{m,f}$: $91^{m,f}$ (o = 3,p = 13)[99]
Lacrimal glands	
DC (%)	67 ± 10 (o = 2,p = 24,•),[60] 63 ± 13 (o = 2,p = 75)[60]
SC (%)	sDC: 93.9 ± 4.7 (τ = 2.5mm,o = 2,p = 24,•)[60]
Mandible	
DC (%)	95 (o = 2,p = 74),[60] 94 ± 2 (o = 2,p = 24,•),[60] 94 ± 3,[32] 92 (o = 3,p = 50),[41] 89 ± 2 (o = 8,p = 16),[145] 87 ± 7 (o = 18, p = 1),[143] 85 ± 4 (o = 3,p = 12)[58]
SC (%)	sDC: 98 ± 2 (τ = 1.01mm,o = 2,p = 24,•)[60]
HD (mm)	$HD91^{m,f}$: 8.9 ± 3.2 (o = 3,p = 12)[58]; $HD91^{m,f}$: 3.9 ± 1.6 (o = 3,p = 12)[58]
ASD (mm)	ASSD: 1.2 ± 0.2[32]; $ASD91^{m,f}$: 0.9 ± 0.5 (o = 3,p = 12)[58]
Oral cavity	
DC (%)	94 ± 5,[32] 81 ± 4 (o = 8,p = 16)[145]
ASD (mm)	ASSD: 2.9 ± 0.6[32]
Temporo-mandibular joints	
DC (%)	50 ± 18 (o = 8,p = 16)[145]
Pharyngeal constrictor muscles	
DC (%)	76 ± 8 (inf),[32] $91^{m,f}$ (o = 5,p = 10,S),[77] 72 ± 7 (mid),[32] 54 ± 8 (inf),[32] 50 ± 8 (middle,o = 8,p = 16),[145] 50 ± 9 (inferior, o = 8,p = 16),[145] 44 ± 7 (superior,o = 8,p = 16)[145]
HD (mm)	$DTA91^{m,f}$: $91^{m,f}$ (o = 5,p = 10,S)[77]
ASD (mm)	ASSD: 1.5 ± 0.2 (mid),[32] 1.7 ± 0.3 (inf),[32] 2.1 ± 0.3 (sup)[32]; $DTA91^{m,f}$: $91^{m,f}$ (o = 5,p = 10,S)[77]
Cervical esophagus	
DC (%)	64 ± 15[32]
ASD (mm)	ASSD: 2.0 ± 0.6[32]
Thyroid	
DC (%)	$91^{m,f}$ (o = 3,p = 13),[99] 84(71,92) (intra,o = 4,p = 7),[98] 82 ± 3 (o = 8,p = 16),[145] 76(53,89) (o = 4,p = 7)[98]
HD (mm)	$HD91^{m,f}$: $91^{m,f}$ (o = 3,p = 13)[99]
ASD (mm)	$ASD91^{m,f}$: 0.8(0.4,1.8) (intra,o = 4,p = 7),[98] $91^{m,f}$ (o = 3,p = 13),[99] 1.9(0.5,4.7) (o = 4,p = 7)[98]
Larynx	
DC (%)	86 ± 11 (supraglottic),[32] $91^{m,f}$ (o = 5,p = 10,S),[77] 73 ± 18 (glottic),[32] 60 ± 5 (supraglottic,o = 8,p = 16),[145] 49 ± 9 (glottic, o = 8,p = 16)[145]

TABLE V. Continued.

Observer variability	
HD (mm)	DTA91^{m,f}: 91^{m,f} (o = 5,p = 10,S)[77]
ASD (mm)	ASSD: 1.4 ± 0.4,[32] 1.8 ± 0.4 (supraglottic)[32]; DTA91^{m,f}: 91^{m,f} (o = 5,p = 10,S)[77]
Trachea	
DC (%)	91^{m,f} (o = 2,p = 12)[63]
Cochlea	
DC (%)	78 ± 8 (o = 2,p = 24,•),[60] 76 ± 9 (o = 2,p = 8),[60] 91^{m,f} (o = 5,p = 10,S),[77] 50 ± 13,[32] 37 ± 10 (o = 8,p = 16)[145]
SC (%)	sDC: 96 ± 4 (τ = 1.25mm,o = 2,p = 24,•)[60]
HD (mm)	DTA91^{m,f}: 91^{m,f} (o = 5,p = 10,S)[77]
ASD (mm)	ASSD: 1.1 ± 0.4[32]; DTA91^{m,f}: 91^{m,f} (o = 5,p = 10,S)[77]
Brachial plexus	
DC (%)	26 (o = 5,p = 1,S*)[71]
VC (%)	TPR: 36 (o = 5,p = 1,S*)[71]
HD (mm)	HD91^{m,f}: 22.2 (o = 5,p = 1,S*)[71]

Legend: m — median, average not reported; f — value estimated from a figure, exact value not reported; o — number of observers; p — number of patients; intra — intra-observer variability; S — compared against the STAPLE consensus among other physicians; S* — comparison of trainee contours against the STAPLE consensus among four other expert physicians; P — compared against the probability map consensus among other physicians; • — evaluated on the TCIA-RT database;[60] +eye muscles — the eyes and eye muscles were segmented as one organ; τ — size of the volumetric neighborhood.

adjacent structures, esophagus, spinal cord, and trachea),[121,123,127] and optic neuropathy (i.e., the optic chiasm).[128]

3.F. Performance metrics

The agreement between the ground truth and the resulting auto-segmentation is quantitatively evaluated by various overlap and distance metrics,[129] computed over the corresponding binary segmentation masks (Table VI). The overlap metrics originate from the statistical measures of the performance of a binary classification test, and the *Dice coefficient* is the standard and widely accepted metrics for volumetric mask overlap that measures the harmonic average of the classification precision and recall (i.e., the F_1 score). Variations of the volumetric coefficient include the *sensitivity* and *positive predictive value* (often referred to as the *inclusion*), which measure the ratio of correctly segmented voxels, while the *specificity* measures the ratio of correctly nonsegmented voxels and the *false discovery rate* measures the ratio of incorrectly segmented voxels. On the other hand, surface coefficients measure the overlap of the corresponding mask surfaces.

Contrary to the overlap metrics, the distance metrics evaluate the mutual proximity of the segmentation mask surfaces. Within this group, the most established are the *Hausdorff distance* and its variations, which measure the maximal distance between any voxel on the mask surface to the other mask surface, as well as variations of the *average surface distance*, which measure the distance between voxels on the mask surface to the closest voxels on the other mask surface.

3.G. Segmentation performance

The performance of different auto-segmentation methods from the perspective of different metrics and OARs is presented in Table VII, which summarizes the comparisons of auto-segmentation results to the corresponding ground truth obtained by manual delineation*. A systematic and relatively unbiased evaluation of different methods can be obtained through computational challenges, which have in the past decade gained increased popularity and become the standard for validation of methods in the field of biomedical image analysis.[130] In such a competition-oriented setting, the challenge organizers first release images with the ground truth that are used by the participating teams for method development, and then the methods are evaluated on images for which the ground truth is known to organizers only.

To this date, five *H&N auto-segmentation challenges* have been organized. In 2009†, five different teams attempted to segment the mandible and brainstem from 25 CT images (10 for training, 15 for testing).[92] The second challenge was organized by the same group in 2010‡, when the same image database was used but six different teams attempted to segment the parotid glands instead.[91] In 2015§, six different teams participated in a challenge to segment the brainstem, mandible, optic chiasm, optic nerves, parotid glands, and

*Table VII does not report comparisons to the ground truth that was obtained by manually corrected or merged auto-segmentation results.[32,80,96] In the case the results were reported separately for multiple versions of a method Table VII reports only the results for the best performing method version.

†*The Head and Neck Auto-segmentation Challenge* was part of the workshop *3D Segmentation in the Clinic: A Grand Challenge* during the conference on *Medical Image Computing and Computer Assisted Interventions - MICCAI 2009*.

‡The *Head and Neck Auto-segmentation Challenge: Segmentation of the Parotid Glands* was part of the workshop *Medical Image Analysis in the Clinic: A Grand Challenge* during *MICCAI 2010*.

§The *Head and Neck Auto-Segmentation Challenge 2015* was held as a standalone satellite event during *MICCAI 2015*.

TABLE VI. Performance metrics applied for measuring the performance of auto-segmentation of organs at risk in the head and neck region for the purpose of radiotherapy planning, and the corresponding references and mathematical definitions.

Metrics label – name and definition		
Overlap metrics, reported in percents (%)		
Standard volumetric coefficient		
DC	Dice coefficient (F_1 score)[22–63,92–95,97–99]	$\frac{2\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
Variations of the volumetric coefficient (VC)		
TPR	Sensitivity[24,31,40,41,50,55,56,59,67,68,71,90,93,94,96]	$\frac{\lvert A\cap B\rvert}{\lvert A\rvert}$
TNR	Specificity[41,93,94,96]	$\frac{\lvert(A\cup B)^C\rvert}{\lvert A^C\rvert}$
PPV	Positive predictive value (inclusion)[24,31,40,55,56,68]	$\frac{\lvert A\cap B\rvert}{\lvert B\rvert}$
FDR	False discovery rate (segmented volume)[50,59]	$\frac{\lvert B\backslash A\rvert}{\lvert B\rvert}$
Variations of the surface coefficient (SC)		
sDC	Surface overlap[60]	$\frac{\lvert\partial A\cap\partial^\tau B\rvert+\lvert\partial B\cap\partial^\tau A\rvert)}{\lvert\partial A\rvert+\lvert\partial B\rvert}$
sPPV	Surface positive predictive value (inclusion)[78,89]	$\frac{\lvert\partial B\cap\partial^\tau A\rvert}{\lvert\partial B\rvert}$
Distance metrics, reported in millimeters (mm)		
Variations of the Hausdorff distance (HD)		
HD^{reg}	Hausdorff distance, regular[25,36,41,43,44,48,52,53,58,66,70,73,76,79,84,88,99]	$\max\limits_{\substack{\boldsymbol{a}\in\partial A\\ \boldsymbol{b}\in\partial B}}\left\{d(\boldsymbol{a},\partial B),d(\boldsymbol{b},\partial A)\right\}$
DTA^{max}	Maximum distance to agreement[27,77]	$\max_{\boldsymbol{b}\in\partial B} d(\boldsymbol{b},\partial A)$
HD_{95}	95-percentile Hausdorff distance[22,23,29–31,35,37–40,46,49,55,58,66,69,71,97]	$K^{95}_{\substack{\boldsymbol{a}\in\partial A\\ \boldsymbol{b}\in\partial B}}\left\{d(\boldsymbol{a},\partial B),d(\boldsymbol{b},\partial A)\right\}$
HD_{95}^{mid}	95-percentile Hausdorff distance, mid-value[24,54,62]	$\frac{1}{2}\left(K^{95}_{\boldsymbol{a}\in\partial A}d(\boldsymbol{a},\partial B)+K^{95}_{\boldsymbol{b}\in\partial B}d(\boldsymbol{b},\partial A)\right)$
HD^{sw}	Slice-wise Hausdorff distance[81,82,85,86,91,92]	<HD^{reg} aggregated over two dimensions>
Variations of the average surface distance (ASD)		
ASSD	Average symmetric surface distance[26,53,57]	$\frac{\sum_{\boldsymbol{a}\in\partial A} d(\boldsymbol{a},\partial B)+\sum_{\boldsymbol{b}\in\partial B} d(\boldsymbol{b},\partial A)}{\lvert\partial A\rvert+\lvert\partial B\rvert}$
ASD^{max}	Average surface distance, maximum[35,64,66,72,75,76,98,99]	$\max\left\{\frac{\sum_{\boldsymbol{a}\in\partial A} d(\boldsymbol{a},\partial B)}{\lvert\partial A\rvert},\frac{\sum_{\boldsymbol{b}\in\partial B} d(\boldsymbol{b},\partial A)}{\lvert\partial B\rvert}\right\}$
ASD^{mid}	Average surface distance, mid-value[24,32,40,55,56,61,81]	$\frac{1}{2}\left(\frac{\sum_{\boldsymbol{a}\in\partial A} d(\boldsymbol{a},\partial B)}{\lvert\partial A\rvert}+\frac{\sum_{\boldsymbol{b}\in\partial B} d(\boldsymbol{b},\partial A)}{\lvert\partial B\rvert}\right)$
$ASD^{n/a}$	Average surface distance, unspecified[39,58,75,79,88]	<unspecified>
DTA^{avg}	Average distance to agreement[27,42,68,77,84,87]	$\frac{\sum_{\boldsymbol{b}\in\partial B} d(\boldsymbol{b},\partial A)}{\lvert\partial B\rvert}$
Variations of the signed surface distance (SSD)		
SSD^{avg}	Signed surface distance, average[45]	$\frac{\sum_{\boldsymbol{a}\in\partial A} d^s}{(}\boldsymbol{a},\partial B)x02010;\sum_{\boldsymbol{b}\in\partial B} d^s(\boldsymbol{b},\partial A)\lvert\partial A\rvert+\lvert\partial B\rvert$
$SDTA^{avg}$	Signed distance to agreement, average[89]	$\frac{\sum_{\boldsymbol{b}\in\partial B} d^s}{(}\boldsymbol{b},\partial A)\lvert\partial B\rvert$
$SDTA^{min}$	Signed distance to agreement, minimum[89]	$\min_{\boldsymbol{b}\in\partial B} d^s(\boldsymbol{b},\partial A)$
$SDTA^{max}$	Signed distance to agreement, maximum[89]	$\max_{\boldsymbol{b}\in\partial B} d^s(\boldsymbol{b},\partial A)$

Legend: $|A|$ and $|B|$ are the number of voxels in volumetric masks A (e.g., ground truth) and B (e.g., auto-segmentation), respectively, and $|\partial A|$ and $|\partial B|$ are the number of voxels in the corresponding subsets of surface voxels ∂A and ∂B, respectively. The Euclidean distances of voxels $\boldsymbol{a}$ and $\boldsymbol{b}$ to surfaces ∂B and ∂A, respectively, are defined as $d(\boldsymbol{a},\partial B)=\min_{\boldsymbol{b}\in\partial B}\|\boldsymbol{a}x02010;\boldsymbol{b}\|$ and $d(\boldsymbol{b},\partial A)=\min_{\boldsymbol{a}\in\partial A}\|\boldsymbol{b}x02010;\boldsymbol{a}\|$, respectively. The signed Euclidean distance $d^s(\boldsymbol{a},\partial B)$ is defined as $d(\boldsymbol{a},\partial B)$ if $\boldsymbol{a}\in B^C$ and as $-d(\boldsymbol{a},\partial B)$ if $\boldsymbol{a}\in B$, and the signed Euclidean distance $d^s(\boldsymbol{b},\partial A)$ is defined as $d(\boldsymbol{b},\partial A)$ if $\boldsymbol{b}\in A^C$ and as $-d(\boldsymbol{b},\partial A)$ if $\boldsymbol{b}\in A$. The volumetric neighborhoods within distance τ from surfaces ∂A and ∂B are defined as $\partial^\tau A=\{\boldsymbol{x}\in\mathbb{R}^3;\exists\boldsymbol{a}\in\partial A,\|\boldsymbol{x}x02010;\boldsymbol{a}\|\le\tau\}$ and $\partial^\tau B=\{\boldsymbol{x}\in\mathbb{R}^3;\exists\boldsymbol{b}\in\partial B,\|\boldsymbol{x}x02010;\boldsymbol{b}\|\le\tau\}$, respectively.

submandibular glands from 40 CT images (25 for training, 15 for testing).[66] In July 2019[|], 10 teams attempted to segment the parotid glands, submandibular glands and lymph nodes from 55 MR images (31 for training, 24 for testing)[131], however, detailed results of this challenge have yet not been published and are not publicly available. The last auto-segmentation challenge was carried out in October 2019[¶], where 12 teams attempted to segment 13 OARs (i.e., the eyes, lens, optic nerves, optic chiasm, pituitary gland, brainstem, temporal lobes, spinal cord, parotid glands, inner ear, middle ear, temporo-mandibular joints, and mandible) as well as the

[|]The *AAPM RT-MAC challenge* was part of the *2019 American Association of Physicists in Medicine (AAPM) Annual Meeting* (https://www.aapm.org/GrandChallenge/RT-MAC/; http://aapmchallenges.cloudapp.net/competitions/34).

[¶]The *StructSeg2019: Automatic Structure Segmentation for Radiotherapy Planning Challenge* was held as a standalone satellite event during *MICCAI 2019* (https://structseg2019.grand-challenge.org; http://www.structseg-challenge.org).

TABLE VII. Performance of auto-segmentation for the purpose of radiotherapy planning, and the corresponding references (cf. Table VI for the list of metrics).

Results	
Parotid glands	
DC (%)	92 ± 4,[37] 91 ± 2,[75] 88 ± 2,[46] 91^{m,f} (▲),[45] 88,[53] 87 ± 3(▲),[60] 87 ± 4 (▲),[24] 87,[64] 86 ± 2 (▲),[40] 86 ± 3,[48] 86 ± 4,[24] 86 ± 5 (▲),[31] 86 ± 5,[42] 86 ± 5,[40] 86 ± 7,[93] 91^{m,f},[72] 85 ± 2,[83] 85 ± 3,[26] 85 ± 4,[91] 85 ± 4,[30] 85 ± 5,[47] 91^{m,f} (DL),[29] 85,[60] 84 ± 3,[34] 84 ± 3 (▲),[55] 84 ± 4 (•),[60] 84 ± 7 (▲,IM),[66] 84,[76] 91^{m,f},[22] 91^{m,f},[23] 83 ± 2,[50] 83 ± 3,[58] 83 ± 5 (•),[36] 83 ± 5,[86] 83 ± 6,[36] 83 ± 6 (▲),[56] 91^{m,f},[95] 91^{m,f},[45] 81 ± 4 (▲),[70] 81 ± 5,[28] 81 ± 8 (▲),[49] 81 ± 8,[27] 81 (▲),[54] 91^{m,f},[52] 91^{m,f} (ABAS),[29] 79 (MR),[68] 91^{m,f},[77] 79,[87] 79,[57] 77 ± 6,[65] 91^{m,f} (▲),[69] 91^{m,f} (▲),[35] 76 ± 6,[63] 76 (CT),[68] 91^{m,f},[35] 75,[51] 72 ± 10,[90] 72 ± 12,[82] 91^{m,f},[84] 91^{m,f},[78] 91^{m,f}[73]
VC (%)	TPR: 97 ± 4,[40] 91 ± 9,[93] 88 ± 5 (▲),[40] 86 ± 7,[24] 85 ± 5 (▲),[24] 85 ± 7 (▲),[31] 85 ± 7,[50] 84 (MR),[68] 83 ± 10 (▲),[56] 82 ± 5 (▲),[55] 72 ± 9,[90] 71 (CT)[68]; TNR: 91 ± 7[93]; PPV: 88 ± 5 (▲),[24] 87 ± 3 (▲),[40] 87 ± 6 (▲),[31] 86 ± 2 (▲),[55] 84 ± 7 (▲),[56] 83 ± 7,[40] 83 (CT),[68] 80 ± 6,[24] 77 (MR)[68]; FDR: 18 ± 6[50]
SC (%)	sDC: 95 ± 3 (τ = 2.85mm,▲),[60] 90 ± 6 (τ = 2.85mm,•)[60]
HD (mm)	HD91^{m,f}: 1.4 ± 0.6,[36] 1.7 ± 0.7 (•),[36] 91^{m,f},[73] 5.1 ± 1.1,[48] 91^{m,f},[76] 10.7,[53] 91^{m,f},[84] 12.1 ± 3.9,[58] 91^{m,f},[52] 91^{m,f} (▲,IM),[66] 14.2 ± 6.6 (▲)[70]; DTA91^{m,f}: 6.8 ± 2.5,[27] 91^{m,f} (▲),[35] 91^{m,f},[35] 91^{m,f}[77]; HD91^{m,f}: 91^{m,f} (▲),[69] 2.6 ± 1.4,[40] 2.7 ± 1.1 (▲),[31] 3.2 ± 0.6,[37] 3.8 ± 1.1 (▲),[40] 4.0 ± 2.2(▲),[55] 91^{m,f},[22] 4.6 ± 1.2,[58] 5.0 ± 2.4 (▲,IM),[66] 91^{m,f},[23] 5.2 ± 1.8 (▲),[49] 91^{m,f} (DL),[29] 6.6 ± 3.3,[30] 91^{m,f},[35] 91^{m,f} (▲),[35] 91^{m,f} (ABAS),[29] 9.3 ± 3.3[46]; HD91^{m,f}: 3.3 ± 1.0 (▲),[24] 3.9 ± 2.0,[24] 3.9 (▲)[54]; HD91^{m,f}: 5.0 ± 1.0,[91] 5.8 ± 1.6,[86] 91^{m,f}[82]
ASD (mm)	ASSD: 0.9 ± 0.3,[26] 1.2,[53] 1.6[57]; ASD91^{m,f}: 91^{m,f},[76] 91^{m,f},[64] 91^{m,f},[72] 91^{m,f} (▲,IM),[66] 3.6 ± 1.4[75]; ASD91^{m,f}: 1.0 ± 0.3 (▲),[55] 1.0 ± 0.4,[40] 1.2 ± 0.3 (▲),[24] 1.3 ± 0.4,[24] 1.4 ± 0.4 (▲),[40] 1.8 ± 0.6 (▲)[56]; ASD91^{m,f}: 0.3 ± 0.1,[75] 1.4 ± 0.4[58]; DTA91^{m,f}: 91^{m,f},[77] 1.6 ± 0.6,[27] 1.7 ± 1.1,[42] 91^{m,f},[84] 2.5 ± 2.8,[87] 4.8 (MR),[68] 6.2 (CT)[68]
SSD (mm)	SSD91^{m,f}: 91^{m,f},[45] 91^{m,f} (▲)[45]
Submandibular glands	
DC (%)	91^{m,f},[22] 85 ± 10,[42] 85,[60] 84 ± 6,[24] 83,[53] 82 ± 5,[86] 82 ± 5 (▲),[40] 82 ± 7,[30] 82 ± 7,[50] 81 ± 4,[46] 81 ± 6 (▲),[55] 80 ± 7 (▲),[24] 80 ± 7,[26] 80 ± 8 (•),[60] 91^{m,f},[77] 91^{m,f},[23] 78 ± 7 (▲),[60] 78 ± 8 (▲,IM),[66] 77 ± 6,[34] 75 ± 13 (▲),[31] 73,[51] 71 ± 12,[65] 91^{m,f},[95] 70 ± 12,[82] 70,[87] 65 ± 8 (▲),[70] 91^{m,f} (▲),[69] 91^{m,f} (▲),[35] 91^{m,f},[35] 91^{m,f}[78]
VC (%)	TPR: 87 ± 5,[24] 85 ± 6 (▲),[55] 80 ± 11,[50] 79 ± 8 (▲),[24] 79 ± 9 (▲),[40] 72 ± 16 (▲)[31]; PPV: 85 ± 9 (▲),[40] 83 ± 11,[24] 82 ± 9 (▲),[24] 82 ± 11 (▲),[31] 80 ± 8 (▲)[55]; FDR: 14 ± 8[50]
SC (%)	sDC: 84 ± 10 (τ = 2mm,•),[60] 82 ± 10 (τ = 2mm,▲)[60]
HD (mm)	HD91^{m,f}: 6.6,[53] 91^{m,f} (▲,IM),[66] 9.7 ± 4.8 (▲)[70]; DTA91^{m,f}: 91^{m,f},[77] 91^{m,f} (▲),[35] 91^{m,f}[35]; HD91^{m,f}: 91^{m,f},[22] 3.2 ± 1.6 (▲),[31] 4.0 ± 2.7 (▲),[40] 91^{m,f},[23] 4.8 ± 1.8 (▲,IM),[66] 4.8 ± 1.7 (▲),[55] 91^{m,f} (▲),[69] 91^{m,f} (▲),[35] 6.0 ± 1.8,[46] 6.2 ± 4.3,[30] 91^{m,f}[35]; HD91^{m,f}: 3.2 ± 2.3,[24] 3.9 ± 1.2 (▲)[24]; HD91^{m,f}: 3.8 ± 1.0,[86] 91^{m,f}[82]
ASD (mm)	ASSD: 1.2,[53] 1.3 ± 1.2[26]; ASD91^{m,f}: 91^{m,f} (▲,IM)[66]; ASD91^{m,f}: 0.9 ± 0.5 (▲),[55] 1.2 ± 0.7,[24] 1.4 ± 1.0 (▲),[40] 2.0 ± 1.9 (▲)[24]; DTA91^{m,f}: 91^{m,f},[77] 1.2 ± 1.3,[42] 1.9 ± 1.4[87]
Brainstem	
DC (%)	93 ± 1,[97] 93 ± 3,[27] 92 ± 3,[40] 92,[53] 91 ± 1,[86] 91 ± 3,[43] 90 ± 1,[26] 90 ± 2,[48] 90 ± 2,[24] 90 ± 3,[47] 90 ± 4 (▲),[56] 89 ± 3,[42] 88 ± 2 (▲),[24] 88 ± 2 (▲),[31] 88 ± 3,[92] 88,[60] 88 ± 3 (•),[36] 87 ± 3(▲),[55] 87 ± 3(▲),[40] 87 ± 4 (▲,IM),[66] 91^{m,f},[22] 91^{m,f} (DL),[29] 86 ± 4,[30] 86 ± 8,[36] 86,[76] 85 (80,88),[94] 91^{m,f},[38] 91^{m,f},[95] 84 (▲),[54] 91^{m,f},[23] 83 ± 6,[89] 91^{m,f},[73] 91^{m,f},[52] 82 ± 4 (▲),[49] 91^{m,f} (ABAS),[29] 80 ± 8 (▲),[60] 79 ± 6,[59] 79 ± 10 (•),[60] 91^{m,f} (▲),[35] 78,[99] 91^{m,f} (▲),[69] 77 ± 7,[93] 77 ± 8,[90] 91^{m,f},[35] 76(68,81),[98] 91^{m,f},[84] 75 ± 12,[82] 73 (MR),[68] 69 (CT),[68] 67 ± 2,[46] 64 ± 16[50]
VC (%)	TPR: 95 ± 3,[40] 91 ± 4,[24] 90 ± 4 (▲),[56] 90 ± 4,[40] 89 ± 3 (▲),[24] 88 ± 3 (▲),[55] 88 ± 6 (▲),[40] 86 ± 14,[50] 87 ± 5 (▲),[31] 79 ± 9,[59] 75 ± 14,[90] 69 (CT),[68] 64 (MR),[68] 63 ± 10[93]; TNR: 98 ± 2[93]; PPV: 91 ± 4 (▲),[56] 89 ± 4,[24] 89 ± 6 (▲),[31] 89 (MR),[68] 88 ± 4 (▲),[40] 87 ± 5 (▲),[24] 85 ± 2 (▲),[55] 74 (CT)[68]; FDR: 15 ± 8,[59] 42 ± 23[50]
SC (%)	sDC: 83 ± 13 (τ = 2.5mm,▲),[60] 83 ± 14 (τ = 2.5mm,•)[60]; sPPV: 91^{m,f} (τ = 2mm)[89]
HD (mm)	HD91^{m,f}: 0.6 ± 0.1 (•),[36] 0.9 ± 2.0,[36] 2.7 ± 0.9,[43] 2.9 ± 0.3,[48] 91^{m,f},[73] 91^{m,f},[52] 6.5,[53] 91^{m,f} (▲,IM),[66] 91^{m,f},[84] 8.7,[99] 91^{m,f}[76]; DTA91^{m,f}: 3.5 ± 1.2,[27] 91^{m,f} (▲),[35] 91^{m,f}[35]; HD91^{m,f}: 1.3 ± 0.5,[40] 2.0 ± 0.3 (▲),[31] 91^{m,f},[97] 3.6 ± 0.8 (▲),[40] 91^{m,f},[22] 91^{m,f},[23] 4.0 ± 0.9 (▲),[55] 4.0 ± 2.0 (▲,IM),[66] 91^{m,f} (ABAS),[29] 4.8 ± 1.6,[30] 91^{m,f},[38] 91^{m,f} (▲),[69] 91^{m,f} (DL),[29] 91^{m,f},[35] 91^{m,f} (▲),[35] 6.4 ± 2.4,[46] 12.4 ± 26.3 (▲)[49]; HD91^{m,f}: 2.6 ± 0.8,[24] 2.9 (▲),[54] 3.0 ± 0.6 (▲)[24]; HD91^{m,f}: 2.8 ± 0.5,[92] 2.8 ± 0.5,[86] 91^{m,f}[82]
ASD (mm)	ASSD: 0.6 ± 0.1,[26] 0.8[53]; ASD91^{m,f}: 91^{m,f},[76] 91^{m,f} (▲,IM),[66] 2.1,[99] 2.2(1.7,3.1)[98]; ASD91^{m,f}: 0.7 ± 0.3,[40] 0.9 ± 0.3 (▲),[56] 1.0 ± 0.2,[24] 1.2 ± 0.6 (▲),[55] 1.2 ± 0.2 (▲),[24] 1.4 ± 0.3 (▲)[40]; DTA91^{m,f}: 0.9 ± 0.4,[27] 1.0 ± 0.5,[42] 91^{m,f},[84] 3.2 (MR),[68] 4.3 (CT)[68]
SSD (mm)	SDTA91^{m,f}: 0.2[89]; SDTA91^{m,f}: −4.3[89]; SDTA91^{m,f}: 5.4[89]
Brain, cerebrum (CBR) and cerebellum (CBE)	
DC (%)	99 ± 0.2 (•),[60] 99,[60] 98 ± 0.3,[36] 98 (CBR),[99] 97 ± 0.5 (•),[36] 91^{m,f} (CBR),[23] 96 ± 1 (CBR),[97] 96 ± 2,[82] 94 ± 1 (CBE),[97] 94(93,95) (CBR),[98] 91^{m,f} (CBE),[23] 92 (CBE),[99] 87(80,91) (CBE),[98] 84(79,86) (CBE)[94]
SC (%)	sDC: 95 ± 2 (τ = 1mm,•)[60]
HD (mm)	HD91^{m,f}: 1.2 ± 1.5,[36] 3.6 ± 0.2 (•),[36] 10.8 (CBE),[99] 18.4 (CBR)[99]; HD91^{m,f}: 91^{m,f} (CBR),[97] 91^{m,f} (CBE),[97] 91^{m,f} (CBR),[23] 91^{m,f} (CBE)[23]; HD91^{m,f}: 91^{m,f}[82]
ASD (mm)	ASD91^{m,f}: 0.8 (CBR),[99] 1.2 (CBE),[99] 1.9(1.3,3.4) (CBE),[98] 2.9(2.5,3.2) (CBR)[98]
Temporal lobes	
DC (%)	93 ± 4,[27] 84 ± 3[30]
HD (mm)	DTA91^{m,f}: 4.7 ± 2.2[27]; HD91^{m,f}: 12.5 ± 4.1[30]
ASD (mm)	DTA91^{m,f}: 1.1 ± 0.6[27]

TABLE VII. Continued.

Results	
Hippocampus	
DC (%)	91[m,f][38]
HD (mm)	HD91[m,f]: 91[m,f][38]
Pituitary gland	
DC (%)	90,[33] 64 ± 9,[30] 30(0,72)[94]
HD (mm)	HD91[m,f]: 3.2 ± 0.8[30]
Spinal cord and spinal canal	
DC (%)	96,[53] 95 (canal),[60] 92 ± 2 (canal,•),[60] 91 ± 1,[48] 88 ± 2,[27] 88 ± 7,[47] 88,[60] 91[m,f],[23] 87 ± 3,[65] 87 ± 3,[42] 86 ± 6,[30] 86 ± 9,[97] 85 ± 2,[28] 85,[99] 91[m,f],[35] 91[m,f],[52] 83 ± 6,[36] 91[m,f],[22] 82 ± 5,[34] 80 ± 5,[58] 80 ± 5,[63] 80 ± 8 (•),[60] 80 (CT),[68] 79 ± 8 (•),[36] 78 (+brainstem),[87] 76 ± 8,[90] 76 (66,82),[98] 91[m,f],[95] 75,[51] 74 ± 8,[82] 74 ± 8,[26] 91[m,f],[73] 37 (MR)[68]
VC (%)	TPR: 80 (CT),[68] 76 ± 12,[90] 26 (MR)[68]; PPV: 93 (MR),[68] 81 (CT)[68]
SC (%)	sDC: 99 ± 1 (τ = 2.93mm,•)[60] , 93 ± 3 (canal,τ = 1.17mm,•)[60]
HD (mm)	HD91[m,f]: 0.5 ± 0.1 (•),[36] 0.7 ± 1.3,[36] 1.7 ± 0.2,[48] 91[m,f],[73] 91[m,f],[52] 4.3,[53] 6.6,[99] 10.4 ± 3.8[58]; DTA91[m,f]: 3.3 ± 0.3,[27] 91[m,f][35]; HD91[m,f]: 91[m,f],[22] 91[m,f],[35] 4.3 ± 1.4,[58] 91[m,f],[97] 91[m,f],[23] 6.9 ± 22.0[30]; HD91[m,f]: 91[m,f][82]
ASD (mm)	ASSD: 0.4,[53] 2.6 ± 1.6[26]; ASD91[m,f]: 0.8,[99] 1.5(0.8,2.4)[98]; ASD91[m,f]: 1.2 ± 0.4[58]; DTA91[m,f]: 0.9 ± 0.1,[27] 1.6 ± 0.9,[42] 2.3 ± 1.4 (+brainstem),[87] 3.5 (CT),[68] 17.5 (MR)[68]
Cerebrospinal fluid	
DC (%)	82 ± 7[97]
HD (mm)	HD91[m,f]: 91[m,f][97]
Eyeballs and vitreous humor (VH)	
DC (%)	96 ± 1 (VH),[97] 95,[60] 95 ± 2,[43] 94,[33] 91[m,f] (DL),[29] 93 ± 1,[48] 93 ± 4,[47] 92 ± 2 (•),[60] 92 ± 2,[30] 91 ± 2 (•),[36] 91[m,f] (ABAS),[29] 91 (MR),[68] 89 ± 4,[36] 91[m,f],[22] 88 ± 3,[65] 87 (CT),[68] 91[m,f],[38] 85 ± 8,[82] 84 ± 5,[59] 84 ± 7,[89] 84(19) (+eye muscles),[79] 81 ± 5,[62] 81 (VH),[99] 81(78,85),[94] 80 (72,84) (VH),[98] 91[m,f][73]
VC (%)	TPR: 93 (MR),[68] 91 (CT),[68] 83 ± 8[59]; PPV: 89 (MR),[68] 84 (CT)[68]; FDR: 10 ± 8[59]
SC (%)	sDC: 95 ± 3 (τ = 1.65mm,•)[60]; sPPV: 91[m,f] (τ = 2mm)[89]
HD (mm)	HD91[m,f]: 0.3 ± 0.1 (•),[36] 0.4 ± 1.0,[36] 1.3 ± 0.3,[43] 1.7 ± 0.3,[48] 91[m,f] (DL),[29] 91[m,f] (ABAS),[29] 5.0 (VH),[99] 5.3(4.7) (+eye muscles),[79] 91[m,f][73]; HD91[m,f]: 91[m,f] (VH),[97] 91[m,f],[22] 91[m,f][38]; HD91[m,f]: 2.4 ± 0.5,[62] 2.4 ± 1.0[30]; HD91[m,f]: 91[m,f][82]
ASD (mm)	ASD91[m,f]: 1.0 (VH),[99] 1.2(0.9,1.8) (VH)[98]; ASD91[m,f]: 0.6(0.8) (+eye muscles)[79]; DTA91[m,f]: 2.0 (MR),[68] 3.3 (CT)[68]
SSD (mm)	SDTA91[m,f]: 0.8[89]; SDTA91[m,f]: −2.3[89]; SDTA91[m,f]: 3.8[89]
Optic chiasm	
DC (%)	91[m,f],[88] 71 ± 9,[43] 64 ± 16,[30] 62 ± 17,[27] 61 ± 6 (▲),[24] 59 ± 7,[40] 59 ± 10 (▲),[40] 59 ± 14,[24] 58 ± 10 (▲),[55] 58 ± 17 (▲),[54] 57 ± 13 (▲, UB),[66] 91[m,f],[73] 53 ± 15,[46] 52 ± 11 (▲),[70] 91[m,f],[22] 45 ± 17 (▲),[31] 42 ± 17 (▲),[49] 91[m,f],[38] 41(0,58),[94] 41 ± 14,[36] 37 ± 13,[65] 37 ± 18,[89] 91[m,f] (▲),[35] 24 ± 15[59]
VC (%)	TPR: 68 ± 8 (▲),[40] 64 ± 11 (▲),[24] 64 ± 15,[24] 61 ± 5,[40] 61 ± 10 (▲),[55] 50 ± 25 (▲),[31] 48 ± 31[59]; PPV: 65 ± 8,[40] 61 ± 12 (▲),[24] 56 ± 10 (▲),[55] 56 ± 11 (▲),[40] 56 ± 16,[24] 47 ± 18 (▲)[31]; FDR: 77 ± 24[59]
SC (%)	sPPV: 91[m,f] (τ = 2mm)[89]
HD (mm)	HD91[m,f]: 91[m,f],[88] 1.0 ± 0.4,[36] 2.5 ± 1.0,[43] 91[m,f] (▲,UB),[66] 5.6 ± 1.6 (▲),[70] 91[m,f][73]; DTA91[m,f]: 3.7 ± 1.4,[27] 91[m,f] (▲)[35]; HD91[m,f]: 2.1 ± 1.4,[40] 2.2 ± 1.0 (▲),[55] 2.6 ± 0.8 (▲,UB),[66] 2.8 ± 1.4 (▲),[31] 3.8 ± 1.2 (▲),[40] 4.4 ± 3 (▲),[49] 4.6 ± 2.4,[30] 91[m,f],[22] 91[m,f] (▲),[35] 5.8 ± 2.5,[46] 91[m,f][38]; HD91[m,f]: 2.7 ± 0.5 (▲),[24] 2.8 ± 1.6 (▲),[54] 3.9 ± 2.2[24]
ASD (mm)	ASD91[m,f]: 91[m,f] (▲,UB)[66]; ASD91[m,f]: 0.7 ± 0.2 (▲),[55] 0.8 ± 0.4,[40] 0.9 ± 0.2 (▲),[24] 1.3 ± 0.3 (▲),[40] 1.5 ± 0.7[24]; ASD91[m,f]: 91[m,f][88]; DTA91[m,f]: 1.1 ± 0.7[27]
SSD (mm)	SDTA91[m,f]: 0.04[89]; SDTA91[m,f]: −2.4[89]; SDTA91[m,f]: 3.0[89]
Optic nerves	
DC (%)	90 ± 4,[37] 82 ± 6,[43] 81,[33] 79 ± 6,[62] 91[m,f],[88] 78 ± 5 (•),[60] 77 ± 6,[60] 76 ± 7,[30] 76(73,82),[74] 75 ± 5 (•),[36] 74 ± 6,[24] 74 ± 8 (▲),[31] 74(41),[79] 72 ± 4,[40] 72 ± 5 (▲),[24] 72 ± 6,[34] 72 ± 6 (▲),[60] 72 ± 6,[46] 71 ± 8 (▲),[54] 70 ± 4 (▲),[40] 69 ± 5 (▲),[55] 69 ± 9,[36] 69 ± 10,[47] 91[m,f] (ABAS),[29] 64 ± 7,[65] 64 ± 8 (▲),[49] 63 ± 10 (▲,UB),[66] 62,[99] 60 ± 12,[27] 91[m,f],[22] 58(49,63),[98] 91[m,f],[38] 52 ± 14,[89] 91[m,f] (DL),[29] 48 ± 11,[59] 91[m,f] (▲),[69] 38(0,53),[94] 91[m,f][35]
VC (%)	TPR: 85 ± 8 (▲),[40] 80 ± 8 (▲),[24] 77 ± 11 (▲),[31] 74 ± 6 (▲),[55] 71 ± 10,[24] 70 ± 6,[40] 64 ± 16[59]; PPV: 80 ± 9,[24] 76 ± 7,[40] 72 ± 9 (▲),[31] 70 ± 8 (▲),[40] 66 ± 8 (▲),[24] 64 ± 6 (▲)[55]; FDR: 57 ± 12[59]
SC (%)	sDC: 98 ± 3 (τ = 2.5mm,•),[60] 92 ± 6 (τ = 2.5mm,▲)[60]; sPPV: 91[m,f] (τ = 2mm)[89]
HD (mm)	HD91[m,f]: 0.5 ± 0.3 (•),[36] 0.7 ± 0.8,[36] 91[m,f],[88] 1.8 ± 0.7,[43] 3.8(6.9),[79] 91[m,f] (▲,UB),[66] 6.5[99]; DTA91[m,f]: 3.7 ± 1.0,[27] 91[m,f] (▲)[35]; HD91[m,f]: 1.4 ± 0.4,[40] 2.0 ± 0.5 (▲),[40] 2.1 ± 0.3,[37] 2.3 ± 2.4 (▲),[31] 2.5 ± 1.0 (▲),[55] 2.6 ± 0.4 (▲),[49] 3.0 ± 1.0 (▲,UB),[66] 91[m,f] (ABAS),[29] 91[m,f] (▲),[69] 3.7 ± 1.1,[30] 4.8 ± 4.3,[46] 91[m,f] (DL),[29] 91[m,f],[38] 91[m,f],[22] 91[m,f][35]; HD91[m,f]: 1.9 ± 1.9 (▲),[24] 1.9 ± 1.3,[24] 2.2 ± 0.9 (▲),[54] 3.3 ± 1.6[62]
ASD (mm)	ASD91[m,f]: 91[m,f] (▲,UB),[66] 1(0.8,1.4),[98] 1.0[99]; ASD91[m,f]: 0.4 ± 0.3,[40] 0.6 ± 0.3,[24] 0.7 ± 0.2 (▲),[24] 0.7 ± 0.2 (▲),[40] 1.1 ± 0.8 (▲)[55]; ASD91[m,f]: 91[m,f],[88] 0.6(2.0)[79]; DTA91[m,f]: 1.2 ± 0.5[27]
SSD (mm)	SDTA91[m,f]: −0.4[89]; SDTA91[m,f]: −2.7[89]; SDTA91[m,f]: 2.4[89]

TABLE VII. Continued.

Results	
Lens	
DC (%)	88 ± 5,[97] 84 ± 7,[47] 84[33], 82 ± 6[30], 81 ± 12[60], 91[m,f] (DL)[29], 80 ± 18 (•)[60], 79 ± 11 (•)[36], 72 ± 14[36], 67[99], 50(37,66)[98], 91[m,f] (ABAS)[29], 35 ± 25[59]
VC (%)	TPR: 50 ± 32[59]; FDR: 73 ± 21[59]
SC (%)	sDC: 93 ± 20 (τ = 0.98mm,•)[60]
HD (mm)	HD91[m,f]: 0.2 ± 0.1 (•)[36], 0.4 ± 0.9[36], 3.7[99]; HD91[m,f]: 91[m,f][97], 2.0 ± 1.1[30], 91[m,f] (DL)[29], 91[m,f] (ABAS)[29]
ASD (mm)	ASD91[m,f]: 1.0[99], 1.6(0.7,2.9)[98]
Sclera	
DC (%)	69 ± 5[97], 46[99], 38(24,55)[98]
HD (mm)	HD91[m,f]: 5.9[99]; HD91[m,f]: 91[m,f][97]
ASD (mm)	ASD91[m,f]: 1.1[99], 1.8(1.0,3.8)[98]
Cornea	
DC (%)	43[99]
HD (mm)	HD91[m,f]: 6.4[99]
ASD (mm)	ASD91[m,f]: 1.7[99]
Lacrimal glands	
DC (%)	70 ± 12[60], 62 ± 13 (•)[60]
SC (%)	sDC: 92 ± 7 (τ = 2.5mm,•)[60]
Extraocular muscle	
DC (%)	76 ± 6[62]
HD (mm)	HD91[m,f]: 2.1 ± 0.5[62]
Mandible	
DC (%)	96[60], 96[53], 91[m,f][23], 94 ± 1 (▲)[56], 94 ± 1 (▲)[55], 94 ± 1 (▲)[40], 94 ± 1 (▲)[24], 94 ± 2 (•)[60], 94 ± 2 (▲)[60], 94 (▲)[41], 94[41], 93 ± 1[30], 93 ± 1[92], 93 ± 1[86], 93 ± 1 (▲,IM)[66], 93 ± 1 (▲)[39], 93 ± 1[24], 93 ± 2[46], 93 ± 2 (▲)[31], 92 ± 1[44], 92 ± 2[48], 92 ± 2[26], 92 (▲)[54], 91 ± 2 (▲)[49], 91 ± 4[47], 91 ± 9[42], 90 ± 2 (•)[36], 90 ± 4[65], 91[m,f][95], 89 ± 4[82], 88 ± 3[28], 89[51], 88[39], 91[m,f] (▲)[69], 87 ± 3[36], 85 ± 2[34], 91[m,f][35], 91[m,f][78], 91[m,f][52], 91[m,f] (▲)[35], 82 ± 4[93], 82 ± 4[40], 80 ± 4[58], 78 ± 8[90]
VC (%)	TPR: 95 ± 2 (▲)[56], 95 (▲)[41], 93 ± 2 (▲)[24], 93[41], 92 ± 2 (▲)[55], 92 ± 3[24], 92 ± 3 (▲)[31], 91 ± 3 (▲)[40], 87 ± 5[40], 83 ± 13[93], 79 ± 11[90]; TNR: 100 (▲)[41], 100[41], 95 ± 3[93]; PPV: 97 ± 2 (▲)[40], 95 ± 2 (▲)[24], 95 ± 2 (▲)[31], 95 ± 5 (▲)[55], 94 ± 2 (▲)[56], 94 ± 3[24], 79 ± 4[40]
SC (%)	sDC: 97 ± 2 (τ = 1mm,•)[60], 97 ± 2 (τ = 1mm,▲)[60]
HD (mm)	HD91[m,f]: 1.3 ± 1.0[36], 1.3 ± 0.4 (•)[36], 2.4 ± 0.4[48], 91[m,f] (▲,IM)[66], 4.6 (▲)[41], 6.4[41], 6.5[53], 6.7 ± 1.3[44], 91[m,f][52], 10.9 ± 2.1[58]; DTA91[m,f]: 91[m,f][35], 91[m,f](▲)[35]; HD91[m,f]: 91[m,f][23], 1.3 ± 0.5 (▲)[31], 1.4 ± 0.6 (▲)[39], 1.5 ± 0.3 (▲)[55], 1.7 ± 0.6 (▲,IM)[66], 1.9 ± 0.6 (▲)[40], 2.4 ± 0.6 (▲)[49], 2.5 ± 0.8[30], 2.7 ± 1.7[40], 91[m,f][35], 91[m,f] (▲)[35], 91[m,f] (▲)[69], 4.3 ± 1.1[58], 6.3 ± 2.2[46]; HD91[m,f]: 1.3 ± 0.1[24], 1.4 ± 0.02 (▲)[24], 1.9 (▲)[54]; HD91[m,f]: 2.1 ± 0.1[92], 2.6 ± 0.6[86], 91[m,f][82]
ASD (mm)	ASSD: 0.2 ± 0.1[26], 0.6[53]; ASD91[m,f]: 91[m,f] (▲,IM)[66]; ASD91[m,f]: 0.4 ± 0.1 (▲)[55], 0.4 ± 0.1 (▲)[56], 0.5 ± 0.1 (▲)[24], 0.5 ± 0.1[24], 0.5 ± 0.1 (▲)[40], 1.1 ± 0.7[40]; ASD91[m,f]: 0.6[39], 1.1 ± 0.3[58]; DTA91[m,f]: 0.7 ± 0.3[42]
Oral cavity	
DC (%)	93 ± 3[47], 91 ± 2[30], 89 ± 2[26], 89 ± 2[28], 91[m,f][35], 87 ± 5[42], 91[m,f][52], 78 ± 7[50]
VC (%)	TPR: 68 ± 11[50]; FDR: 5 ± 3[50]
HD (mm)	HD91[m,f]: 91[m,f][52]; DTA91[m,f]: 91[m,f][35]; HD91[m,f]: 91[m,f][35], 7.4 ± 2.1[30]
ASD (mm)	ASSD: 1.0 ± 0.3[26]; DTA91[m,f]: 0.8 ± 0.4[42]
Temporo-mandibular joints	
DC (%)	87 ± 3[30], 87 ± 6[42], 85 ± 5[47]
HD (mm)	HD91[m,f]: 2.8 ± 0.9[30]
ASD (mm)	DTA91[m,f]: 0.4 ± 0.3[42]
Mastoids	
DC (%)	82 ± 6[47]
Chewing muscles	
DC (%)	91[m,f] (pterygoid)[95], 91[m,f] (masseter)[95], 71[87]
ASD (mm)	DTA91[m,f]: 1.6 ± 1.4[87]
Pharyngeal constrictor muscles (PCM), cricopharynx (CP), orohypopharynx constrictor muscle (OPCM)	
DC (%)	81 ± 4 (PCM)[28], 73 ± 11 (CP)[50], 71 ± 8 (PCM)[40], 69 ± 6 (PCM)[65], 68 ± 9 (PCM)[50], 91[m,f] (PCM)[23], 91[m,f][52], 61 (middle) & 58 (inferior) & 46 (superior)[53], 58 (OPCM)[51], 54 ± 26 (inferior) & 58 ± 18 (middle) & 52 ± 11 (superior) (PCM)[26], 91[m,f] (PCM)[77], 50 (PCM)[87]
VC (%)	TPR: 78 ± 7 (PCM)[40], 70 ± 11 (CP)[50], 66 ± 9 (PCM)[50]; PPV: 69 ± 8 (PCM)[40]; FDR: 20 ± 16 (CP)[50], 29 ± 9 (PCM)[50]
HD (mm)	HD91[m,f]: 9.6 (inferior) & 12.7 (middle) & 14.7 (superior)[53], 91[m,f][52]; DTA91[m,f]: 91[m,f] (PCM)[77]; HD91[m,f]: 2.8 ± 1.3 (PCM)[40], 91[m,f] (PCM)[23]

TABLE VII. Continued.

Results	
ASD (mm)	ASSD: 1.6 ± 1.7 (inferior) & 1.9 ± 1.7 (middle) & 3.7 ± 5.2 (superior) (PCM)[26], 2.0 (middle) & 2.0 (inferior) & 2.1 (superior)[53]; ASD91^{m,f}: 1.0 ± 0.5 (PCM)[40]; DTA91^{m,f}: 91^{m,f} (PCM)[77], 2.0 ± 1.9 (PCM)[87]
Cervical esophagus with the cricopharyngeal inlet, upper esophageal sphincter (UES)	
DC (%)	86 ± 3[42], 82 ± 6[28], 81 ± 14 (UES)[50], 81 ± 7[36], 70 ± 7[61], 69 ± 10[26], 62[51], 60 ± 11[50], 91^{m,f}[52], 91^{m,f}[23], 35[53]
VC (%)	TPR: 80 ± 16 (UES)[50], 50 ± 15[50]; FDR: 15 ± 14 (UES)[50], 21 ± 14[50]
HD (mm)	HD91^{m,f}: 1.1 ± 1.1[36], 91^{m,f}[52], 35.8[53]; HD91^{m,f}: 91^{m,f}[23]
ASD (mm)	ASSD: 1.3 ± 0.6[26], 7.7[53]; ASD91^{m,f}: 1.9 ± 0.7[61]; DTA91^{m,f}: 1.0 ± 0.7[42]
Thyroid	
DC (%)	92 ± 3.7[37], 86 ± 5[30], 91^{m,f}[23], 80[85], 79 ± 6[44], 68[99], 57(37,80)[98]
HD (mm)	HD91^{m,f}: 10.2 ± 2.9[44], 17.5[99]; HD91^{m,f}: 2.7 ± 0.6[37], 91^{m,f}[23], 3.9 ± 2.4[30]; HD91^{m,f}: 91^{m,f}[85]
ASD (mm)	ASD91^{m,f}: 2.5[99], 5.1(1.1,9.3)[98]
Larynx	
DC (%)	89 ± 3[30], 87 ± 4[47], 86 ± 4[65], 86 ± 7[42], 83 ± 8[28], 80 ± 5[40], 78 ± 4[50], 91^{m,f}[52], 77 ± 7[26], 74[51], 91^{m,f}[35], 71[53], 91^{m,f}[77]
VC (%)	TPR: 88 ± 6[40], 83 ± 8[50]; PPV: 77 ± 6[40]; FDR: 25 ± 10[50]
HD (mm)	HD91^{m,f}: 11.1[53], 91^{m,f}[52]; DTA91^{m,f}: 91^{m,f}[35], 91^{m,f}[77]; HD91^{m,f}: 3.2 ± 2.7[40], 6.2 ± 5.8[30], 91^{m,f}[35]
ASD (mm)	ASSD: 1.0 ± 0.4[26], 2.2[53]; ASD91^{m,f}: 1.7 ± 1.6[40]; DTA91^{m,f}: 1.3 ± 1.0[42], 91^{m,f}[77]
Trachea	
DC (%)	84 ± 8[63], 81 ± 5[30], 91^{m,f}[52]
HD (mm)	HD91^{m,f}: 91^{m,f}[52]; HD91^{m,f}: 20.9 ± 9.0[30]
Cochlea	
DC (%)	95 ± 10[60], 82 ± 7 (•)[60], 74[53], 66 ± 13[36], 65 ± 7[26], 41 ± 8 (•)[36], 91^{m,f}[77]
SC (%)	sDC: 99 ± 2 (τ = 1.25mm,•)[60]
HD (mm)	HD91^{m,f}: 0.5 ± 0.4[36], 0.7 ± 0.1 (•)[36], 1.7[53]; DTA91^{m,f}: 91^{m,f}[77]
ASD (mm)	ASSD: 0.4[53], 0.6 ± 0.2[26]; DTA91^{m,f}: 91^{m,f}[77]
Brachial plexus	
DC (%)	77[81], 56 ± 11[30], 53 ± 12[67], 32[71]
VC (%)	TPR: 49[71], 47 ± 12[67]
HD (mm)	HD91^{m,f}: 15.4[71]; HD91^{m,f}: 91^{m,f}[81]
ASD (mm)	ASD91^{m,f}: 1.6[81]
Carotid artery	
DC (%)	91[25], 91^{m,f}[23]
HD (mm)	HD91^{m,f}: 0.9[25]; HD91^{m,f}: 91^{m,f}[23], 18.3 ± 14.5[30]

Legend: m – median, average not reported; f – value estimated from a figure, exact value not reported; o1/o2 — compared against observer 1/observer 2; ▲ — evaluated on the PDDCA database;[66] • — evaluated on the TCIA-RT database;[60] CT, MR — the results in[68] are obtained from CT or MR images; IM, UB — winning teams of the 2015 computational challenge[66]; +brainstem — the spinal cord and brainstem were segmented as one organ; +eye muscles — the eyes and eye muscles were segmented as one organ; +chiasm — optic nerves and optic chiasm were segmented as one organ; τ — size of the volumetric neighborhood.

tumor gross target volumes of the nasopharyngeal cancer from 60 CT images (50 for training, 10 for testing). While detailed results for this challenge are yet to be published, the publicly available data indicate that best ranking method achieved an average Dice coefficient of 81% and 95-percentile Hausdorff distance of 2.8 mm across all OARs. Moreover, a new edition of this challenge is scheduled for October 2020**.

4. DISCUSSION

The field of RT planning in the H&N region expands beyond auto-segmentation of OARs that was presented in this review, for example to (auto-)segmentation of target volumes (including gross target volume, clinical target volume, and planning target volume), analysis of commercial solutions for RT planning, dosimetric evaluations, and longitudinal studies. For additional information, we kindly refer the reader to specific reviews that include the topics of segmentation methodology,[8,21] target volume segmentation,[20] ABAS,[19,132] commercial segmentation tools,[5,66,119] MR-only RT[133] and observer variability in OAR delineation[3].

In this review, we focused on auto-segmentation of OARs in the H&N region, and provided a comprehensive and systematic overview with a complete list of relevant references from 2008 to date along with a systematic analysis from different perspectives that we consider relevant: *image modality*, *OAR*, *image database*, *methodology*, *ground truth*, *performance metrics*, and *segmentation performance*. In this section we discuss the advantages and limitations of the

**The *Automatic Structure Segmentation for Radiotherapy Planning Challenge 2020* is planned as a standalone satellite event during *MICCAI 2020* (https://miccai2020.org/en/MICCAI-2020-CHALLENGES.html).

reviewed methods, and provide corresponding recommendations from the relevant perspectives.

4.A. Image modality

For the purpose of RT planning, CT images are always acquired because they contain information about the electron density that is required to calculate the interaction of radiation beams with tissues, and further used to define radiation dose distribution maps. Although MR images proved to be advantageous for RT planning because they can provide anatomical information complementary to CT images, especially in the case of soft tissues, they are not commonly used in clinical practice. Moreover, the structures in MR images may be subjected to geometrical distortions,[134] for example, due to the magnetic field inhomogeneities.[101] However, as MR imaging has become more accessible in the past decade, it can be expected that its utilization will increase toward making MR images an integral part of RT planning, and that auto-segmentation approaches exploring both CT and MR image modalities simultaneously will be further developed. The start of this trend is already indicated by the recent increase in the number of studies that include the MR image modality.[38,40,43,57-59] In a single study where OARs were independently auto-segmented from CT and MR images of the same patients, the results for MR images outperformed those for CT images in the case of the parotid glands, eyeballs, and brainstem.[68]

Although methods for MR-only RT planning are being developed,[135] their routine clinical implementation is still very limited, as challenges remain of how to assign data on electron density to MR images for the purpose of dose calculation[133] by means of synthetic CT image generation[136] or MR-to-CT image registration.[105,137] In general, better performance is achieved by applying deformable (i.e., nonrigid) image registration and using rigid registration as the first step,[103,104] however, this may not always be the case.[105] To further improve the registration process, DL approaches have recently started to emerge.[137]

Complementary information can be obtained from PET-CT and PET-MR scanners, which combine the CT or MR with the PET modality and acquire coregistered images. However, as PET images enable functional investigation through the radiolabeling of tissues with a high metabolic activity (i.e., cancerous cells), they are more appropriate for target volume than for OAR segmentation.[118,138] On the other hand, monoenergetic images generated from DECT were shown to be adequate for H&N OAR segmentation[108] because they can exhibit superior image quality in comparison to classical 120 keV CT, especially in terms of a better contrast-to-noise ratio, reduced influence of the beam hardening phenomenon and metal artifact suppression. For several OARs, it was shown that ABAS and DL-based auto-segmentation can be successfully applied to monoenergetic images of 40 and 70 keV.[29] However, a study on a larger DECT database with a complete set of OARs and comparison to classical CT images needs to be performed in order to objectively assess and identify eventual advantages.

To conclude, both CT and MR image modalities are being explored for H&N OAR auto-segmentation, but the potential of the MR image modality for auto-segmentation of several soft tissues should be explored more in the future.

4.B. Organ at risk

The relatively small area of the H&N region comprises a large number of OARs with a relatively complex and variable anatomy. The decision of which OAR needs to be delineated is based on a number of factors, including the proximity of the OAR to the tumor, its susceptibility to the radiation and importance for life functions. Auto-segmentation was therefore commonly performed for OARs whose RT-induced damage proved to be linked to post-RT complications that may endanger the life of the patient or notably jeopardize its quality.[109-111]

Due to the potentially devastating morbidity resulting from over-irradiation of the spinal cord and brainstem, delineation of these two anatomical structures is a mandatory part of any segmentation process in the H&N region.[102] The parotid and submandibular glands are by far the most represented of the remaining OARs, although their poor boundary distinction in CT images makes segmentation very challenging. On the other hand, the optic chiasm and optic nerves are also demanding to segment because of their small size and tubular geometry. The mandible is the only well visible bony structure, and due to its excellent visibility in CT images it can act as a spatial reference for segmenting other neighboring OARs.[51,66] As the definition of exact OAR boundaries is subjected to observer interpretation, new studies should adhere to existing delineation guidelines.[102] Nevertheless, with the introduction of additional image modalities, such as the MR, the boundaries of OARs should become easier to interpret.

To conclude, the spinal cord, brainstem and major salivary glands (the parotid and submandibular glands) are the most studied OARs in the H&N region, however, more experiments should be conducted in the future for auto-segmentation of the pharyngeal constrictor muscles, larynx and cervical esophagus with the cricopharyngeal inlet that are important for RT planning.

4.C. Image database

To account for the anatomical and disease-related variability among different patients as well as for the variability in the image acquisition settings, auto-segmentation methods must be validated on a preferably large number of images and patients to ensure reliable statistical results. In general, the current trend shows an increasing number of cases being included in evaluation databases, which is mostly due to the application of state-of-the-art machine learning methods, such as DL, which require relatively large training datasets. Image databases should include representative clinical

samples, with images from various acquisition setups and of patients with different tumors according to their localization and stage. However, images should retain certain common characteristics (e.g., imaging sequence, field of view, image noise), otherwise auto-segmentation may become too challenging. Still, objective comparison of different auto-segmentation methods is often difficult, because they were evaluated on different image databases, or on a different set of annotations representing reference OAR delineations. As the construction of a representative set of samples requires a lot of effort, many such databases remain proprietary and represent a valuable research advantage.

Besides using proprietary databases, evaluation should be performed also on publicly available image databases to ensure an objective comparison to existing approaches. Among the publicly available CT image databases, PDDCA[66] has been already used in several studies[45,54–56,60,69,70] because it was devised for a computational challenge that set benchmarks for auto-segmentation of OARs in the H&N region, while TCIA-RT[60] and StructSeg have yet to gain visibility. As it was shown that MR images provide valuable support to CT image auto-segmentation, or can be treated as standalone in the case of MR-only RT planning, public MR image databases have recently surfaced, such as the RT-MAC[116] or MRI-RT,[105] which is augmented with CT images of the same patients.

To conclude, several image databases with the corresponding ground truth are currently publicly available and should be used for an independent performance evaluation of OAR auto-segmentation approaches. In the future, there is a need for such databases to evolve, that is, to include a large number of cases and reference delineations, preferably performed by multiple observers from different institutions and at multiple times, so as to enable a proper evaluation of multimodal auto-segmentation methods.

4.D. Methodology

For OAR auto-segmentation in the H&N region, ABAS is still the prevailing methodological approach, and has been as such implemented in several commercial tools for RT planning.[5,66,119] However, its segmentation performance highly depends on the range of anatomical variations that can be observed in the library of atlases, which can be built up from previously treated patients or, if used, built into the commercial software. As a result, ABAS may perform poorly for cases that differ from the library of atlases,[5] therefore making the selection of the most appropriate atlases a challenging task. For most OARs, perfect ABAS results cannot be reasonably expected, however, the performance of a level corresponding to clinical quality can be consistently expected given a large atlas database under the assumption of perfect atlas selection.[139] It was shown that ABAS reaches its upper performance limit with the inclusion of 10–20 atlases,[23,67,140] and that it generally underperforms for small and/or thin OARs (e.g., swallowing muscles).[87] Another drawback is its long execution time due to atlas registration, which limits on-line clinical applications.

Recently, the focus has shifted toward machine learning, with DL approaches for H&N OAR auto-segmentation starting to emerge as early as in 2016,[70] and have been considerably increasing in number since (Fig. 1). When compared to ABAS, DL-based auto-segmentation requires considerably less time for on-line applications, but is associated with a high computational burden in the off-line training phase, where currently up to a few days or more may be required to complete the model training. Moreover, the training set of images has to be quite large, but the actual number depends on image quality and representativeness, and can be reduced by applying different training set augmentation techniques (e.g., intensity and geometrical transformations of original images). The underlying DL model is, in comparison to ABAS, also more robust because it can be trained with all available data, including patients with metal artifacts and diverse anatomy.[7] The main advantage of DL-based auto-segmentation is in its ability to systematically learn the most adequate features for segmentation from a set of annotated training images, and then automatically search for the same features in a previously unseen image. Although this proved to result in the best overall segmentation performance,[49] it is not without drawbacks. For example, the most popular DL-based medical image auto-segmentation architecture, the U-Net,[8] can result in many false positives if the approximate location and size of the observed OAR is not constrained beforehand. As a result, state-of-the-art techniques from the field of artificial intelligence (e.g., attention learning,[24] adversarial learning[40]) are constantly being explored and utilized to improve its performance.[141]

Both ABAS and DL-based auto-segmentation are based on reference OAR delineations in the given image database, which may, however, not represent the ground truth. If the cases included in the image database are not representative for the actual OAR segmentation task, or if the corresponding manual delineations are of low quality and inconsistent, the underlying DL model will either fail to train or produce inconsistent segmentations. Therefore, attention needs to be given to the choice of image database and to reduce the intra- and interobserver variability of reference delineations, for example, by including publicly available databases[112,113] and adhering to OAR delineation guidelines.[102]

To conclude, while ABAS was the dominating approach for segmenting OARs in the H&N region in the past, current approaches have shifted to DL, resulting in a superior segmentation performance. Moreover, DL-based auto-segmentation is expected to become even more sophisticated through the inclusion of methodological advances in the field of artificial intelligence,[142] and even more powerful from the perspective of being trained on larger and more diverse image databases.

4.E. Ground truth

To generate the ground truth, manual delineation of OARs by human experts is still the most common approach,

although it has been recognized as a very tedious and time-consuming task. For the delineation of ground truth contours, it is strongly recommended to follow the recently introduced guidelines,[102] which have been formed as a consensus of different professional associations and groups,[††] and also incorporate guidelines that have been introduced in the past.[124,125,127] However, even if guidelines are followed, the delineation is still biased by subjective observer interpretation, and therefore it is strongly recommended to perform basic observer training with joint delineation review sessions,[143,144] and to include additional modalities to improve the visibility of structure boundaries.[144]

Moreover, to increase the reliability of statistical results related to the methodology testing in the clinical context, the ground truth should be provided from multiple experts performing the delineation on multiple time occasions, therefore enabling the evaluation of the variability among and within the observers, that is, the inter- and intraobserver variability, respectively. In a study where manual H&N OAR delineations of eight different observers from CT and MR images of 20 subjects were compared to ABAS, it was reported that manual delineations and ABAS generated structures of similar volume with no statistically significant difference in volume overlap, however, the observers exhibited higher variation with respect to tubular structures (e.g., optic chiasm, optic nerves).[89] On the other hand, a different study evaluated 32 multi-institution delineations of six OARs from a single CT image, and reported a significant delineation variability among observers that consequently caused large differences in the planned radiation doses, with the most variable organs being the brainstem and the two parotid glands.[143] Similarly, in a multi-institutional study where eight observers manually delineated 20 OARs from 16 CT images, statistically significant interobserver delineation variability as well as differences in dosimetric parameters were reported for all OARs, however, both could be reduced for most OARs by manually editing the results of ABAS, in particular for the brainstem, spinal cord, cochleae, temporo-mandibular joints, larynx, and pharyngeal constrictor muscles.[145] On the other hand, a high agreement was reported for auto-segmentations of 13 OARs from 125 CT images that were independently obtained at seven different institutions with the same commercial RT planning system but with different institution-specific settings.[82]

Nevertheless, the variability in manual as well as auto-segmentation results cannot be completely eliminated because each individual observer is exposed to his/her subjective bias that is conditioned by experience (i.e., novice vs expert), and because imaging protocols and setups as well as RT protocols and planning systems vary greatly across institutions.[146] For a particular OAR, the observer variability imposes the upper limit for auto-segmentation performance, as we cannot expect any auto-segmentation result to overcome the obtained consensus among the ground truth delineations. Although manual correction of auto-segmentation boundaries is a less labor intensive approach for ground truth generation, it contains auto-segmentation bias and is therefore not the most appropriate reference for performing auto-segmentation evaluation. On the other hand, the ground truth can be relatively easily obtained by using phantom objects, synthetic images, or cadaver sections,[67,89,121,147] however, they represent unrealistic surrogates for patient imaging and were in fact not present in the reviewed studies.

To conclude, delineation guidelines should be followed for the ground truth generation, and participation of multiple experts from multiple institutions is recommended for a reliable reporting of the intra/interobserver variability.

4.F. Performance metrics

When reporting the geometric accuracy of auto-segmentation results, there is unfortunately no universal consensus about the corresponding performance metrics. Moreover, various mutually incompatible definitions and different nomenclatures make the comparison of auto-segmentation results relatively difficult.[129] As there is a strong need for an agreed-upon metrics, which would allow an exact comparison of results and eliminate the need for specifying its definition in each new study, we would recommend the nomenclature and definitions presented in Table VI.

For reporting the volumetric overlap of two segmentation masks, we advise a mandatory use of the Dice coefficient. Although the Jaccard index is an established volumetric coefficient and has been reported in a few studies,[59,67,96] it is redundant because it can be calculated from the Dice coefficient[‡‡]. Other variations of the volumetric coefficient provide additional insight into the segmentation performance from the perspective of binary classification, specifically the degree of over- or under-segmentation, but their interpretation may be ambiguous. For example, in the case of reporting the specificity, a dilemma about the calculation of true negatives (the set complement in its definition in Table VI) may arise.[94] On the other hand, sensitivity is the metrics of choice in the case we want to reduce the number of voxels that are missing from the resulting segmentation (i.e., false negatives), even if at the expense of adding voxels (i.e., false positives). Although volumetric metrics may result in a high overlap, clinically relevant differences between segmentation boundaries may still exist, which are important in RT planning because they are used to compute the radiation dose distribution. The mismatches in boundary segments that encompass

[††]Radiotherapy Oncology Group for Head and Neck (GORTEC), France; The Danish Head and Neck Cancer Group (DAHANCA), Denmark; Head and Neck Cancer Group of the European Organization for Research and Treatment of Cancer (EORTC), European Union; Hong Kong Nasopharyngeal Cancer Study Group (HKNPCSG), Hong Kong; National Cancer Research Institute (NCRI), UK; National Cancer Institute of Canada Clinical Trials Group (NCIC CTG), Canada; NRG Oncology Group (NRG), USA; Trans Tasman Radiation Oncology Group (TROG), Australia.

[‡‡]Jaccard index: JI = $|A \cap B|/|A \cup B|$; Dice coefficient: DC = $2|A \cap B|/(|A|+|B|)$; DC = 200%· JI/(100%+JI); JI = DC/(200%−DC).

a volumetrically small but eventually important regions of interest can be, to a certain degree, captured by surface coefficients,[60] which measure the overlap of the corresponding mask surfaces. While surface coefficients may gain a wider adoption among the overlap metrics in the future, especially if different values of the neighborhood distance τ are explored simultaneously, a consensus needs to be made about their usage, with the surface Dice coefficient being the most appropriate due to its bidirectional (i.e., symmetric) properties.

Any overlap metrics should be accompanied with at least one distance metrics, which provides complementary information about the segmentation boundaries by measuring the spatial separation between the corresponding surfaces. The Hausdorff distance measures the maximum point-to-point distance between two segmentation masks, and it originates from a proper mathematical metrics to measure the distance between two subsets in a metric space. However, because it is very sensitive to outliers, the 95-percentile version of this metrics may be alternatively used to robustly suppress their influence. On the other hand, two-dimensional computation of metrics, such as in the case of the slice-wise Hausdorff distance, is not appropriate for volumetric segmentation. In the case of the average surface distance, we recommend to report the average symmetric surface distance because it equally takes into account all possible point-to-surface distances and is bidirectional (i.e., symmetric). On the other hand, both the maximum and mid-value versions of the average surface distance unnecessarily use two different point-to-surface weighting factors, while the average distance to agreement is unidirectional. The variations of the signed surface distance can be used to deduce consistent over- or under-segmentation, however, they are unable to detect the overall boundary mismatch when either over- or under-segmentation regions are present in an approximately equal quantity, because they cancel out. In general, distance metrics perform better when the observed structures are small, and are especially efficient for structures with a high surface-to-volume ratio (e.g., tubular structures such as the spinal cord, optic nerve and optic chiasm, and the pharyngeal constrictor muscles) and cases where otherwise acceptable small boundary variations result in a large relative volume discrepancy (e.g., the pharyngeal constrictor muscles). Other reported metrics, such as the volume difference[35,93,94] or distance/variation of mass centers,[29,52,94] do not represent meaningful overlap or distance measurements, and are therefore not proper to evaluate segmentation results.

It has to be noted that, for a specific OAR, the reported performance metrics only evaluate how close is the obtained segmentation mask to its corresponding ground truth. Although they represent a powerful tool for general method comparison, they overlook the potential consequences of segmentation errors from the clinical perspective. However, a method named LinSEM[148] has been recently developed from the premise that an ideal segmentation metrics should reflect the degree of clinical acceptability directly from its values, and show the same acceptability meaning with the same value for structures of different shape, size, and form. The method combines, in a linear manner, the commonly used segmentation performance metrics (i.e., the Dice coefficient, Jaccard index, and Hausdorff distance) with the clinical acceptability, which was provided by an expert observer (i.e., a subjective score from 1 to 5). By performing experiments on CT images including OARs from the H&N region (i.e., the right parotid gland, mandible, and cervical esophagus), it was concluded that the Jaccard index has the most linear relationship with the acceptability before actual linearization, while the Dice coefficient and Hausdorff distance exhibit a significant improvement in acceptability meaning from the perspective of an ideal metrics-to-acceptability relationship.[148]

To conclude, the Dice coefficient is the standard volumetric coefficient for reporting the overlap of two segmentation masks, and it should be always accompanied with at least one distance metrics, preferably the Hausdorff distance (or its 95-percentile version) and the average symmetric surface distance. Future research should focus on combining existing geometrical performance metrics with clinical acceptability scores and risk assessments into a new class of metrics for the purpose of augmenting the quantitative evaluation of segmentation performance.

4.G. Segmentation performance

Although the auto-segmentation methods do not always provide clinically acceptable results, their performance is constantly improving due to the application of new technologies. The auto-segmentation of OARs and subsequent manual corrections require considerably less time than direct manual delineation[19,119] and reduce the intra/interobserver variability.[145] However, a direct comparison of the segmentation performance among different methods is difficult, mostly because they were, in general, not evaluated on the same image databases. The comparison is therefore often affected by different image acquisition setups (e.g., imaging sequence, field of view), image properties (e.g., size, resolution, noise), manual delineation guidelines and patient cohorts. Moreover, the studies report different performance metrics, focus on different OARs or even do not provide a detailed statistical description of the corresponding ground truth.

The results reported by state-of-the-art techniques indicate that auto-segmentation of OARs in the H&N region is feasible to be clinically implemented into an automated RT planning system. However, from the perspective of RT, both target volume and OAR segmentation has direct clinical implications. Apart from the geometrical agreement with the corresponding ground truth, auto-segmentation results have to be evaluated also from the perspective of their dosimetric impact, because even if the geometric differences are small, the impact on the final dose distribution may still be clinically relevant. As a result, the geometrical performance metrics are not sufficient to predict the dosimetric impact of auto-segmentation inaccuracies. For example, it was shown that the interobserver variability in manual delineations of OARs from the H&N region (e.g., the brainstem, brain, parotid

glands, mandible, and spinal cord) can lead to substantially different dosimetric plans.[143,145,149] However, for several OARs (e.g., the brainstem, spinal cord, cochlea, temporomandibular joint, larynx and pharyngeal constrictor muscles), the consistency in dosimetric plans can be improved by reducing the interobserver variability, for example, by manually editing the results of ABAS,[90,145,150] which was shown to produce clinically acceptable RT plans from the perspective of dosimetric impact.[58] Similar conclusions were drawn in a study that applied DL-based auto-segmentation,[50] and reported little effect on the OAR dose despite the variation in the Dice coefficient, indicating that imperfect geometrical performance metrics do not necessarily result in inferior OAR dosimetry.[50] Although the average radiation dose was, for specific OARs (i.e., the pharyngeal constrictor muscles), significantly higher for the DL-based than for manually defined RT plans, these differences were not considered to be clinically relevant.[50] On the other hand, a study evaluated RT plans, obtained from expert manual delineations of several H&N OARs, against those obtained by a knowledge-based planning system, which is based on a preconfigured model inferred from a cohort of past RT plans that were judged as optimal.[151,152] A weak correlation between the geometric performance metrics (i.e., the Dice coefficient, Hausdorff distances, volume differences, and centroid distances) and dosimetric indices (i.e., dose to the hottest 98% of the planning target volume and mean OAR dose) was reported, indicating that the geometric performance metrics are not appropriate for estimating the dosimetric impact.[152] However, besides observer variability in manual delineation, other factors may affect the RT plan, such as the changes in the location and size of the observed OARs due to RT effects, or the random and systematic patient setup errors due to multiple RT sessions. In a study where reference manual delineations were randomly perturbed to simulate delineation variability and combined with simulated patient setup variability at random magnitudes, it was concluded that the dosimetric impact of the delineation variability is overstated when considered in isolation from the setup variability, and that it depends largely on the OAR distance from the target volume.[153] Nevertheless, it has to be noted that the dosimetric impact of OAR auto-segmentation is always compared to the dosimetric impact of manual OAR delineation, which is inherently subjected to observer variability. Future studies on H&N OAR auto-segmentation should therefore report, besides multiple geometric performance metrics, also metrics related to the dosimetric impact to encompass clinically relevant endpoints for RT planning.

Nevertheless, the analysis of the reported results indicates that the performance of OAR auto-segmentation in the H&N region is, if we consider as clinically acceptable the results with the Dice coefficient above 90% and average surface distance below 1.5 mm, currently adequate for several OARs, including the parotid glands, brainstem, brain, cerebrum and cerebellum, temporal lobes, spinal cord, eyeballs and vitreous humor, mandible, oral cavity, and cochlea (Table VII).[48,60,97] According to the reported interobserver variability, there may still be room for improvements in auto-segmentation of the salivary glands, especially if performed on MR images.[68] On the other hand, the eyeballs can be segmented relatively accurately due to their spherical geometry, while the optic nerves and optic chiasm can come close to the ground truth in terms of the distance but not overlap metrics.[66,88] For the pharyngeal constrictor muscles, larynx and cervical esophagus with the cricopharyngeal inlet, unfortunately not enough studies have been conducted to draw relevant conclusions. Therefore, it is expected that these OARs will receive more focus in the future, especially because of their importance in the process of the H&N RT planning. On the other hand, it has to be again pointed out that all auto-segmentation results are compared to corresponding reference segmentations, and their definition is subjected to observer variability, meaning that the reasonably achievable performance is not ideal segmentation, for example, it is not realistic to expect that the Dice coefficient will reach 100% or that the Hausdorff and average surface distance will drop to zero.

To conclude, the best performing methods achieve clinically acceptable auto-segmentation for several H&N OARs, even if manual corrections may still be needed, but certainly they reduce the overall delineation time and observer variability. To better evaluate the segmentation performance, future studies should focus also on the dosimetric impact to provide clinically relevant endpoints for RT planning.

5. CONCLUSIONS

We performed a systematic review of OAR auto-segmentation for H&N RT planning from 2008 to date. Besides outlining, analyzing and categorizing the relevant publications within this field, we have provided also a critical discussion of the corresponding advantages and limitations. The main conclusions that may not only assist in the introduction to the field but also be a valuable resource for studying existing or developing new methods and evaluation strategies are as follows: (a) *Image modality* — Both CT and MR image modalities are being exploited for the task, but the potential of the MR image modality for auto-segmentation of several soft tissues should be explored more in the future. (b) *OAR* — The spinal cord, brainstem, and major salivary glands (the parotid and submandibular glands) are the most studied OARs, however, more experiments should be conducted for auto-segmentation of the pharyngeal constrictor muscles, larynx, and cervical esophagus with the cricopharyngeal inlet that are important for RT planning. (c) *Image database* — Several image databases with the corresponding ground truth are currently publicly available and should be used for an independent performance evaluation of OAR auto-segmentation approaches, however, they should be augmented with data from multiple observers and multiple institutions. (d) *Methodology* — While ABAS was dominating in the past, current approaches have shifted to DL, which resulted in superior performance, and are expected to become even more methodologically sophisticated and trained on larger image databases. (e) *Ground truth* — Delineation guidelines should be followed for the ground truth generation, and participation

of multiple experts from multiple institutions is recommended for a reliable reporting of the intra/inter-observer variability. (f) *Performance metrics* — The Dice coefficient as the standard volumetric overlap metrics should be always accompanied with at least one distance metrics, preferably the Hausdorff distance (or its 95-percentile version) and the average symmetric surface distance, and future research should focus on combining them with clinical acceptability scores and risk assessments. (g) *Segmentation performance* — The best performing methods achieve clinically acceptable auto-segmentation for several OARs, even if manual corrections may still be needed, but certainly they reduce the overall delineation time and observer variability, however, future studies should focus also on the dosimetric impact to provide clinically relevant endpoints for RT planning.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency (ARRS) under grants J2-1732, P2-0232 and P3-0307.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

[a]Author to whom correspondence should be addressed. Electronic mail: tomaz.vrtovec@fe.uni-lj.si.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel R, Torre L, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394–424.
2. Borras J, Barton M, Grau C, et al. The impact of cancer incidence and stage on optimal utilization of radiotherapy: methodology of a population based analysis by the ESTRO-HERO project. *Radiother Oncol*. 2015;116:45–50.
3. Vinod S, Jameson M, Min M, Holloway L. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol*. 2016;121:169–179.
4. Chaney E, Pizer S. Autosegmentation of images in radiation oncology. *J Am Coll Radiol*. 2009;6:455–458.
5. Sharp G, Fritscher K, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41:050902.
6. Sahiner B, Pezeshk A, Hadjiiski L, et al. Deep learning in medical imaging and radiation therapy. *Med Phys*. 2019;46:e1–e36.
7. Seo H, Khuzani M, Vasudevan V, et al. Machine learning techniques for biomedical image segmentation: an overview of technical aspects and introduction to state-of-art applications. *Med Phys*. 2020;47:e148–e167.
8. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional neural networks for biomedical image segmentation. In:*Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Volume 9351 of *LNCS*. Springer; 2015:234–241.
9. Çiçek O, Abdulkadir A, Lienkamp S, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, volume 9901 of *LNCS*. Springer; 2016:424–432.
10. Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *Fourth International Conference on 3D Vision - 3DV 2016*. IEEE; 2016:565–571.
11. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:2481–2495.
12. Kamnitsas K, Ledig C, Newcombe V, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017 36:61–78.
13. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille A. DeepLab: semantic image segmentation with deep convolutional nets. Atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell*. 2018;40:834–848.
14. Chen H, Dou Q, Yu L, Qin J, Heng P-A. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage*. 2018;170:446–455.
15. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:386–397.
16. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. *Comput Biol Med*. 2018;98:126–146.
17. Thompson R, Valdes G, Fuller C, et al. Artificial intelligence in radiation oncology imaging. *Int J Radiat Oncol Biol Phys*. 2018; 102:1159–1161.
18. Boldrini L, Bibault J-E, Masciocchi C, Shen Y, Bittner MI. Deep learning: a review for the radiation oncologist. *Front Oncol*. 2019; 9:977.
19. Lim J, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol*. 2016;55: 799–806.
20. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol*. 2019;135:130–140.
21. Cardenas C, Yang J, Anderson B, Court L, Brock K. Advances in auto-segmentation. *Semin Radiat Oncol*. 2019;29:185–197.
22. Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother Oncol*. 2020;144:152–158.
23. van Dijk L, Van den Bosch L, Aljabar P et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115–123.
24. Gou S, Tong N, Qi S, Yang S, Chin R, Sheng K. Self-channel-and-spatial-attention neural network for automated multi-organ segmentation on head and neck CT images. *Phys Med Biol*. 2020.
25. de Ruijter J, van Sambeek M, van de Vosse F, Lopata R. Automated 3D geometry segmentation of the healthy and diseased carotid artery in free-hand, probe tracked ultrasound images. *Med Phys*. 2020;47:1034–1047.
26. Vandewinckele L, Willems S, Robben D, etal. Segmentation of head-and-neck organs-at-risk in longitudinal CT scans combining deformable registrations and convolutional neural networks. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2020.
27. Fung N, Hung W, Sze C, Lee M, Ng W. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: time, geometrical, and dosimetric analysis. *Med Dosim*. 2020;45:60–65.
28. Lei Y, Harms J, Dong X, et al. Organ-at-risk (OAR) segmentation in head and neck CT using U-RCNN. In: *SPIE Medical Imaging 2020: Computer-Aided Diagnosis*. Volume 11314. SPIE; 2020:1131444.
29. van der Heyden B, Wohlfahrt P, Eekers D, et al. Dual-energy CT for automatic organs-at-risk segmentation in brain-tumor patients using a multi-atlas and deep-learning approach. *Sci Rep*. 2019;9:4126.
30. Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Sci Rep*. 2019;1:480–491.
31. Wang Y, Zhao L, Song Z, Wang M. Organ at risk segmentation in head and neck CT images by using a two-stage segmentation framework based on 3D U-Net. *IEEE Access*. 2019;7:144591–144602.
32. van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol*. 2019;138:68–74.
33. Sun Y, Shi H, Zhang S, Wang P, Zhao W, Zhou X, Yuan K. Accurate and rapid CT image segmentation of the eyes and surrounding organs for precise radiotherapy. *Med Phys*. 2019;46:2214–2222.
34. Huang C, Badiei M, Seo H, et al. Atlas based segmentations via semi-supervised diffeomorphic registrations. arXiv 1911.10417; 2019.

35. Haq R, Berry S, Deasy J, Hunt M, Veeraraghavan H. Dynamic multi-atlas selection based consensus segmentation of head and neck structures from CT images. *Med Phys*. 2019;46:5612–5622.
36. Rhee D, Cardenas C, Elhalawani H, et al. Automatic detection of contouring errors using convolutional neural networks. *Med Phys*. 2019;46:5086–5097.
37. Zhong T, Huang X, Tang F, Liang S, Deng X, Zhang Y. Boosting-based cascaded convolutional neural networks for the segmentation of CT organs-at-risk in nasopharyngeal carcinoma. *Med Phys*. 2019;46:5602–5611.
38. Agn M, Rosenschöld P, Puonti O, et al. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Med Image Anal*. 2019;54:220–237.
39. Qiu B, Guo J, Kraeima J. Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network. *Phys Med Biol*. 2019;64:1750.
40. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med Phys*. 2019;46:2669–2682.
41. Torosdagli N, Liberton D, Verma P, Sincan M, Lee J, Bagci U. Deep geodesic learning for segmentation and anatomical landmarking. *IEEE Trans Med Imaging*. 2019;38:919–931.
42. Chan J, Kearney V, Haaf S, et al. A convolutional neural network algorithm for automatic segmentation of head and neck organs-at-risk using deep lifelong learning. *Med Phys*. 2019;46:2204–2213.
43. Chen H, Lu W, Chen M, et al. A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy. *Phys Med Biol*. 2019;64:025015.
44. Lee H, Lee E, Kim N, et al. Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region. *Front Oncol*. 2019;9:239.
45. Hänsch A, Schwier M, Gass T, et al. Evaluation of deep learning methods for parotid gland segmentation from CT images. *J Med Imaging*. 2019;6:011005.
46. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46:576–589.
47. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol*. 2019;29:1961–1967.
48. Men K, Geng H, Cheng C, et al. Technical note: more accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades. *Med Phys*. 2019;46:286–292.
49. Tappeiner E, Pröll S, Hönig M, et al. Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. *Int J Comput Assist Radiol Surg*. 2019;14:745–754.
50. van Rooij W, Dahele M, Ribeiro Brandao Het al. Deep learning-based delineation of head and neck organs-at-risk: geometric and dosimetric evaluation. *Int J Radiat Oncol Biol Phys*. 2019;104:677–684.
51. Wu X, Udupa J, Tong Y, et al. AAR-RT – a system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med Image Anal*. 2019;54:45–62.
52. Ayyalusamy A, Vellaiyan S, Subramanian S, et al. Auto-segmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity. *Radiat Oncol J*. 2019;37:134–142.
53. Willems S, Crijns W, La Greca Saint-Esteven A, et al. Clinical implementation of DeepVoxNet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy. In: *Clinical Image-Based Procedures: Translational Research in Medical Imaging - CLIP 2018*, volume 11041 of *LNCS*. Springer; 2018:223–232.
54. Ren X, Xiang L, Nie D, et al. Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images. *Med Phys*. 2018;45:2063–2075.
55. Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys*. 2018;45:4558–4567.
56. Wang Z, Wei L, Wang L, Gao Y, Chen W, Shen D. Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. *IEEE Trans Image Process*. 2018;27:923–937.
57. Močnik D, Ibragimov B, Xing L, et al. Segmentation of parotid glands from registered CT and MR images. *Phys Med*. 2018;52:33–41.
58. Kieselmann J, Kamerling C, Burgos N, et al. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. *Phys Med Biol*. 2018;63:145007.
59. Meillan N, Bibault J-E, Vautier J, et al. Automatic intracranial segmentation: is the clinician still needed? *Technol Cancer Res Treat*. 2018;17:1–7.
60. Nikolov S, Blackwell S, Mendes R, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv 1809.04430; 2018.
61. Yang J, Haas B, Fang R, et al. Atlas ranking and selection for automatic segmentation of the esophagus from CT scans. *Phys Med Biol*. 2017;62:9140–9158.
62. Aghdasi N, Li Y, Berens A, Harbison R, Moe K, Hannaford B. Efficient orbital structures segmentation with prior anatomical knowledge. *J Med Imaging*. 2017;4:034501.
63. Urban S, Tanács A. Atlas-based global and local RF segmentation of head and neck organs on multimodal MRI images. In: *International Symposium on Image Signal Processing Analysis - ISPA 2017*. IEEE; 2017:99–103.
64. Wachinger C, Brennan M, Sharp G, Golland P. Efficient descriptor-based segmentation parotid glands with nonlocal means. *IEEE Trans Biomed Eng*. 2017;64:1492–1502.
65. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys*. 2017;44:547–557.
66. Raudaschl P, Zaffino P, Sharp GC, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys*. 2017;44:2020–2036.
67. Van de Velde J, Wouters J, Vercauteren T, et al. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. *Radiat Oncol*. 2016;11:1.
68. Wardman K, Prestwich R, Gooding M, Speight R. The feasibility of atlas-based automatic segmentation of MRI for H&N radiotherapy planning. *J Appl Clin Med Phys*. 2016;17:146–154.
69. Zaffino P, Raudaschl P, Fritscher K, Sharp G, Spadea M. Technical note: plastimatch mabs, an open source tool for automatic image segmentation. *Med Phys*. 2016;43:5155.
70. Fritscher K, Raudaschl P, Zaffino P, Spadea M, Sharp G. Deep neural networks for fast segmentation of 3D medical images. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, volume 9901 of *LNCS*. Springer; 2016:158–165.
71. Awan M, Dyer B, Kalpathy-Cramer J, et al. Auto-segmentation of the brachial plexus assessed with TaCTICS – a software platform for rapid multiple-metric quantitative evaluation of contours. *Acta Oncol*. 2015;54:562–566.
72. Wachinger C, Fritscher K, Sharp G, Golland P. Contour-driven atlas-based segmentation. *IEEE Trans Med Imaging*. 2015;34:2492–2505.
73. Hoang DA, Eminowicz G, Mendes R, et al. Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Med Phys*. 2015;42:5027–5034.
74. Dolz J, Leroy H, Reyns N, Massoptier L, Vermandel M. A fast and fully automated approach to segment optic nerves on MRI and its application to radiosurgery. In: *International Symposium on Biomedical Imaging - ISBI 2015*, pages 1102–1105. IEEE; 2015.
75. Yang X, Wu N, Cheng G, et al. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy. *Int J Radiat Oncol Biol Phys*. 2014;90:1225–1233.
76. Fritscher K, Peroni M, Zaffino P, Spadea M, 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;41:051910.

77. 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;9:173.
78. Sjöberg C, Johansson S, Ahnesjö A. How much will linked deformable registrations decrease the quality of multi-atlas segmentation fusions? *Radiat Oncol*. 2014;9:251.
79. Harrigan R, Panda S, Asman A, et al. Robust optic nerve segmentation on clinically acquired computed tomography. *J Med Imaging*. 2014;1:034006.
80. Walker G, Awan M, Tao R, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother Oncol*. 2014;112:321–325.
81. Yang J, Amini A, Williamson R, et al. Automatic contouring of brachial plexus using a multi-atlas approach for lung cancer radiation therapy. *Pract Radiat Oncol*. 2013;3: 139–e147.
82. Zhu M, Bzdusek K, Brink C, et al. Multi-institutional quantitative evaluation and clinical validation of smart probabilistic image contouring engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male pelvis areas. *Int J Radiat Oncol Biol Phys*. 2013;87:809–816.
83. Cheng G, Yang X, Wu N, Xu Z, Zhao H, Wang Y, Liu T. Multi-atlas-based segmentation of the parotid glands of MR images in patients following head-and-neck cancer radiotherapy. In: *Medical Imaging 2013: Computer-Aided Diagnosis*, volume 8670, SPIE; 2013:86702Q.
84. Daisne J-F., Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiat Oncol*. 2013;8:154.
85. Chen A, Niermann K, Deeley M, Dawant B. Evaluation of multiple-atlas-based strategies for segmentation of the thyroid gland in head and neck CT images for IMRT. *Phys Med Biol*. 2012;57:93–111.
86. Qazi A, Pekar V, Kim J, Xie J, Breen S, Jaffray D. 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.
87. Teguh D, Levendag P, Voet P, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *Int J Radiat Oncol Biol Phys*. 2011;81:950–957.
88. Noble J, Dawant B. An atlas-navigated optimal medial axis and deformable model algorithm (NOMAD) for the segmentation of the optic nerves and chiasm in MR and CT images. *Med Image Anal*. 2011;15:877–884.
89. Deeley M, 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;56:4557–4577.
90. Tsuji S, Hwang A, Weinberg V, Yom S, Quivey J, Xia P. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. *Int J Radiat Oncol Biol Phys*. 2010;77:707–714.
91. Pekar V, Allaire S, Qazi A, Kim J, Jaffray D. Head and neck auto-segmentation challenge: segmentation of the parotid glands. In: *Medical Image Analysis for the Clinic: A Grand Challenge 2010*, MICCAI; 2010:273–280.
92. Pekar V, Allaire S, Kim J, Jaffray D. Head and neck auto-segmentation challenge. *MIDAS J*. 2009;5:5.
93. Sims R, Isambert A, Grégoire V, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiother Oncol*. 2009;93:474–478.
94. 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.
95. 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*, volume 5242 of *LNCS*, Springer; 2008:434–441.
96. Bekes G, Máté E, Nyúl L, Kuba A, Fidrich M. Geometrical model-based segmentation of the organs of sight on CT images. *Med Phys*. 2008;35:735–743.
97. Fortunati V, Verhaart R, Niessen W, Veenland J, Paulides M, van Walsum T. Automatic tissue segmentation of head and neck MR images for hyperthermia treatment planning. *Phys Med Biol*. 2015;60:6547–6562.
98. Verhaart R, Fortunati V, Verduijn G, van Walsum T, Veenland J, Paulides M. CT-based patient modeling for head and neck hyperthermia treatment planning: manual versus automatic normal-tissue-segmentation. *Radiother Oncol*. 2014;111:158–163.
99. Fortunati V, Verhaart R, van der 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: 071905.
100. Schneider U, Pedroni E, Lomax A. The calibration of CT Hounsfield units for radiotherapy treatment planning. *Phys Med Biol*. 1996;41: 111–124.
101. Pereira G, Traughber M, Muzic R. The role of imaging in radiation therapy planning: past, present, and future. *Biomed Res Int*. 2014; 2014:231090.
102. Brouwer C, Steenbakkers R, 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.
103. Leibfarth S, Mönnich D, Welz S, et al. A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning. *Acta Oncol*. 2013;52:1353–1359.
104. Fortunati V, Verhaart R, Angeloni F, et al. Feasibility of multimodal deformable registration for head and neck tumor treatment planning. *Int J Radiat Oncol Biol Phys*. 2014;90:85–93.
105. Joint Head and Neck MRI-Radiotherapy Development Cooperative. Prospective quantitative quality assurance and deformation estimation of MRI-CT image registration in simulation of head and neck radiotherapy patients. *Clin Transl Radiat Oncol*. 2019;18:120–127.
106. Peroni M, Ciardo D, Spadea M, et al. Automatic segmentation and online virtualCT in head-and-neck adaptive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2012;84:e427–e433.
107. Hvid C, Elstrxxxom C, Jensen K, Alber M, Grau C. Accuracy of software-assisted contour propagation from planning CT to cone beam CT in head and neck radiotherapy. *Acta Oncol*. 2016;55:1324–1330.
108. Wang T, Bradshaw GB, Beitler J, et al. Optimal virtual monoenergetic image in "TwinBeam" dual-energy CT for organs-at-risk delineation based on contrast-noise-ratio in head-and-neck radiotherapy. *J Appl Clin Med Phys*. 2019;20:121–128.
109. Bhandare N, Mendenhall W. A literature review of late complications of radiation therapy for head and neck cancers: incidence and dose response. *J Nucl Med Radiat Ther*. 2012;S2:009.
110. Siddiqui F, Movsas B. Management of radiation toxicity in head and neck cancers. *Semin Radiat Oncol*. 2017;27:340–349.
111. Strojan P, Hutcheson K, Eisbruch A, et al. Treatment of late sequelae after radiotherapy for head and neck cancer. *Cancer Treat Rev*. 2017;59:79–92.
112. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045–1057.
113. Prior F, Smith K, Sharma A, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. *Sci Data*. 2017;4:170124.
114. Vallières M, Kay-Rivest E, Perrin L, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7:10117.
115. Grossberg A, Mohamed A, Elhalawani H, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci Data*. 2018;5:180173.
116. Cardenas C, Mohamed A, Yang J, et al. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. *Med Phys*. 2020;47:2317–2322.
117. Fedorov A, Clunie D, Ulrich E, et al. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. *PeerJ*. 2016;4:e2057.

118. Beichel R, Smith BJ, Bauer C, et al. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. *Med Phys*. 2017;44:479–496.
119. La Macchia M, Fellin F, Amichetti M, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol*. 2012;7:160.
120. Kearney V, Chan J, Valdes G, Solberg T, Yom S. The application of artificial intelligence in the IMRT planning process for head and neck cancer. *Oral Oncol*. 2018;87:111–116.
121. Van de Velde J, Audenaert E, Speleers B, et al. An anatomically validated brachial plexus contouring method for intensity modulated radiation therapy planning. *Int J Radiat Oncol Biol Phys*. 2013;87:802–808.
122. Sun Y, Yu XL, Luo W, et al. Recommendation for a contouring method and atlas of organs at risk in nasopharyngeal carcinoma patients receiving intensity-modulated radiotherapy. *Radiother Oncol*. 2014;110:390–397.
123. Kong F, Ritter T, Quint D, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *Int J Radiat Oncol Biol Phys*. 2011;81:1442–1457.
124. Christianen M, Langendijk J, Westerlaan H, van de Water T, Bijl H. Delineation of organs at risk involved in swallowing for radiotherapy treatment planning. *Radiother Oncol*. 2011;101:394–402.
125. van de Water T, Bijl H, Westerlaan H, Langendijk J. Delineation guidelines for organs at risk involved in radiation-induced salivary dysfunction and xerostomia. *Radiother Oncol*. 2009;93:545–552.
126. Pacholke H, Amdur R, Schmalfuss I, Louis D, Mendenhall W. Contouring the middle and inner ear on radiotherapy planning scans. *Am J Clin Oncol*. 2005;28:143–147.
127. Hall W, Guiou M, Lee N, et al. Development and validation of a standardized method for contouring the brachial plexus: preliminary dosimetric analysis among patients treated with IMRT for head-and-neck cancer. *Int J Radiat Oncol Biol Phys*. 2008;72:1362–1367.
128. Chen W, Zhang H, Zhang W, et al. Development of a contouring guide for three different types of optic chiasm: a practical approach. *J Med Imaging Radiat Oncol*. 2019;63:657–664.
129. Taha A, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
130. Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun*. 2018;9:5217.
131. Armato S, Tahir B, Sharp G. AAPM grand challenges symposium. *Med Phys*. 2019;46:e485–e486.
132. Iglesias J, Sabuncu M. Multi-atlas segmentation of biomedical images: a survey. *Med Image Anal*. 2015;24: 205–219.
133. Edmund J, Nyholm T. A review of substitute CT generation for MRI-only radiation therapy. *Radiat Oncol*. 2017;12:28.
134. Adjeiwaah M, Bylund M, Lundman J, et al. Dosimetric impact of MRI distortions: a study on head and neck cancers. *Int J Radiat Oncol Biol Phys*. 2019;103:994–1003.
135. Raaymakers BW, Jürgenliemk-Schulz IM, Bol GH, et al. First patients treated with a 1.5 T MRI-Linac: clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment. *Phys Med Biol*. 2017;62:L41–L50.
136. Lei Y, Harms J, Wang T, et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med Phys*. 2019;46:3565–3581.
137. Klages P, Benslimane I, Riyahi S, et al. Patch-based generative adversarial neural network models for head and neck MR-only planning. *Med Phys*. 2020;47:626–642.
138. Comelli A, Stefano A, Bignardi S, et al. Active contour algorithm with discriminant analysis for delineating tumors in positron emission tomography. *Artif Intell Med*. 2019;94:67–78.
139. Schipaanboord B, Boukerroui D, Peressutti D, et al. Can atlas-based auto-segmentation ever be perfect? Insights from extreme value theory. *IEEE Trans Med Imaging*. 2019;38:99–106.
140. Larrue A, Gujral D, Nutting C, Gooding M. The impact of the number of atlases on the performance of automatic multi-atlas contouring. *Phys Med*. 2015;31:e30.
141. Ibtehaz N, Rahman M. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw*. 2020;121:74–87.
142. Zhang X, Wang L, Yang D, etal. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging;*. 2020;(in press).
143. Nelms B, Tomé W, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys*. 2012;82:368–378.
144. Brouwer C, Steenbakkers R, van den Heuvel E, et al. 3D variation in delineation of head and neck organs at risk. *Radiat Oncol*. 2012;7:32.
145. Tao C-J, Yi J-L, Chen N-Y, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. *Radiother Oncol*. 2015;115:407–411.
146. Krayenbuehl J, Zamburlini M, Ghandour S, et al. Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer. *Radiat Oncol*. 2018;13:170.
147. Graves Y, Smith AA, McIlvena D, et al. A deformable head and neck phantom with in-vivo dosimetry for adaptive radiotherapy quality assurance. *Med Phys*. 2015;42:1490–1497.
148. Li J, Udupa J, Tong Y, Wang L, Torigian D. LinSEM: linearizing segmentation evaluation metrics for medical images. *Med Image Anal*. 2020;60:101601.
149. Loo S, Martin W, Smith P, Cherian S, Roques T. Interobserver variation in parotid gland delineation: a study of its impact on intensity-modulated radiotherapy solutions with a systematic review of the literature. *Br J Radiol*. 2012;85:1070–1077.
150. Voet P, Dirkx M, Teguh D, Hoogeman M, Levendag P, Heijmen B. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother Oncol*. 2011;98:373–377.
151. Delaney A, Dahele M, Slotman B, Verbakel W. Is accurate contouring of salivary and swallowing structures necessary to spare them in head and neck VMAT plans? *Radiother Oncol*. 2018;127:190–196.
152. Lim T, Gillespie E, Murphy J, Moore K. Clinically oriented contour evaluation using dosimetric indices generated from automated knowledge-based planning. *Int J Radiat Oncol Biol Phys*. 2019;103:1251–1260.
153. Aliotta E, Nourzadeh H, Siebers J. Quantifying the dosimetric impact of organ-at-risk delineation variability in head and neck radiation therapy in the context of patient setup uncertainty. *Phys Med Biol* 2019;64:135020.