Validation plan

Jared Bowden

Hippocampal Volume Quantification in Alzheimer's Progression

Algorithm Description
The algorithm described here is intended to assist a radiologist with the quantification of hippocampal volume, as measured from a series of T2 MRI scan images.

Indications for Use
The algorithm is indicated for the purposes of improving the efficiency with which a radiologist is able to quantify hippocampal volume from longitudinal T2 MRI scans, and when measured longitudinally, quantify the progression hippocampal volume decreases associated with Alzheimer’s disease.

Description of Training Dataset
The algorithm was built from 263 T2 MRI images taken from the Medical Decathlon dataset, cropped to a region surrounding the hippocampus, and saved as a series of NIFTI files. Training labels of hippocampal volumes were provided within the dataset on behalf of staff from Vanderbilt University Medical Center. Prior to training, images were preprocessed to remove files with incomplete information and/or spurious values. The final number of image series included in training was 260.

Validation plan
Dice and Jaccard scores were used to inform algorithm performance through the process of training and development. Through hold out validation, mean Dice and Jaccard coefficient values on the final model were measured as ~0.85 and ~0.75, respectively.

Prior to more widespread deployment, the ability of the algorithm’s performance to generalize should be evaluated against a novel dataset, preferably one which captures the demographic and medical complexity which can be expected from a relevant clinical environment. A team of radiologists will be required to label hippocampal volumes within this validation datasets. Ideally, these label traces will be averaged across practitioners, weighting for experience and/or expertise. The dataset should also be documented to account for basic demographic information, medical history, and other features that could serve to bias the interpretation of hippocampal volumes beyond a clinically relevant range. Initial performance on this validation set should be reviewed by way of comparison to the Dice and Jaccard values specified above; however, final validation should take into account the feedback of clinicians and practitioners with relevant expertise. Here, feedback would be particularly valuable on failure cases. Realistically, both training and validation must take place on a finite set of available data. The degree to which the algorithm will generalize in the presence of comorbidities, differences in age (developmental stages of infants and adolescents vs. geriatiric) therapeutic treatments (pharmacological confounds), or previous injuries is largely unknown. Operation outside these norms has the potential to result in spurious results.