Medical Image Analysis 73 (2021) 102161

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

The predictive value of segmentation metrics on dosimetry in organs at risk of the brain

Robert Poel[a,b], Elias Rüfenacht[b], Evelyn Hermann[a,c], Stefan Scheib[d], Peter Manser[e], Daniel M. Aebersold[a], Mauricio Reyes[b,*]

[a] Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Bern, Switzerland
[b] ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland
[c] Radiotherapy Department, Riviera-Chablais Hospital, Rennaz, Switzerland
[d] Varian Medical Systems Imaging Laboratory, GmbH, Switzerland
[e] Division of Medical Radiation Physics and Department of Radiation Oncology, Inselspital, Bern University Hospital, and University of Bern, Bern, Switzerland

ARTICLE INFO

Article history:
Received 29 January 2021
Revised 29 June 2021
Accepted 2 July 2021
Available online 13 July 2021

Keywords:
Segmentation parameters
Clinical validation
Radiotherapy
Brain OARs
Automatic segmentation
Dose distribution

ABSTRACT

Background: Fully automatic medical image segmentation has been a long pursuit in radiotherapy (RT). Recent developments involving deep learning show promising results yielding consistent and time efficient contours. In order to train and validate these systems, several geometric based metrics, such as Dice Similarity Coefficient (DSC), Hausdorff, and other related metrics are currently the standard in automated medical image segmentation challenges. However, the relevance of these metrics in RT is questionable. The quality of automated segmentation results needs to reflect clinical relevant treatment outcomes, such as dosimetry and related tumor control and toxicity. In this study, we present results investigating the correlation between popular geometric segmentation metrics and dose parameters for Organs-At-Risk (OAR) in brain tumor patients, and investigate properties that might be predictive for dose changes in brain radiotherapy.

Methods: A retrospective database of glioblastoma multiforme patients was stratified for planning difficulty, from which 12 cases were selected and reference sets of OARs and radiation targets were defined. In order to assess the relation between segmentation quality -as measured by standard segmentation assessment metrics- and quality of RT plans, clinically realistic, yet alternative contours for each OAR of the selected cases were obtained through three methods: (i) Manual contours by two additional human raters. (ii) Realistic manual manipulations of reference contours. (iii) Through deep learning based segmentation results. On the reference structure set a reference plan was generated that was re-optimized for each corresponding alternative contour set. The correlation between segmentation metrics, and dosimetric changes was obtained and analyzed for each OAR, by means of the mean dose and maximum dose to 1% of the volume (Dmax 1%). Furthermore, we conducted specific experiments to investigate the dosimetric effect of alternative OAR contours with respect to the proximity to the target, size, particular shape and relative location to the target.

Results: We found a low correlation between the DSC, reflecting the alternative OAR contours, and dosimetric changes. The Pearson correlation coefficient between the mean OAR dose effect and the Dice was -0.11. For Dmax 1%, we found a correlation of -0.13. Similar low correlations were found for 22 other segmentation metrics. The organ based analysis showed that there is a better correlation for the larger OARs (i.e. brainstem and eyes) as for the smaller OARs (i.e. optic nerves and chiasm). Furthermore, we found that proximity to the target does not make contour variations more susceptible to the dose effect. However, the direction of the contour variation with respect to the relative location of the target seems to have a strong correlation with the dose effect.

* Corresponding author at: University of Bern ARTORG Center, Murtenstrasse 50 CH-3008 Bern, Switzerland.
E-mail address: Mauricio.reyes@med.unibe.ch (M. Reyes).

https://doi.org/10.1016/j.media.2021.102161

Conclusions: This study shows a low correlation between segmentation metrics and dosimetric changes for OARs in brain tumor patients. Results suggest that the current metrics for image segmentation in RT, as well as deep learning systems employing such metrics, need to be revisited towards clinically oriented metrics that better reflect how segmentation quality affects dose distribution and related tumor control and toxicity.

1. Introduction

For radiotherapy (RT) planning it is important to have accurate contours of the target as well as the organs that need to be spared. Contouring in clinical practice is predominantly performed by manual segmentation. Unfortunately manual segmentation is subject to inconsistencies which are known as inter- and intra-observer variability (Mazzara et al., 2004; Deeley, 2011; Visser et al., 2019). Manual contouring will thus be subject to inaccuracies which are known to have a high impact on treatment quality (Jameson et al., 2010; Marks, 2013; Stanley et al., 2013; Sandström et al., 2016; Vinod et al., 2016; Cloak et al., 2019). In addition, manual segmentation of targets and organs at risk (OARs) is a very time consuming task, varying from 1 to 4 h depending on location and tumor extent (Bondiau et al., 2005; Harari et al., 2010; Deeley, 2011; Voet et al., 2011). In the current RT era where daily adaptive treatment finds its way into the clinic (Brock, 2019), the need of fast and automated segmentation is increasing. Full automatic segmentation has therefore been one of the "holy grails" in RT.

Recent publications have shown that auto-segmentation can yield consistent and time efficient contours for different tumor sites, which is summarized by Cardenas, 2019. Besides, for the common clinical practice, auto-segmentation can be a useful tool creating data for retrospective studies. With the large amount of digital imaging and dosimetric data, large retrospective studies on treatment outcome and toxicity can be performed..

The current state of the art auto-segmentation methods are based on deep learning (DL) and more particular on convolutional neural networks (Meyer et al., 2018). Ever more deep learning based approaches are developed and are becoming clinically available through commercial products (Brunenberg, 2020; van Dijk et al., 2020). This new generation of auto-segmentation methods has outperformed the quality of atlas based and traditional machine learning based auto-segmentation approaches. The first published deep learning based auto-segmentation studies already showed results in terms of dice similarity coefficient (DSC) of well above 0.8 (Roth et al., 2015; Ben-Cohen, 2016; Hu et al., 2016; Milletari et al., 2016; Zhou, 2016; Litjens, 2017). Recent and more sophisticated DL methods show DSCs in the range over 0.8, with some reported cases exceeding 0.9, depending on the type of the OAR (Cardenas, 2019). Most recently Mlynarski et al. published impressive results in OARs of the brain by combining deep learning with sophisticated post processing methods (Mlynarski et al., 2020).

Although these results are promising, a DSC of 0.8 still leaves a lot of room for errors, especially in larger OARs, that might have a substantial impact on the treatment. More importantly, it is unknown when and where such an error occurs. Consequently, auto-segmentation results require thorough visual inspection by a trained professional, which again requires additional valuable time.

To solve this issue, one can aim to improve automatic segmentation results in terms of geometrical similarity parameters up to the point it reaches perfection (i.e. a dice of 1.0). This is an ambitious task that is pursued by many. Since progress over the last years have been incremental, it is uncertain if this goal can ever be achieved. Instead, in this study we focus on how we can validate auto-segmentation results in such a way that they can predict the quality of the treatment.

The practical standard for validating automatic segmentation is based on geometrical similarity indices of which the DSC and the Hausdorff distance are the most popular. In the case of RT we can think of other parameters that are perhaps more clinically relevant. Current methods for validating auto-segmentation results have been criticized before (Gooding et al., 2018; Maier-Hein et al., 2018; Nikolov, 2018; Vaassen et al., 2020). Gooding et al. (2018) suggest a qualitative measure of experts in the field being able to distinct auto-segmented contours from manually drawn contours. More recently, Kofler et al. suggested new parameters for loss functions based on quality assessment by experts (Kofler, 2021). Other parameters have been suggested, like added path length (Vaassen et al., 2020) or surface dice (Nikolov, 2018), to determine how valuable contours are for RT, in terms of manual adjustments time. Although the time required to adjust auto-segmented results is important, the most relevant parameter to look at in RT is treatment outcome. Treatment outcome in general is quantified by tumor control and toxicity, both reflected by the dose distribution. Dose distribution is a readily available measure, provided one has access to an RT treatment planning system (TPS).

Describing the correlation between contour variation and dosimetry, to our knowledge has only been explored by Xian et al. (Xian and Chen, 2020). They studied the effect of systematic geometrical transformation to several c-shaped targets, and concluded that dosimetric indices should be included in the assessment of contour accuracy. However, in their assessment they only provided the plan of the reference contour and determined the dose parameters of the geometrical transformations on the dose distribution. Obviously, systematically moving the alternative target contour away from the reference target will decrease both geometrical similarity, as well as dose coverage. This does not exactly reflect the dose effect of an incorrect contour, since for this matter you need to calculate a dose distribution for both the reference target and the transformed target, and then determine the differences these both distributions have on the reference contour volume.

In this study, we analyzed the correlation between the geometric similarity parameters and the effect a specific change on an OAR contour has on the dose distribution. To do so we focus on radiotherapy for intracranial diseases. A large amount of cancers situated in the brain, such as metastasis, but also primary diseases as gliomas, are being treated with RT. The brain is a location with a large amount of critical structures that are important to spare, and thus accurate delineation is of importance (Scoccianti et al., 2015). Most of the structures are small and can only be distinguished on magnetic resonance imaging (MRI). Contouring is therefore a tedious process. Deep learning methods for intracranial OAR segmentation are under development, but are up to date not yet commercially available.

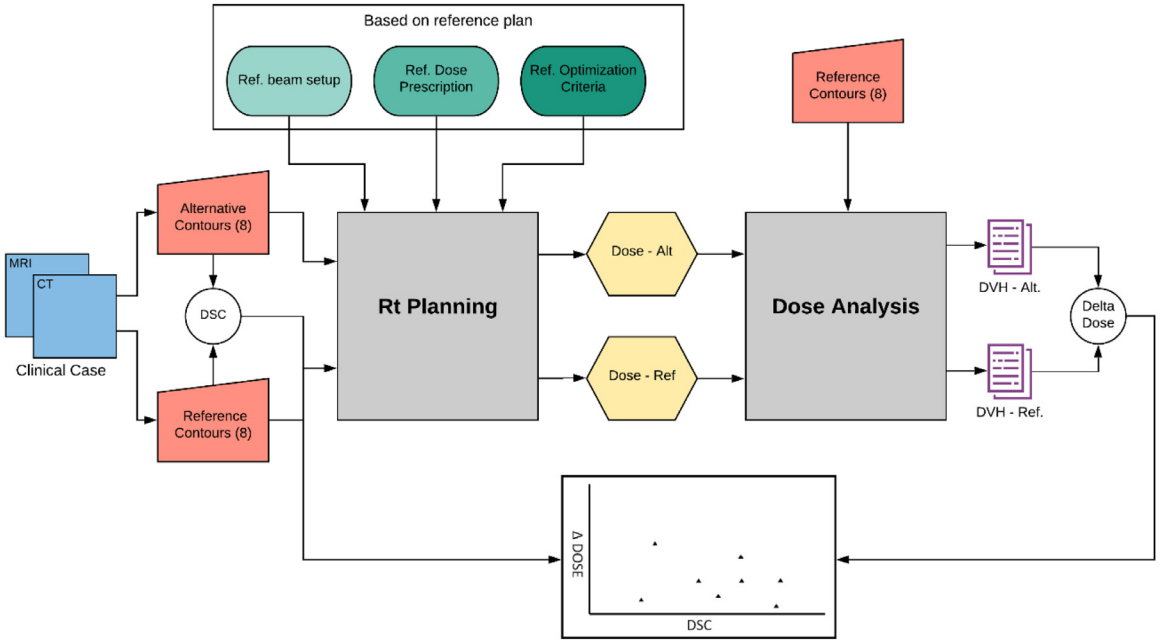

Fig. 1. Graphical description of the methodology as performed on a single alternative contour set containing 8 OARs. On the computed tomography (CT) and MRI imaging of a clinical case the OAR contours are defined (i.e., reference structures), as well as an alternative structure set defined by a second physician, from an auto segmentation method, or manual manipulation of the reference. The geometric metric (in this case DSC) is determined for the alternative OARs with respect to the reference. The two structure sets are used as input for an RT plan. The beam setup, dose prescription and the optimization criteria are set based on the reference plan. This generates two different dose distributions. The reference structure set is overlaid on the output dose distributions, and the dose volume histograms (DVH) of the reference and alternative plans are determined. The difference in dose between the alternative plan and the reference plan is plotted against their respective DSC.

It is our hypothesis that currently used geometrical indices are not a good predictor of the quality of a segmentation for the purpose of intracranial RT. We analyzed the level of correlation between dosimetry and geometrical metrics used to assess segmentation quality. As geometrical metrics, we have selected a set of 23 commonly used parameters. As geometrical similarity approaches a perfect metric (i.e. a DSC of 1.0), it is expected that dose effects will be minimal. On the other hand, if there is barely geometrical similarity, it is questionable whether this information is clinically relevant at all. Consequently, we want to focus on analyzing contour variations that could present itself in a clinical situation, regardless of how the contours are obtained. For this purpose, we want to stay away from contour alternatives that are near perfect or on the other side, are obviously wrong. In terms of DSCs however, the values depend heavily on the respective OAR, mainly influenced by its size. For readability, we focus on one specific parameter throughout this manuscript, the DSC. We specifically choose this metric since it is still the most used parameter and is well interpreted by many professionals in the field. Furthermore, the DSC is a widely used parameter in loss functions in deep learning based auto-segmentation methods. We will come back to the other parameters in the results section.

Additionally, we performed synthetic experiments to find what other characteristics, that are not depicted by these geometric parameters, have an effect on the dose distribution. Since we are focusing on OARs, the goal of RT is to avoid dose as much as possible. The amount of dose an OAR receives is therefore dependent on its location relative to the target, dose constraints and optimization parameters. Furthermore, how the dose will be affected by a change in contour is additionally dependent on the technique of dose delivery and the shape and nature of the specific changes to the contour. Is it an over-segmentation or an under-segmentation? Are errors in the segmentation placing the OAR closer or more distant to the target? Does the size of the OAR have an influence? Are there specific outliers? Consequently, we are investigating characteristics as shape, size, distance to the target and relative location. With these findings, we expect to contribute to a better understanding as to what quality of auto-segmentation is required to obtain clinically acceptable treatments, as well as to foster with implementing auto-segmentation into the clinics in a safe and secure way.

2. Materials and methods

2.1. Correlation on clinical cases

To assess the correlation between the DSC metric and the dose effect in OARs of the brain, we have constructed RT plans for different sets of contours on a selection of cases from a cohort of glioblastoma multiforme (GBM) patients. Fig. 1 presents a schematic overview of the methodology, which is detailed in the subsections below. From the left to right: 1) Selection of clinical imaging data and reference contours. 2) Creation of alternative sets of contours to mimic segmentation and dice metric variability. 3) Calculation of DSC and other geometrical metrics used to assess segmentation. 4) Calculation of dose distributions on contour sets. 5) Assessment of dosimetric differences for the reference and the alternative contour sets. 6) Correlation analysis between dosimetric differences and DSC.

2.1.1. Clinical imaging data

The clinical data for this study was selected from a retrospective database of 100 post-operative GBM cases that have been treated with RT at the Inselspital, University Hospital Bern. All cases contained a planning computed tomography (CT) registered

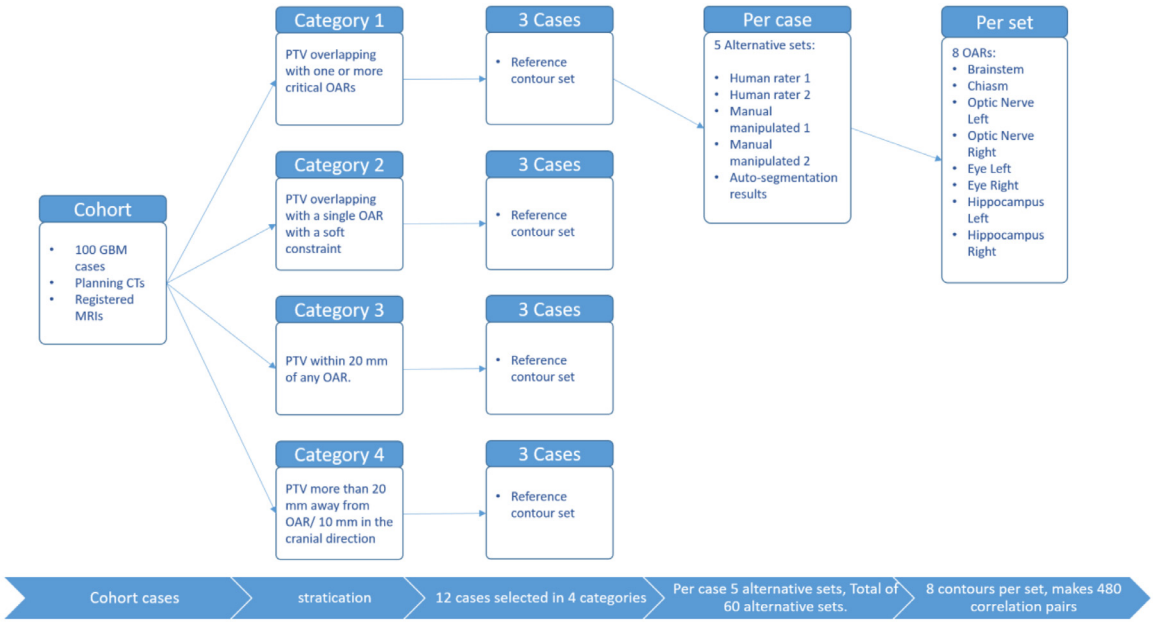

Fig. 2. Selection and acquisition of the study data. From left to right it started with 100 post-operative glioblastoma multiforme cases. The 100 cases were stratified into 4 categories. From each category, 3 cases were selected. For these selected cases, 5 alternative sets of OAR contours were composed. Each of the alternative contour sets contained 8 OARs.

to MRI images and a reference structure set containing OARs as well as the target volumes.

The planning target volume (PTV) was defined according to the ESTRO-ACROP guidelines (Niyazi, 2016). The OARs are contoured according to Scoccianti et al. (2015) and verified by mutual consensus of three experienced radiation oncology experts.

To obtain a representative selection of cases in terms of tumor location, the cohort of GBM cases was divided in 4 categories depending on how demanding a case is for radiotherapy planning in terms of the included OARs (Fig. 2).

- Category 1; is highly challenging, and is defined as the PTV overlapping with one or more critical OARs with a hard constraint (brainstem or optic tract).
- Category 2; is defined as the PTV overlapping with one of the hippocampi.
- Category 3; are those cases where the PTV resides within 20 mm of one or more OARs.
- Category 4; the least challenging cases are defined as the PTV more than 20 mm away from any OAR or more than 10 mm in the cranial direction from any OAR. Since planning is performed with co-planar volumetric modulated arc therapy (VMAT) technique perpendicular to the body axis, OARs residing superior of the target are automatically spared.

From each category, three cases were included to complete our study set of 12 cases. No additional analysis is performed on the stratification categories.

2.1.2. Alternative contours

The reference structure sets comprise 8 selected OARs; the brainstem, the optic chiasm, the optic nerves (left and right), the eyes (left and right) and the hippocampi (left and right). Other smaller and peripheral located OARs such as the cochlea, lenses and lacrimal glands were not included since the impact on the resultant dose distribution is typically limited due to size and location.

Each of the 12 included cases received next to the 8 reference OAR structures, five sets of alternative OAR contours. Within these alternative contours, we want to have realistic data from different sources that does provide sufficient variety in relation to the reference contours. Two radiation oncology physicians manually contoured the OARs resulting in alternative contours modeling inter-rater variability. Furthermore, for each case an alternative structure set was obtained by a standard version of an in-house developed deep learning based auto-segmentation method based on the U-net architecture (Isensee, 2021). We have specifically chosen for a standard version of the auto-segmentation method that did not provide state of the art results, but instead provides us with a wider range of segmentation quality results.

A version of the U-Net (Ronneberger et al., 2015) was adjusted to meet the needs of multi-organ automatic segmentation on multiple MRI sequences. In order to incorporate recent improvements we interleaved batch normalization [33] and a 10%-dropout [75] layer after each convolution layer. The resulting feature maps of the up-sampling layer are then concatenated with the feature maps from the contractive path, which are provided by the skip connections. The ending sequence of the expanding path consists of a 1×1 convolution and a softmax layer to get the probabilities for each OAR and the background. For training we used focal Loss (gamma $= 2$) [48] in combination with an ADAM optimizer (betas $= (0.99, 0.999)$) [38]. The initial learning rate was 10e-3, which reduced to $4 \times$ 10e-4 after 150 epochs, and to $1.6 \times$ 10e-4 after 250 epochs. The model was trained for 300 epochs in total, with a mini-batch size of 20 training examples.

Additionally, all 12 cases received two sets of alternative OAR structures by means of controlled manual manipulation of the reference contours. These manual manipulations were designed to further increase the range of geometrical similarity, and study the patterns of correlations at a low regime of segmentation performance. This data will complement the data of the human raters and the auto-segmentation results in order to obtain a wide distribution of possible alternatives. All structures were contoured in a research environment of the clinical version of Eclipse TPS (Eclipse, version 15.6, Varian, Palo Alto, United States of America). In summary, every case had a set of reference OARs and 5 sets of alternative OAR contours. In total 60 alternative contour sets were created, resulting in 480 alternative OAR contours.

Table 1
Structures and dose prescription.

Dose prescription		
PTV (Reference only*)		60 Gy
Constraint doses		
Brainstem Surface	Max dose to 1%	≤ 60 Gy
Brainstem Center	Max dose to 1%	≤ 54 Gy
Eye ($L + R$)	Max dose to 1%	≤ 10 Gy
Chiasm	Max dose to 1%	≤ 55 Gy
Optic Nerve ($L + R$)	Max dose to 1%	≤ 55 Gy
Hippocampus ($L + R$)	Dose to 40% of volume	≤ 7.3 Gy
Reference only*		
Lens ($L + R$)	Max dose to 1%	≤ 10 Gy
Lacrimal gland ($L + R$)	Mean dose	≤ 25 Gy
Cochlea ($L + R$)	Hard: Mean Dose	≤ 45 Gy
	Soft: Mean Dose	≤ 32 Gy
Retina ($L + R$)	Max dose to 1%	≤ 45 Gy
Pituitary	Hard: Mean Dose	≤ 45 Gy
	Soft: Mean Dose	≤ 20 Gy

*The structures labeled under reference only, do not have alternative versions and are therefore not interchanged during the different dose calculations, since the dosimetric effect due to size and location is typically limited.

2.1.3. Geometrical similarity indices

To determine the DSC of each alternative structure with respect to the reference contour, the structure sets were exported from the TPS in RT-Dicom format. They were converted to Nifti format in 3D slicer software (www.slicer.org). With the open-source python software pymia (Jungo et al., 2021), the DSC for each alternative - reference contour pair was determined. Additionally, another set of 22 alternative segmentation parameters was determined using evaluation tools provided by the Visual Concept Extraction Challenge in Radiology (VISCERAL, www.visceral.eu) project. The list is supplemented with current popular measures as the average distance and the normalized surface dice (NSD) (Nikolov, 2018).

2.1.4. RT plan calculation

For every case, a reference RT plan was generated based on the reference structures. The Clinical Target Volume (CTV) was defined as the resection cavity and remaining GBM, including peritumoral edema, as per ESTRO-ACROP guidelines (Niyazi, 2016). A 3 mm margin was added, to form the PTV. According to clinical standard, the prescription dose for the PTV was set to 60 Gray (Gy) in a conventional scheme (30 × 2.0 Gy). The defined OARs and their respective hard and soft constraint doses can be found in Table 1.

A co-planar VMAT plan was set up, with a double full arc, and 6 megavolt X-ray flattening filter free beams, and optimized with the anisotropic analytical algorithm. The plan was accepted when all constraints were met. The plans were normalized on the PTV so that 100% of the prescribed dose covered 50% of the PTV.

For the alternative structure sets, we wanted to create a new plan while keeping all treatment parameters except the OAR structures the same. To do so we duplicated the reference plan and substituted the reference OARs with the alternative OARs. The beam orientation, prescription, constraints and optimization weights, remained unchanged from the reference plan. Thereafter, the plan was re-optimized. This would result in a slightly different dose distribution because of the different orientation of the defined OARs. These plans are also normalized so that 100% of the prescribed dose covered 50% of the PTV.

2.1.5. Dose parameter analysis

For all constructed RT plans (1 reference, 5 alternatives per case), the dose to the OARs of the reference structure set was analyzed. This reflects the dose the actual organ (i.e., reference) would receive, when it is incorrectly contoured (i.e., alternative). The difference in dose between the alternative plan and the reference plan is referred hereafter as the dose effect or delta dose. We determined the delta dose for both the mean OAR dose and the maximum dose to 1% of the OAR volume (Dmax 1%). These are typical metrics used to determine dose constraints to specific OARs (Emami, 2013).

2.1.6. Data analysis and statistics

We analyzed the data in two ways: I). By the nature of how the segmentation variability was established, divided into three groups: intra-rater variability, manual adjustments, and auto-segmentation results. This is to show the variability in contour similarity with respect to the reference for each of these groups. II). Per specific organ type. Since segmentation metrics are influenced by the volume of the segmentation, and inter-rater variability is OAR dependent, results might differ among different sizes of OARs. The specific organ types were divided in five groups; brainstem, optic chiasm, optic nerves, eyes and hippocampi.

The correlation for each of the groups was determined by the Pearson correlation coefficient. Additionally, the correlation with 22 alternative segmentation parameters, listed in Table 3, was determined. The calculations of the metrics are performed with the open-source python software pymia (Jungo et al., 2021) and the open source implementation of the surface DSC (Nikolov, 2018), available from https://github.com/deepmind/surface-distance. All the distance parameters are computed while considering the voxel spacing.

2.2. Possible characteristics predictive for the dose effect

Additional to the clinical data, synthetic experiments were performed to assess the correlation between the effect of alternative OAR contours and (i) the distance with respect to the target, (ii) the size of the OAR, (iii) their relative location with respect to the target and the radiation beams, (iv) their specific shape.

2.2.1. Dice versus distance

A synthetic spherical target and one reference OAR were defined in the center of the brain in the planning CT of one of the included subjects. Based on the reference OAR, 8 alternative contours were constructed with different shapes and sizes. This resulted in a variety of DSC with respect to the reference OAR (Fig. 3). This set of 9 different OARs (reference plus alternatives) were duplicated at 5 different distances from the target starting from 1.5 cm, up to 6.5 cm, with 1.5 cm increments. For each of the 5 resulting distances a reference plan was constructed. The goal of the reference plan was to obtain the lowest possible dose to the reference OAR, without compromising the prescription dose to the target of 60 Gy. For each of the 8 alternative OARs, the reference plan was duplicated while substituting the reference OAR for each of the alternative ones in the dose optimization step, in the same way as described in Section 2.1.4.

The obtained DSC of the alternative OARs with respect to the reference, are plotted against the dose-effect. The dose effect is determined by the dose difference to the reference OARs between the reference and alternative plans similar as in Section 2.1.5.

2.2.2. Dice versus size

It is well known that the size of an OAR has influence on voxel wise similarity segmentation metrics such as the DSC metric. We wanted to determine if the size of an OAR would correlate with the dose effect given a specific fixed DSC. For this purpose, we synthetically created 7 spherical reference OARs ascending in size from 1.0 cc to 64.4 cc, on the planning CT of an actual subject. All OARs had the same minimum distance to the target. For each of the reference OARs, we produced two alternative OARs, obtained by displacements in two different directions, with a DSC with respect

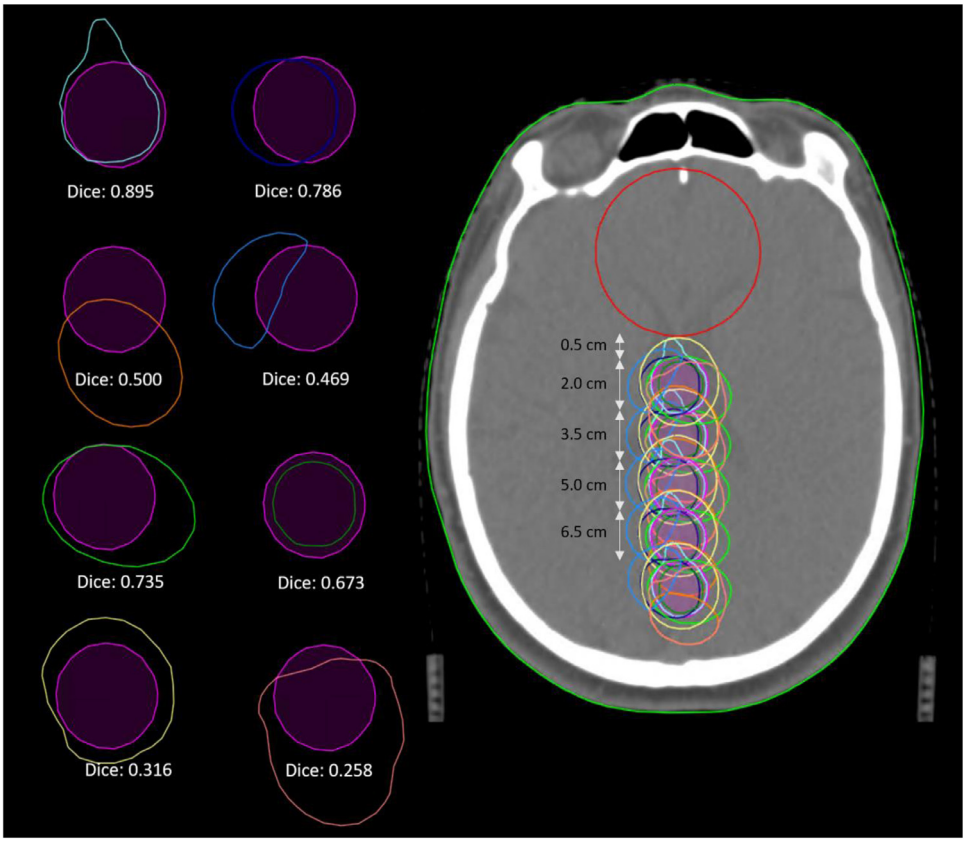

Fig. 3. Synthetic experiment to assess the relationship of distance to the target on the dice-dose effect. On the left, axial slice representations of the 8 synthetic variations on the spherical reference contour with their respective dice similarity coefficient. On the right, the reference OAR and the alternatives are located at 5 different distances from the target (PTV, red circle), leading to a total of 45 synthetically generated alternative contours. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

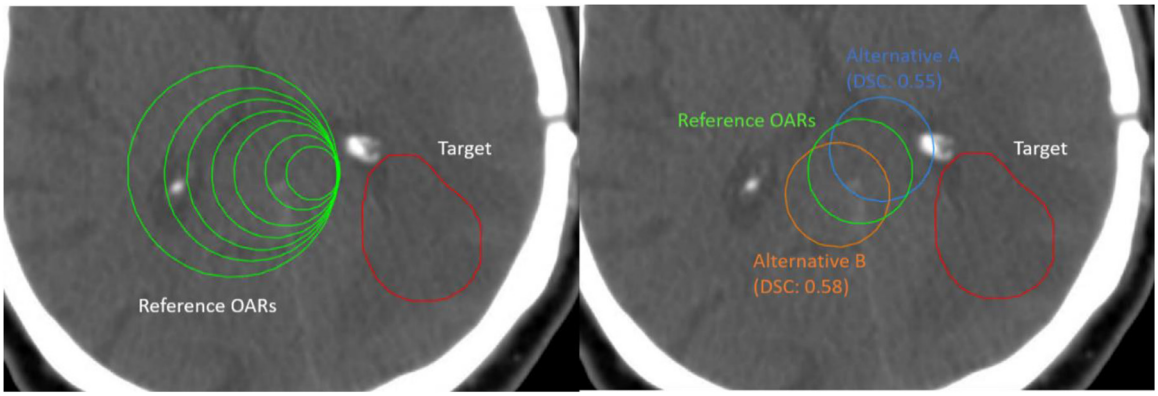

Fig. 4. Synthetic experiment to assess the influence of the size of the OAR on the dose effect with respect to the DSC metric. On the left we see the 7 reference OARs with different sizes (in green). On the right, an example of a reference OAR is shown accompanied with the respective alternative contours in blue and orange. The target is shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

to the reference OAR of respectively 0.55 and 0.58 (see Fig. 4). A reference plan was constructed on each of the reference OARs. The goal of the plan is the same as in Section 2.2.1. This reference plan was duplicated and re-optimized for the alternative OAR contours. The dose difference for the reference OAR between reference and alternative plan was determined for each of the 7 sizes.

2.2.3. Dice versus location

To determine the effect a specific location might have on the dose-effect, another synthetic experiment was designed. A spherical target and a single OAR reference contour, as well as an alternative OAR were defined on the planning CT of an actual subject. The alternative OAR had a DSC of 0.46 with respect to the reference OAR. The two OARs were duplicated to different locations with respect to the target, while keeping the same distance from the target (Fig. 5). The locations are posterior, medial, lateral and superior of the target. A reference plan was constructed for each of the reference OARs. This plan was duplicated and re-optimized on the alternative OAR contours, similar as in Section 2.2.1. The dose difference for the reference OAR between reference and alternative plan was determined for each location.

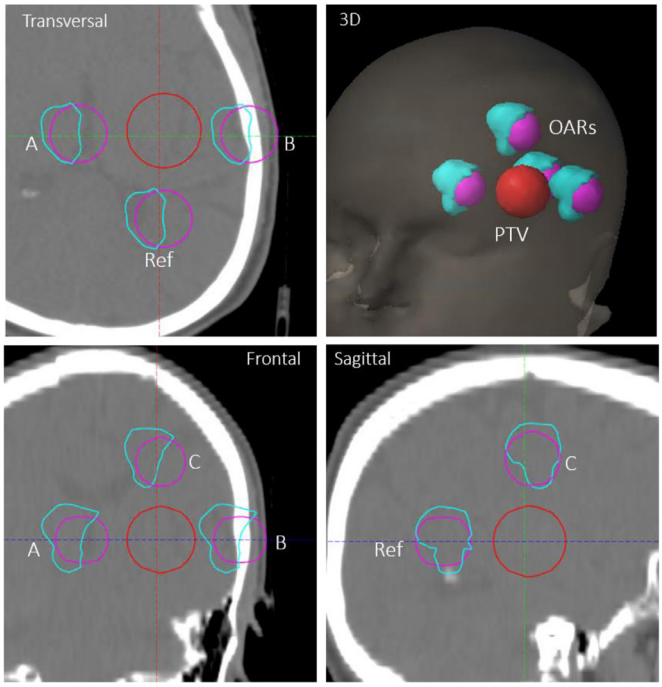

Fig. 5. Assessing the relationship of location relative to the target, on the dice-dose effect. Transversal, frontal sagittal and a 3D view of a human subject's head are depicted. The red circle represents the PTV. The pink circle is the reference OAR and the blue structure is the alternative OAR structure. The pair of OARs is duplicated in locations A, B and C. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

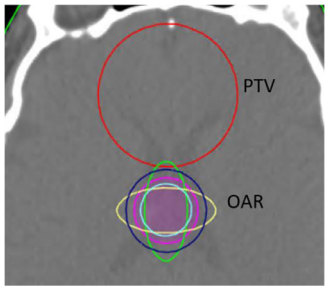

Fig. 6. Representation of the dice versus shape synthetic experiment. The red circle represents the PTV. The pink circle represents the reference OAR. At the same location, 4 alternative OARs with similar DSC to the reference were constructed with different shapes and size then the reference OAR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

2.2.4. Dice versus shape

The fourth synthetic experiment consisted of a spherical target and one reference OAR. Four alternative OARs were constructed with different shape or size, but with the same arbitrary DSC of 0.66 with respect to the reference OAR (Fig. 6). A reference plan was calculated and optimized based on the target and the reference OAR. This plan was duplicated and re-optimized while substituting the reference OAR for any of the alternative OARs. The difference in dose to the reference OAR and target, between the alternative and the reference plan, are compared to assess the correlation of different shapes of alternative OAR contours on the dose distribution.

3. Results

3.1. Segmentation dose-effect correlation on clinical cases

In total, we have constructed 60 alternative structure sets for the 12 reference cases. Including the 12 reference plans we calculated 72 plans, and created 480 single pairs of DSC and their corresponding Δ mean dose and Δ Dmax 1%. Depending on the method the alternative contours were produced, results are presented in three categories: (i) human-rater variability, (ii) explicit manual manipulation and (iii) auto-segmentation results. Additionally, an organ specific analysis is performed.

3.1.1. Human rater variability

For the human rater variability, the median DSC was 0.83 (interquartile range [IQR]: 0.13) The median Δ mean dose to the reference OAR was 0.25 (IQR: 0.80) Gy, and the median Δ Dmax 1% was 0.4 (IQR: 1.1) Gy. The Pearson's correlation coefficient for the mean dose difference and the maximum dose difference with the DSC was −0.11 and −0.08, respectively. The Δ mean dose, and the Δ Dmax 1%, are plotted against their corresponding DSC in Fig. 7A and D.

3.1.2. Explicit manual manipulations

The manual manipulations resulted in a median DSC of 0.68 (IQR: 0.22). The median Δ mean dose was 0.30 (IQR: 0.8) Gy and the median Δ Dmax 1% was 0.50 (IQR: 1.22) Gy. The Pearson's correlation coefficient for the Δ mean dose and the Δ Dmax 1% with the DSC was −0.17 and −0.13, respectively. The Δ mean dose, and the Δ Dmax 1%, are plotted against their corresponding DSC, and shown in Fig. 7B and E.

3.1.3. Auto-segmentation results

The auto-segmentation results had a median DSC of 0.70 (IQR: 0.33) with respect to the reference contours. The median Δ mean dose was 0.40 (IQR: 1.6) Gy and the median Δ Dmax 1% was 0.75 (IQR: 1.95) Gy. The Pearson's correlation coefficient for the Δ mean dose and the Δ Dmax 1% with the DSC, was −0.31 and −0.13 respectively. The Δ mean dose and the Δ Dmax 1% are plotted against their corresponding DSC, and shown in Fig. 7C and F.

3.1.4. Segmentation dose-effect per OAR type

The segmentation results differ slightly over the different OARs. The results are summarized in Table 2 and displayed as scatterplots in Fig. 8. The similarity for the chiasm and optic nerves were relatively low with a median DSC of 0.67 and 0.66 respectively. The brainstem and the eyes showed relatively better similarity with a median DSC of 0.85 and 0.84 respectively (Table 2). The dose effects among the different OARs did not show much difference. The highest observed median Δ mean dose was 0.70 Gy for the optic chiasm and for the Δ Dmax 1% dose 0.75 for the Hippocampi. The Pearson correlation coefficient is very low for the smaller OARs as the optic nerves and optic chiasm. However, it can be a lot higher for larger OARs as the brainstem and the eyes. Interestingly the Pearson correlation for the brainstem is very low for the delta mean dose, but relatively high for the delta max dose (Table 2).

3.1.5. Correlation of all alternative contours combined

The correlation of the DSC and the dose effect of the three categories combined, as well as for 22 additional segmentation parameters can be found in Table 3.

3.2. Possible characteristics predictive for the dose effect

3.2.1. Dice versus distance

A total of 40 plans were calculated, on the eight alternative OARs, at five different distances from the target (Fig. 3). The mean

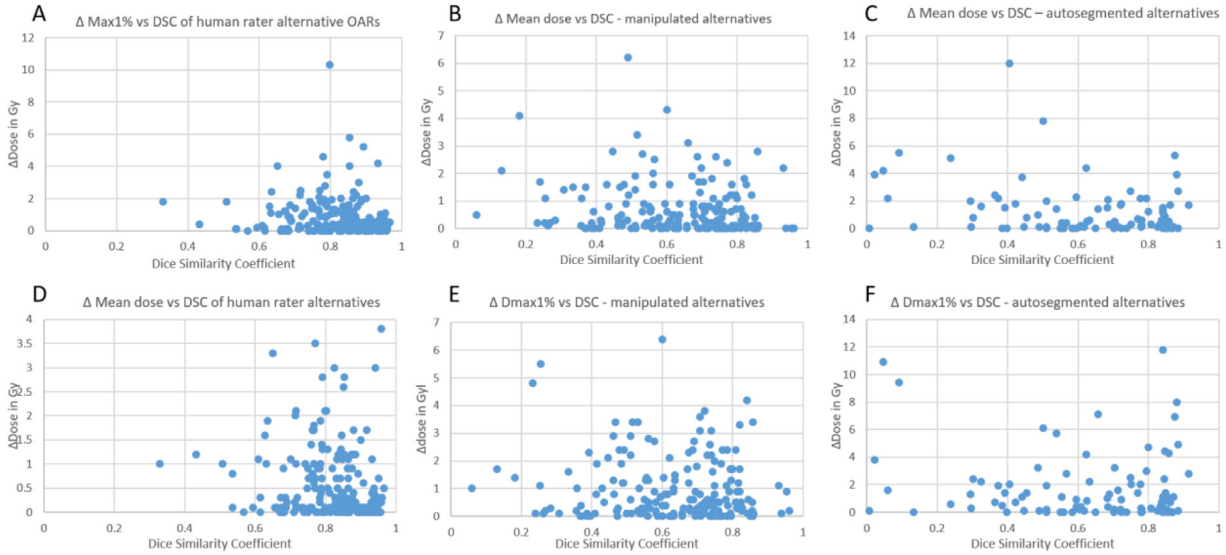

Fig. 7. Scatter plots of the DSC versus the dose effect. The dose effects of the three different natures of alternative contours are plotted against their respective DSC. From left to right, the human-rater alternatives, the manually manipulated alternatives and the auto-segmented alternatives. The Δ mean dose results are located in the upper plots while the Δ Dmax 1% results are shown below.

Table 2
Results of the organ specific analysis of the correlation of the DSC and the dose effect (Median and IQR).

	Mean volume (cc)	DSC	Delta mean dose (Gy)	Pearson correlation - Δ Mean dose and DSC	Delta max dose (Gy)	Pearson correlation - Δ Max dose and DSC
Brainstem	26.5	0.849 (0.097)	0.2 (0.8)	−0.013	0.5 (1.3)	−0.387
Eyes	8.38	0.843 (0.145)	0.2 (0.6)	−0.396	0.3 (0.7)	0.312
Optic Chiasm	0.24	0.674 (0.207)	0.7 (1.5)	−0.04	0.4 (1.8)	−0.072
Hippocampi	1.85	0.733 (0.268)	0.3 (0.8)	−0.289	0.75 (1.3)	−0.147
Optic Nerves	0.36	0.659 (0.245)	0.4 (1.0)	−0.063	0.6 (1.5)	−0.006

Table 3
Pearson correlation coefficients for additional segmentation parameters. The used metrics are a collection of segmentation metrics composed by the VISCERAL evaluation software (www-visceral.eu) complemented by some new popular metrics as the average distance and NSD (Nikolov, 2018).

	Correlation coefficient with:	Δ Mean dose	Δ Maximum dose
Similarity measures	Dice	−0.112	−0.137
	Jaccard	−0.134	−0.152
	Area under curve	−0.117	−0.140
	Cohen kappa	−0.134	−0.164
	Rand index	0.015	−0.102
	Adjusted rand index	−0.134	−0.164
	Interclass correlation	−0.134	−0.164
	Volumetric Similarity Coefficient	−0.055	−0.035
	Mutual information	−0.102	0.054
	Normalized Surface Dice	0.075	0.010
Distance measures	Hausdorff distance	0.186	0.184
	Average HDD	0.160	0.175
	Average Distance	−0.011	0.080
	Mahanbolis Distance	0.083	0.168
	Variation of info	−0.031	0.091
	Global consistency error	−0.023	0.097
	Probabilistic distance	0.103	0.202
Classic Measures	Sensitivity	−0.117	−0.140
	Specificity	0.071	−0.027
	Precision	−0.108	−0.141
	F-Measure	−0.134	−0.164
	Accuracy	0.015	−0.102
	Fallout	−0.071	0.027

dose and Dmax 1% received by the reference OARs are plotted against the DSC, for each distance, in Fig. 9. From Fig. 9A and C we observed that the dose effect to the OAR, does not seem to be directly influenced by the distance between target and OAR. The dose versus dice plots do not seem to lead to more variation as the distance to the target is decreased. The absolute dose differences (Fig. 9B and D), show that proximity to the target does not necessarily lead to a larger dose effect. Where we expect to see increasing dose effects with decreasing distance to the target, we actually see that specific alternative contours, characterized by their DSC on the x-axis, show a lot of dose variation (indicated by the asterisks in Fig. 9B). On the other hand, the other alternative con-

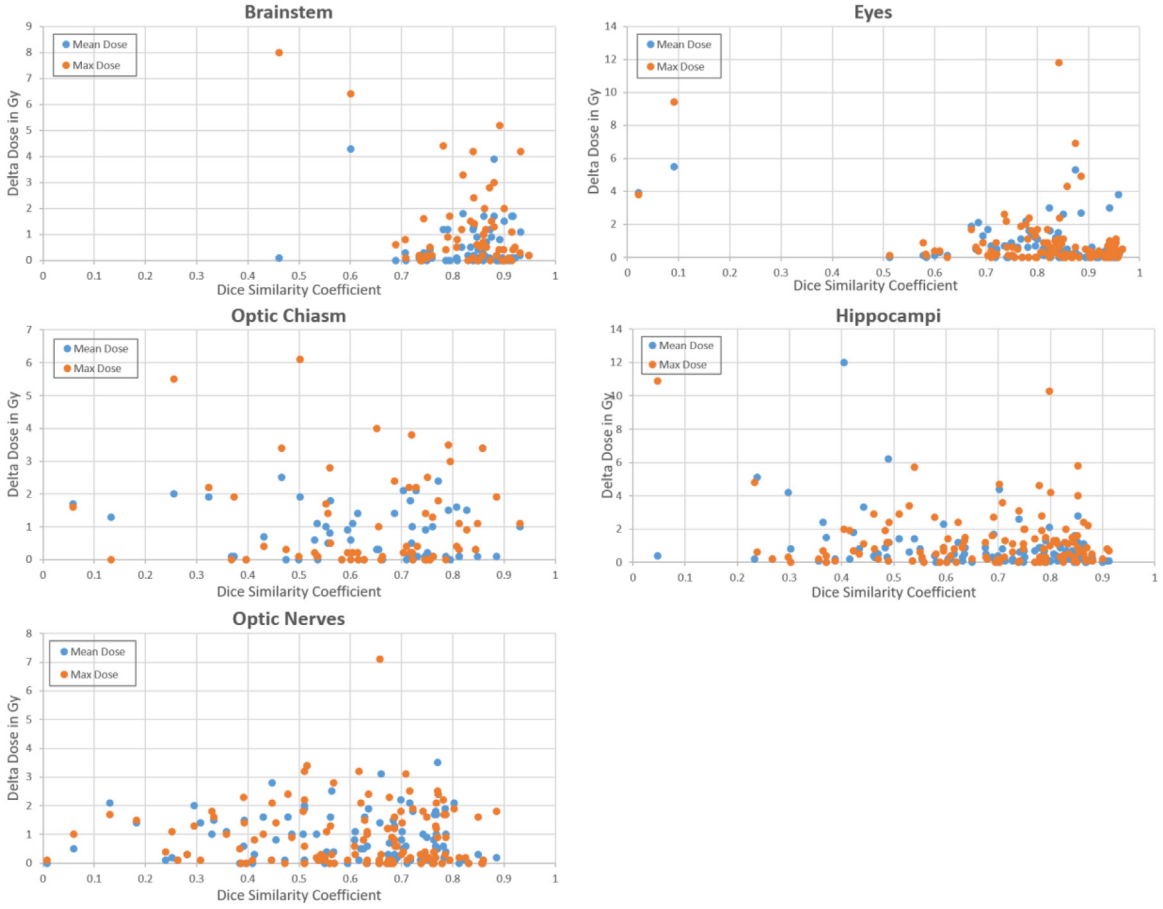

Fig. 8. Scatter plots of the DSC versus the dose effect for the 5 different organ types. The mean dose (blue dots) and the max dose (orange dots) effects, in Gy, are plotted against their respective DSC on the x-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

tours show almost no dose variation at all, regardless of the distance to the target.

3.2.2. Dice versus size

The different reference OAR sizes and their alternative contours with DSC values of 0.55 and 0.58 respectively, show different results in the mean dose effect and the Dmax 1% dose effect. It is observed that when the size of the OAR increases, the maximum dose increases and the mean dose decreases. This is a logical consequence, since a larger OAR results in less room for the dose to avoid the OAR near the target and simultaneously the volume receiving less dose is increasing due to the increased size of the OAR. The dose-effect seems to follow a different trend, which increases with increasing size of OAR but seems to stabilize and slowly decreases as a specific size is reached (Fig. 10).

3.2.3. Dice versus location

The mean dose and the Dmax 1% to the reference OARs, are shown in Fig. 11A and B. The difference in dose due to the planning on the alternative OAR, differs with the respective location to the target. This data shows that one single specific contour deviation of an OAR can lead to both an increase in dose, a decrease in dose, or to no change in dose at all, depending on its relative location to the target. The same effect can also be seen for the coverage of the PTV (Fig. 11C and D).

3.2.4. Dice versus shape

The five plans, optimized for the 5 different OARs, have been analyzed. In Fig. 12, the mean dose and Dmax 1% to the reference OAR are depicted for each of the plans. Despite having the same DSC with respect to the reference OAR, the mean dose to the reference OAR can vary up to 7.7 Gy between different alternatives. The largest difference in Dmax 1% among the alternative plans was 11.2 Gy. The target coverage is stable among all plans (Fig. 12, center).

4. Discussion

This study shows the correlation between current segmentation parameters and dosimetric effects in a selection of GBM cases treated with VMAT RT. It was our hypothesis that geometric similarity might not be a good method to validate, or qualify contours, for the purpose of radiotherapy. The same question has previously been investigated by other authors, but with a slightly different motivation. Gooding et al. used an adapted Turing test for the clinical validation of auto-segmented contours (Gooding et al., 2018). This approach was motivated by the benchmark trap, which is created by comparing results to a ground truth that does not actually exist. Vaassen et al. also proposed a different contouring validation scheme by claiming that correction time is clinically more important than geometrical similarity (Vaassen et al., 2020). This resulted in a new parameter that is better able to predict the amount of manual adjustment time. Although manual adjustment time is clinically relevant, it assumes that all contours require correction. However, our data suggest that many OAR contours do not need correction at all.

In this study, we looked at contour validation through a more clinical end goal perspective of radiotherapy. Hence, we looked at

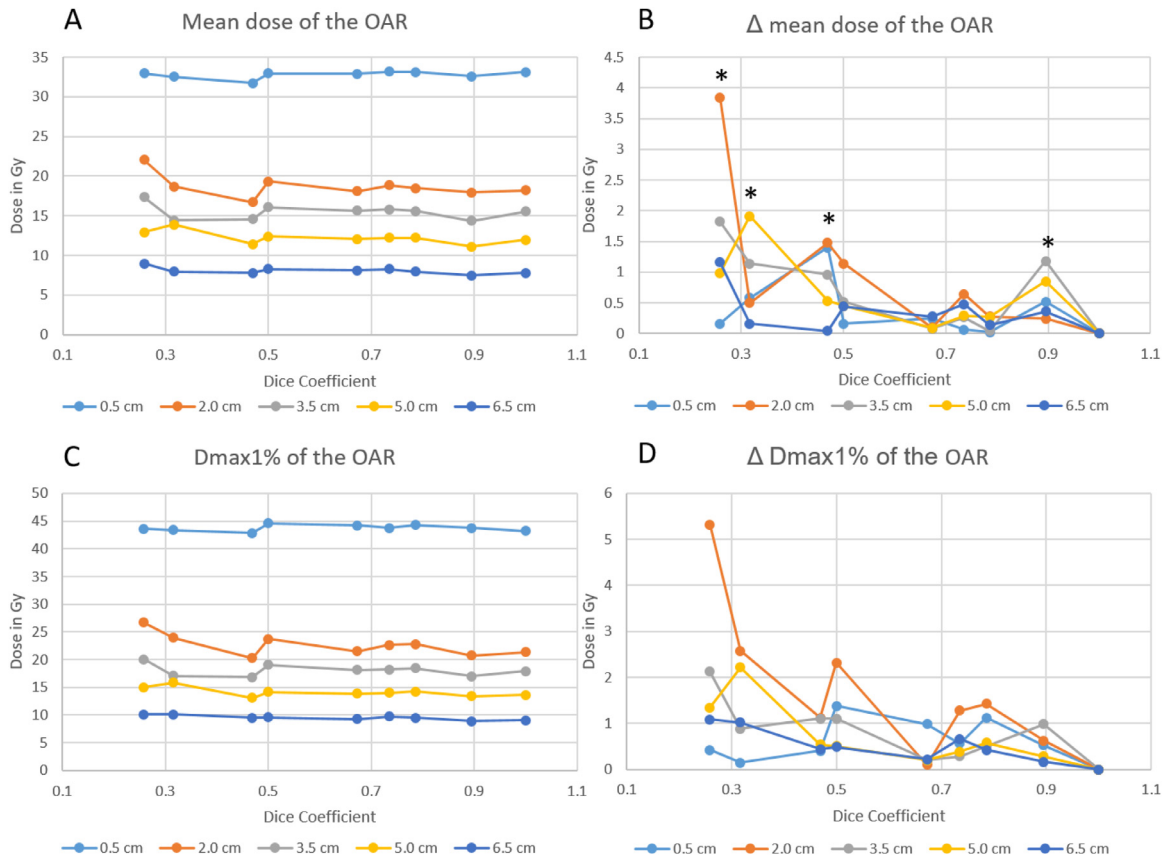

Fig. 9. Represented are the doses to the reference OARs at different distances as shown in Fig. 3. The dots represent the plan based on the specific alternative OAR with the corresponding DSC on the x-axis. The upper plots show the results for the mean dose to the OAR (A) and the absolute difference with respect to the reference plan (B). The lower plots show the Dmax 1% of the OAR (C) and the absolute difference in Dmax 1% (D). The asterisks in B indicate cases that show a lot of dose effect variation among the different distances.

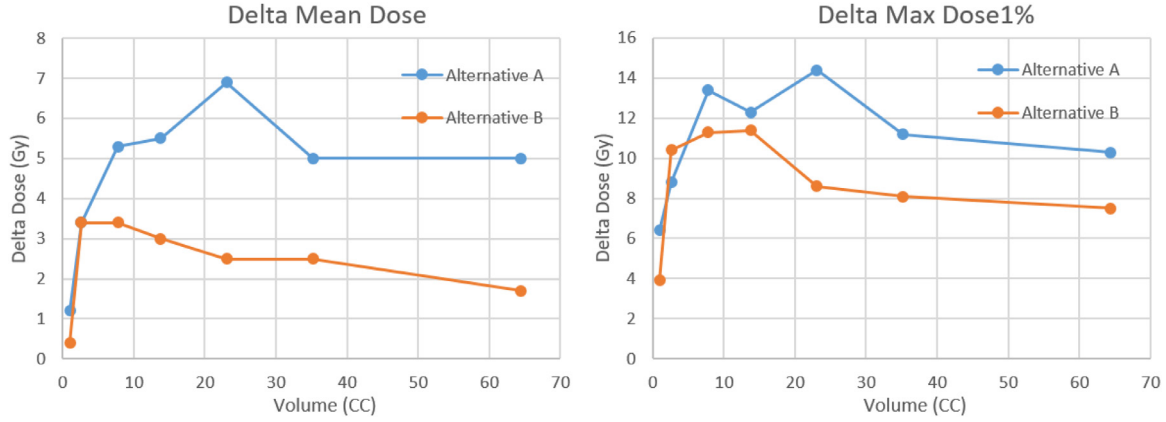

Fig. 10. The influence on the size of an OAR on the dosimetric effect for a two fixed alternative contours with a respective DSC of 0.55 (alternative A) and 0.58 (alternative B). The dose difference of the reference plan and the plan optimized on the specific alternative is plotted against the size of the volume in cubic centimeter (CC). The absolute dose difference is given in Gy.

the dosimetric effects, which are directly related to the treatment outcome. We asked ourselves the questions: How incorrect does a contour need to be for it to start influencing the dose? Are there parameters able to predict the dose effect? In these regards, we assessed the correlation between the geometrical similarity and the dose effect in OARs of the brain, and found a low correlation. Not only for the DSC metric but also for other well-known segmentation assessment metrics as well as recently introduced improved metrics. As expected, some amount of correlation was found. However, if the geometrical similarity gets worse, we found a low correlation to a certain dose effect. In conclusion, the predictive value of current segmentation parameters for corresponding dose effect is inadequate for segmentation tasks in brain radiation therapy planning. It cannot be determined if a specific contour would be clinically unacceptable based on the analyzed segmentation metrics.

These results are different than the conclusions made by Xian et al. (Xian and Chen, 2020) who also looked at the correlation of contour variation and dose. We see some significant differences in the experimental design of our study and theirs. They specifi-

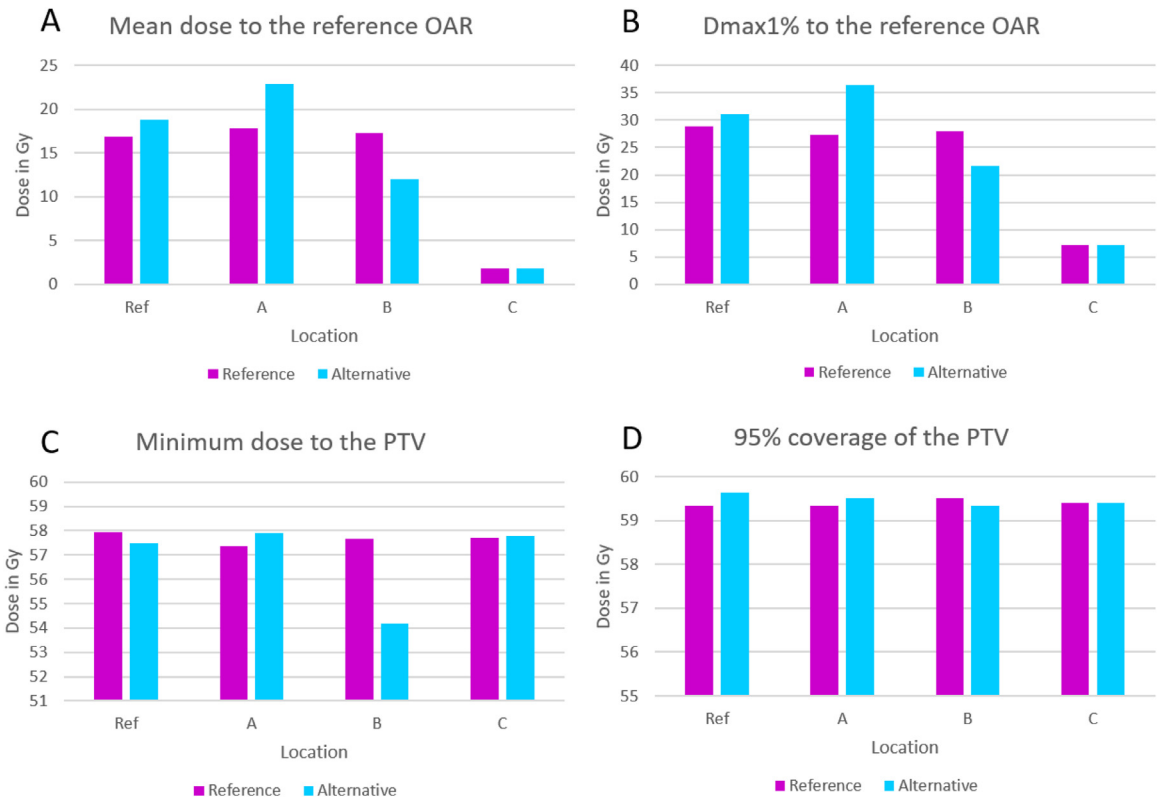

Fig. 11. Influence of location on the dose effect to a specific alteration in OAR contour. Bar graphs represent the dose effect to the OAR (A, B) and the PTV (C, D) for the reference plans (pink) and the alternative plans (blue), for the 4 different locations shown in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

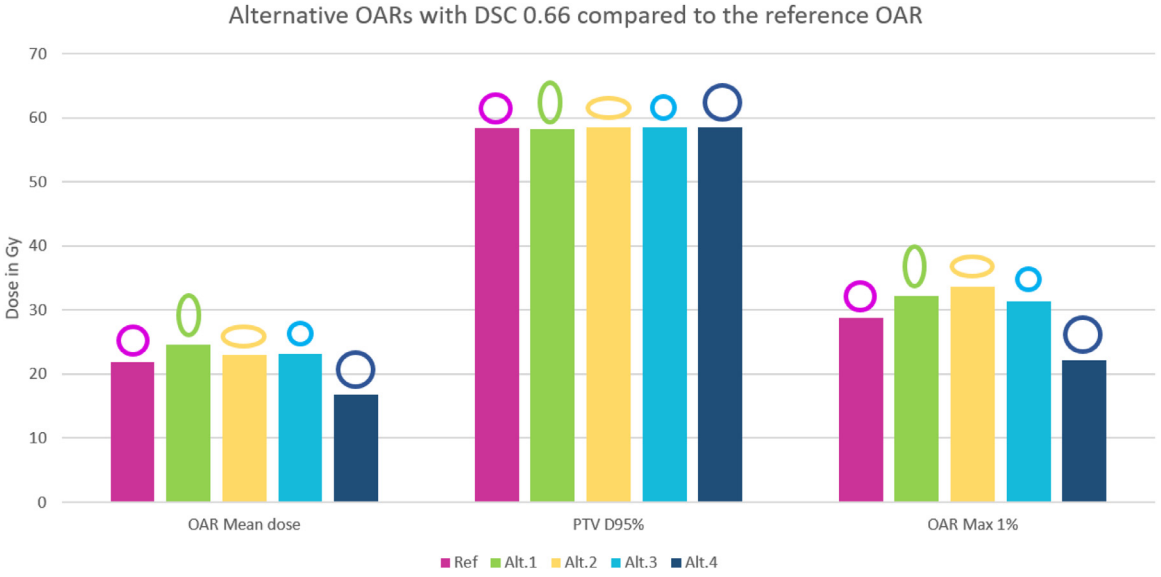

Fig. 12. Influence of shape and volume of an alternative OAR contour on the dose effect. The bar plots represent the dose received by the reference OAR (pink, in Fig. 5) for the reference plan and the 4 alternative plans. The colors and the shape above the bars correlate with the shapes and colors from Fig. 5. The middle bar graph represents the dose coverage (D95%) to the PTV.

cally looked at the correlation of geometrical similarity and dosimetric indices to target structures, while we focused on OARs. Even though they concluded that geometrical similarity is not sufficient for clinical contour validation, they showed strong correlations in their results. A good reason for this could be the fact that they analyzed targets. The target is a structure where all the dose is pointed towards and has a steep dose fall-off. Consequently, systematically transforming this target contour over the existing dose distribution will lead to a correlating effect.

This method was also used by Beasley et al. when they looked at the correlation of DSC between "ground truth" and auto-segmented parotid and larynx contours (Beasley et al., 2016). They did not re-optimize the plans based on the alternative contour but rather overlaid both "ground truth" and alternative contour on the

existing dose distribution to determine the dose effect. In line with our results, they did find a weak correlation between segmentation metrics and dose effect that was OAR dependent. However, it should be mentioned that they only had 10 data pairs per OAR to determine this correlation.

In this study, we are looking at large number of alternative OAR contours in the brain, where the target area varies in location. Many more factors are involved in determining the dose effect to a specific OAR, Increasing the complexity of predicting which factors will have an effect and to what extent. We investigated a few characteristics that have an influence on the dose effect to OARs in the brain. Notably, we found that the distance from a target does not directly influence the dose effect; however, change in a specific direction does seem to have a more prominent effect. In addition, the relative location of the OAR and its contour variation with respect to the target could be of large influence. Furthermore, we noticed that the size of a specific OAR could have an influence on the dose effect and shows that the dosimetric parameters will depend on the size as well. I.e. the difference in mean and max dose is substantially small for small OARs but can be large for larger OARs. Nevertheless, size is a disturbing factor in volume based similarity metrics. Variability of these metrics can differ largely among different types of OARs. Fig. 8 shows that this variability is also present in our test data. The variability of smaller longitudinal structures as the optic nerves and chiasm is larger as that of larger more spherical structures as the eyes and the brainstem.

Although we cannot attribute any conclusions due to the synthetic nature of the experiments, it does show the complexity of how dose is affected in OARs. Moreover, it presents how many parameters and factors are involved determining the final dose distribution. Since many critical OARs are in close proximity within the brain, they can also influence the dose to each other (i.e. a change in contour to OAR A could lead to a dose effect in OAR B). Which is something we did not account for in the current study and necessitates more investigation.

As mentioned, the dose effect is also dependent on how the dose is delivered and how the optimization is performed. In this study we worked with a clinical protocol delivering a co-planar VMAT technique using a dose prescription where the constraint doses to the OARs where prioritized. Different delivery techniques and different optimization approaches will therefore lead to different dosimetric outcomes. Although we used stratification to have a diverse distribution of cases, it has to be mentioned that dosimetry is very case-specific and many specific situations are not covered by our study data. The results from this study are therefore only valid for this particular type of RT delivery to the brain region. On the other hand, the general rationale and experimental setup to investigate whether geometrical similarity metrics are not a good predictor of RT quality, could be valid for different types of RT in different regions, and is worth investigating.

As additional follow up, we believe it is important to find characteristics that do reflect treatment quality. In other words, to find good predictor parameters of the dose effect. A parameter like this would be very helpful for clinical validation of contours that are derived from manual contouring or any type of auto-contouring, as long as there is a reference to compare against. Additionally, such a parameter would be very useful as part of the cost function for designing and optimizing deep learning based auto-segmentation methods (Ma et al., 2021). Kofler et al. proposed incorporating qualitative measures into the loss function of a tumor segmentation method (Kofler, 2021). This can lead to several improvements. First, validating a contouring system on a robust treatment quality is expected to improve the clinical implementation of such tools. Secondly, if one is able to determine that changes to dose effects are negligible despite geometrical differences, one can establish a more clinically oriented performance objective for an auto-segmentation method.

For clinical RT it could mean that we do not have to visually inspect and manually adjust all OAR contours. If we can predict that segmented outcomes do not have a dose effect, we can skip the inspection and correction part for these cases. Another scenario could be to predict which specific contours have an effect on the dose distribution. In this case, the visual inspection and manual correction step, which is often required when using auto-segmentation, could be made significantly more efficient.

Looking at the results from our data, we observed that a large number of alternative contours do not lead to a significant dose effect (Figs. 7 and 8). However, the question is if this is also clinically insignificant. This is not an easy question to answer. In general, any increase in dose to an OAR is undesired. However, due to the optimization process in RT, an increase in dose to a specific region often results in a decrease somewhere else. This can be beneficial if this region is a critical organ as well, however, it will be detrimental if it comes at the expense of the target coverage. Therefore, it is difficult to say that a certain increase to a specific organ is affecting the overall treatment quality. Furthermore, an absolute increase in dose is difficult to quantify. At what increase, either in absolute or relative numbers is a change significant. Should it be absolute dose or relative dose or should it be relative to its specific dose constraint? Besides, it is important which parameter, mean dose or Dmax 1%, is used. For instance, if one looks at the arbitrary threshold of 2.0 Gy absolute dose effect, from the 960 parameters analyzed in this study, 79 exceeded this threshold. Of these 79, in 46 the dose increased, while in the other 33 the dose decreased.

A solution for this problem might be found in normal tissue complication probability models (Yorke, 2001). Provided that valid models are available for the specific OARs in the brain, one is able to determine the trade-off between sub-optimal contours and the increase in chance of developing a specific complication.

Even though our data included a wide variety of geometrical similarity values, i.e. an average DSC of 0.71 ± 0.19, the dose effect to the large majority of cases showed to be limited. This information is indeed encouraging for exploring new approaches to improve and implement auto-segmentation methods.

In conclusion, currently used segmentation assessment parameters, which are mainly based on geometrical similarity, are not well correlated with dosimetric changes on OARs in the brain. Our results also show that in the brain the majority of imperfect contours, whether resulting from manual segmentation, auto-segmentation or deliberate manipulations, do not lead to clinically relevant dose changes. In order to find specific contour variations that do lead to dose changes, other characteristics, such as relative distance and orientation to the target and the shape and nature of the contour variation seem to have an influence. These results suggest that the current evaluation metric for medical image segmentation in radiation therapy, as well as the training of deep learning systems employing such metrics, need to be revisited towards clinically oriented metrics that more accurately reflect how segmentation quality affects dosimetry and related tumor control and toxicity.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stefan Scheib is a full time employee of Varian Medical Systems, Imaging Laboratory GmbH, Dättwil, Switzerland. The other Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Robert Poel: Conceptualization, Methodology, Validation, Formal analysis, Investigation. **Elias Rüfenacht:** Software. **Evelyn Hermann:** Data curation. **Stefan Scheib:** Resources. **Peter Manser:** Methodology, Supervision. **Daniel M. Aebersold:** Methodology, Supervision. **Mauricio Reyes:** Conceptualization, Methodology, Supervision.

Acknowledgement

This work was supported by Innosuisse grant number 31274.1 and the Swiss Cancer League.

References

Beasley, W.J., et al., 2016. The suitability of common metrics for assessing parotid and larynx autosegmentation accuracy. J. Appl. Clin. Med. Phys. 17 (2), 41–49. doi:10.1120/jacmp.v17i2.5889.

Bondiau, P.Y., et al., 2005. Atlas-based automatic segmentation of MR images: validation study on the brainstem in radiotherapy context. Int. J. Radiat. Oncol. Biol. Phys. 61 (1), 289–298. doi:10.1016/j.ijrobp.2004.08.055.

Brock, K.K., 2019. Adaptive radiotherapy : moving into the future. Semin. Radiat. Oncol. 29 (3), 181–184 doi: 10.1016/j.semradonc.2019.02.011.Adaptive.

Brunenberg, E.J.L., et al., 2020. External validation of deep learning-based contouring of head and neck organs at risk. Physics and Imaging in Radiation Oncology 15 (June), 8–15. doi:10.1016/j.phro.2020.06.006, Elsevier.

Cardenas, C.E., et al., 2019. Advances in auto-segmentation. Seminars in Radiation Oncology 29 (3), 185–197. doi:10.1016/j.semradonc.2019.02.001, Elsevier Inc..

Cloak, K., et al., 2019. Contour variation is a primary source of error when delivering post prostatectomy radiotherapy: results of the trans-Tasman radiation oncology group 08.03 radiotherapy adjuvant versus early salvage (RAVES) benchmarking exercise. J. Med. Imaging Radiat. Oncol. 63 (3), 390–398. doi:10.1111/1754-9485.12884.

Nikolov, S. et al. (2018). Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. ArXiv, pp. 1–31. Available at: http://arxiv.org/abs/1809.04430.

Ben-Cohen, A. et al. (2016). Fully Convolutional Network for Liver Segmentation and Lesions Detection, in Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol. 10008. Springer, Cham., pp. 77–85. doi: 10.1007/978-3-319-46976-8.

van Dijk, L.V., et al., 2020. Improving automatic delineation for head and neck organs at risk by deep learning contouring. Radiother. Oncol. 142, 115–123. doi:10.1016/j.radonc.2019.09.022, The Authors.

Emami, B. (2013). Tolerance of normal tissue to therapeutic radiation', 1(1), pp. 35–48. Available at: https://cdn.neoscriber.org/cdn/serve/eb/27/eb27adb334594d3093f4ed1b7d088c0a7a390f0b/4316-13810-1-PB.pdf.

Deeley, M.A., et al., 2011. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. Physics in Medicine & Biology 56 (14), 4557.

Gooding, M.J., et al., 2018. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. Med. Phys. 45 (11), 5105–5115. doi:10.1002/mp.13200.

Harari, P.M., Song, S., Tome, W.A., 2010. Treatment planning in head and neck cancer. Int. J. Radiat. Oncol. Biol. Phys. 77 (3), 950–958. doi:10.1016/j.ijrobp.2009.09.062.

Hu, P., et al., 2016. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. Phys. Med. Biol. 61 (24), 8676–8698. doi:10.1088/1361-6560/61/24/8676.

Isensee, F., et al., 2021. nnU-Net: self-adapting framework for U-Net-based medical image segmentation. Nat. Methods (2) 203–2011.

Jameson, M.G., et al., 2010. A review of methods of analysis in contouring studies for radiation oncology. J. Med. Imaging Radiat. Oncol. 54 (5), 401–410. doi:10.1111/j.1754-9485.2010.02192.x.

Jungo, A., et al., 2021. pymia: a Python package for data handling and evaluation in deep learning-based medical image analysis. Comput. Methods Prog. Biomed. 198, 105796. doi:10.1016/j.cmpb.2020.105796, Elsevier B.V.

Litjens, G., et al., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88. doi:10.1016/j.media.2017.07.005, Elsevier B.V.(December 2012).

Ma, J., et al., 2021. Loss odyssey in medical image segmentation. Med. Image Anal. 71. doi:10.1016/j.media.2021.102035.

Maier-Hein, L., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (5217), 1–13. doi:10.1038/s41467-018-07619-7.

Marks, L.B., et al., 2013. Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary. Practical Radiation Oncology 3 (3), 149–156. doi:10.1016/j.prro.2012.11.010, American Society for Radiation Oncology.

Mazzara, G.P., et al., 2004. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. Int. J. Radiat. Oncol. Biol. Phys. 59 (1), 300–312. doi:10.1016/j.ijrobp.2004.01.026.

Meyer, P., et al., 2018. Survey on deep learning for radiotherapy. Comput. Biol. Med. 98 (May), 126–146. doi:10.1016/j.compbiomed.2018.05.018, Elsevier Ltd.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, pp. 565–571. doi:10.1109/3DV.2016.79.

Mlynarski, P., et al., 2020. Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy. J. Med. Imaging 7 (1). doi:10.1117/1.JMI.7.1.014502.

Kofler, F. et al. (2021). Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient., Arxiv Preprint. *arXiv:2103.06205*.

Niyazi, M., et al., 2016. ESTRO-ACROP guideline target delineation of glioblastomas. Radiother. Oncol. 118 (1), 35–42. doi:10.1016/j.radonc.2015.12.003.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 9351, 234–241. doi:10.1007/978-3-319-24574-4_28, (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).

Roth, H.R., et al.Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.), 2015. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, 9349, 556–564. doi:10.1007/978-3-319-24553-9.

Sandström, H., et al., 2016. Assessment of organs-at-risk contouring practices in radiosurgery institutions around the world – the first initiative of the OAR Standardization Working Group. Radiother. Oncol. 121 (2), 180–186. doi:10.1016/j.radonc.2016.10.014.

Scoccianti, S., et al., 2015. Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice. Radiother. Oncol. 114 (2), 230–238. doi:10.1016/j.radonc.2015.01.016, Elsevier Ireland Ltd.

Stanley, J., et al., 2013. The effect of contouring variability on dosimetric parameters for brain metastases treated with stereotactic radiosurgery. Int. J. Radiat. Oncol. Biol. Phys. 87 (5), 924–931. doi:10.1016/j.ijrobp.2013.09.013, Elsevier Inc..

Vaassen, F., et al., 2020. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys. Imaging Radiat. Oncol. 13, 1–6. doi:10.1016/j.phro.2019.12.001, Elsevier(December 2019).

Vinod, S.K., et al., 2016. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. Radiother. Oncol. 121 (2), 169–179. doi:10.1016/j.radonc.2016.09.009, Elsevier Ireland Ltd.

Visser, M., et al., 2019. Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage 22, 101727. doi:10.1016/j.nicl.2019.101727, Elsevier(July 2018).

Voet, P.W.J., et al., 2011. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother. Oncol. 98 (3), 373–377. doi:10.1016/J.RADONC.2010.11.017, Elsevier.

Yorke, E.D., 2001. Modeling the Effects of Inhomogeneous Dose Distributions in Normal Tissues. Seminars in Radiation Oncology 11 (3), 197–209.

Xian, L. and Chen, L. (2020). Clinically oriented contour evaluation using geometric and dosimetric indices based on simple geometric transformations. Research Square; 2020. DOI: 10.21203/rs.3.rs-19265/v3.

Zhou, X. et al. (2016) . Three-Dimensional CT Image Segmentation by Combining 2D Fully Convolutional Network with 3D Majority Voting. In Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol. 10008. Springer, Cham., pp. 111–120. doi: 10.1007/978-3-319-46976-8.