Validation Plan

Name of the device:

HippoVolume.AI

Intended Use:

Assist a radiologist with labeling the hippocampus on MRI head images and calculating its volume.

Training dataset information:

The training dataset consists of MRI head images cropped to hippocampus area.

The labels have been produced manually. Such method has some imperfections. The factors affecting the accuracy of the method include for example years of experience of the labelling radiologist or rush/amount of available time intended for the labelling task.

Patient Population Description for Validation Dataset:

Patient characteristics:

- patient age: >20
- patient gender: male or female

Image characteristics:

- imaging modality: MRI
- body part examined: head

Patient condition:

- patients may be in any condition, including neurodegenerative and other diseases which may affect hippocampal volume

Ground Truth Acquisition Methodology:

The ground truth should be obtained via manually labeling the images by the radiologists. Three radiologists would label the hippocampus pixels. The produced masks would be added and the pixels intensity divided by 3. Then, with selected threshold 0.5, each pixel would be considered as true classs – hippocampus – or negative class – background. Thus, the final hippocampus mask is an average from three radiologists and should be more accurate and more robust to mistakes, especially in the hippocampus edge area, than the mask produced by one person.

As the algorithm was trained on labels with anterior/posterior distinction, the radiologist would label the pixels as anterior/posterior as well.

Algorithm Performance:

The algorithm performance is evaluated using 4 metrics:

- Dice score
- Jaccard score
- Sensitivity
- Specificity

Once the ground truth mask is ready, both the DICOM image and the mask are sent to HippoVolume.AI. The algorithm produces its own masks, which are then compared to the ground truth and the metrics are calculated. While Jaccard score, Sensitivity and Specificity serve as additional metrics to help with evaluation (to confirm if the prediction is reliable or accidental), the main metric Dice score should be >= 0.9. Apart from the numerical results, the outline of the predicted mask is overlaid on the image to further help with assessment of the algorithm.

The goal of the algorithm is to calculate the hippocampus volume. Once the mask is predicted, the algorithm calculates the anterior, posterior and total volume of the hippocampus. The result is checked by the radiologist and should be compared to the literature (eg. HippoFit calculator[1]).

[1] http://www.smanohar.com/biobank/calculator.html