Deep learning-based delineation of head and neck organs-at-risk: geometric and dosimetric evaluation

Ward van Rooij  
w.vanrooij@vumc.nl
Max Dahele  
m.dahele@vumc.nl
Hugo Ribeiro Brandao  
h.ribeirobrandao@vumc.nl
Alex R. Delaney  
a.delaney@vumc.nl
Berend J. Slotman  
bj.slotman@vumc.nl
Wilko F. Verbakel  
w.verbakel@vumc.nl

Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Radiation Oncology, Cancer Center Amsterdam, De Boelelaan 1117, Amsterdam

Editors: Under Review for MIDL 2019

Keywords: Deep Learning, Segmentation, Radiotherapy

1. Introduction

Delineation of target volumes and organs-at-risk (OARs) in scan data is a key step in radiotherapy treatment planning in order to assure sufficient dose to the tumor while sparing healthy tissue. However, it is often time consuming and therefore resource intensive. It is also subject to interobserver variability and may require considerable anatomical knowledge for which there is a learning curve (Brouwer et al., 2012, 2015). Such limitations have driven attempts at automation. Because traditional automated methods do not perform well when there is low gray-scale contrast, deep learning (DL)-based techniques merit investigation. While early results from DL-based delineation (DLD) have generally been promising, the number of OARs in these studies has often been limited and evaluation was typically only presented in terms of volumetric comparative parameters (e.g. Sørensen-Dice similarity coefficient (SDC)) or by multi-observer rating of the contours (Fritscher et al., 2016; Ibragimov et al., 2017; Mocnik et al., 2018; Nikolov et al., 2018; Ren et al., 2018; Tong et al., 2018; Zhu et al., 2018). However, this gives little information about the clinical impact of different approaches to delineation. We investigated DLD, using head-and-neck cancer (HNC) as the paradigm because of the large number of OARs, with low gray-scale contrast, and different OAR image characteristics, located close to the planning target volumes (PTV). In addition to the SDC, we also looked at the impact of DLD on treatment plan dosimetry. In doing so, we can investigate whether we need to keep striving for perfect model performance or that deep learning performance is already at a level at which it can be used in practice.

2. Methods

The dataset consisted of 3D CT images and structure sets of 157 patients who had previously been treated for HNC. 142 were used to train a convolutional neural network (CNN) and 15 as the test set. The experimental procedure was as follows. A DL-model was created...
Dosimetry analysis of deep learning delineation for each OAR by training on the manually delineated (MD) structures. The resulting DL-models were used to predict the location of OARs in a new set of patients. The objective was to generate DL-based contours of the following OARs: left and right submandibular gland (SMG), left and right parotid gland, larynx, cricopharynx, pharyngeal constrictor muscle (PCM), upper esophageal sphincter (UES), brain stem, oral cavity and esophagus. Next, two treatment plans were created per patient: one based on the deep learning delineations (DLD) and one based on the MD. Finally, the resulting dose distributions were mapped on the MD, to see whether using DLD for treatment planning would have an effect on dose received by manually contoured OARs.

<table>
<thead>
<tr>
<th>Structure</th>
<th>SDC ± SE</th>
<th>Sensitivity ± SE</th>
<th>PFP ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMG (r)</td>
<td>.81 ± .13</td>
<td>.79 ± .16</td>
<td>.16 ± .09</td>
</tr>
<tr>
<td>SMG (l)</td>
<td>.82 ± .07</td>
<td>.80 ± .11</td>
<td>.14 ± .08</td>
</tr>
<tr>
<td>Parotid (r)</td>
<td>.83 ± .02</td>
<td>.85 ± .07</td>
<td>.18 ± .06</td>
</tr>
<tr>
<td>Parotid (l)</td>
<td>.83 ± .03</td>
<td>.81 ± .06</td>
<td>.14 ± .07</td>
</tr>
<tr>
<td>Larynx</td>
<td>.78 ± .05</td>
<td>.83 ± .08</td>
<td>.25 ± .10</td>
</tr>
<tr>
<td>PCM</td>
<td>.68 ± .09</td>
<td>.66 ± .09</td>
<td>.29 ± .09</td>
</tr>
<tr>
<td>Cricopharynx</td>
<td>.73 ± .11</td>
<td>.70 ± .11</td>
<td>.20 ± .16</td>
</tr>
<tr>
<td>UES</td>
<td>.81 ± .14</td>
<td>.80 ± .16</td>
<td>.15 ± .14</td>
</tr>
<tr>
<td>Brain stem</td>
<td>.64 ± .16</td>
<td>.86 ± .14</td>
<td>.42 ± .23</td>
</tr>
<tr>
<td>Esophagus</td>
<td>.60 ± .11</td>
<td>.50 ± .15</td>
<td>.21 ± .14</td>
</tr>
<tr>
<td>Oral cavity</td>
<td>.78 ± .07</td>
<td>.68 ± .11</td>
<td>.05 ± .03</td>
</tr>
</tbody>
</table>

Figure 1: Performance of the models. \( SDC = \frac{2tp}{2tp + fp + fn} \), sensitivity = \( \frac{tp}{tp + fn} \) and \( PFP = \frac{fp}{tp + fp} \), where \( tp \) is number of true positive voxels, \( fp \) is number of false positive voxels and \( fn \) is number of false negative voxels.

### 3. Results

Model performance can be seen in figure 1. Lower model performance was generally the result of small amounts of training data and/or (highly) variable contouring in the train set. Figure 2 shows scatterplots of dose with DLD plans versus dose with MD plans for each OAR. The average dose was significantly higher for DLD-plans for the inferior PCM (36.5±13.7Gy vs. 35.2±13.5Gy; \( t(14)=2.95 \), \( p=0.005 \)) and the esophagus (22.7±5.6Gy vs. 20.4±6.1Gy; \( t(8)=4.08 \), \( p=0.002 \)). Although statistically significant, these mean dose increases were not considered clinically significant, because increases of that magnitude have little to no toxic impact on the patient.

### 4. Conclusions

We have shown that, even though our dataset was limited in size and subject to limited curation, automated DLD performs well enough to form the basis for clinically acceptable
Figure 2: Scatterplots of the dose (in Gy) with DLD plans (DLD-D) versus dose with MD plans (MD-D) for each OAR.

plans made using a model-based automated planning solution. Despite the variation in model performance, there was little effect on OAR dose, indicating that (1) acceptable plans can be generated with DL-based contours and (2) imperfect SDC scores do not necessarily result in inferior OAR dosimetry. We have therefore shown that in principle, automated OAR segmentation combined with automated planning is feasible. This has implications for efficiency gains in routine clinical care, for facilitating adaptive radiotherapy, and for automated contour quality assurance. However, we saw that in individual patients/OARs some contours may be sub-standard. If the OAR is a critical structure (e.g., one where excessive dose to part of the OAR missed by the contour may result in serious toxicity), then that OAR should be manually checked. For non-critical OARs, checking may be less important. Nonetheless, with the help of DL, largely automated clinical workflows for radiotherapy are within reach. Even for complex scenarios such as HNC.
References

CL Brouwer et al. 3d variation in delineation of head and neck organs at risk. 2012.


