Validation plan

Intended use:

The intended use of the product is to help clinicians determine the total (anterior and posterior) volume of a head MRI to determine the evolution of Alzheimer disease. The measure of the evolution allows doctors to determine the therapy which fits best with the patient and try to find new strategies.

The algorithm works as a helpful tool for the clinician, but it won't substitute it because the accuracy is not too high but would help to make their job faster and more consistently.

Training data collection:

The training data was collected from MRI scans with two images, one for the volume and the other from the mask. Images are T2 MRI scans of the full brain, but images have been cut to the regions of the hippocampus to save space and make the algorithm faster and more precise.

The dataset was taken from Medical Decathlon Competition.

How training data was labeled:

Training data was labeled with ground truth. Software like Slice 3D allow to select and mask regions of interest and create a label. However, this task was done by expert people like radiologists who have seen thousands of images.

All data was labeled and verified by an expert human rater, and with the best effort to mimic the accuracy required for clinical use as it is said in Medical Decathlon competition website.

How was the training performance of the algorithm measured and how is the real-world performance going to be estimated?

In case of the training performance of the algorithm will be done with two metrics (Dice Score (DSC) and a Jaccard Score which compare the obtained result of our model with the labeled ones of the dataset.

In real world, to estimate the performance we would need a ground truth which could be created by an expert like was done in the Medical Decathlon competition and see how it differs from the result we obtained. However, we are going to say that the performance in real-world is the one obtained with the test dataset used.

What data will the algorithm perform well in the real world and what data it might not perform well on?

Data which is like the one used in the training dataset, cropped T2 MRI in the hippocampus region, with a volume between ~2200 and ~4500 and with certain dimensions (~30, ~50, ~30) will work fine. However, if we work with data which does not follow those parameters, because the model has been trained in that conditions, the performance of the algorithm will decrease significantly.