Independent Testing of a Publicly Available CNN tool for Hippocampal Image Segmentation

Independent Testing of a Publicly Available CNN tool for Hippocampal Image Segmentation

FN Sneden[1[0000-0003-3773-529X]], KJ Ferguson[1[0009-0007-9621-9734]], S Muñoz Maniega[1[0000-0001-5185-6384]], M Valdés Hernández[1[0000-0003-2771-6546]], JM Wardlaw[1[0000-0002-9812-6642]]

[1] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
fsneden@ed.ac.uk

Abstract. Hippocampal atrophy is a common measure in ageing and dementia. Automated segmentation methods are often computationally expensive. HippMapp3r is a CNN-based algorithm trained on patients with significant hippocampal degeneration. We tested HippMapp3r in the Lothian Birth Cohort 1936. We ran HippMapp3r with 2 preprocessing methods: 1) on skull stripped images (recommended) and 2) transformed to standard space on 602 participants with T1-weighted imaging and manually corrected ground truth masks. We found significant differences between using these two preprocessing methods (mean Dice = 0.594 versus 0.676, $p<0.001$). We found a proportional bias where the greatest variation between our measurements at smaller hippocampal volumes. We suggest automated techniques require multicenter data training and testing on different datasets prior to deployment.

Keywords: Ageing, Hippocampus, Image Segmentation.

1 Introduction

The hippocampus plays a crucial role in many cognitive processes (Valdés Hernández et al., 2017). Hippocampal volumes are an important biomarker of ageing and dementia but manual segmentation of this structure from brain images is time consuming and often prone to inter-rater variability issues (Goubran et al., 2019). Automated segmentation methods of the brain are often limited by many factors, including training on a locally available population, using a limited number of imaging protocols and some being computationally demanding (Goubran et al., 2019). HippMapp3r is a publicly available CNN-based segmentation tool that attempts to correct these issues by being trained on patients from various samples, including scans with significant atrophy/disease and being computationally inexpensive. HippMapp3r outperforms other automatic segmentation methods, with an average Dice coefficient of 0.89 when compared to manual segmentations (Goubran et al., 2019). The aim of this study is to evaluate the use of HippMapp3r in an ageing population, the Lothian Birth Cohort 1936 (LBC1936) (Deary et al., 2007). We hypothesize:

1. Our use of HippMapp3r will perform differently than described by those who developed it due to differences between the populations studied.
2. Different pre-processing methods will differentially impact in the results. For example skull stripped imagines will have different impact than transforming images to the standard space.

2 Methods

In this study, we ran HippMapp3r on the first MRI brain scan from participants in LBC1936 that had T1-weighted images (FOV = 256 x 256mm; Matrix = 192 x 192; Slices = 160; Thickness = 1.3mm; voxel size = 1 x 1 x 1.3mm; TR = 10ms; TE = 4ms; TI = 500ms), manually corrected ground truth hippocampal masks (methods by MacLullich et al., 2002) and intracranial volume (ICV) mask. We repeated hippocampal mask generation using HippMapp3r under different circumstances:

1. Preprocessing method 1: T1 images in native image space and skull stripped (recommended by HippMapp3r).
2. Preprocessing method 2: T1 images linear (affine) transformed to MNI152 space using RNiftyReg.

We compared these newly generated hippocampal masks with our ground truth measurements, using Dice similarity scores and positive predictive value, calculated in Matlab 2018. Visual quality control was performed independently in a sample of 40 randomized participants.

3 Results

MRI brain scans from 602 participants in the LBC1936 study were included in our analysis. When comparing to ground truth masks, we found HippMapp3r performed significantly better ($p<0.001$) when images were transformed to MNI152 space before image segmentation; Preprocessing method 1 demonstrated a median Dice score 0.623 versus the preprocessing method 2 with median Dice score 0.760.

Table 1. Dice similarity scores and other related performance metrics between the two methods. Dice = Dice coefficient; Spec = Specificity; Sens = Sensitivity; PPV = Positive Predictive Value. Assessed using Wilcoxon Signed T.

	Preproc. Method 1 (n = 602)			Preproc. Method 2 (n = 602)			
	Mean	Median	SD	Mean	Median	SD	Sig.
Dice	0.594	0.623	0.157	0.676	0.760	0.226	<0.001
Spec	0.999	0.999	0.000	0.997	0.998	0.002	<0.001
Sens	0.494	0.502	0.173	0.807	0.838	0.153	<0.001
PPV	0.787	0.801	0.106	0.619	0.722	0.242	<0.001

We compared the volumes obtained from HippMapp3r masks using Preprocessing method 2 to our ground truth measurements to assess for any bias. We found the greatest variation where hippocampi are smaller (Figure 1).

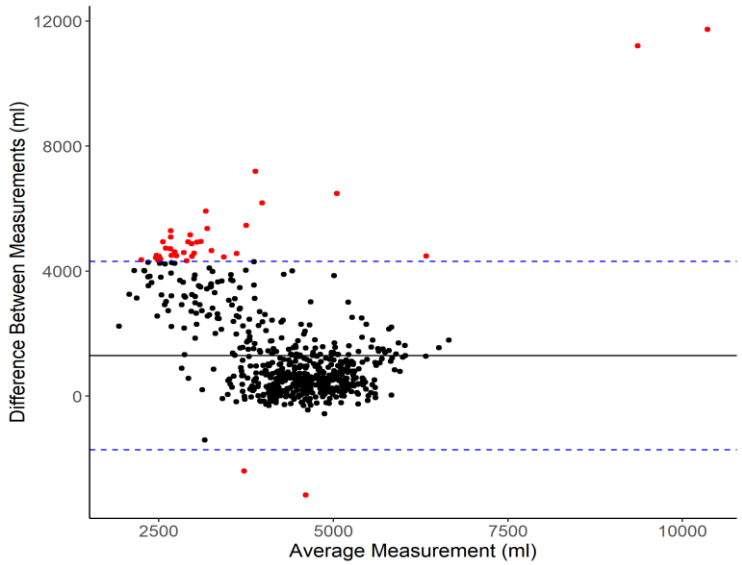

Fig. 1. Bland-Altman plot of preprocessing method 2.

4 Discussion

These data demonstrate differences between the performance of HippMapp3r in the original study, and our performance on a different population-based dataset. Our analysis found that normalizing the LBC1936 images to MNI152 space significantly increased performance metrics. This may be due to alterations in image resolution after co-registration or changes in direction of head tilt within the images. Additionally, our data suggests a greater mismatch between ground truth and HippMapp3r labels, with a mean Dice of 0.68 achieved in this sample, compared with the previously reported Dice of 0.89. Bland-Altman analysis showed a likely proportional bias where HippMapp3r consistently failed to assess smaller hippocampal volumes. Further testing of HippMapp3r is required to fully assess reasons for differences between ground truth measurements and CNN-based masks. HippMapp3r generated masks sometimes excluded anterior portions of the hippocampus and detection of true hippocampal boundaries was affected by the presence of CSF in the hippocampal fissure in some participants. Our results, hence, reinforce the necessity for collaborative training and developing CNN-based algorithms if they are to be widely used, and multi-center testing prior to their deployment in publicly available repositories.

References

Deary, I.J., Gow, A.J., Taylor, M.D., Corley, J., Brett, C., Wilson, V., Campbell, H., Whalley, L.J., Visscher, P.M., Porteous, D.J., Starr, J.M., 2007. The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr 7, 28. https://doi.org/10.1186/1471-2318-7-28

Goubran, M., Ntiri, E.E., Akhavein, H., Holmes, M., Nestor, S., Ramirez, J., Adamo, S., Ozzoude, M., Scott, C., Gao, F., Martel, A., Swardfager, W., Masellis, M., Swartz, R., MacIntosh, B., Black, S.E., 2019. Hippocampal segmentation for brains with extensive atrophy using three-dimensional convolutional neural networks. Hum Brain Mapp 41, 291–308. https://doi.org/10.1002/hbm.24811

MacLullich, A.M.J., Ferguson, K.J., Deary, I.J., Seckl, J.R., Starr, J.M., Wardlaw, J.M., 2002. Intracranial capacity and brain volumes are associated with cognition in healthy elderly men. Neurology 59, 169–174. https://doi.org/10.1212/wnl.59.2.169

Valdés Hernández, M. del C., Cox, S.R., Kim, J., Royle, N.A., Muñoz Maniega, S., Gow, A.J., Anblagan, D., Bastin, M.E., Park, J., Starr, J.M., Wardlaw, J.M., Deary, I.J., 2017. Hippocampal morphology and cognitive functions in community-dwelling older people: the Lothian Birth Cohort 1936. Neurobiology of Aging, Roadmap to Alzheimer's Biomarkers in the Clinic 52, 1–11. https://doi.org/10.1016/j.neurobiolaging.2016.12.012