

## Validation Plan

Device Name: *Hippocampus Volume Segmentation Algorithm*

### Algorithm Description

The proposed Hippocampus Volume Segmentation Algorithm is intended to be used as a supplemental aide for doctors to gain high accuracy data describing the volume of a patient's hippocampus from a CT scan. This algorithm is expected to be executed at regular intervals during a longitudinal study of a single patient's hippocampus volume. Regular analysis of the hippocampal volume estimated by this algorithm may then be interpreted by an expert doctor to quantify the progression of Alzheimer's disease. Reduction in hippocampal volume over time has been shown to be an indicator of Alzheimer's progression.

### Data Acquisition Patient Population and Target Patient Population

The algorithm uses a machine learning method to analyze a large body of labeled sample data to produce accurate estimations of hippocampus volume. Data was acquired from a large study (1 million participants) comprising equal numbers of men and women aged between 50 and 100. This study recorded full-brain CT scan data, so the right and left hippocampal volumes were available from all participants. 1000 expert radiologists provided hand-labeled hippocampus segmentation analysis of the full dataset, aided by the Microsoft InnerEye segmentation software. Each segmentation was reviewed and verified by two additional radiologists. The segmentation labeling resulted in a rectangular bounding volume around the hippocampus, with a maximum size of 64x64x64 voxels. Within the bounding volume, each voxel was labeled 0 to indicate the voxel did not contain any part of the hippocampus, 1 to indicate the hippocampus anterior, and 2 to indicate the hippocampus posterior.

The algorithm is valid only when applied to CT scan data acquired from patients with demographics resembling the sample data population. CT scan data must first be pre-processed to 64x64x64 mm bounding volume around the hippocampal volume. Patients must be aged between 50 and 100, and may be men or women. This algorithm may not perform well when used on patients outside of the sample demographics, such as younger patients and children. Additionally, the algorithm may not be applied to patients with additional conditions exhibited in CT scans, rare brain deformities, or poorly acquired images (e.g. patients moving too much during scanning).

### Algorithm Performance

The trained machine learning algorithm is measured against unseen volumes from the sample dataset using two standard (and related) measures of segmentation algorithm performance, dice coefficient and Jaccard coefficient. Both dice and Jaccard coefficients are scores between 0 and 1 indicating the similarity between two sets of samples, in this case the segmentation mask produced by the machine learning algorithm and the hand-labeled segmentation mask provided by the expert radiologists. For each coefficient, a higher value indicates higher similarity. When tested against the withheld test samples

from the study sample data, the algorithm achieves a mean dice coefficient of 0.9145 and a mean Jaccard coefficient of 0.8244.

These values are sufficient for us to have high confidence in the real-world utility of the algorithm, especially when used to detect changes in volume over time for a single patient. When used in a clinical environment, the performance of the algorithm will be estimated by a required process to track the device usage and enable doctors to share aggregate anonymized feedback including expert interpretation and validation of the segmentation result and record failure cases of the algorithm. It is expected that expert doctors will be able to rapidly gauge the quality of the resulting segmentation and determine whether or not to incorporate the results into their diagnosis of Alzheimer's progression.

### **Device Limitations**

The algorithm does not achieve 100% accuracy. Therefore, an expert radiologist must review the volume segmentation, identify anomalies in segmentation, and only use the algorithm result as supplemental information to determine the final diagnosis of disease progression. Algorithm results may not be trusted individually for final diagnosis.