HippoVolume Validation Plan

Ryan Case

June 13, 2020

1 HippoVolume Intended/Indicated Use

When ultimately seeking FDA approval HippoVolume will be marked as a support tool for radiologists to be used in concert with the HippoCrop algorithm to localize and then segment/measure the volume of the hippocampus region. While done automatically, this will require the monitoring and approval of the radiologist and will contain features for the radiologist to change or reject the algorithm's output. This should hopefully result in approval via the 510(k) process for Class II devices.

2 Validation Plan

2.1 Ground Truth

The gold standard for measuring hippocampus volume would likely be direct measurement via dissection. Since that is hardly possible in a living patient we fall back to a silver standard of labeling 3D medical images (primarily MRI) by several different, experienced radiologists. By getting labels for multiple radiologists we avoid biasing the platform towards the tendencies of a single professional, we also minimize the impact of mistakes. These can be turned into a weighted label for the purposes of validating the algorithm. This should be accepted as a ground truth with little contest as radiologist reports are the current standard for measuring and labeling.

2.2 Observed Performance

The following averages were observed on the unseen validation set during the training process:

Metric	Score
Dice Coefficient	0.904
Jaccard Similarity	0.824
Sensitivity	0.902
Specificity	0.998

Specificity was exceptionally high, which is not surprising given that most of the voxels in the 3D volume are *not* part of the hippocampus. In regards to the correct labeling of the desired region the **Dice coefficient** shows greater than 90% overlap between the two regions, **sensitivity** similarly shows us that we are, on average, capturing 90% of the desired zone.

These numbers are very promising, and for the purposes of measuring *changes* in volumes we can ideally demonstrate that not only is the system accurate, but that it errs consistently on the same patients. In other words if the algorithm achieves an 0.85 Dice coefficient on a particular patient in one series, we'd expect another series from the same patient to have similar performance. This way we can be confident that a smaller measurement is the result of genuine, clinical changes and not variance on the part of the algorithm.

2.3 Performance Metric

Primarily we will use the Dice coefficient. This metric is penalized by both a failure to identify positive voxels, as well as incorrectly identifying negative voxels as positive. We are hoping to achieve Dice scores in our target demographic of around 0.90, which we achieved in testing.

Additionally, the patients for whom we have multiple studies will be analyzed to ensure that the variance of the algorithm from study to study is not significant enough to throw off the volume measurements. In other words we want to establish statistically that a decrease of size n in the hippocampal volume is enough to reasonably conclude that the structure has, in fact, shrunk. This can be achieved by comparing Dice and sensitivity scores in subsequent studies.

2.4 Validation Data Needed

To validate this algorithm we will need from our clinical partners MRI brain images from a wide range of patients. We are primarily targeting older individuals who are susceptible to age-related dementia so we want a significant number of those, focusing on patients with multiple studies done over periods of months/years. Beyond that we should gather an additional set from all age groups to ensure that there are no demographics that the algorithm performs notably worse on. Should this be the case we can add an indication for use that precludes this group.