FDA Validation Plan

Name of Applicant: Nivedita Bijlani

Device: Assistive end-to-end AI algorithm to measure right hippocampal volumes represented as T2 MRI scans of the full brain that have been cropped to include the region around the hippocampus.

Intended Use Statement:

This end-to-end algorithm is intended for assisting the radiologist in segmenting and computing the right hippocampal volume from T2 MRI scans of volumes of the full human brain that have been cropped to include the region around the hippocampus. The system integrates into a clinical-grade viewer and automatically measures hippocampal volumes of new patients, as their studies are committed to the clinical imaging archive. The algorithm is intended to be used to quantify the progression of Alzheimer's disease, a progressive neurodegenerative disorder.

Collection of Training Data:

The training data is the "Hippocampus" dataset from the 2018 Medical Segmentation Decathlon competition held during the 2018 Medical Image Computing and Computer Aided Interventions Conference in Granada, Spain, obtained from the collection of annotated medical imaging datasets detailed at this source: https://arxiv.org/pdf/1902.09063.pdf. The authors of this paper collected these data as part of a multi-institution collaboration. All images were deidentified and reformatted so that these and their corresponding segmentation masks can be viewed via tools accessible to the general machine learning community and not only to medical specialists. The training dataset consists of 263 training images in the NIFTI format. These are "structural images acquired with a 3D T1-weighted MPRAGE sequence (TI/TR/TE, 860/8.0/3.7 ms; 170 sagittal slices; voxel size, 1.0 mm3). All images were collected on a Philips Achieva scanner (Philips Healthcare, Inc., Best, The Netherlands)" by the Vanderbilt University Medical Center. All subjects were free from significant medical or neurological illness, head injury, and active substance use or dependence.

Labelling of Training Data:

The labels or segmentations for the training data were manually provided by expert radiologists enlisted by the scientific institutions noted in the link provided in the above section. The labelled files are provided as NIFTI files, each named identically to the corresponding training image file, so that the training image and label files are easy to match up. The labels are 0 – background, 1 – anterior, 2 – posterior.

Algorithm Performance Standard:

The training performance of the algorithm was measured using two metrics that are commonly used to evaluate segmentation performance – Dice score and Jaccard score. The model was trained on 50% of the overall dataset, validated on an internal hold-out 25% validation dataset and scored on the remaining internal partition of 25% of the dataset.

The Dice and Jaccard metrics both measure the similarity or overlap between the actual label or mask provided by the expert radiologist for the hippocampus scan and the label or mask predicted by the AI algorithm for the same. The scores vary between 0 (no overlap) and 1 (perfect overlap). These are computed as explained below:

Dice score – This is computed as: 2 * |X intersection Y|/(|X| + |Y|)

Here:

|X| and |Y| represent the number of elements in the ground truth and predicted label masks respectively,

|X intersection Y| represents the number of common elements in the ground truth and predicted label masks

trained model on an internal test set

Jaccard score – This is computed as: |X intersection Y|/(|X| + |Y| - |X intersection Y|

The individual terms in the Jaccard computation above have the same semantics as the corresponding elements in the Dice score. It also lies between 0 and 1.

On the given dataset, we obtained an average Dice score of 0.81 and average Jaccard score of 0.68. To address class imbalance between background and non-background voxels, these were computed based on the non-zero labels of voxels only.

In the real world, the performance is likely to be estimated by the above metrics but also other metrics such as sensitivity and specificity. These are defined as follows:

Sensitivity = TP/(TP + FN)

Specificity = TN/(TN + FP)

In segmentation terms, the terms above can be understood as follows:

TP = Number of voxels that have the same non-background (i.e. positive) label in the ground truth as in the predicted mask

TN = Number of voxels that have the same background label (i.e. negative) in the ground truth as in the predicted mask

FN = Number of voxels in the predicted mask that are marked as background, but are non-background per the ground truth

FP = Number of voxels in the predicted mask that are marked as non-background, but are background per the ground truth

Clinicians will also usually look at TP, TN, FP and FN scores, and so, we can redefine the Dice and Jaccard scores in terms of these as follows:

Dice = (2 * TP)/(TP + FP) + (TP + FN)

Jaccard score = TP/(TP + FP + FN)

Algorithm Data:

The model has been trained on rectangular cropped images of right hippocampal data of human adults based on T2 MRI scans of the full brain where all subjects were free from significant medical or neurological illness, head injury, and active substance use or dependence. The model is expected to perform at the Dice and Jaccard score benchmarks stated above, on this type of data only. It is **NOT** expected to perform on other kinds of data with the performance figures shown above.