Deep Learning Segmentation in RayStation

This webinar consisted of two presentations: (1) how a clinic collaborated with RaySearch to create a new DL segmentation for RayStation, and (b) and overview of DL segmentation in RayStation.

# Training, Validation, and Clinical Implementation of a DL-Segmentation Model for Breast Cancer

## Presenter

Sigrun Saur Almberg, medical physicist at St. Olavs Hospital in the Netherlands

## Intro

Sigrun’s clinic and another clinic worked with RaySearch’s ML Imaging team to train RaySearch’s DLS model for breast. Sigrun was especially interested in automating breast planning because the two clinics see over 400 breast patients annually.

## Training Data

The model was created using 200 (100 from each clinic) Breast\_L CTs with DIBH: 170 training images and 30 validation images. All structures were manually delineated specifically for this project. Eight targets were delineated according to ESTRO guidelines, 11 OARs, and eight support structures according to PROCAB.

## Project Timeline

From planning and ethics approval to full implementation in both clinics, the project took from spring 2020 to December 2021. The hardest part was finding MD time to delineate training structures.

## Model Evaluation

Model performance was evaluated against manual contours from five MDs who were not involved in training set creation.

### Geometric Evaluation

For each test case, the difference between each DLS contour and any of the five MD contours\* was compared to the inter-observer variability (IOV) among the five MDs. The median of each of the following metrics across all test cases was used:

* 95% Hausdorff distance (HD95): The 95th percentile of the shortest distance between a point on X and a point on Y. A smaller value indicates greater similarity.
* Dice similarity coefficient (DSC): Relative volumetric overlap. A biased metric because it performs better for large or spherical targets. A larger value indicates greater similarity.

Based on both HD95 and DSC, the model has less variability than IOV for the vast majority of targets and OARs. There was no IOV for lungs, humeral head, or sternum.

### Clinical Evaluation (Manual Review)

Two MDs reviewed 15 contours each. They rated each contour on a five-point scale, from “useless” to “no corrections necessary.”

72 percent of OARs required no corrections. None of the lungs or sternum contours needed corrections. As for support structures, 73 percent of vessels required no corrections, 82 percent of breast clips required only minor corrections, and no breast string contours required corrections.

## Dosimetric Evaluation

Dosimetric evaluation is more clinically relevant than geometric evaluation. Each VMAT plan was automatically optimized using a script that is also used clinically. Each plan was 2.67 Gy × 15 = 36.05 Gy. For each target, the percent of plans meeting the CTV goal D98% > 95% Rx was reported. All CTVp\_Breast geometries met this goal along with 89 percent of CTVn geometries. For OARs, there were small statistically significant dosimetric differences, but none of them were clinically relevant.

## Demo

Without support structures, the model runs in about 1.5 minutes in RayStation. Structure review/correction takes 10–15 minutes. Compared to manual delineation, which takes around an hour, DLS segmentation provides around a 75 percent time reduction. Most corrections are to caudal and cranial ends in the transverse direction even though the original structures look fine in 3D. MDs love and are surprised by DLS segmentation’s speed and accuracy.

The results in the actual RayStation release are even better than in the demo due to postprocessing, including overlap removal.

## Future Work

* Publish the paper
* Breast\_R model (targets and humeral head will be different from the Breast\_L model)
* Quantify the effect on efficiency and consistency
* ML planning model for the same patient group
* DLS pelvic model

# Deep Learning Segmentation in RayStation

## Presenter

Grietsje Schregardus-Abma, Senior Application Specialist at RaySearch. Grietsje also works with RaySearch’s DL segmentation team and has two years’ clinical delineation experience.

## Overview of DLS in RayStation

RayStation’s DLS models are 3D u-net convolutional neural network (CNN) (I don’t know what that is, either). Reasons to use DLS include increased efficiency and consistency, decreased variability, and automation. Starting in 12A, DLS can be part of templates and protocols, and it can be scripted in all Raystation versions. DLS comes free with versions 11B and up.

Models are trained and validated by RaySearch and must be commissioned by the clinic before clinical use. New and updated models are released regularly for the latest few RayStation versions. Help structures are not yet included in DLS models.

## Demo

* Thorax-Abdomen-Breast:
  + Targets based on ESTRO guidelines, Frehudman (?) for vessels.
  + Breasts, lungs, heart, IMN, LAD, esophagus, spinal canal, pancreas, spleen, stomach, kidneys, gallbladder…
* Head and Neck:
  + Based on Brower guidelines. For deviations from Brower and for structures not included in Brower, see model data sheets.
  + Brain, brainstem, eyes, lenses, lacrimal glands, nasolacrimal duct, mandible, optic nerves, optic chiasm, oropharynx, submandibular glands, larynx, oral cavity, thyroid, lungs (!), …
  + June 2022 update: Optic nerves go through optical canal

Each of these models has heretofore only been trained on data from a single clinic. They need more training before release in the software.

## Clinic Customization

You can choose custom subsets of RSL’s DLS models, including from multiple models. You can of course use DLS, ABS, and MBS in the same case, and automatically perform ROI algebra on DLS-generated contours.

## Getting Started in Your Clinic

Steps to start using a DLS model:

1. Get the model from RaySearch. Either upgrade your RayStation version. If you have version 11B or higher, you may also directly request the model from RaySearch.
2. Commission the model for your clinic. Use your own internal workflow or RaySearch’s recommended workflow:
   1. Participate in training. Use videos on Ray Community. DLS is also included in RaySearch’s “intro” training course.
   2. Create a commissioning dataset of ~10 patients from your clinic.
   3. Run the model on your dataset. Evaluate the differences between the DLS contours and your manual contours:
      1. Qualitative
      2. Quantitative: HD95, HD99, HD100, DCS, other HD percentiles

RaySearch provides scripts that output HD and DCS values to a CSV file.

Document results.

* 1. Go live! Continue to monitor model performance.
  2. Provide feedback to RaySearch about how you use the model in your clinic.

1. Join the Deep Learning Segmentation group on Ray Community.

## Ray Community DLS Resources

* Deep Learning Segmentation group
* Training videos
* Commissioning guidelines
* Model data sheets. Describe “where a model comes from,” including deviations from any guidelines or standards.

## Improvements

Known issues that RaySearch is working on:

* Structure overlap
* Near the boundary between structures, part of the structure is “missing”

Please provide your suggestions for new and updated DLS models!

* Not all clinics use the contouring guidelines that the models were built on. RaySearch would like to know what models you use.

# Audience Q&A

**Q:** Sigrun, did you use a model provided by RaySearch, or does your model comply with local contouring protocols?

**A:** Sigrun’s model *became* the RaySearch model. It was trained from scratch on contours from her clinic.

**Q:** Sigrun, nearly all of your model’s heart contours needed corrections. How confident are you in the “ground truth” of your manual contours, given the heart’s movement in the body. Are the manually corrections really worth it?

**A:** Dosimetric evaluations revealed very little difference with the corrections, so the corrections have no clinical consequence.

Also, a big reasons that the DLS heart contours differ from the MDs’ is that the MDs’ contours are often only the partial organ: they abruptly cut off on a transverse slice. The model is thus anatomically better, but the MDs will have to accept the power of computing at some point. :)

**Q:** Sigrun which types of model validation—geometric, clinical, or dosimetric—is most important?

**A:** For a model to be used clinically, clinicians must be on board, so involve them! Geometric measurements, such as HD or DSC, and dosimetric metrics won’t do the trick; they must engage in manual review (e.g., on a five-point scale).

**Q:** Grietsje, what guidelines were used to train the RSL models?

**A:** This information is in the model data sheets, but:

* Head & Neck: Brower with some deviations
* Breast: ESTRO for targets
* Abdomen: RTOG atlas. There are two options for the liver: simple and detailed.
* Thorax: RTOG 1106, RTOG 2014 Jebbul (?) for the heart
* Male pelvic: RTOG

**Q:** Does RaySearch refine models as guidelines are updated?

**A:** The moderator misunderstood this question and said that no, RaySearch does not retrain models just because a specific clinic does not like the model results.

**Q:** Can clinics train models?

**A:** RaySearch trains; clinics use. But obviously, RaySearch must collaborate with clinics (such as Sigrun’s) to build models.

**Q:** Grietsje, there is esophagus-trachea overlap in the demo’d Breast-Thorax-Abdomen contours. How do RSL models handle structure overlap?

**A:** The models are trained in 3D, so an overlap in 2D doesn’t tell the whole picture. Regardless, postprocessing includes some overlap removal.

**Q:** How many patients were used to train the Head and Neck model?

**A:** The moderator guesses ~100, but it varies by organ and is difficult to determine. There are obviously more test cases than training cases, and testing is often the source of difficult cases that are included in the model.

**Q:** Sigrun, does your model work on patients with breast expanders?

**A:** The model is not trained on such patients, but testing reveals that it works for all structures except breast CTVs.

**Q:** How do RSL models handle imaging artifacts?

**A:** There is no special handling. How the model perform when they are present depends on whether they are present in the training data. Test it out and see if you like the results.

**Q:** Do RSL models work on MR images?

**A:** They’re trained on CT, not MR, and MR would be more difficult due to greater variability among MR images.

**Q:** Grietsje, is version 11B FDA approved?

**A:** Yes.

**Q:** Sigrun, your model was trained on breath-hold images. Would you need a separate model for free breathing?

**A:** Based on test results, no. The variation in lung volume across patients dwarfs the variation in DIBH versus free-breathing lung volume for a single patient, so the model probably wouldn’t even “notice” the latter.

**Q:** Are there morphotypes, particular patients, or particular organs that generally require more corrections?

**A:** The moderator did not get to this question, but it was already answered in the first presentation.

\*This is how I interpreted Sigrun’s explanation, but it could be that the DLS contour was compared to an average of the five MD contours.