Automatic contouring of normal tissues with deep learning for preclinical radiation studies

Georgios Lappas1,4, Cecile J A Wolfs1,4, Nick Staut1,2, Natasja G Lieuwes1, Rianne Biemans3, Stefan J van Hoof6, Ludwig J Dubois4,5 and Frank Verhaegen1,2,5

1 Department of Radiation Oncology (Maastro), GROW—School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
2 SmART Scientific Solutions BV, Maastricht, The Netherlands
3 The M-Lab, Department of Precision Medicine, GROW—School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands
4 Indicates equal first authorship.
5 Indicates equal senior authorship.

E-mail: frank.verhaegen@maastro.nl

Keywords: deep learning, artificial intelligence, autocontouring, normal tissue, preclinical

Supplementary material for this article is available online

Abstract

Objective. Delineation of relevant normal tissues is a bottleneck in image-guided precision radiotherapy workflows for small animals. A deep learning (DL) model for automatic contouring using standardized 3D micro cone-beam CT (μCBCT) volumes as input is proposed, to provide a fully automatic, generalizable method for normal tissue contouring in preclinical studies. Approach. A 3D U-net was trained to contour organs in the head (whole brain, left/right brain hemisphere, left/right eye) and thorax (complete lungs, left/right lung, heart, spinal cord, thorax bone) regions. As an important preprocessing step, Hounsfield units (HUs) were converted to mass density (MD) values, to remove the energy dependency of the μCBCT scanner and improve generalizability of the DL model. Model performance was evaluated quantitatively by Dice similarity coefficient (DSC), mean surface distance (MSD), 95th percentile Hausdorff distance (HD95p), and center of mass displacement (ΔCoM). For qualitative assessment, DL-generated contours (for 40 and 80 kV images) were scored (0: unacceptable, manual re-contouring needed - 5: no adjustments needed). An uncertainty analysis using Monte Carlo dropout uncertainty was performed for delineation of the heart. Main results. The proposed DL model and accompanying preprocessing method provide high quality contours, with in general median DSC > 0.85, MSD < 0.25 mm, HD95p < 1 mm and ΔCoM < 0.5 mm. The qualitative assessment showed only few contours needed manual adaptations (40 kV: 20/155 contours, 80 kV: 3/155 contours). The uncertainty of the DL model is small (within 2%). Significance. A DL-based model dedicated to preclinical studies has been developed for multi-organ segmentation in two body sites. For the first time, a method independent of image acquisition parameters has been quantitatively evaluated, resulting in sub-millimeter performance, while qualitative assessment demonstrated the high quality of the DL-generated contours. The uncertainty analysis additionally showed that inherent model variability is low.

1. Introduction

Preclinical radiation studies are of great benefit for clinical radiation therapy (RT), as they allow for large scale evaluation of radiation effects on a shorter time scale compared to clinical studies (Schlaak et al 2020). Over the years, image-guided precision radiotherapy for preclinical rodent models has become possible as a result of developments in both hardware, e.g. image-guided radiation platforms, and software, e.g. advanced treatment planning systems (Verhaegen et al 2011, van Hoof et al 2013, Tillner et al 2014, Verhaegen et al 2018).
The preclinical image-guided precision radiotherapy workflow closely mimics the clinical radiotherapy workflow (Schlaak et al. 2020). It starts with the animal set-up, followed by imaging, for which in most cases high precision X-ray micro cone-beam CT (μCBCT) is used. The next step is treatment planning, which consists of a repetitive process including tissue contouring, beam configuration planning and dose calculation. The aim is to achieve an optimal dose to be used during subsequent irradiation. In preclinical radiotherapy, these various steps are usually performed in one session while the animal is under anesthesia. Lately, there is a need to align with the Refine, Reduce, Replace (3-R) principle for animal experimentation, with the goal of minimizing animal burden.

A bottleneck in this image-guided precision radiotherapy workflow for small animals is the delineation of the relevant organs. Delineation of these structures can be useful for various applications. For instance, for longitudinal assessment of organ response to the received radiation dose. In case of tumor irradiation, organ-at-risk (OAR) delineation is needed to evaluate and minimize the radiation dose to normal tissues, while ensuring optimal target coverage. However, manual contouring is a time-consuming and laborious task requiring expertise from animal technicians, which limits workflow efficiency. It involves precise delineation of relevant tissue volumes using 2D polygons adjacent to the borders of these tissues, until the whole tissue is covered, introducing variances dependent on several factors. Furthermore, inter-observer variability can introduce bias into the treatment plan because of anatomical complexities and lack of guidelines for murine organ delineations (Schoppe et al. 2020, Lappas et al. 2022).

With automatic segmentation methods, the time needed to contour normal tissues can be decreased substantially, increasing workflow efficiency, and decreasing the time animals need to be anesthetized. Several (semi-)automatic contouring methods have been proposed in clinical radiotherapy, which can be divided into two main categories: methods that do or do not use prior knowledge. The latter category includes methods such as adaptive thresholding, edge detection, graph cuts and active contours (Boykov and Jolly 2000), while the former includes atlas-based, statistical, machine learning and hybrid methods (Yang et al. 2012, Liu et al. 2018, Yuzhen and Barrett 2019).

Although significant advances have been made, most of these methods suffer from limited flexibility and versatility (e.g. they are organ shape and size specific) making them prone to topological errors and only applicable to specific tasks. Furthermore, the high computational complexity and the need for manual pre- and/ or postprocessing (e.g. pre-alignment on expert contours) hamper their application in preclinical radiotherapy. Literature on automatic contouring in preclinical studies using the aforementioned methods is limited and mostly use magnetic resonance (MR) or fan-beam μCT images as input (Scheenstra et al. 2009, Baiker et al. 2010, Lancelot et al. 2014, Yan et al. 2017). Multi-atlas based automatic segmentation for μCBCT is a relatively new development (van der Heyden et al. 2019).

Recently, deep learning (DL) has been employed for automatic segmentation. DL is a novel approach that has the advantage of aggregating structural features such as intensities and gradient changes by training convolutional kernels in neural networks. This approach has been shown to be powerful for segmentation in clinical radiotherapy (Lustberg et al. 2018, Elguindi et al. 2019, Zeleznik et al. 2021), and has been recently introduced in preclinical studies as well (Schoppe et al. 2020, van der Heyden et al. 2020). While DL-based automatic contouring is becoming well established in the clinical field, this is not yet the case for the preclinical field, where fast calculation is essential to obtain contours while the animal is sedated.

In this work, a 3D DL model for fast automatic contouring for preclinical studies in mice and rats is proposed. The goal is to use standardized 3D μCBCT volumes as input, and as such provide a fully automatic, generalizable method for normal tissue contouring in preclinical studies. Additionally, a method for assessing the uncertainty of the proposed DL model is shown, as a first step towards quantifying the overall uncertainty of the automated DL-based contouring workflow.

### 2. Methods

#### 2.1. Datasets

For training the DL model, two datasets which were retrospectively obtained from two larger rodent radiation studies (Granton et al. 2014, Mowday et al. 2020) were used. The first is a thorax dataset, consisting of 115 sets of 3D μCBCT images with corresponding manual contours of 6 organs of 95 mice. The second is a head dataset, consisting of 50 sets of contrast-enhanced 3D μCBCT images with corresponding manual contours of 5 organs of 30 rats. The μCBCT images were acquired with an image-guided precision irradiator (X-RAD 225Cx, Precision X-Ray Inc., North Branford, CT, USA). Detailed information about these datasets is provided in table 1.

A third dataset was generated for final external testing of the DL model for the head organs. This contained similar cases as the head training dataset, but without ground truth segmentations. Details of this dataset are also
While the training data only contained μCBCT images taken at an energy of 80 kV, this validation dataset also contained μCBCT images taken at 40 kV. Due to the lower imaging energy, slightly more artefacts were present in the 40 kV images compared to the 80 kV images. The DL model generated 3D contours for these cases, and the automatically generated contours were scored by two experienced animal technicians on a scale from 0 to 5. A score of 0 means that the generated contour is unacceptable and would need to be re-delineated from scratch manually, whereas a score of 5 means that the generated contour does not need any manual adjustments. The scoring scale is further explained in table 2.

### Table 1. Specifications of the datasets used in this work. μCBCT = micro cone-beam CT.

| Body site | Thorax | Head | Head validation |
|-----------|--------|------|-----------------|
| Study     | Granton et al (2014) | Mowday et al (2020) | — |
| Imaging modality | μCBCT | Contrast enhanced μCBCT | Contrast enhanced μCBCT |
| Energy (kV) | 80 | 80 | 40 and 80 kV |
| Spatial resolution (mm) | 0.1 | 0.1 | 0.1 |
| Rodent strain | C57BL/6 mice | F344/IcoCrl rats | F344/IcoCrl rats |
| Age/sex | 12 weeks/male | 8–12 weeks/female | 11–13 weeks/female |
| Organs | Left and right lung, complete lungs, heart, spinal cord, thorax bone | Left and right brain hemisphere, whole brain, left and right eye | Left and right brain hemisphere, whole brain, left and right eye |
| Number of animals | 95 | 30 | 31 |
| Number of CT images | 95 | 30 | 62 |
| Number of manual contour sets | 115 | 50 | — |
| Tumor presence | No | Not visible/early stage of tumor growth | Not visible/early stage of tumor growth |

As a subset of the head and thorax datasets (20 cases of each) was used for an inter-observer variability study (Lappas et al 2022), manual contours from both animal technicians were available for the cases in this subset. For the cases not in this subset, only one set of manual contours was available. Having manual contours from different observers in the training data set was shown to positively influence DL-based automatic contouring performance (Schoppe et al 2020). Therefore, all available manual contour sets were used during model training.

### Table 2. Ranking used for expert evaluation of DL generated contours on the head validation dataset.

| Score | Description |
|-------|-------------|
| 0     | The contour is unacceptable and needs to be re-delineated from scratch |
| 1     | Very major manual changes to the contour are needed |
| 2     | Major manual changes to the contour are needed |
| 3     | Minor manual changes to the contour are needed |
| 4     | Very minor manual changes to the contour are needed |
| 5     | No changes to the contour are needed, it can be used as generated |

displayed in table 1 (Head validation). While the training data only contained μCBCT images taken at an energy of 80 kV, this validation dataset also contained μCBCT images taken at 40 kV. Due to the lower imaging energy, slightly more artefacts were present in the 40 kV images compared to the 80 kV images. The DL model generated 3D contours for these cases, and the automatically generated contours were scored by two experienced animal technicians on a scale from 0 to 5. A score of 0 means that the generated contour is unacceptable and would need to be re-done from scratch manually, whereas a score of 5 means that the generated contour does not need any manual adjustments. The scoring scale is further explained in table 2.

#### 2.2. Manual organ segmentation

Manual contouring of the relevant organs was performed by two experienced animal technicians using the SmART-ATP software (Precision X-ray Inc., North Branford, CT, USA & SmART Scientific Solutions BV, Maastricht, the Netherlands) (van Hoof et al 2013). Due to the different anatomical complexities of the various organs, different delineation planes, techniques and drawing tools were applied for each organ. The lungs, spinal cord, thorax bone and brain hemispheres were contoured semi-automatically, while the other organs were delineated fully manually. Further details of the delineation techniques used are provided by Lappas et al (2022).

As a subset of the head and thorax datasets (20 cases of each) was used for an inter-observer variability study (Lappas et al 2022), manual contours from both animal technicians were available for the cases in this subset. For the cases not in this subset, only one set of manual contours was available. Having manual contours from different observers in the training data set was shown to positively influence DL-based automatic contouring performance (Schoppe et al 2020). Therefore, all available manual contour sets were used during model training.

#### 2.3. Image standardization

To standardize the data before using it as input into the DL model, several preprocessing steps were performed (figure 1). First, the μCBCT volumes and the corresponding ground truth segmentations were resized using linear and nearest-neighbor interpolation, respectively, to obtain the same μCBCT volume size (i.e. slice thickness, pixel spacing and image dimensions) for the whole dataset. Next, the resized volumes were converted from Hounsfield units (HUs) to mass density (MD) values, to remove the energy dependency of the μCBCT imager. This allows for application of the DL model to μCBCT images acquired with different acquisition...
parameters (e.g. 40 kV versus 80 kV images), making the model robust to variations in these parameters and generalizable to institutions using different imaging protocols. This step involves conversion from the reconstructed \( \mu \text{CBCT} \) voxel values to 3D mass density \( \rho \) matrices \((g \text{ cm}^{-3})\) using the appropriate HU-\( \rho \) calibration curve for the specific photon energy. To obtain this calibration curve per dataset, a cylindrical 30 mm diameter phantom, dedicated for preclinical research (SmART Scientific Solutions BV, Maastricht, the Netherlands), was scanned (figure 1, bottom left). The phantom consists of a solid water bulk with 10 tissue-mimicking inserts including three different types of cortical bone and 2 air holes paired with verified mass density values (e.g. \( \rho_{\text{solid water}} = 1.02 \text{ g cm}^{-3}, \rho_{\text{adipose}} = 0.95 \text{ g cm}^{-3}, \rho_{\text{cortical bone}} = 1.33–1.82 \text{ g cm}^{-3} \)). As a last preprocessing step, the MD volumes were resized to 128\(^3\) voxels to fit the hardware limitations, normalized such that the voxel values were in the range [0,1] and finally used as input to the model.

2.4. Deep learning model

The DL model was adapted from a publicly available 3D U-net architecture (https://github.com/xf4j/aapm_thoracic_challenge), which was originally designed for the 2017 AAPM Thoracic Auto-Segmentation Challenge in humans (Yang et al. 2018). In this study, initially the 3D U-net was applied to both the rodent head and thorax data in 2 steps: the first to obtain a bounding box around the organ of interest, and the second to segment the organ. However, considering practical issues, such as training time and computational complexity, in the present work the model is applied in only one step to segment the organs directly. This decreases the number of network parameters which minimizes hardware needs. This was determined by analyzing the network’s behavior for the two steps separately. Increasing the training iterations for the 1st step already resulted in the network precisely segmenting the borders of the organ of interest, thereby avoiding having to maintain the feature map for the 2nd step and consequently handling less parameters. The differences in performance between the 2-step and the simplified 1-step 3D U-net were small (supplementary material A (available online at stacks.iop.org/PMB/67/044001/mmedia)).

The 3D U-net architecture (figure 2) consists of 3 encoding and 3 decoding layers. The encoding layers extract features from the \( \mu \text{CBCT} \) data, while the decoding layers perform the organ segmentation. Each encoding layer takes 3D volumes as input and then applies a block of 3D convolutions followed by batch normalization, rectified linear (ReLU) activation and a downsampling step (max pooling) halving the spatial resolution at each layer. The decoding layers perform the same block of processes as the encoding layers, but instead of downsampling for feature detection, there is an upsampling step (bilinear interpolation) to expand the feature maps back to the original resolution of the volumes. Skip connections are used to combine the features extracted in the encoding path with the upsampled output of the decoding layers for a higher precision prediction.

The DL model was trained per organ on a 24 GB Quadro P6000 GPU (NVIDIA, Santa Clara, CA, USA). For each organ, \( k \)-fold cross validation with \( k = 5 \) was performed, i.e. 80% of the data was used for training, and 20% for internal testing, and this process was repeated \( k = 5 \) times, rotating the cases in the 20% test split such that each case was included for testing only once. Throughout this process, a fixed set of hyperparameters was used, which are summarized in table 3. These are based on the findings of Çiçek et al. (2016) and van der Heyden et al. (2020), relative to the task complexity and the hardware limitations. During training, 2/3rd of the input
volumes were used to train the network weights, while the remaining 1/3rd was used for evaluation of those weights on unseen data and to monitor potential overfitting. In addition, on-the-fly data augmentation was applied, consisting of a combination of random translations between $-8$ and $+8$ voxels in the vertical and lateral directions, rotations between $-5$° and $+5$° around the longitudinal axis and scaling between 0.9 and 1 for the vertical and lateral axes of the input volumes (Ronneberger et al 2015, Çiçek et al 2016, Yang et al 2018).

2.5. Evaluation metrics

To evaluate the performance of the DL autocontouring model versus manual delineations, four evaluation metrics are used: the Dice similarity coefficient (DSC), mean surface distance (MSD), the 95th percentile of the Hausdorff distance ($HD_{95\text{p}}$), and the displacement of the center of mass ($\Delta \text{CoM}$). The DSC is a measure of how similar two segmented areas are, and is computed by calculating the overlapping area of two delineations. Its values range from 0 (no overlap) to 1 (perfect overlap). For the MSD, the average Euclidean distance between two contours is calculated, while for the HD the maximum distance between a point in contour A and the nearest point in contour B is computed. The 95th percentile of the HD is calculated to avoid the impact of outliers. The ΔCoM represents the distance between the centers of mass (in 3D space) between two segmented volumes. The MSD, $HD_{95\text{p}}$, and $\Delta \text{CoM}$ values are expressed in millimeters (mm). All evaluation metrics are calculated on the original \(\mu\)CBCT image size, i.e. the preprocessing steps are reverted before calculation of the evaluation metrics.

2.6. Model uncertainty

There are various sources of uncertainty in a DL-based automatic contouring workflow that can influence the quality of the generated segmentations. On the one hand, uncertainty can arise in the input, e.g. in the form of

**Table 3.** Hyperparameters used for training the 3D U-net.

| Hyperparameter           | Value/method               |
|-------------------------|----------------------------|
| Layers                  | 3 encoding, 3 decoding     |
| # Epochs                | 500                        |
| # Convolutional filters in the first layer | 24                        |
| Convolutional kernel size | 3                         |
| Max pooling kernel size  | 3                         |
| Stride kernel size       | 2                         |
| Weight initialization    | He et al (2015)            |
| Optimizer               | Adam (Kingma and Ba 2015) with default parameters |
| Learning rate           | $1 \times 10^{-3}$        |
| Loss function           | Weighted cross entropy     |
| Batch size              | 1                         |
| Dropout                 | 0.5                       |

Figure 2. Schematic overview of the 3D U-net architecture used in this work. Conv3D: 3D convolution; BN: batch normalization; ReLu: rectified linear unit activation; Deconv3D: 3D deconvolution.
inter-observer variability, potentially resulting in disagreement regions as shown by Schoppe et al. (2020), or from the use of different image acquisition parameters (Huang et al. 2021). The influence of Gaussian noise in μCBCT images on the DL segmentation quality is provided in Supplementary Material B. In short, DL model performance is only influenced when Gaussian noise with standard deviation $\geq 500$ HU is present in the images. On the other hand, due to the various sources of randomness in DL models, there is an uncertainty associated with the DL model itself (Gal and Ghahramani 2016, van Rooij et al. 2021). The uncertainty in the input data is addressed in this work by using segmentations from multiple observers as gold standard and by standardizing the input volumes. As a first example of quantifying the uncertainty of the proposed DL model, the latter is investigated for one organ, namely the heart, which is one of the more difficult organs to segment.

The uncertainty of the DL model was assessed by applying Monte Carlo dropout (MCD) (Gal and Ghahramani 2016). When using dropout in a DL model, it is commonly only applied in the training phase, and switched off when applying the trained model to the test dataset (the inference phase). The idea of MCD is to leave dropout on, also during inference. Inference can then be repeated a number of times, and due to the random nature of dropout, the output of the model will be slightly different every time. In this work, inference with dropout on was performed for delineation of the heart, 100 times for all cases (115) in the dataset, resulting in 11,500 heart segmentations.

To assess the effect of applying MCD, these 11,500 DL-generated segmented volumes were compared to the manual segmentations, and differences between the values of the evaluation metrics with and without dropout were calculated. Additionally, the 100 model predictions for each case were summed. As the model predictions are bitmasks, this resulted in summed volumes that can be considered probability masks, where a voxel-value of 0 meant that none of the predictions included that voxel in the heart segmentation, whereas a voxel-value of 100 meant that all of the predictions included that voxel in the heart segmentation. Isolines of these summed volumes at the values 10, 50, 80 and 95 were assessed qualitatively.

3. Results

Figure 3 shows examples of μCBCT slices with DL generated contours, compared to their manual ground truth contours. For four organs, the best, average and worst case are displayed. For the best cases, the manual and DL contours overlap almost perfectly, while in the worst cases there are regions of disagreement. For the spinal cord, it is clear that this disagreement is mostly in the longitudinal direction, meaning that the number of slices taken into account in the contour differs between manual and DL contours. Slice-wise animations for an average case of the organs shown in figure 3 are available in supplementary material C.

The boxplots in figure 4 display the results of the quantitative comparison between manual and DL generated segmentations for all five test folds from the 5-fold cross validation and all four evaluation metrics. Overall performance is high for both thorax and head organs. For most organs, the median DSC is higher than 0.85. Spinal cord (0.79) and thorax bone (0.74) showed slightly lower median DSC values with a higher variance. Higher variance in the DSC was also visible for the left lung. For the two distance metrics (MSD and HD95p), generally, sub-millimeter median performance was achieved. The whole brain and thorax bone showed larger differences and more variance compared to the other organs, although with a median MSD and HD95p smaller than 0.25 mm and 1.4 mm, respectively, the DL model performance can still be considered good. The ΔCoM also resulted in sub-millimeter performance, with a median value less than 0.5 mm for almost all organs. Again, the highest values relative to the other organs were reached for spinal cord and thorax bone, although only thorax bone had a median ΔCoM larger than 0.5 mm.

Moreover, figure 4 also indicates the median and interquartile range inter-observer variability (IOV) for manual contouring that was previously calculated by Lappas et al. (2022). Generally, for the brain (whole as well as the hemispheres separately) and spinal cord, the median DL model performance is within the range of the IOV. An exception to this is the HD95p for the whole brain, which has a higher median and a larger spread for the DL model than the IOV. For the eyes, lungs and thorax bone, the median DL performance is generally worse than the IOV range, although the differences are small ($<0.1$ mm for the distance metrics and $<0.05$ for the DSC). For the heart, the DL model scores better (higher DSC, lower MSD, HD95p, and ΔCoM) than the median IOV.

The average and standard deviation per test fold of the 5-fold cross validation is provided in supplementary material D. In general, the results were stable across all five folds, meaning that the model is robust to changes in the training data. Moreover, these results confirm those of figure 4, revealing the highest degree of variance for the spinal cord and thorax bone for the DSC and ΔCoM, and for whole brain and thorax bone for the MSD and HD95p.

The results of the expert evaluation of the DL generated contours for the head validation dataset are shown in figure 5, both for the 40 kV and the 80 kV μCBCT images. Overall, these results show the high quality of the DL generated contours. For the brain (whole, left and right), only one case needed minimal changes, while for all other cases, regardless of the imaging energy, the contours would be usable without the need for manual
adaptations. For the eyes, there are slightly more contours that would need manual adaptations, although most cases only need minor changes. The same holds when comparing between the different imaging energies: when using 40 kV, more contours need manual adaptation (40 kV: 20/155 contours, 80 kV: 3/155 contours), but very few major adaptations (40 kV: 1/155, 80 kV: 2/155) are required.

Figure 6 displays the results of the MCD uncertainty analysis for the heart as an example of this method. From the quantitative comparison in figure 6(a), it can be derived that the differences in evaluation metrics with and without dropout are very small (<0.01 mm for the distance metrics, <0.001 for the DSC). Expressing these differences as percentages, the uncertainty is generally within 2%. The qualitative analysis of the isolines of the summed predictions confirms these results. In an average case, shown in figure 6(b), the isolines are mostly almost on top of each other, indicating little disagreement. Only small areas, indicated by the blue circles, show larger disagreement. Similarly, the largest observed difference (figure 6(c)) is isolated in a single area and still small compared to the organ’s total size.

4. Discussion

Overall, the proposed DL model and accompanying preprocessing method provide high quality contours for normal tissues in preclinical studies, that are generated fully automatically after training on manual expert
segmentations. For most of the organs evaluated in this work, sub-millimeter performance was achieved, which was within the range of the previously studied IOV (Lappas et al 2022). Relative to the organ’s sizes and the voxel resolution, the sub-millimeter differences between manual and DL generated contours can be considered small. The decrease in performance for the thorax bone is due to insufficient semi-automatic segmentation of the ribs in the manual ground truth contours. As the ribs are not always completely included in the manual contours, but are segmented by the DL algorithm, higher differences in the evaluation metrics can be observed. Similarly, due to variability in the number of slices included in the manual delineation of the spinal cord by the two annotators, higher variances in the evaluation metrics are seen when comparing DL-generated contours and manual contours. A possible solution for this might be to limit the DL-contouring to certain number of slices. For the left lung, larger disagreements were seen at the lower border or close to the heart; areas that are considered ambiguous for the expert annotators as well. Additionally, for the brain hemispheres, voxels belonging to the fissure between them, as well as additional boundary regions segmented only by the DL method.
resulted in higher variance in the distance metrics. However, examining the voxel intensity in those regions showed that the DL method reached higher precision due to the convolutional filters revealing features that were inconspicuous to the bare eye. As there are no guidelines for organ contouring in preclinical studies, the apparently ‘worse’ performance of the DL model for these organs is mainly the result of inter-observer variability in the ground truth.

For the HD95p of the whole brain, a larger discrepancy between the DL model results and the IOV can be observed. There are two main explanations for this discrepancy. First, there are cases where the annotators agree (i.e. IOV is low), but the DL model segmentation does not (i.e. high HD95p for manual segmentation versus DL model). This occurs mostly close to the edges of the brain, i.e. regions where there is generally more variability between cases and observers. An example of this is visualized in Supplementary Material E. Likely, upsampling of the DL model prediction to the original image resolution also has an influence on the higher HD95p, which is not an issue in the IOV calculation. Second, manual segmentations from two expert annotators were available only for a subset of the total dataset. Therefore, not all cases with high HD95p are part of the IOV evaluation.

The expert evaluation of the contours generated by the DL model for the external head validation dataset demonstrates that these contours need few manual adaptations and are therefore highly usable in preclinical studies. This evaluation also shows that the conversion from HUs to MD values works well, as high-quality contours can be generated for both 40 and 80 kV µCBCT images. The slightly lower scores for the 40 kV images compared to the 80 kV images is likely due to some artefacts in the 40 kV images. As DL models are often limited to a very specific input, the ability to process images with different acquisition parameters is a vast improvement.

Figure 6. Results of the uncertainty analysis of the DL model for the heart. (a) Quantitative comparison of the difference in evaluation metrics with and without dropout. Each datapoint represents one DL model prediction (out of 100) for one case (out of 115), resulting in a total of 11,500 datapoints per evaluation metric. DSC: Dice similarity coefficient; MSD: mean surface distance; HD95p: 95th percentile of the Hausdorff distance; ΔCoM: displacement of center of mass. (b) Visualization of the DL model uncertainty for an average case. (c) Visualization of the DL model uncertainty for the worst case. Red, orange, yellow and green lines represent the 10, 50, 80 and 95 isolines of the summed model predictions, respectively. Blue circles indicate areas of disagreement.
towards generalization and inter-institutional usage. This HU-MD conversion is already performed routinely for accurate dose calculations for KV X-rays, and using it for automatic contouring therefore does not add additional workload. Moreover, if a measured curve for a specific imaging energy is not available, one can be calculated from the knowledge of the X-ray spectrum (which can be derived from the imaging energy and filtration used (Poludniowski et al 2009, van der Heyden et al 2018)) and the material characteristics of the phantom. Although a measured calibration curve is more accurate, a calculated curve may be sufficient for the purpose of automatic contouring.

The results from this preclinical study agree with previous findings in this field, indicating the advantage of DL-based models compared to (multi-)atlas and conventional approaches (Baiker et al 2010, van der Heyden et al 2019, Wang et al 2019, Schoppe et al 2020, van der Heyden et al 2020). To the best of the authors knowledge, only Schoppe et al (2020) recently published a DL-based automated contouring method dedicated to normal tissue contouring for preclinical studies. Even though their dataset was larger (220 rodents, 2 annotators), model performance in terms of median DSC for heart and complete lungs segmentation is similar as presented here (0.92 and 0.93, respectively). Additionally, this work shows that DL is also valuable for segmentation of smaller organs, such as the brain and the eyes. These were not taken into account by Schoppe et al (2020). Furthermore, the DL architecture used by Schoppe et al (2020) was a U-net employing 2D convolutions. This limits the capabilities of their network regarding feature extraction in adjacent voxels in the longitudinal axis (i.e. slices). The 3D U-net with 3D convolutions that was trained in this work overcomes this issue.

Whereas manual contouring in preclinical studies can take up to 10 minutes per organ, and therefore up to an hour per animal, generating contours with the DL model proposed in this work only takes approximately 5 seconds per organ. DL-based automatic contouring can thus vastly reduce the time needed for contouring. This will clearly increase the efficiency of the image-guided precision radiotherapy workflow and the effective use of animals, without compromising the accuracy of the extracted information. Animal discomfort will also be kept at a minimum this way. This ensures that preclinical radiation studies adhere to the 3-R principle.

The MCD uncertainty analysis for the prediction of heart contours showed that the DL model is very robust, with differences smaller than 0.01 mm, or 2%. This shows that the DL model does not rely on certain connections between neurons in the model, i.e. it is not overfitted. For organs with more complex shapes, such as the thorax bone, the uncertainty may be slightly higher. The analysis in this work only considered MCD, however, in future work, other methods for assessing DL model uncertainty, such as slightly changing the model weights as proposed by van Rooij et al (2021), should be included as well. Combining this with further quantification of uncertainty in the model’s input can ultimately provide spatial confidence estimates for the generated contours, allowing users to quickly determine the quality and usability of these contours, as well as identify potential problematic areas and their relevance. Evaluating the uncertainty of DL models has only recently gained attention in clinical radiotherapy, but is of equal importance for preclinical studies.

In future work, dose calculations should be performed to assess if there are differences in normal tissue doses when using manual versus DL-based automatic contours. However, due to the small quantitative contour differences combined with relatively large margins and limited beam modulation possibilities in preclinical radiotherapy, dose differences are expected to be small. Vaassen et al (2021) recently showed that dose differences due to automatic contour variations in clinical radiotherapy are indeed small, although they increase when an organ is closer to the target volume. Moreover, not all preclinical studies that employ imaging and contouring of normal tissues are radiation studies. Hence, geometric accuracy of the DL-generated contours is equally important as dosimetric accuracy.

5. Conclusion

In this work, a DL-based model dedicated to preclinical studies has been developed for multi-organ segmentation in two body sites (head and thorax). For the first time, a method independent of image acquisition parameters has been quantitatively evaluated, resulting in DSC scores over 0.85 and sub-millimeter performance for the MSD, HD_{95p} and ∆CoM. Furthermore, the model was qualitatively assessed on an external head dataset by two experienced animal technicians. This demonstrated the high quality of DL-generated contours. The uncertainty analysis additionally showed that the inherent model variability is low. This approach can vastly decrease the time needed to contour normal tissues in preclinical studies, moving towards further standardization, automation and higher efficiency of the workflow and adherence to the 3-R principle.

Acknowledgments

The authors would like to thank Dr Brent van der Heyden for fruitful discussion and the NVIDIA Corporation for the donation of the NVIDIA Quadro P6000 GPU used for this research. The OMEGA collaboration project is
co-funded by the PPP allowance made available by Health-Holland, Top Sector Life Sciences & Health to stimulate public–private partnerships. This manuscript reflects the authors’ view only, and the Stichting LSH-TKI or the Ministry of Economic Affairs and Climate Policy (Netherlands) are not responsible for any use that may be made of the information it contains.

**Ethical statement**

$\mu$CBCT images were re-used from previous animal experiments, which were all in accordance with local institutional guidelines for animal welfare and approved by the Animal Ethical Committee of Maastricht University (protocol number 2012-006 and 2017-012).

**ORCID iDs**

Cecile J A Wolfs  [orcid.org/0000-0001-6428-2762]
Nick Staut  [orcid.org/0000-0001-8568-6037]

**References**

Baiker M, Milles J, Dijkstra J, Henning T D, Weber A W, Que I, Kajziel E L, Lowik C W G M, Reiber J H C and Lelieveldt B P F 2010 Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data Med. Image Anal. 14 723–37

Boykov Y and Jolly M-P 2000 Interactive object segmentation using graph cuts Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000 pp 276–86

Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D U-net: learning dense volumetric segmentation from sparse annotation Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016 pp 424–32

Elgindy S, Zelefsky M J, Jiang J, Veeraraghavan H, Deasy J O, Hunt M A and Tyagi N 2019 Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy Phys. Imaging Radiat. Oncol. 12 80–6

Gal Y and Graharnani Z 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning Proc. 33rd Int. Conf. on Machine Learning vol 48, pp 1050–9

Granton P V, Dubois L, van Elmp W, van Hoof S J, Lieuwes N G, De Ruysscher D and Verhaegen F 2014 A longitudinal evaluation of partial lung irradiation in mice by using a dedicated image-guided small animal irradiator Int. J. Radiat. Oncol. Biol. Phys. 90 696–704

He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification Proc.IEEE Int. Conf. on Computer Vision pp 1026–34

Huang K, Rhee D J, Ger R, Layman R, Yang J, Cardenas C E and Court L E 2021 Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms J. Appl. Clin. Med. Phys. 22 168–74

Kingma D P and Ba J 2015 Adam: a method for stochastic optimization 3rd Int. Conf. for Learning Representations

Lancelot S, Roche R, Slimen A, Boullot C, Levigoureux E, Langlois J-B, Zimmer I and Costes N 2014 A multi-atlas based method for automated anatomical rat brain MRI segmentation and extraction of PET activity PLoS One 9 e109113

Lappas G, Staut N, Lieuwes N G, Biemans R, Wolfs C J A, van Hoof S J, Dubois L J and Verhaegen F 2022 Inter-observer variability of organ contouring for preclinical studies with cone beam Computed Tomography imaging Physics and Imaging in Radiation Oncology 21 11–7

Liu X et al 2018 Automatic organ segmentation for CT scans based on super-pixel and convolutional neural networks J. Digit. Imaging 31 748–60

Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoop J, van Elmp W and Dekker A 2018 Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer Radiother. Oncol. 126 512–7

Mowday A M, Lieuwes N G, Biemans R, Marcus D, Rezaeifar B, Reniers B, Verhaegen F and Dubois L J 2020 Use of a luciferase-expressing orthotopic rat brain tumor model to optimize a targeted irradiation strategy for efficacy testing with temozolomide Cancers 12 1585

Poludnioswski G, Landry G, Delhoss F, Evans P M and Verhaegen F 2009 SpekCalc: a program to calculate photon spectra from tungsten anode x-ray tubes Med. Biol. 54 N433–8

Ronneberger O, Fischer P and Brox T 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (Berlin: Springer) pp 234–41 vol Series

Scheenastra A E H, van de Vecht C G, van der Weerd L, van den Maagdenberg A M J M, Dijkstra J and Reiber J H C 2009 Automated segmentation of in vivo and ex vivo mouse brain magnetic resonance images Mol. Imaging 8 35–44

Schlack R A, Senthilkumar G, Boerma M and Bergom C 2020 Advances in preclinical research models of radiation-induced cardiac toxicity Cancers 12 415

Schope O et al 2020 Deep learning-enabled multi-organ segmentation in whole-body mouse scans Nat. Commun. 11 5626

Tillner F, Thute P, Bütof R, Krause M and Enghardt W 2014 Pre-clinical research in small animals using radiotherapy technology—a bidirectional translational approach Z. Med. Phys. 24 335–51

Vaassen F, Hazelaar C, Canters R, Peeters S, Petit S and van Elmp W 2021 The impact of organ–at-risk contour variations on automatically generated treatment plans for NSCLC Radiother. Oncol. 163 156–42

van der Heyden B, Podesta M, Ekers D B, Vanqui A, Almeida P, Schyns L E, van Hoof S J and Verhaegen F 2019 Automatic multi-atlas based organ at risk segmentation in mice Br. J. Radiol. 92 20180364

van der Heyden B, Schyns L E J R, Podesta M, Vanqui A, Almeida P, Landry G and Verhaegen F 2018 VOXSI: a voxelized single- and dual-energy CT scenario generator for quantitative imaging Phys. Imaging Radiat. Oncol. 6 47–52

van der Heyden B, van de Worp W R P H, van Helvoort A, Theys J, Schols A M W J, Langen R C J and Verhaegen F 2020 Automated CT-derived skeletal muscle mass determination in lower hind limbs of mice using a 3D U-Net deep learning network J. Appl. Physiol. 128 42–9
van Hoof S J, Granton P V and Verhaegen F 2013 Development and validation of a treatment planning system for small animal radiotherapy: SmART-plan Radiother. Oncol. 109 361–6
van Rooij W, Verbakel W F, Slotman B J and Dahele M 2021 Using spatial probability maps to highlight potential inaccuracies in deep learning-based contours: facilitating online adaptive radiation therapy Adv. Radiat. Oncol. 6 100658
Verhaegen F et al 2018 ESTRO ACROP: technology for precision small animal radiotherapy research: optimal use and challenges Radiother. Oncol. 126 471–8
Verhaegen F, Granton P and Tryggestad E 2011 Small animal radiotherapy research platforms Phys. Med. Biol. 56 R55–83
Wang H, Han Y, Chen Z, Hu R, Chatrianonnou A F and Zhang B 2019 Prediction of major torso organs in low-contrast micro-CT images of mice using a two-stage deeply supervised fully convolutional network Phys. Med. Biol. 64 245014
Yan D, Zhang Z, Luo Q and Yang X 2017 A novel mouse segmentation method based on dynamic contrast enhanced micro-CT images PLoS One 12 e0169424
Yang J et al 2018 Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017 Med. Phys. 45 4568–81
Yang J, Wei C, Zhang L, Zhang Y, Blum R S and Dong L 2012 A statistical modeling approach for evaluating auto-segmentation methods for image-guided radiotherapy Comput. Med. Imaging Graph. 36 492–500
Yuzhen N W and Barrett S 2019 A review of automatic lung tumour segmentation in the era of 4DCT Rep. Pract. Oncol. Radiotherapy 24 204–20
Zeleznik R et al 2021 Deep-learning system to improve the quality and efficiency of volumetric heart segmentation for breast cancer npj Digital Med. 4 43