





## RESEARCH ARTICLE

## OPEN ACCESS

# Evaluating Traditional, Deep Learning and Subfield Methods for Automatically Segmenting the Hippocampus From MRI

Sabrina Sghirripa<sup>1,2</sup>  | Gaurav Bhalerao<sup>3,4</sup> | Ludovica Griffanti<sup>3,4</sup>  | Grace Gillis<sup>4</sup>  | Clare Mackay<sup>4</sup> | Natalie Voets<sup>3</sup>  | Stephanie Wong<sup>5</sup> | Mark Jenkinson<sup>1,2,3</sup> | For the Alzheimer's Disease Neuroimaging Initiative

<sup>1</sup>Australian Institute for Machine Learning, School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide, South Australia, Australia | <sup>2</sup>Hopwood Centre of Neurobiology, Lifelong Health Theme, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia | <sup>3</sup>Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK | <sup>4</sup>Oxford Centre for Human Brain Activity, Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, Oxford, UK | <sup>5</sup>College of Education, Psychology and Social Work, Flinders University, Adelaide, Australia

**Correspondence:** Sabrina Sghirripa ([sabrina.sghirripa@adelaide.edu.au](mailto:sabrina.sghirripa@adelaide.edu.au))

**Received:** 8 August 2024 | **Revised:** 10 March 2025 | **Accepted:** 16 March 2025

**Funding:** Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co. Inc.; Meso Scale Diagnostics LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Oxford Brain Health Clinic data collection and analysis is supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316)—a partnership between the University of Oxford and Oxford Health NHS Foundation Trust, the NIHR Oxford Cognitive Health Clinical Research Facility, and the Wellcome Centre for Integrative Neuroimaging (201319/Z/16/Z and 201319/A/16/Z). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work is supported by NHMRC Medical Research Future Funding (MRFF) Dementia, Ageing and Aged Care Mission grant (2023947).

**Keywords:** hippocampus | MRI | neuroimaging | segmentation

## ABSTRACT

Given the relationship between hippocampal atrophy and cognitive impairment in various pathological conditions, hippocampus segmentation from MRI is an important task in neuroimaging. Manual segmentation, though considered the gold standard, is time-consuming and error-prone, leading to the development of numerous automatic segmentation methods. However, no study has yet independently compared the performance of traditional, deep learning-based and hippocampal subfield segmentation methods within a single investigation. We evaluated 10 automatic hippocampal segmentation methods (FreeSurfer, SynthSeg, FastSurfer, FIRST, e2dhipseg, Hippmapper, Hippodeep, FreeSurfer-Subfields, HippUnfold and HSF) across 3 datasets with manually segmented hippocampus labels. Performance metrics included overlap with manual labels, correlations between manual and automatic volumes, volume similarity, diagnostic group differentiation and systematically located false positives and

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [http://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Human Brain Mapping* published by Wiley Periodicals LLC.

negatives. Most methods, especially deep learning-based ones that were trained on manual labels, performed well on public datasets but showed more error and variability on clinical data. Many methods tended to over-segment, particularly at the anterior hippocampus border, but were able to distinguish between healthy controls, MCI, and dementia patients based on hippocampal volume. Our findings highlight the challenges in hippocampal segmentation from MRI and the need for more publicly accessible datasets with manual labels across diverse ages and pathological conditions.

## 1 | Introduction

Using structural magnetic resonance imaging (MRI) to analyse the morphology of the hippocampus is crucial in both clinical and research neuroimaging. The correlation between hippocampal atrophy and cognitive impairment has long been established (Nadel and O'Keefe 1978), implicating the hippocampus in various neurological and psychiatric conditions such as Alzheimer's disease (AD) (Barnes et al. 2009), mild cognitive impairment (MCI) (Jack et al. 1999), epilepsy (Thom 2014) and schizophrenia (Csernansky et al. 1998). For example, hippocampal atrophy stands as a well-established biomarker of AD, with volumetric measurements demonstrating the capability to distinguish between healthy older adults and patients at mild to advanced stages of the disease (Frisoni et al. 2008; Jack et al. 1997). In the interest of early detection and treatment of neurodegenerative disorders such as AD, it is unsurprising that extensive research effort has been directed towards developing efficient and accessible methods to accurately and reliably segment the hippocampus from MRI.

Manual segmentation of the hippocampus from MRI is widely regarded as the 'gold standard' for volumetric measurement (Dill et al. 2015). However, manual segmentation is a time-consuming process, rendering it impractical for analysing large datasets. Likewise, issues with inter- and intra-rater reliability are prevalent, and results often vary significantly depending on the manual segmentation protocol employed (Boccardi et al. 2011). In response to these limitations, numerous automatic hippocampal segmentation methods have been proposed. Traditional methods of subcortical segmentation, such as FreeSurfer (Fischl 2012) and FIRST (Patenaude et al. 2011), rely on model-based approaches. Each of these methods employs a Bayesian framework, where FreeSurfer uses deformable registration to determine subcortical labels, while FIRST uses active shape and appearance models. However, these methods, in practice, are time-consuming and resource-intensive, making them less efficient for processing large datasets.

Recently, there has been a surge in publicly available, supervised deep-learning-based segmentation algorithms, offering shorter runtimes and heightened accuracy compared to traditional methods (Goubran et al. 2020; Henschel et al. 2020; Thyreau et al. 2018). For example, methods using convolutional neural networks (CNN) learn from existing hippocampus labels in a set of training data to identify and extract features to segment the hippocampus, with the intention of learning how to segment unseen images, measured using a separate test set. However, these methods face challenges due to the limited number of datasets containing manually labelled ground truth images for training. Moreover, the variance within training sets—such

as sample demographics and scanner sequences—is quite limited, which underscores the importance of assessing the transfer of the learned segmentation approach to novel datasets and datasets that may more accurately reflect real-world, clinical populations.

Most studies introducing a new hippocampal segmentation technique benchmark the new technique against traditional methods such as FreeSurfer and FIRST, and other, newer techniques such as recently published deep learning-based methods (Goubran et al. 2020). Likewise, hippocampal segmentation methods that segment subfields (DeKraker et al. 2022; Poirer et al. 2023) benchmark across other subfield segmentation methods, so it is unclear how these methods compare to traditional sub-cortical segmentation and deep learning-based methods that segment the whole hippocampus. Currently, no study has aimed to independently assess the performance of traditional, deep learning-based and hippocampal subfield segmentation methods within a single investigation.

Our objective was to assess the performance of established and recent techniques for segmenting the hippocampus from T1-weighted MR images across three distinct datasets containing manual hippocampus labels. Broadly, our inclusion criteria encompassed segmentation methods that are publicly available, freely downloadable and accept a single T1-weighted (T1w) image as input. For methods focusing on hippocampal subfields, we collapsed over subfields and solely evaluated the entire hippocampal mask. We were interested in a range of performance metrics including overlap with manual labels, correlations between manual and automatically segmented volumes, the ability to separate diagnostic groups based on volume and the location and number of systematically located false positives and false negatives within a method.

## 2 | Method

### 2.1 | Datasets

Each segmentation method was tested against manual labels in three datasets: ADNI HarP, MNI-HISUB25 and an in-house dataset of a subset of images collected by the Oxford Brain Health Clinic (OBHC). Table 1 presents the demographics from each dataset.

The HarP dataset (Boccardi, Bocchetta, Morency, et al. 2015) uses data from the Alzheimer's disease Neuroimaging Initiative (ADNI) database ([adni.loni.usc.edu](https://adni.loni.usc.edu)), and consists of 135 T1-weighted MRI scans. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD, and is primarily focused on investigating the progression of MCI and early AD. The dataset

## Summary

- Key messages
  - We evaluated 10 automatic hippocampal segmentation methods, including traditional and deep learning-based approaches, across 3 datasets with manually segmented hippocampus labels.
  - While deep learning-based methods trained on manual labels perform well on public datasets, they show more errors and variability on unseen data that are more reflective of a clinical population.
  - Based on our investigation, Hippodeep and FastSurfer emerge as particularly attractive options for researchers looking to segment the hippocampus, based on reliability, accuracy and computational efficiency.
- Practitioner points
  - Although deep learning based automatic hippocampal segmentation methods offer faster processing times—a requirement for translation to clinical practice—the lack of variance within training sets (such as sample demographics and scanner sequences) currently prevents transfer of learning to novel data, such as those acquired clinically.
  - More training data with varying demographics, scanner sequences and pathologies are required to adequately train deep learning methods to quickly, accurately and reliably segment the hippocampus for use in clinical practice.
  - Hippodeep and FastSurfer emerge as the most attractive options for use by practitioners due to reliable and accurate performance on data more closely resembling clinical samples, ease of use and computational efficiency.

contains 1.5T and 3T-MRI scans acquired at a resolution of  $1.0 \times 1.0 \times 1.2$  mm with scanners from multiple MRI manufacturers (GE, Philips and Siemens). The images were acquired using an MP-RAGE sequence with parameters optimised for different scanners—for more details, see (Jack Jr. et al. 2008). Manual labels were created using the HarP protocol that is described in detail elsewhere (Boccardi, Bocchetta, Apostolova, et al. 2015), but briefly, a HarP segmentation includes the whole hippocampal head, body and tail, the alveus and whole subiculum based on the boundary with the entorhinal cortex. The dataset consists of 45 subjects with Alzheimer's disease (AD), 46 mildly cognitively impaired subjects (MCI) and 44 older adult control subjects (CN). Of the 135 subjects, 68 were scanned on a 1.5T machine (23 Siemens, 24 GE, 21 Philips) and 67 were scanned on a 3T machine (23 Siemens, 22 GE, 22 Philips).

The MNI-HISUB25 (Kulaga-Yoskovitz et al. 2015) dataset contains manual hippocampal subfield labels traced from T1 and T2 weighted images collected from 25 healthy subjects. Data were acquired on a 3T Siemens TimTrio scanner using a 32 phased-array head coil. T1w images were acquired at a resolution of  $1.0 \times 1.0 \times 1.0$  mm, and to increase the signal-to-noise ratio, two scans were acquired, motion corrected, and averaged into a single volume. The manual protocol was guided by intensities and morphological characteristics of the hippocampal molecular

layer and divided the hippocampus into 3 subregions: subicular complex, Cornu Ammonis (merged CA1, 2 and 3 regions) and CA-4-dentate gyrus.

The Oxford Brain Health Clinic (OBHC) is a joint clinical-research service for patients of the UK National Health Service (NHS) who have been referred to a memory clinic (O'Donoghue et al. 2023). At the OBHC, patients are offered high-quality assessments not routinely available, including a multimodal brain MRI scan acquired on a Siemens 3T Prisma scanner using a protocol matched with the UK Biobank imaging study (Griffanti et al. 2022; Miller et al. 2016). Patients are invited to participate in research and over 90% consented to their clinical data, including subsequent diagnosis, being used for research purposes, representing a real-world clinical dataset. The data are stored on the OBHC Research Database, which was reviewed and approved by the South Central-Oxford C research ethics committee (SC/19/0404). Manual labels for the hippocampi were created for 29 subjects (17 patients who received a dementia diagnosis, 8 MCI patients, and 4 with no dementia-related diagnosis—NDRD).

Manual delineation of the hippocampus for the OBHC dataset followed the approach described by Cook et al. (1992) and Mackay et al. (2000), with reference to the Duvernoy atlas (Duvernoy 1998). High intra-rater test-retest reliability was previously established on a separate dataset segmented twice at least 2 weeks apart (left hippocampus ICC 0.96, 95% confidence interval: 0.92–0.98; right hippocampus ICC 0.87, 95% confidence interval: 0.77–0.93) (Voets et al. 2015). To minimize variations in the segmentation of the hippocampus across subjects, the intensity range of the structural images was set to the same value for all the T1-weighted images. The hippocampus was delineated in all three planes (axial, sagittal and coronal). Anteriorly, the hippocampus could be distinguished from the amygdala in the sagittal view by locating the white matter of the alveus. The posterior aspects of the hippocampus continued up to the splenium of the corpus callosum. Inferiorly, the white matter of the parahippocampal gyrus was excluded, and the lateral boundary extended to the alveus separating the hippocampus from the temporal horn of the lateral ventricle.

## 2.2 | Automatic Segmentation Methods

We evaluated the performance of 10 segmentation algorithms. To be included in our study, the algorithm was required to be freely and publicly available for download. There were six algorithms that performed hippocampus-only segmentations (Hippodeep, Hippmapper, e2dhipseg, HippUnfold, HSF, FreeSurfer-Subfields) and four that created segmentations of a range of brain structures (FreeSurfer, SynthSeg, FastSurfer, FIRST). For comprehensive information on each of the segmentation methods, please see the related publications. All algorithms were run using the default or recommended settings with a raw, full head T1w image as input, on either a MacBook Pro (Apple M2 Pro 2023, 16GB RAM), a machine with a NVIDIA GeForce GTX 1070, running Ubuntu V22.04 or on a dedicated cluster composed of CPU servers (16GB RAM per core), GPU servers (K-series NVidia with CUDA) and Grid Engine queuing software.

**TABLE 1** | Demographic information and hippocampal volume from manual labels for ADNI HarP, MNI-HISUB25 and Oxford Brain Health Clinic (OBHC) datasets.

	ADNI HarP			MNI-HISUB25	OBHC		
	AD	MCI	CN	CN	Dementia	MCI	NDRD
<i>N</i>	44	45	42	25	17	8	4
Age (years)							
Mean (SD)	74 (8)	75 (8)	76 (7)	31 (7)	81 (7)	79 (5)	75 (6)
Range	63–90	60–87	61–90	21–53	72–101	73–86	66–82
Sex							
Female	23 (52%)	19 (42%)	22 (52%)	13 (52%)	7 (41%)	5 (63%)	3 (75%)
Male	21 (48%)	26 (58%)	20 (48%)	12 (48%)	10 (59%)	3 (38%)	1 (25%)
Hippocampal volume (mm <sup>3</sup> )							
Mean (SD)	2417 (530)	2678 (471)	3165 (518)	4386 (246)	2012 (669)	2163 (479)	2844 (554)
Range	1054–3953	1794–3878	2007–5140	3823–5036	1050–3490	1249–2920	2193–3711

*Hippodeep* (Thyreau et al. 2018) was run using the PyTorch version ([https://github.com/bthyreau/hippodeep\\_pytorch](https://github.com/bthyreau/hippodeep_pytorch)) on a CPU (MacBook Pro or cluster). Hippodeep is based on a CNN trained on hippocampus outputs from FreeSurfer that were derived from 2500 images spanning 4 large cohort studies, as well as a small in-house sample of manually labelled ground truth images. The input to the algorithm was a T1w image, and the output of the algorithm was a probabilistic segmentation map, which was thresholded at 0.5 as recommended.

*Hippmapper* (Goubran et al. 2020) version 0.1.1 was run using (<https://github.com/AICONSlab/Hippmapper/tree/master>) on a GPU (GeForce GTX 1070) or via Singularity (cluster). Hippmapper is based on a 3D CNN trained on manual segmentations from 259 older adults with a range of diagnoses leading to extensive atrophy, vascular disease and lesions, including the 135 images from the ADNI HarP dataset. The input to the algorithm was a T1w image, which can either be whole or skull stripped. For these analyses, a whole T1w was provided. The output was a single file containing a binarised, bilateral hippocampal mask, which was then separated into a left and right hippocampus mask file. As Hippmapper was trained using ADNI HarP data, it was excluded from evaluation on the ADNI HarP dataset.

*e2dhipseg* (Carmo et al. 2021) was run on a CPU (MacBook Pro or cluster) using the recommended affine registration option (FLIRT) (<https://github.com/MICLab-Unicamp/e2dhipseg>). The architecture of e2dhipseg consists of an ensemble of 2D CNNs that were trained on the ADNI HarP dataset and 190 images collected locally for an epilepsy study. The algorithm creates a set of segmentations that are then post-processed into a final hippocampal mask. The input to the algorithm was a T1w image, and the output was a single mask file containing both left and right hippocampal masks. The output mask was divided into a left and right hippocampus mask file and then thresholded at 0.5, as recommended. As e2dhipseg was trained using ADNI HarP data, it was excluded from evaluation on the ADNI HarP dataset.

*HippUnfold* (DeKraker et al. 2022) version 1.3.0 is a BIDS app that was run using Docker or Singularity (cluster) (<https://github.com/khanlab/hippunfold>). HippUnfold uses CNNs and topological constraints to generate folded surfaces that correspond to an individual subject's hippocampal morphology and was designed and trained with the Human Connectome Project 1200 young adult data release. The data are first gathered via snakebids before pre-processing, tissue class segmentation, post-processing and unfolding (see DeKraker et al. 2020). As this algorithm is a BIDS app, data from all datasets were first renamed and organised into the BIDS file structure format before running HippUnfold using the default settings. The input is either a T1w image, a T2w image, or both, and several output files are generated including hippocampal masks with subfields, surface metrics and warps. Here, HippUnfold was run with a T1w image as input, and the resulting hippocampal masks with subfields were binarised and combined into a whole hippocampus mask.

*Hippocampal Segmentation Factory* (HSF) version 4.0.0 was run on a GPU (MacBook Pro or cluster) (<https://github.com/clementpoiret/HSF>). The HSF pipeline includes brief pre-processing of raw T1w or T2w images and segmentation of hippocampal subfields through deep learning models trained on images from 12 datasets (411 subjects in total) with manual labels spanning across the chronological age range and multiple pathological groups, including the 25 manual labels from the MNI-HISUB25 dataset. HSF was run using the default and/or recommended settings, including using the ROILoc package (<https://github.com/clementpoiret/ROILoc>) to centre and crop T1w images. The input was a T1w image, and the output includes hippocampal masks with subfields, which were then binarised into a whole hippocampus mask. As HSF was trained using MNI-HISUB25 data, it was excluded from evaluation on the MNI-HISUB25 dataset.

*FIRST* (Patenaude et al. 2011) was run on a CPU (MacBook Pro or cluster). FIRST is a Bayesian appearance model-based tool that incorporates both shape and intensity information



for segmentation, derived from a set of images from 336 subjects, each of which was manually segmented by the Center for Morphometric Analysis (CMA). FIRST was run using the *run\_first\_all* command specifying *L\_Hipp* and *R\_Hipp* as structures to segment. The output mask was then binarised using the recommended threshold.

*FreeSurfer* (Fischl 2012) version 7.4.1 was used. FreeSurfer was also developed using 336 subjects, each manually segmented by the CMA. The recon-all pipeline with default settings was run on all T1 images using parallel processing (computation time approximately 4h per subject on CPU). The steps in the recon-all pipeline are reported in detail elsewhere (Fischl et al. 2002). Hippocampal masks were extracted from the *aseg.auto* output and binarised.

*SynthSeg* (Billot et al. 2023) was run on a CPU (cluster). SynthSeg is a whole brain segmentation CNN trained with synthetic data sampled from a generative model conditioned on segmentations, using a domain randomisation strategy that fully randomises the contrast and resolution of the synthetic training data. As the training images are generated with fully randomised parameters, the network is exposed to combinations of morphological variability, resolution, contrast, noise and artefact, allowing the method to be used without retraining or fine-tuning. Hippocampal masks were extracted from the whole brain segmentation and binarised.

*FreeSurfer Hippocampal Subfields and Nuclei of Amygdala* (Iglesias et al. 2015) (from now on referred to as ‘FreeSurfer-Subfields’) was developed using manual labels from ex and in vivo samples that were combined using a Bayesian inference-based atlas building algorithm to form a single computational atlas. The data included 15 autopsy samples scanned at 0.13mm isotropic resolution that were manually segmented into 13 different hippocampal structures, and a separate in vivo dataset of T1w whole brain MRI (1 mm resolution) containing annotations for neighbouring structures such as the amygdala and cortex. FreeSurfer-Subfields requires a whole brain T1w image that has been processed with recon-all as input. The hippocampus component of the output masks (head, body, tail mask) was then binarised to create a whole hippocampus mask.

*FastSurfer* (Henschel et al. 2020) version 2.2.0 (<https://github.com/Deep-MI/FastSurfer/>) was run using Docker or Singularity (cluster). FastSurfer is a deep learning-based method that aims to present an alternative to FreeSurfer with a shorter run time. The FastSurferCNN architecture contains 3 CNNs operating on coronal, axial and sagittal 2D slices and is trained on FreeSurfer parcellation following the Desikan–Killiany–Tourville protocol atlas. The *seg\_only* function was used to generate subcortical segmentations from a T1w image, and from these, hippocampal masks were extracted and binarised.

For FreeSurfer, FreeSurfer-Subfields and FastSurfer, post-processing steps were required to allow the hippocampal masks to be compared with other segmentation algorithms. To do this, hippocampal masks were transferred back to the original image space using *mri\_label2vol* and converted from mgz to NiFTI

format using the function *mri\_convert*. For these methods, hippocampal volumes were extracted from the relevant *aseg.stats* output file.

## 2.3 | Validation Metrics

Segmentation performance of each segmentation method was evaluated using common medical imaging segmentation metrics: Dice coefficient, 95th percentile Hausdorff distance (HD) and Pearson correlation between manual and automatically segmented volumes. Dice coefficients, HD and volume similarity were implemented using the *seg-metrics* Python package (Jia et al. 2024).

Dice coefficient measures the overlap between two binary sets. It ranges from 0 to 1, where 1 indicates a perfect overlap between the ground truth (manually segmented) and predicted (automatically segmented) masks. HD was used as a metric to assess similarity in shape between the manual and automatic segmentations and is defined as the 95th percentile of the distances between the closest points in an automatic segmentation mask and a manual segmentation mask. HD95 is measured in millimetres, with a value of 0mm indicating perfect prediction. Volume similarity is a measure of the volume size difference between the automatic segmentation mask and the manual segmentation mask. A positive volume similarity value indicates an overestimation of volume in the automatic segmentation mask, a negative value indicates underestimation, and a 0 value indicates perfect volume prediction.

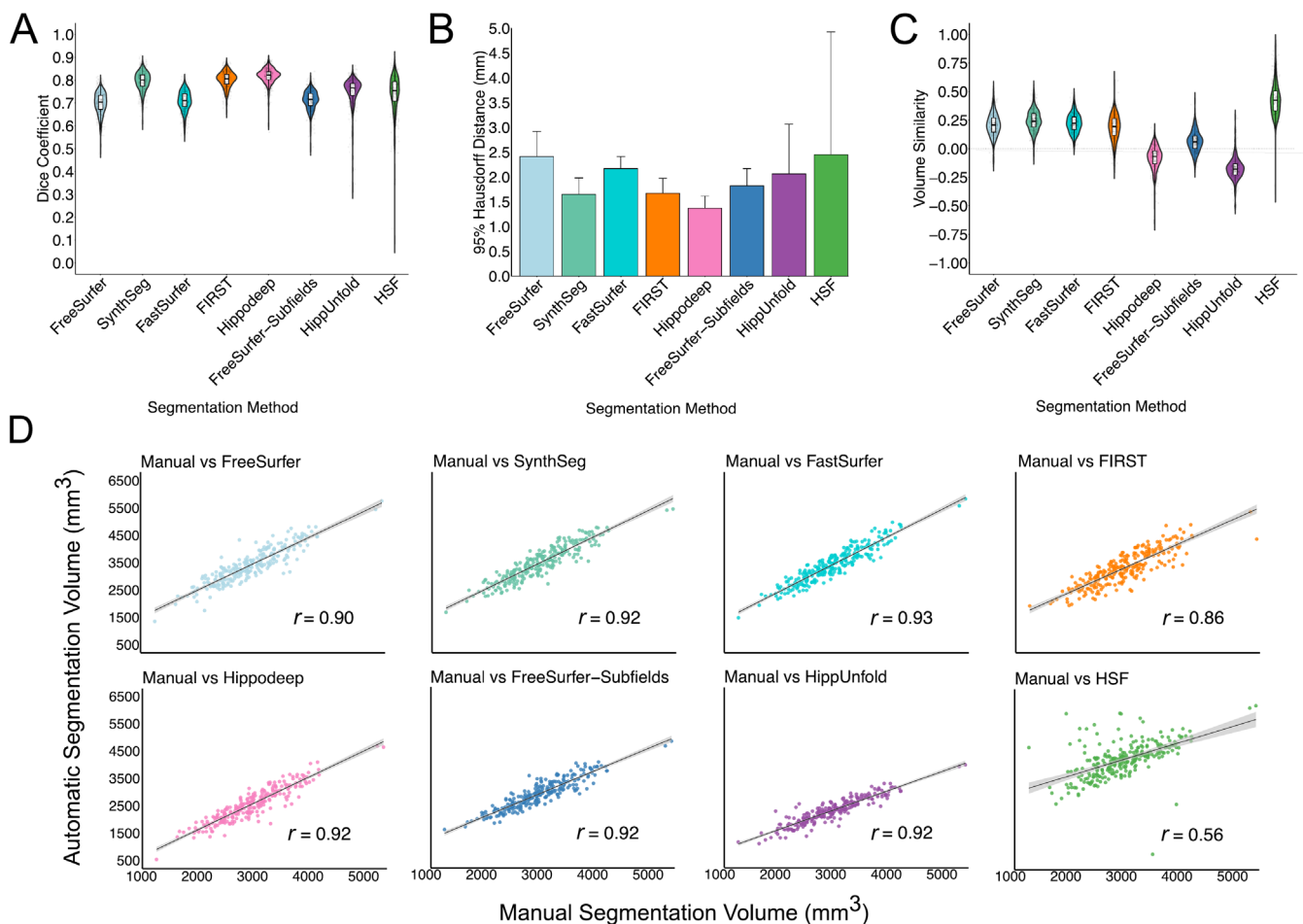
## 2.4 | Error Maps

Average error maps for each segmentation method were also created to understand where each method was systematically incorrect in specific regions of the segmentation. False positive error maps were calculated by subtracting the manual mask from the automatic mask, using *fslmaths*, and keeping only positive values, while for false negatives the negative values were kept. Individual T1w images were linearly (*FLIRT*) and non-linearly (*FNIRT*) registered to MNI space, and the resulting warp fields were applied to transform individual false positive and false negative error maps into MNI space. The false negatives and false positives were then averaged and shown on the T1 1mm MNI template for visualisation.

To calculate the number of false positive and false negative voxels in the anterior and posterior hippocampus, manual labels for each dataset were registered to MNI space and averaged for each hemisphere. The y-coordinate of the centre of gravity was computed from these average masks. Voxels with a y-coordinate greater than the centre of gravity were classified as anterior, while those with a y-coordinate less than the centre were classified as posterior.

## 2.5 | Statistics

Statistical analyses and data visualisations were performed using R (R Version 4.3.1) and ggplot2 (Wickham 2016).



**FIGURE 1** | (A) Dice coefficients, (B) mean 95% Hausdorff distance and (C) volume similarity for automatic segmentation methods compared with manual hippocampal masks from the ADNI HarP dataset. (D) Correlations between manually segmented and automatically segmented volumes for each segmentation method, averaged over diagnostic groups and hemispheres.

Linear mixed effects models run using lme4 (Bates et al. 2015) were used to determine whether there were diagnostic group differences in hippocampal volume across segmentation methods. In these analyses, hippocampal volume was the outcome variable, chronological age, sex, segmentation method and diagnostic group were fixed effects, and subject was the random effect. Post hoc pairwise *t*-tests were performed to explore group differences within segmentation methods. In these tests, a *p* value of less than 0.05 was considered statistically significant. Associations between manually and automatically segmented volumes were performed using Pearson's correlation.

Data are presented as mean  $\pm$  SD in text, mean (SD) in tables, and mean  $\pm$  SD in figures unless otherwise stated.

### 3 | Results

#### 3.1 | ADNI HarP

##### 3.1.1 | Segmentation Metrics

Hippodeep, FIRST and SynthSeg demonstrated comparable performance, yielding mean Dice scores of  $0.82 \pm 0.03$ ,  $0.80 \pm 0.03$  and  $0.79 \pm 0.04$ , respectively (Figure 1A). The hippocampal

subfield methods demonstrated similar performance based on mean Dice, but HippUnfold ( $0.75 \pm 0.07$ ) and HSF ( $0.74 \pm 0.09$ ) showed larger variance than FreeSurfer-Subfields ( $0.71 \pm 0.04$ ). FastSurfer ( $0.71 \pm 0.04$ ) and FreeSurfer ( $0.70 \pm 0.05$ ) demonstrated the poorest mean Dice scores, but relatively tight Dice distributions. For 95th percentile HD, Hippodeep demonstrated the smallest ( $1.37 \pm 0.25$ ) while HSF demonstrated the largest value ( $2.45 \pm 2.47$ ) (Figure 1B).

Detailed results, including Dice coefficients and 95th percentile HD for each group and hemisphere, per segmentation method, are provided in Table 2. For a breakdown of segmentation metrics per field strength (1.5T and 3T) and scanner vendors (Philips, Siemens and GE) for each diagnosis group, see Tables S1–S3.

##### 3.1.2 | Volumes

FreeSurfer-Subfields and Hippodeep demonstrated volume similarities close to 0, and strong correlations with manual volumes, indicating accurate volume estimates ( $0.06 \pm 0.08$  and  $-0.07 \pm 0.10$ , respectively) (Table 3). HSF displayed the weakest correlations, with coefficients ranging from 0.43 to 0.82 across groups and hemispheres, indicating variable discrepancies between manual and automatic volume estimates, particularly for

**TABLE 2** | Approximate run times (CPU) and the approximate output file size for each segmentation method.

Segmentation method	Approximate run time (CPU)	Approximate output file size (single run)	Notes
FreeSurfer	4h	300MB	Parallel flag added to recon-all call
SynthSeg	2min	35MB	
FastSurfer	5min	30MB	Segmentation only mode, no_cereb flag used
FIRST	1min	<1MB	Specified hippocampus only
e2dhipseg	5min	<1MB	
Hippmapper	3min	60MB	
Hippodeep	7s	<1MB	
FreeSurfer-Subfields	35min	300MB	Requires full recon-all output
HippUnfold	60min	2GB	Requires data to be in BIDS format
HSF	3min	<1MB	

the AD group. Similarly, the volume similarity values ranged between 0.32 and 0.53 across groups and hemispheres for HSF, indicating significant overestimation of volumes, particularly in the AD group. Apart from HSF, the segmentation methods demonstrated relatively consistent performance across diagnostic groups and hemispheres based on correlations, but FreeSurfer, SynthSeg, FastSurfer and FIRST tended to overestimate volumes, while HippUnfold tended to underestimate volumes.

### 3.1.3 | Error Maps

Consistent with the positive volume similarity values, all segmentation methods, except for HippUnfold, primarily showed false positives in the error maps, suggesting a systematic over-segmentation, particularly along the hippocampus-amygdala border. This tendency is most pronounced in the anterior hippocampal region in the axial slice (Figure 2A,B). HippUnfold displayed relatively high numbers of consistently located false negatives, particularly in the anterior hippocampus region (Figure 2C).

## 3.2 | MNI-HISUB25

### 3.2.1 | Segmentation Metrics

Hippmapper ( $0.86 \pm 0.020$ ) achieved the highest Dice coefficient, followed by Hippodeep ( $0.85 \pm 0.02$ ) and FIRST ( $0.83 \pm 0.021$ ). HippUnfold ( $0.79 \pm 0.03$ ), FastSurfer ( $0.79 \pm 0.02$ ), FreeSurfer-Subfields ( $0.78 \pm 0.03$ ) and FreeSurfer ( $0.76 \pm 0.03$ ) performed comparably, while e2dhipseg ( $0.68 \pm 0.12$ ) yielded the poorest results (Figure 3A). Similarly, Hippmapper exhibited the smallest 95th percentile HD ( $1.40 \pm 0.17$ ), followed by FreeSurfer-Subfields ( $1.52 \pm 0.25$ ). Consistent with the Dice value and distribution, e2dhipseg performed the poorest ( $4.94 \pm 6.20$ ) (Figure 3B). Overall, there was far less variability in performance between segmentation methods in this dataset compared to ADNI HarP and OBHC, likely due to the lack of patient

groups. Segmentation metrics separated by hemisphere are presented in Table 4.

### 3.2.2 | Volumes

Correlations between manually and automatically segmented volumes notably weaken in the MNI-HISUB25 dataset compared to the ADNI HarP and OBHC datasets (Figure 3D). The strongest correlation between manual and automatic volumes was observed for HippUnfold ( $r=0.82$ ), while no correlation was evident for e2dhipseg ( $r=-0.07$ ). In this dataset, considerable differences in correlation coefficients between hemispheres were observed for certain methods, particularly in FIRST (right hemisphere:  $r=0.61$ , left hemisphere:  $r=0.43$ ), FreeSurfer (right hemisphere:  $r=0.61$ , left hemisphere:  $r=0.48$ ) and FreeSurfer-Subfields (right hemisphere:  $r=0.55$ , left hemisphere:  $r=0.34$ ) (Table 3). Interestingly, there is a disconnect between Dice coefficients and volume correlations. Though FIRST was a strong performer based on Dice, the correlation between manual and automatic volumes was relatively weak ( $r=0.54$ ) compared to HSF ( $r=0.88$ ), Hippmapper ( $r=0.79$ ) and Hippodeep ( $r=0.78$ ).

Volume similarity scores also vary within this dataset (Figure 3C). SynthSeg ( $0.22 \pm 0.05$ ), FastSurfer ( $0.21 \pm 0.05$ ) and FIRST ( $0.17 \pm 0.07$ ) demonstrated overestimation of volumes, while HippUnfold ( $-0.20 \pm 0.04$ ) and e2dhipseg ( $-0.54 \pm 0.27$ ) demonstrated underestimation. Hippodeep ( $-0.05 \pm 0.05$ ), Hippmapper ( $-0.05 \pm 0.05$ ), FreeSurfer ( $0.08 \pm 0.08$ ) and FreeSurfer-Subfields ( $0 \pm 0.08$ ) demonstrated volume similarity scores close to 0.

### 3.2.3 | Error Maps

Figure 4 illustrates the systematic false positives and negatives for each segmentation method in the MNI-HISUB25 dataset. Hippmapper exhibited the fewest consistently located false positives and negatives, although all methods display false negatives

**TABLE 3** | Mean volume, correlation coefficients, mean volume similarity, mean Dice coefficient and mean 95% Hausdorff distance across diagnostic groups and hemisphere for the ADNI HarP dataset.

	FreeSurfer	SynthSeg	FastSurfer	FIRST	Hippodeep	FreeSurfer-Subfields	HippUnfold	HSF
Left hemisphere volume (mm <sup>3</sup> )								
AD	3019 (583)	3078 (564)	2971 (562)	2879 (581)	2096 (547)	2537 (471)	2007 (434)	4021 (633)
MCI	3340 (508)	3405 (501)	3362 (496)	3153 (467)	2454 (452)	2833 (451)	2232 (387)	4042 (557)
CN	3734 (590)	3891 (524)	3841 (578)	3636 (521)	2922 (512)	3213 (489)	2582 (388)	4384 (805)
Right hemisphere volume (mm <sup>3</sup> )								
AD	3083 (646)	3174 (608)	3094 (620)	3041 (613)	2240 (614)	2649 (505)	2077 (448)	4236 (659)
MCI	3420 (542)	3485 (537)	3407 (525)	3300 (535)	2549 (483)	2943 (487)	2283 (390)	4202 (544)
CN	3785 (550)	3946 (545)	3887 (563)	3775 (534)	3044 (514)	3315 (432)	2599 (405)	4488 (589)
Left hemisphere correlation with manual volume ( <i>r</i> )								
AD	0.81	0.90	0.92	0.88	0.89	0.90	0.84	0.43
MCI	0.85	0.93	0.91	0.82	0.89	0.85	0.93	0.82
CN	0.88	0.87	0.88	0.72	0.90	0.85	0.93	0.48
Right hemisphere correlation with manual volume ( <i>r</i> )								
AD	0.90	0.89	0.93	0.76	0.86	0.92	0.88	0.46
MCI	0.92	0.92	0.94	0.84	0.92	0.91	0.92	0.64
CN	0.89	0.84	0.89	0.82	0.90	0.88	0.90	0.48
Left hemisphere volume similarity								
AD	0.25 (0.13)	0.28 (0.10)	0.24 (0.09)	0.21 (0.10)	−0.12 (0.11)	0.09 (0.09)	−0.15 (0.12)	0.53 (0.17)
MCI	0.23 (0.09)	0.26 (0.07)	0.25 (0.08)	0.18 (0.09)	−0.07 (0.09)	0.07 (0.09)	−0.16 (0.07)	0.43 (0.09)
CN	0.17 (0.08)	0.22 (0.08)	0.21 (0.08)	0.15 (0.11)	−0.07 (0.08)	0.03 (0.09)	−0.18 (0.06)	0.32 (0.27)
Right hemisphere volume similarity								
AD	0.21 (0.10)	0.25 (0.11)	0.22 (0.09)	0.20 (0.16)	−0.12 (0.15)	0.07 (0.10)	−0.18 (0.12)	0.52 (0.19)
MCI	0.22 (0.07)	0.25 (0.07)	0.23 (0.06)	0.19 (0.10)	−0.07 (0.07)	0.08 (0.07)	−0.17 (0.07)	0.43 (0.13)
CN	0.16 (0.07)	0.21 (0.08)	0.19 (0.07)	0.16 (0.09)	−0.05 (0.08)	0.04 (0.07)	−0.20 (0.07)	0.33 (0.15)
Left hemisphere Dice coefficient								
AD	0.67 (0.05)	0.77 (0.04)	0.69 (0.04)	0.79 (0.03)	0.80 (0.03)	0.69 (0.04)	0.71 (0.08)	0.69 (0.09)
MCI	0.69 (0.04)	0.79 (0.03)	0.70 (0.04)	0.80 (0.03)	0.82 (0.03)	0.71 (0.04)	0.76 (0.04)	0.75 (0.05)
CN	0.74 (0.03)	0.82 (0.03)	0.74 (0.03)	0.82 (0.03)	0.84 (0.02)	0.74 (0.03)	0.78 (0.02)	0.76 (0.12)
Right hemisphere Dice coefficient								
AD	0.67 (0.05)	0.78 (0.04)	0.68 (0.04)	0.79 (0.04)	0.80 (0.04)	0.69 (0.05)	0.71 (0.10)	0.70 (0.09)
MCI	0.69 (0.04)	0.80 (0.03)	0.70 (0.04)	0.80 (0.03)	0.82 (0.03)	0.71 (0.04)	0.75 (0.06)	0.75 (0.06)
CN	0.73 (0.02)	0.82 (0.02)	0.74 (0.03)	0.82 (0.03)	0.83 (0.02)	0.74 (0.02)	0.78 (0.03)	0.78 (0.06)
Left hemisphere 95% Hausdorff distance (mm)								
AD	2.77 (0.82)	1.81 (0.50)	2.26 (0.32)	1.62 (0.24)	1.41 (0.21)	1.95 (0.49)	2.47 (1.41)	2.69 (1.35)
MCI	2.48 (0.46)	1.73 (0.27)	2.25 (0.23)	1.69 (0.27)	1.38 (0.26)	1.94 (0.32)	1.89 (0.78)	2.08 (0.43)
CN	2.20 (0.20)	1.56 (0.23)	2.11 (0.14)	1.70 (0.37)	1.30 (0.21)	1.82 (0.29)	1.67 (0.31)	2.78 (5.02)

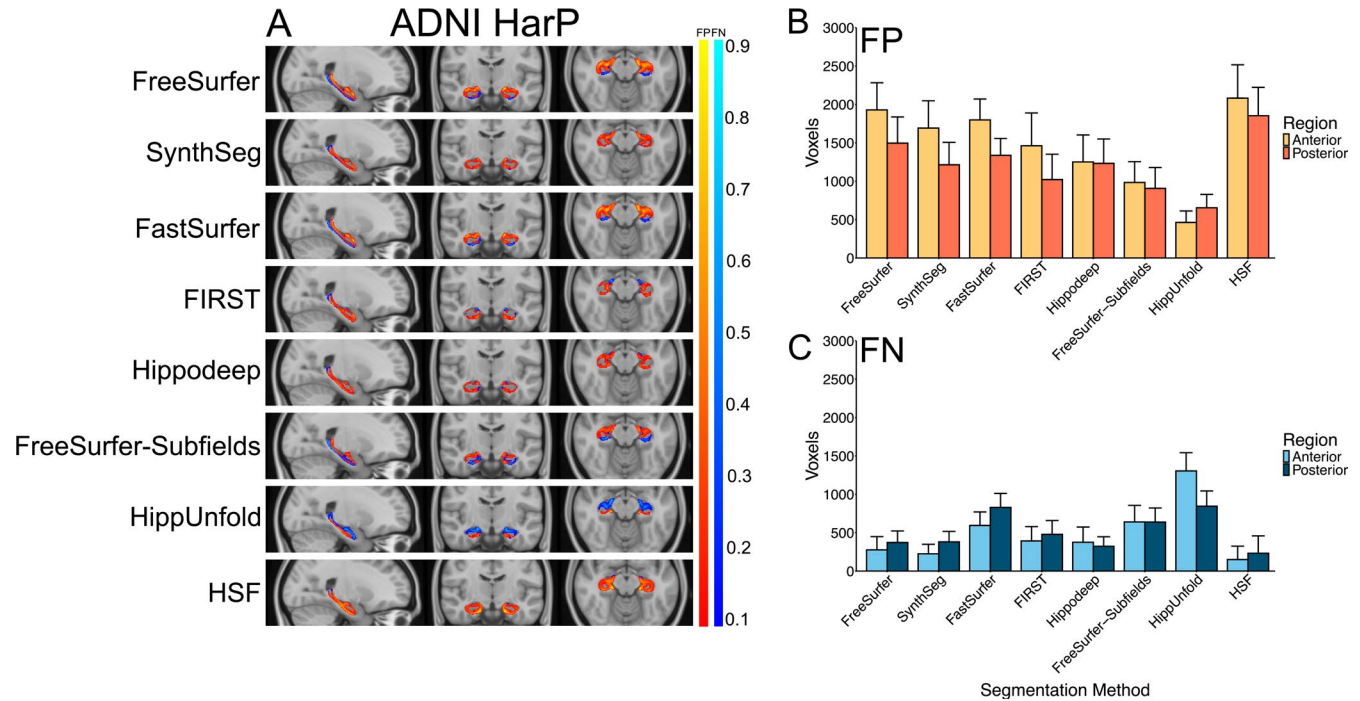
(Continues)



TABLE 3 | (Continued)

	FreeSurfer	SynthSeg	FastSurfer	FIRST	Hippodeep	FreeSurfer-Subfields	HippUnfold	HSF
Left hemisphere 95% Hausdorff distance (mm)								
AD	2.46 (0.42)	1.63 (0.35)	2.18 (0.19)	1.69 (0.37)	1.46 (0.34)	1.79 (0.32)	2.45 (1.25)	2.66 (1.18)
MCI	2.40 (0.41)	1.64 (0.29)	2.18 (0.28)	1.72 (0.29)	1.37 (0.22)	1.74 (0.28)	2.06 (1.01)	2.11 (0.69)
CN	2.12 (0.20)	1.51 (0.18)	2.02 (0.16)	1.60 (0.25)	1.30 (0.19)	1.68 (0.23)	1.78 (0.36)	2.39 (3.03)

Note: Data are displayed as Mean (SD), except for when presenting Pearson correlation coefficients.



**FIGURE 2** | (A) False positive (FP) (red-yellow) and false negative (FN) (blues) heat maps for each segmentation method projected onto the standard MNI152 T1 template image for ADNI HarP. The colour bar represents the proportion of times a given voxel was incorrectly labelled in comparison to the manually segmented hippocampal mask. (B) Number of false positive and (C) false negative voxels in the anterior and posterior hippocampus regions for each segmentation method, averaged over hemispheres and diagnostic groups.

along the superior-medial or inferior-medial border of the hippocampus. Consistent with the positive volume similarity values, FreeSurfer, SynthSeg, FastSurfer, FIRST and Hippodeep all demonstrate systematic over-segmentation, predominantly in anterior regions (Figure 4A,B). In line with its low Dice value, negative volume similarity score and significant performance variation, e2dhipseg exhibits a high number of consistently located false negatives throughout the entire structure, indicating widespread under-segmentation (Figure 4C).

### 3.3 | Oxford Brain Health Clinic

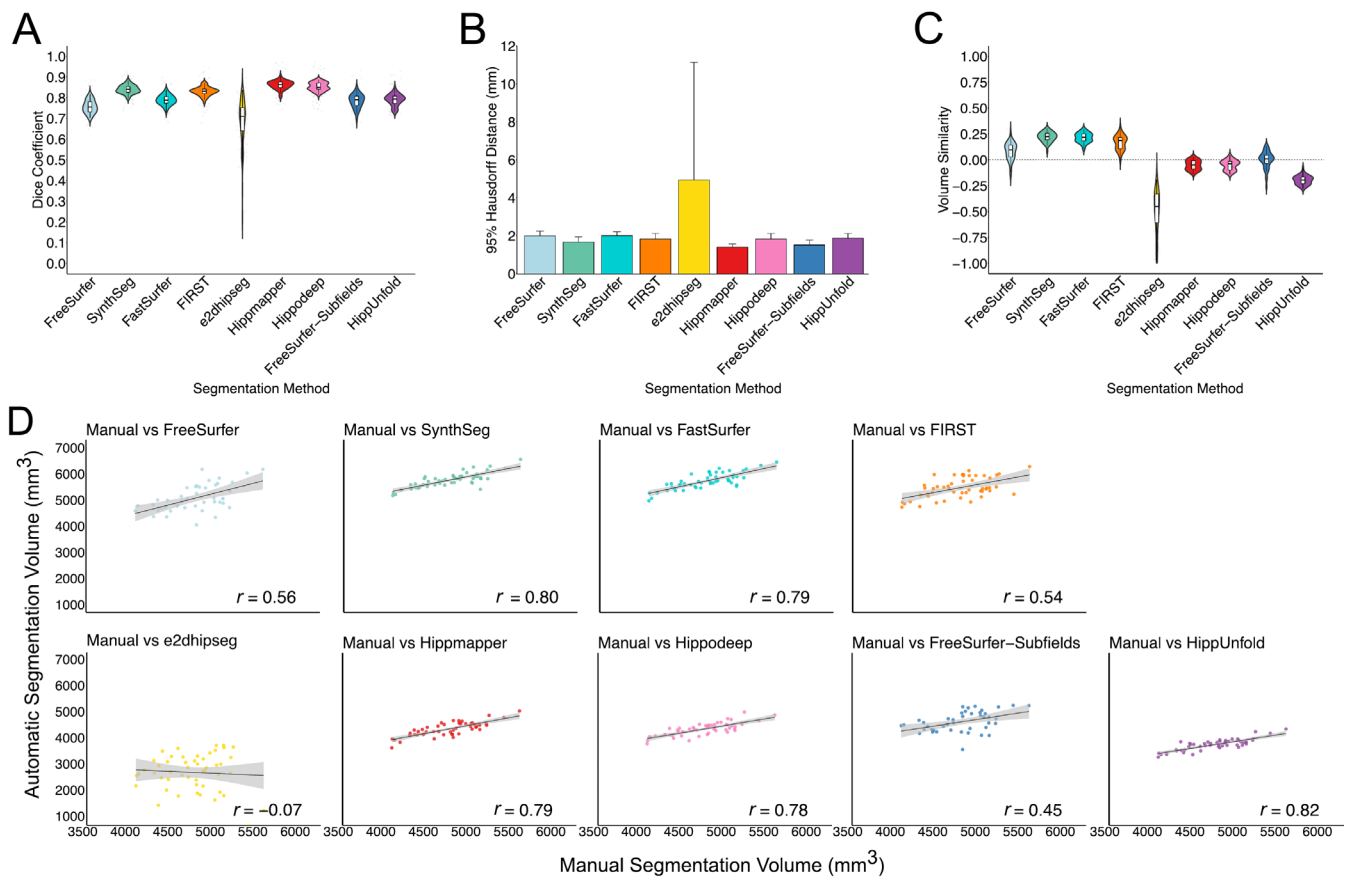
#### 3.3.1 | Segmentation Metrics

On average, Hippodeep ( $0.76 \pm 0.06$ ) and FIRST ( $0.76 \pm 0.12$ ) exhibited the highest Dice coefficients, followed closely by FreeSurfer-Subfields ( $0.74 \pm 0.06$ ), although notable variation was observed in the Dice scores for FIRST. HippUnfold ( $0.72 \pm 0.09$ ), FastSurfer ( $0.71 \pm 0.06$ ), FreeSurfer ( $0.69 \pm 0.07$ ),

Hippmapper ( $0.69 \pm 0.18$ ) and SynthSeg ( $0.67 \pm 0.09$ ) showed similar Dice values, with HippUnfold and Hippmapper displaying large distribution tails (Figure 3A). HSF ( $0.58 \pm 0.15$ ) and e2dhipseg ( $0.44 \pm 0.26$ ) were the poorest performers based on Dice, with both exhibiting substantial distribution tails attributed to failure rates. Conversely, the 95% HD was smallest for HippUnfold ( $2.32 \pm 1.09$ ) and, as expected based on Dice, largest for e2dhipseg ( $6.90 \pm 5.04$ ) (Figure 5B). Additionally, in this dataset, FreeSurfer recon-all failed for two subjects who were removed from subsequent analysis. Overall, segmentation performance was considerably poorer in this dataset compared to ADNI HarP and MNI-HISUB25.

#### 3.3.2 | Volumes

Hippocampal volumes for each group and hemisphere, along with the correlation between manually segmented and automatically segmented volumes, are detailed in Table 4. FastSurfer and Hippodeep demonstrated the strongest



**FIGURE 3** | (A) Dice coefficients, (B) mean 95% Hausdorff distance and (C) volume similarity for automatic segmentation methods compared with manual hippocampal masks from the MNI-HISUB25 dataset. (D) Correlations between manually segmented and automatically segmented volumes for each segmentation method, averaged over hemispheres.

correlation between manual and automatically segmented volumes on average (both  $r=0.93$ ) (Figure 3C), with both methods achieving correlation coefficients  $>0.85$  across hemispheres and diagnostic groups. HSF demonstrated the weakest correlation between manual and automatic volumes ( $r=0.52$ ), despite achieving correlation coefficients of 0.93 and 0.96 for the left and right hemispheres, respectively, in NDRD subjects. There is greater variation in correlation coefficients between diagnostic groups and hemispheres in the OBHC dataset compared to ADNI HarP. For example, FreeSurfer-Subfields demonstrated a correlation of  $r=0.65$  and  $r=0.79$  for subjects with an MCI diagnosis in the left and right hemispheres, respectively, but all other correlations were  $>0.89$ , while e2dhipseg performed particularly poorly in NDRD subjects compared with MCI and dementia groups (Table 5).

Volume similarity scores also highlighted a tendency towards volume over-estimation in the OBHC dataset. e2dhipseg ( $0.62 \pm 0.21$ ) demonstrated the largest degree of volume over-estimation, followed by FreeSurfer ( $0.45 \pm 0.16$ ), SynthSeg ( $0.45 \pm 0.21$ ) and FastSurfer ( $0.44 \pm 0.14$ ). Hippodeep ( $0.32 \pm 0.12$ ) and FreeSurfer-Subfields ( $0.32 \pm 0.14$ ) demonstrated similar and relatively consistent levels of over-segmentation, followed by FIRST ( $0.23 \pm 0.17$ ). HippUnfold demonstrated a relatively tight distribution and a volume similarity score close to zero ( $-0.01 \pm 0.14$ ). There was significant variability in the volume similarity scores for both Hippmapper ( $-0.09 \pm 0.32$ ) and HSF ( $0.38 \pm 0.38$ ), suggesting that both under- and over-segmentation

were present within the dataset, consistent with the variability in Dice scores.

### 3.3.3 | Error Maps

Figure 6 illustrates the systematic false positives and negatives for each segmentation method in the OBHC dataset. Like in ADNI HarP and MNI-HISUB25, most segmentation methods demonstrated systematic over-segmentation, particularly in the anterior hippocampal region (Figure 6A,B). Hippmapper and HippUnfold exhibit the fewest consistently located false positives but showed a larger number of false negatives relative to other methods (Figure 6C). In line with its low Dice value, e2dhipseg exhibits a high number of consistently located false positives and negatives throughout the entire structure, indicating both over-segmentation and under-segmentation (Figure 6C). Consistent with the positive volume similarity values, FreeSurfer, SynthSeg, FastSurfer, FIRST and Hippodeep all demonstrate a similar level of systematic over-segmentation, predominantly in anterior regions (Figure 4A,B).

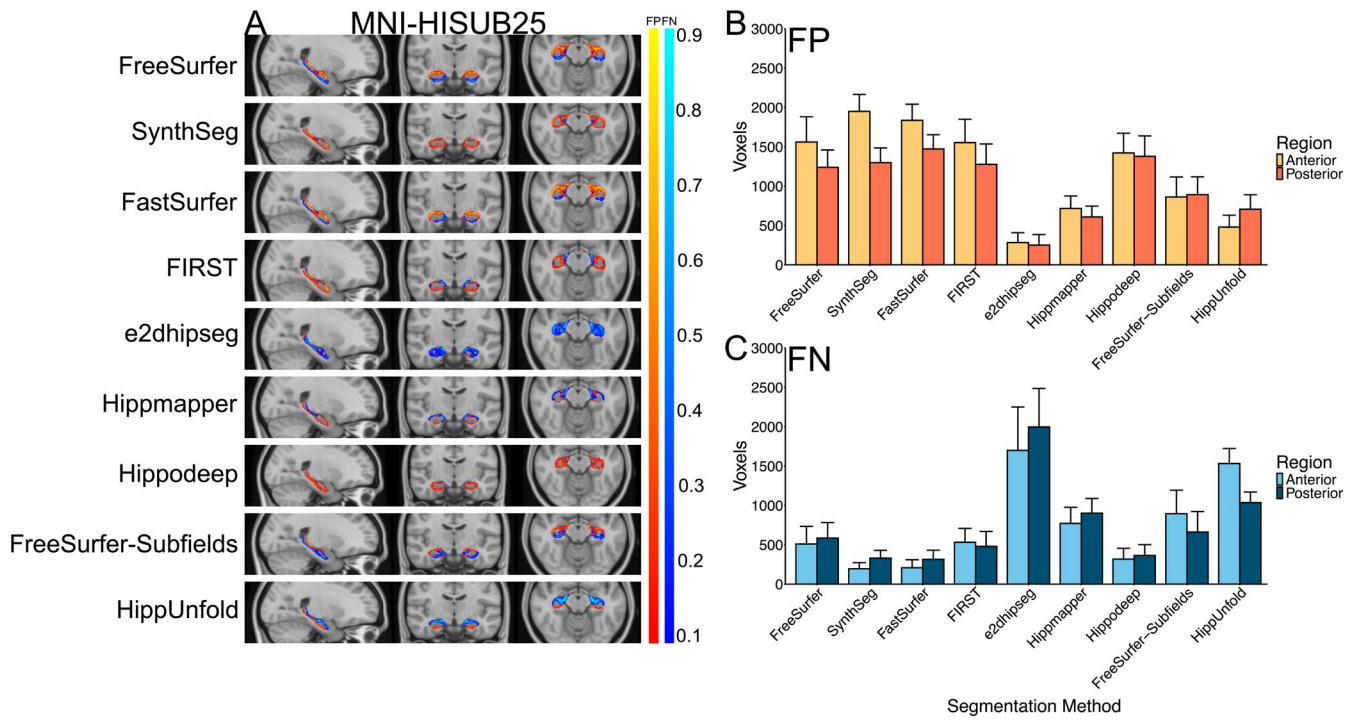
## 3.4 | Sensitivity to Diagnosis Groups

As assessing volumetric changes of the hippocampus is an important target in clinical contexts, we also sought to determine

**TABLE 4** | Mean volume, correlation coefficients, mean Dice coefficient and mean 95% Hausdorff distance across hemisphere for the MNI-HISUB25 dataset.

	FreeSurfer	SynthSeg	FastSurfer	FIRST	e2dhipseg	Hippmapper	Hippodeep	FreeSurfer-Subfields	HippUnfold
Left hemisphere volume (mm <sup>3</sup> )									
CN	5000 (475)	5709 (239)	5701 (293)	5401 (368)	2590 (747)	4331 (263)	4354 (239)	4616 (391)	3767 (199)
Right hemisphere volume (mm <sup>3</sup> )									
CN	5144 (533)	5816 (310)	5745 (333)	5522 (414)	2846 (625)	4443 (271)	4418 (254)	4708 (370)	3820 (235)
Left hemisphere correlation with manual volume ( <i>r</i> )									
CN	0.48	0.80	0.81	0.43	-0.09	0.74	0.79	0.34	0.80
Right hemisphere correlation with manual volume ( <i>r</i> )									
CN	0.61	0.80	0.77	0.61	-0.11	0.84	0.76	0.55	0.84
Left hemisphere volume similarity									
CN	0.08 (0.09)	0.22 (0.05)	0.22 (0.05)	0.17 (0.08)	-0.58 (0.30)	-0.05 (0.05)	-0.05 (0.05)	0.00 (0.10)	-0.19 (0.05)
Right hemisphere volume similarity									
CN	0.09 (0.08)	0.22 (0.05)	0.21 (0.05)	0.17 (0.07)	-0.50 (0.24)	-0.05 (0.04)	-0.05 (0.05)	0.00 (0.07)	-0.20 (0.04)
Left hemisphere Dice coefficient									
CN	0.76 (0.03)	0.84 (0.02)	0.79 (0.02)	0.83 (0.02)	0.66 (0.13)	0.86 (0.02)	0.85 (0.02)	0.78 (0.03)	0.80 (0.02)
Right hemisphere Dice coefficient									
CN	0.76 (0.03)	0.84 (0.02)	0.79 (0.03)	0.83 (0.02)	0.70 (0.11)	0.86 (0.02)	0.85 (0.02)	0.79 (0.03)	0.78 (0.03)
Left hemisphere 95% Hausdorff distance (mm)									
CN	2.04 (0.17)	1.68 (0.28)	2.05 (0.22)	1.87 (0.33)	6.22 (7.85)	1.44 (0.14)	1.87 (0.33)	1.53 (0.18)	1.85 (0.22)
Right hemisphere 95% Hausdorff distance (mm)									
CN	1.95 (0.31)	1.66 (0.29)	1.96 (0.21)	1.78 (0.25)	3.66 (3.70)	1.37 (0.19)	1.78 (0.25)	1.51 (0.31)	1.90 (0.27)

Note: Data are displayed as Mean (SD), except for when presenting Pearson correlation coefficients.



**FIGURE 4** | (A) False positive (FP) (red-yellow) and false negative (FN) (blues) heat maps for each segmentation method projected onto the standard MNI152 T1 template image. The colour bar represents the proportion of times a given voxel was incorrectly labelled in comparison to the manually segmented hippocampal mask from the MNI-HISUB25 dataset. (B) Number of false positive and (C) false negative voxels in the anterior and posterior hippocampus regions for each segmentation method (averaged over hemispheres).

whether each segmentation method could detect changes in hippocampal volume between diagnostic groups in the ADNI HarP and OBHC datasets.

For the ADNI HarP dataset, a linear mixed effects model demonstrated a significant main effect of age ( $F_{1,126} = 19.7$ ,  $p < 0.001$ ), sex ( $F_{3,432} = 629.9$ ,  $p < 0.001$ ), group ( $F_{2,126} = 36.8$ ,  $p < 0.001$ ), segmentation method ( $F_{10,2695} = 1257.7$ ,  $p < 0.001$ ) and a group by segmentation method interaction ( $F_{20,2695} = 7.32$ ,  $p < 0.001$ ).

As a point of comparison, manual segmentation demonstrated that subjects with AD had smaller hippocampal volumes than controls ( $p < 0.001$ ) and MCI subjects ( $p = 0.03$ ), and MCI subjects had smaller hippocampal volumes than controls ( $p < 0.001$ ). All automatic segmentation methods demonstrated smaller volumes in subjects diagnosed with AD compared with controls (all  $p < 0.01$ ). Likewise, all segmentation methods except for HippUnfold ( $p = 0.12$ ) and HSF ( $p = 0.92$ ) demonstrated lower volumes in subjects diagnosed as AD compared with MCI (all  $p < 0.03$ ). Finally, all segmentation methods demonstrated smaller hippocampal volumes in MCI subjects compared to controls (all  $p < 0.01$ ) (Figure 7A).

For the OBHC dataset, the model demonstrated significant main effects of age ( $F_{1,24} = 7.48$ ,  $p = 0.01$ ), sex ( $F_{1,24} = 4.74$ ,  $p = 0.04$ ), group ( $F_{2,24} = 5.00$ ,  $p = 0.02$ ), segmentation method ( $F_{10,567} = 69.04$ ,  $p < 0.001$ ), but no significant group by segmentation method interaction ( $F_{20,567} = 1.15$ ,  $p = 0.29$ ). Though the group by segmentation method interaction was not statistically significant, comparisons between groups for each segmentation method were explored. For manual segmentation, subjects with a dementia diagnosis had smaller hippocampal volumes than

those with no dementia-related diagnosis ( $p = 0.04$ ). However, no significant differences were found in hippocampal volume between subjects with NDRD and MCI ( $p = 0.16$ ), nor between subjects with MCI and dementia patients ( $p = 0.78$ ). All segmentation methods (all  $p < 0.04$ ) except SynthSeg ( $p = 0.38$ ), e2dhipseg ( $p = 0.33$ ) and HippUnfold ( $p = 0.13$ ) found smaller hippocampal volumes for subjects with a dementia diagnosis compared with NDRD. Only FIRST ( $p = 0.04$ ), Hippmapper ( $p = 0.02$ ) and Hippodeep ( $p = 0.04$ ) demonstrated a difference in volume between NDRD and subjects with MCI, while no segmentation methods detected a difference in volume between subjects with a dementia or MCI diagnosis (all  $p > 0.36$ ) (Figure 7B).

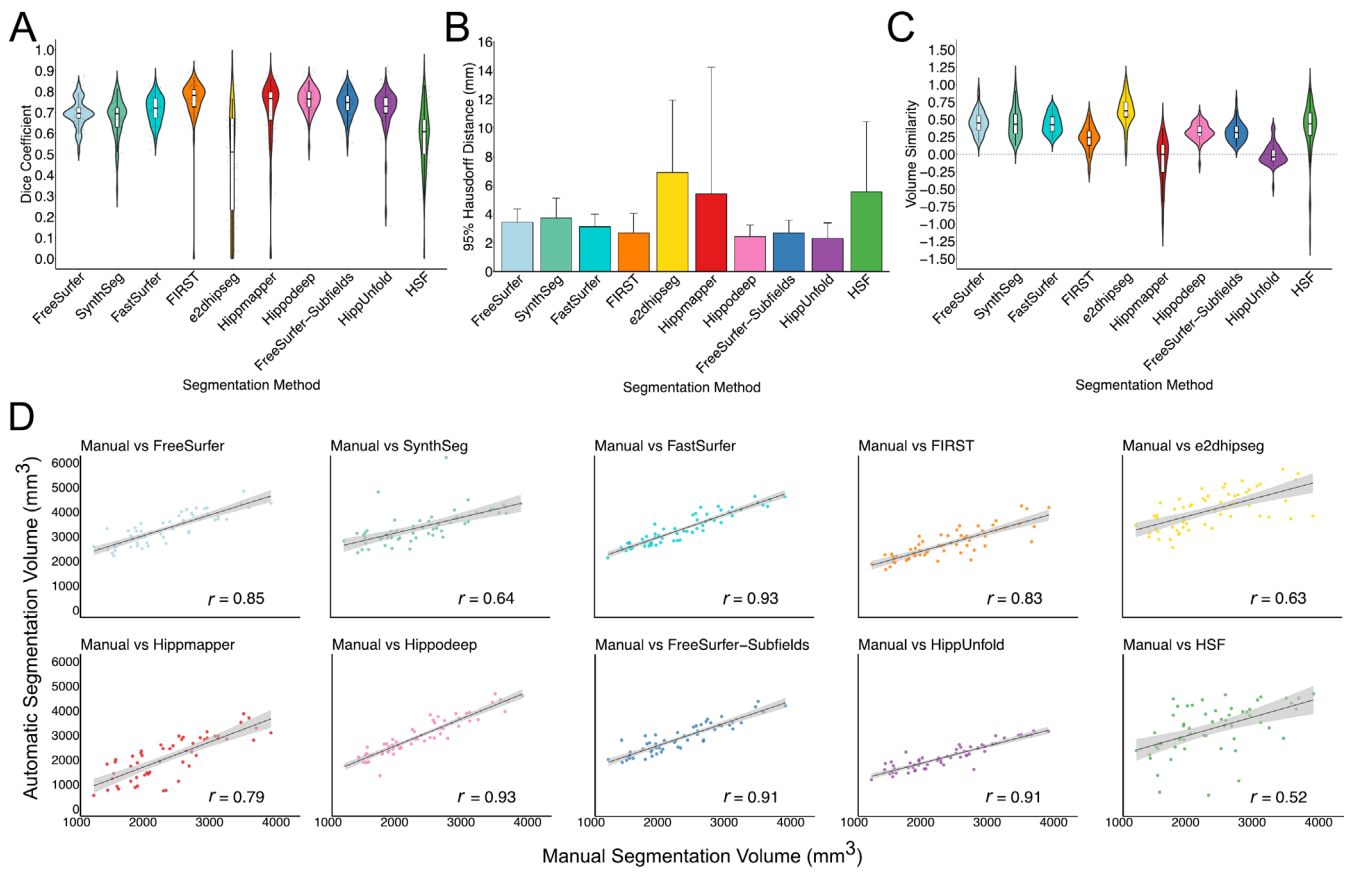
### 3.5 | Performance Summary

Table 6 summarises the performance of the best and worst segmentation methods for each dataset, evaluated across the metrics of Dice, HD95, volume similarity, volume correlation and total voxel error (encompassing both false positives and false negatives), averaged over hemispheres. The best-performing methods for each diagnosis group within a dataset were determined by the highest Dice score, lowest HD95, volume similarity closest to zero, volume correlation closest to 1, and the fewest total voxel errors.

## 4 | Discussion

Here, we investigated the performance of 10 publicly available tools that segment the hippocampus on three datasets with





**FIGURE 5** | (A) Dice coefficients, (B) mean 95% Hausdorff distance for automatic segmentation methods and (C) volume similarity scores compared with manual hippocampal masks from the Oxford Brain Health Clinic dataset. (D) Correlations between manually segmented and automatically segmented volumes for each segmentation method, averaged over diagnostic groups and hemispheres. *Note:* FreeSurfer recon-all failed to run on two subjects.

manual labels: ADNI HarP, MNI-HISUB25 and a real-world memory clinic dataset from the OBHC. Briefly, we found that most segmentation methods: (1) performed consistently and accurately on the two publicly available datasets but were more prone to error and variability on an unseen clinical dataset, (2) are likely to overestimate volumes and systematically over-segment the hippocampus, particularly at the anterior hippocampus border, and (3) can delineate between healthy controls, subjects with a diagnosis of MCI and subjects with a dementia diagnosis based on hippocampal volume.

In evaluating the performance of each of the segmentation methods, we were interested in both accuracy relative to manual labels and reliability, and particularly interested in performance on the OBHC dataset. This dataset presents a more challenging test of generalisation due to its different demographic and clinical characteristics, which was reflected in the generally poorer performance observed across all segmentation methods, compared to that in ADNI HarP and MNI-HISUB25. Being an unselected sample of memory clinic patients, the average age is older than ADNI and most dementia research datasets, and there is a higher amount of atrophy and vascular pathology (Griffanti et al. 2022). Moreover, due to minimal exclusion criteria (related to MR-safety or being too frail to travel to the assessment centre—see O'Donoghue et al. 2023 for details), it is a more representative sample of real-world patients, but also likely more heterogeneous. If the goal is to make automated

measures available in clinical practice, it is important to evaluate the performance of tools in such clinical samples. In this dataset, SynthSeg, FIRST, e2dhipseg, Hippmapper, HippUnfold and HSF exhibited lower mean Dice coefficients and worse tail distributions than in the other datasets, suggesting higher segmentation failure rates. However, Hippodeep, FastSurfer, FreeSurfer-Subfields and HippUnfold performed relatively well on the OBHC data, with strong correlations with manual volumes and better accuracy and consistency than other methods as measured by Dice distributions, potentially suggesting better performance in data collected in real-world clinical settings and populations beyond typical research samples.

Despite the poorer performance on unseen data for deep-learning-based methods that were trained on manual labels, such as e2dhipseg and HSF, not all deep-learning methods exhibited inconsistency across datasets. Hippodeep demonstrated strong performance across all segmentation metrics in all evaluated datasets, with high Dice values, tightly distributed Dice and HD measures, strong volume similarities and robust correlations with manual volumes. Likewise, although slightly underperforming compared with other methods based on mean Dice values, FastSurfer consistently demonstrated relatively tight Dice distributions and strong correlations with manual volumes across all datasets. The consistent performance of these methods can be attributed to training methods that allowed for larger and diverse datasets to be used in the development



**TABLE 5** | Mean volume, correlation coefficients, mean Dice coefficient and mean 95% Hausdorff distance across diagnostic groups and hemisphere for the Oxford Brain Health Clinic dataset.

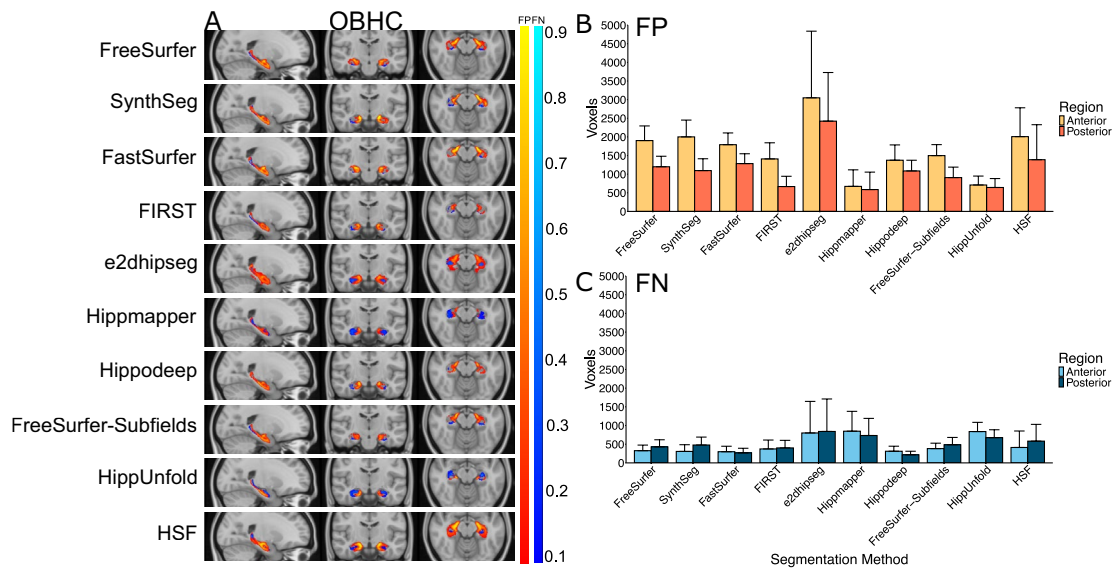
FreeSurfer		SynthSeg	FastSurfer	FIRST	e2dhipseg	Hippmapper	Hippodeep	FreeSurfer-Subfields	HippUnfold	HSF
Left hemisphere volume (mm <sup>3</sup> )										
Dementia	2989 (604)	3600 (911)	3060 (640)	2433 (474)	3862 (699)	1743 (728)	2598 (760)	2541 (539)	1907 (551)	2808 (1092)
MCI	3451 (379)	3284 (275)	3149 (523)	2615 (412)	3845 (580)	2001 (765)	2789 (562)	2983 (421)	1998 (375)	3341 (766)
NDRD	3910 (341)	3830 (352)	3852 (586)	3517 (480)	4349 (920)	2816 (271)	3673 (675)	3453 (527)	2600 (419)	3898 (578)
Right hemisphere volume (mm <sup>3</sup> )										
Dementia	3129 (598)	3042 (551)	3257 (711)	2581 (552)	4084 (748)	1923 (873)	2923 (750)	2715 (624)	2089 (516)	3327 (790)
MCI	3387 (413)	3121 (370)	3340 (501)	2674 (550)	4414 (518)	2328 (778)	3141 (563)	3003 (443)	2184 (308)	3268 (1314)
NDRD	4204 (489)	3790 (360)	4099 (313)	3505 (807)	4542 (623)	3406 (454)	4079 (572)	3819 (496)	2764 (269)	4053 (421)
Left hemisphere correlation with manual volume ( <i>r</i> )										
Dementia	0.88	0.62	0.95	0.78	0.70	0.82	0.91	0.93	0.81	0.43
MCI	0.69	0.82	0.85	0.92	0.91	0.62	0.99	0.65	0.97	0.68
NDRD	0.84	0.89	0.90	0.97	-0.14	0.84	0.96	0.96	0.92	0.96
Right hemisphere correlation with manual volume ( <i>r</i> )										
Dementia	0.85	0.96	0.94	0.90	0.72	0.79	0.94	0.94	0.95	0.43
MCI	0.66	0.55	0.85	0.45	0.48	0.56	0.86	0.79	0.84	0.22
NDRD	0.77	0.93	0.95	0.85	-0.90	0.97	0.95	0.89	0.96	0.96
Left hemisphere volume similarity										
Dementia	0.49 (0.15)	0.63 (0.21)	0.49 (0.14)	0.27 (0.20)	0.68 (0.22)	-0.16 (0.29)	0.31 (0.16)	0.34 (0.14)	0.01 (0.20)	0.32 (0.46)
MCI	0.52 (0.22)	0.48 (0.22)	0.43 (0.17)	0.26 (0.17)	0.61 (0.12)	-0.07 (0.41)	0.31 (0.10)	0.38 (0.23)	-0.01 (0.13)	0.48 (0.21)
NDRD	0.37 (0.16)	0.35 (0.16)	0.35 (0.12)	0.27 (0.11)	0.45 (0.31)	0.05 (0.16)	0.30 (0.08)	0.25 (0.10)	-0.04 (0.11)	0.36 (0.10)
Right hemisphere volume similarity										
Dementia	0.46 (0.16)	0.39 (0.15)	0.45 (0.13)	0.22 (0.15)	0.66 (0.20)	-0.14 (0.33)	0.34 (0.12)	0.32 (0.11)	0.01 (0.14)	0.46 (0.27)
MCI	0.40 (0.11)	0.33 (0.13)	0.39 (0.08)	0.16 (0.20)	0.65 (0.13)	-0.06 (0.41)	0.33 (0.09)	0.29 (0.09)	-0.03 (0.09)	0.25 (0.62)
NDRD	0.35 (0.10)	0.25 (0.08)	0.33 (0.09)	0.16 (0.14)	0.42 (0.28)	0.14 (0.05)	0.32 (0.05)	0.26 (0.07)	-0.06 (0.08)	0.32 (0.07)
Left hemisphere Dice coefficient										
Dementia	0.69 (0.06)	0.60 (0.11)	0.70 (0.07)	0.70 (0.19)	0.51 (0.19)	0.71 (0.09)	0.74 (0.07)	0.74 (0.05)	0.70 (0.14)	0.55 (0.13)

(Continues)

TABLE 5 | (Continued)

	FreeSurfer	SynthSeg	FastSurfer	FIRST	e2dhipseg	Hippmapper	Hippodeep	FreeSurfer-Subfields	HippUnfold	HSF
MCI	0.70 (0.09)	0.67 (0.08)	0.74 (0.06)	0.78 (0.07)	0.55 (0.28)	0.73 (0.16)	0.77 (0.06)	0.75 (0.09)	0.75 (0.06)	0.57 (0.15)
NDRD	0.69 (0.09)	0.74 (0.06)	0.69 (0.07)	0.81 (0.05)	0.20 (0.16)	0.81 (0.02)	0.80 (0.04)	0.73 (0.06)	0.77 (0.03)	0.69 (0.14)
Right hemisphere Dice coefficient										
Dementia	0.71 (0.05)	0.69 (0.06)	0.72 (0.06)	0.77 (0.05)	0.45 (0.27)	0.59 (0.26)	0.75 (0.05)	0.74 (0.05)	0.70 (0.08)	0.58 (0.10)
MCI	0.66 (0.07)	0.72 (0.03)	0.71 (0.05)	0.77 (0.08)	0.42 (0.30)	0.72 (0.14)	0.77 (0.03)	0.72 (0.08)	0.72 (0.04)	0.53 (0.24)
NDRD	0.67 (0.02)	0.76 (0.06)	0.71 (0.06)	0.80 (0.05)	0.19 (0.20)	0.81 (0.02)	0.79 (0.03)	0.73 (0.03)	0.76 (0.04)	0.73 (0.11)
Left hemisphere 95% Hausdorff distance										
Dementia	3.00 (0.83)	4.46 (1.85)	3.01 (1.01)	3.17 (2.08)	4.88 (2.48)	3.86 (3.24)	2.48 (0.95)	2.24 (0.58)	2.56 (1.56)	4.80 (2.14)
MCI	3.21 (1.21)	3.81 (1.22)	2.72 (0.65)	2.64 (1.21)	4.95 (4.53)	4.18 (6.48)	2.21 (0.64)	2.52 (1.09)	2.07 (0.74)	8.31 (8.71)
NDRD	3.32 (0.97)	2.82 (0.82)	3.18 (0.73)	2.37 (0.72)	8.72 (2.69)	1.72 (0.24)	1.97 (0.39)	2.57 (0.44)	1.77 (0.42)	3.84 (2.38)
Right hemisphere 95% Hausdorff distance (mm)										
Dementia	3.47 (0.82)	3.61 (1.00)	3.16 (0.91)	2.52 (0.81)	6.76 (4.66)	9.66 (14.17)	2.54 (0.96)	2.98 (0.87)	2.41 (1.00)	4.55 (1.24)
MCI	4.03 (0.87)	3.34 (0.90)	3.45 (0.88)	2.40 (0.91)	9.65 (8.70)	4.43 (6.01)	2.54 (0.59)	2.99 (1.11)	2.35 (0.51)	8.12 (8.86)
NDRD	4.10 (0.76)	2.78 (1.01)	3.60 (0.61)	2.20 (0.92)	10.13 (3.85)	2.02 (0.56)	2.49 (0.63)	3.05 (0.85)	1.89 (0.75)	3.82 (2.20)

Note: Data are displayed as Mean (SD), except for when presenting Pearson correlation coefficients.



**FIGURE 6** | (A) False positive (FP) (red-yellow) and false negative (FN) (blues) heat maps for each segmentation method projected onto the standard MNI152 T1 template image. The colour bar represents the proportion of times, a given voxel was incorrectly labelled in comparison to the manually segmented hippocampal mask from the Oxford Brain Health Clinic dataset. (B) Number of false positive and (C) false negative voxels in the anterior and posterior hippocampus regions for each segmentation method (averaged over hemispheres).

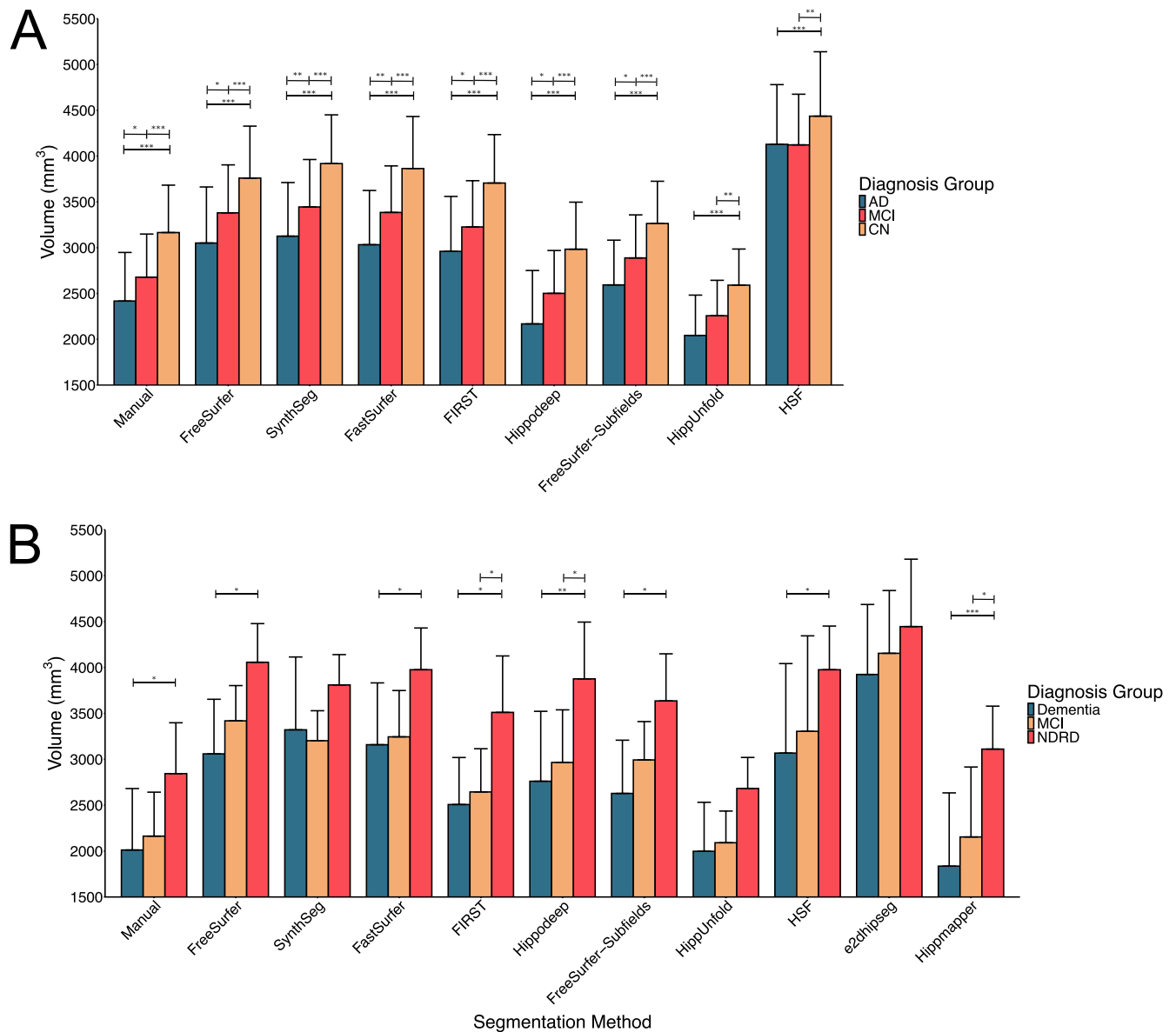
of these tools. FastSurfer implements a CNN that is trained on labels derived from FreeSurfer cortical and subcortical segmentation (Henschel et al. 2020), while Hippodeep was trained in part on hippocampal labels derived from FreeSurfer (Thyreau et al. 2018). By reducing the reliance on manually labelled data, both methods accessed larger training datasets that spanned different ages, disease groups, scanner types, scanner vendors, and field strengths, which likely explains the better performance on unseen data here.

In comparison, traditional methods like FreeSurfer and FIRST, while slightly less effective than deep learning-based methods trained on manual labels in the ADNI and MNI-HISUB25 datasets, demonstrate consistent performance across all datasets when considering mean Dice values and distribution, mean 95th percentile HD, and correlation with manual volumes, despite a tendency to over-estimate volumes. Although FIRST and FreeSurfer showed relatively weaker correlations with manual volumes in the MNI-HISUB25 dataset, and FIRST exhibited a large Dice distribution tail in the OBHC data, their overall performance across datasets was generally moderate to strong, although FIRST outperformed FreeSurfer in both ADNI HarP and MNI-HISUB25. The consistency of traditional segmentation methods compared with deep learning-based methods trained on manual labels highlights that a lack of manual label sources is a notable limitation of the hippocampal segmentation field. The shortage of manual labels that span the entire chronological age range and pathological conditions results in new methods being repeatedly trained on similar data, limiting generalisability even when the unseen data matches the demographic characteristics of common training datasets, such as ADNI HarP.

Another finding of this study is the consistent over-segmentation in the anterior hippocampal region among most segmentation methods, evident through volume similarity values and the number of false positive and false negative voxels, unless they perform

exceptionally poorly (e.g., e2dhipseg on MNI-HISUB25). Even the most reliably performing methods, such as Hippodeep, exhibit systematic over-segmentation at the anterior hippocampus border and a tendency towards over-segmentation, indicating challenges in delineating the boundaries between the hippocampus and amygdala. This difficulty is expected given the lack of visible landmarks to reliably demarcate regions within the medial temporal lobe, particularly between the borders of the hippocampus and amygdala, CA subregions, subiculum and entorhinal cortex (Amunts et al. 2005). Additionally, consistently located false negatives, although less frequent, were seen at the medial and posterior borders of the hippocampus. This region is challenging to segment, as many manual segmentation protocols rely on the appearance of non-hippocampal structures to assist in defining hippocampal borders (Konrad et al. 2009). As cytoarchitecture is not visible in MR images, it is unsurprising that accurately and reliably segmenting the hippocampus both manually and automatically remains an ongoing challenge in the neuroimaging field.

A caveat to consider is that all three datasets in this study were labelled using different manual labelling methods, potentially influencing the comparative results across datasets and segmentation methods (Frisoni and Jack 2011). For example, while the ADNI HarP data were labelled using the extensively tested HarP segmentation protocol (Frisoni and Jack 2015), the MNI-HISUB25 dataset was labelled with the intention to capture three broad regions (subiculum, CA1-3 and CA4-DG) (Kulaga-Yoskovitz et al. 2015). The automatic segmentation methods that were not trained on subfield data showed larger segmented volumes and weak correlations between manual and automatic volumes in the MNI-HISUB25 dataset despite the sample only containing healthy controls, which may indicate differences resulting from manual labelling methods. However, the consistent pattern of false negatives and false positives observed across segmentation methods and datasets may indicate that regardless



**FIGURE 7** | Mean volumes for each segmentation method across diagnostic groups in the (A) ADNI HarP dataset and (B) Oxford Brain Health Clinic dataset. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ .

of manual labelling protocol, automatic segmentation methods systematically fail in similar ways. Finally, it is worth considering that although manual labels are considered the ‘gold standard’ for segmentation, there is inherent subjectivity introduced during this process. For example, identifying landmarks and applying thresholds in the presence of partial volume effects, combined with differences in interpretations of labelling protocols may introduce variability that should be considered when using manual labels as a ground truth.

Selecting the most appropriate hippocampal segmentation method for a given analysis should be guided by research aims and available resources for processing. Based on the data presented, Hippodeep emerges as particularly attractive for solely segmenting the hippocampus, offering high similarity to manual masks based on Dice, particularly in AD or dementia groups, strong correlations with manual volumes, the ability to detect group differences based on volume, and efficient processing times (i.e., in our tests, processing a single participant in

under 10s on a CPU). If whole brain segmentation is of interest, FastSurfer presents a viable, computationally inexpensive alternative to FreeSurfer, demonstrating improved performance over FreeSurfer in all datasets. Although FIRST achieved higher mean Dice values than FastSurfer across all datasets, FastSurfer exhibited greater reliability with tighter Dice distributions and stronger correlations with manual volumes, despite its tendency for over-segmentation and false positives. However, both methods are appropriate options for capturing differences in diagnostic groups.

Alternatively, if hippocampal subfields are of interest, from the methods tested here, the combination of FreeSurfer and FreeSurfer-Subfields is a reliable option, followed closely by HippUnfold, which also provides additional output such as surface data and Laplacian fields. However, both methods are computationally intensive, have larger output sizes and may require high-performance computing resources for large datasets, which may not be an option in all use cases. For studies

**TABLE 6** | Summary of the best and worst performing segmentation methods based on all reported performance metrics across ADNI HarP, MNI-HISUB25 and OBHC datasets.

	Dice coefficient			HD95 (mm)			Volume similarity			Correlation with manual volume (r)			Total errors (FP and FN voxels)							
	Best performer	Worst performer		Best performer	Worst performer		Best performer	Worst performer		Best performer	Worst performer		Best performer	Worst performer						
ADNI HarP																				
AD	Hippodeep	0.80	FreeSurfer	0.67	Hippodeep	1.44	HSF	2.61	FreeSurfer-Subfields	-0.12	HSF	0.23	FastSurfer	0.92	HSF	0.44	FreeSurfer-Subfields	1541.00	HSF	2295.00
MCI	Hippodeep	0.82	FreeSurfer	0.69	Hippodeep	1.37	FreeSurfer	2.44	Hippodeep	-0.07	HSF	0.22	HippUnfold	0.93	HSF	0.73	FreeSurfer-Subfields	1569.00	FastSurfer	2288.00
CN	Hippodeep	0.84	FreeSurfer	0.73	Hippodeep	1.30	HSF	2.16	FreeSurfer-Subfields	-0.06	HSF	0.17	HippUnfold	0.91	HSF	0.48	FreeSurfer-Subfields	1646.00	FastSurfer	2303.00
MNI-HISUB25																				
CN	Hippmapper	0.86	e2dhipseg	0.68	Hippmapper	1.40	e2dhipseg	4.94	FreeSurfer-Subfields	-0.05	e2dhipseg	-0.54	HippUnfold	0.82	e2dhipseg	-0.10	Hippmapper	1499.00	e2dhipseg	2115.00
Oxford Brain Health Clinic																				
Dementia	Hippodeep	0.75	e2dhipseg	0.48	HippUnfold	2.51	Hippmapper	5.91	HippUnfold	0.01	e2dhipseg	0.67	FastSurfer	0.94	HSF	0.43	HippUnfold	1377.00	e2dhipseg	3189.00
MCI	FIRST	0.78	e2dhipseg	0.47	HippUnfold	2.52	HSF	7.64	HippUnfold	-0.02	e2dhipseg	0.63	Hippodeep	0.92	HSF	0.45	FIRST	1404.00	e2dhipseg	3743.00
NDRD	Hippmapper	0.81	e2dhipseg	0.19	HippUnfold	1.87	e2dhipseg	9.42	HippUnfold	-0.05	e2dhipseg	0.44	HSF	0.96	e2dhipseg	-0.52	Hippmapper	1455.00	e2dhipseg	5023.00



investigating group differences in hippocampal volume, most of the segmentation methods evaluated in this study are suitable, even for subtle volume changes such as those between MCI and AD or MCI and control groups. Therefore, the choice of method should also consider other factors such as computational resources and compatibility with study objectives.

One limitation of this study is the relatively small sample size of the OBHC dataset, with a very small number of participants without a dementia-related diagnosis. However, this limitation is not unique to our study but rather reflects the broader challenge in the field of manual label availability. Moreover, the NDRD group in the OBHC dataset cannot be considered a healthy control group in a similar way as other datasets, as all OBHC participants were referred for a memory clinic appointment. We used as our third diagnostic group those patients who did not receive a diagnosis of MCI or dementia, but they may have received other mental health diagnoses or no diagnosis. However, this group is more typical of a general clinical population. Another limitation is that we did not aim to optimise the automatic segmentation methods evaluated here; instead, we used default or recommended settings in all cases. While methods with adjustable parameters or additional input information (e.g., T2w or brain-extracted images) may show improved performance when these are provided, it was beyond the scope of this study to explore the optimisation of each method. Finally, it was also beyond the scope of this study to evaluate the performance of automatic segmentation methods longitudinally, but this is a natural extension of this work that we will focus on in the future.

In this study, we assessed the performance of 10 automatic hippocampal segmentation methods across three datasets with manual labels. While it is challenging to provide a single, definitive recommendation for the most valid method(s) based on this investigation, our findings underscore the ongoing challenge of hippocampal segmentation from MR images within the neuroimaging field. As the field moves towards deep-learning-based segmentation, future efforts should prioritise increasing the availability of publicly accessible manual labels covering a wide range of ages and pathological conditions. This would facilitate adequate training of segmentation methods and enhance their generalisability, both cross-sectionally and longitudinally.

## Acknowledgements

We are grateful to the operations team of the OBHC (see <https://www.psych.ox.ac.uk/research/translational-neuroimaging-group/team/oxford-brain-health-clinic>). Open access publishing facilitated by The University of Adelaide, as part of the Wiley - The University of Adelaide agreement via the Council of Australian University Librarians.

## Ethics Statement

The storage of data on the OBHC Research Database was reviewed and approved by the South Central—Oxford C research ethics committee (SC/19/0404).

## Conflicts of Interest

The authors declare no conflicts of interest.

## Data Availability Statement

The OBHC dataset will be available via the Dementias Platform UK (<https://portal.dementiasplatform.uk/CohortDirectory>) and access will be granted through an application process, reviewed by the OBHC Data Access Group. Data will continue to be released in batches as the OBHC progresses to minimise the risk of participant identification.

## References

- Amunts, K., O. Kedo, M. Kindler, et al. 2005. "Cytoarchitectonic Mapping of the Human Amygdala, Hippocampal Region and Entorhinal Cortex: Intersubject Variability and Probability Maps." *Anatomy and Embryology* 210, no. 5: 343–352. <https://doi.org/10.1007/s00429-005-0025-5>.
- Barnes, J., J. W. Bartlett, L. A. van de Pol, et al. 2009. "A Meta-Analysis of Hippocampal Atrophy Rates in Alzheimer's Disease." *Neurobiology of Aging* 30, no. 11: 1711–1723. <https://doi.org/10.1016/j.neurobiolaging.2008.01.010>.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67: 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Billot, B., D. N. Greve, O. Puonti, et al. 2023. "SynthSeg: Segmentation of Brain MRI Scans of any Contrast and Resolution Without Retraining." *Medical Image Analysis* 86: 102789. <https://doi.org/10.1016/j.media.2023.102789>.
- Boccardi, M., M. Bocchetta, L. G. Apostolova, et al. 2015. "Delphi Definition of the EADC-ADNI Harmonized Protocol for Hippocampal Segmentation on Magnetic Resonance." *Alzheimer's & Dementia* 11, no. 2: 126–138. <https://doi.org/10.1016/j.jalz.2014.02.009>.
- Boccardi, M., M. Bocchetta, F. C. Morency, et al. 2015. "Training Labels for Hippocampal Segmentation Based on the EADC-ADNI Harmonized Hippocampal Protocol." *Alzheimer's & Dementia* 11, no. 2: 175–183. <https://doi.org/10.1016/j.jalz.2014.12.002>.
- Boccardi, M., R. Ganzola, M. Bocchetta, et al. 2011. "Survey of Protocols for the Manual Segmentation of the Hippocampus: Preparatory Steps Towards a Joint EADC-ADNI Harmonized Protocol." *Journal of Alzheimer's Disease* 26, no. s3: 61–75. <https://doi.org/10.3233/JAD-2011-0004>.
- Carmo, D., B. Silva, C. Yasuda, L. Rittner, and R. Lotufo. 2021. "Hippocampus Segmentation on Epilepsy and Alzheimer's Disease Studies With Multiple Convolutional Neural Networks." *Heliyon* 7, no. 2: e06226. <https://doi.org/10.1016/j.heliyon.2021.e06226>.
- Cook, M. J., D. R. Fish, S. D. Shorvon, K. Straughan, and J. M. Stevens. 1992. "Hippocampal Volumetric and Morphometric Studies in Frontal and Temporal Lobe Epilepsy." *Brain* 115, no. 4: 1001–1015. <https://doi.org/10.1093/brain/115.4.1001>.
- Csernansky, J. G., S. Joshi, L. Wang, et al. 1998. "Hippocampal Morphometry in Schizophrenia by High Dimensional Brain Mapping." *Proceedings of the National Academy of Sciences* 95, no. 19: 11406–11411. <https://doi.org/10.1073/pnas.95.19.11406>.
- DeKraker, J., R. A. Haast, M. D. Yousif, et al. 2022. "Automated Hippocampal Unfolding for Morphometry and Subfield Segmentation With HippUnfold." *eLife* 11: e77945. <https://doi.org/10.7554/eLife.77945>.
- DeKraker, J., J. C. Lau, K. M. Ferko, A. R. Khan, and S. Köhler. 2020. "Hippocampal Subfields Revealed Through Unfolding and Unsupervised Clustering of Laminar and Morphological Features in 3D BigBrain." *NeuroImage* 206: 116328. <https://doi.org/10.1016/j.neuroimage.2019.116328>.
- Dill, V., A. R. Franco, and M. S. Pinho. 2015. "Automated Methods for Hippocampus Segmentation: The Evolution and a Review of the State of the Art." *Neuroinformatics* 13, no. 2: 133–150. <https://doi.org/10.1007/s12021-014-9243-4>.

- Duvernoy, H. M. 1998. *The Human Hippocampus*. Springer. <https://doi.org/10.1007/978-3-662-03628-0>.
- Fischl, B. 2012. "FreeSurfer." *NeuroImage* 62, no. 2: 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>.
- Fischl, B., D. H. Salat, E. Busa, et al. 2002. "Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain." *Neuron* 33, no. 3: 341–355. [https://doi.org/10.1016/S0896-6273\(02\)00569-X](https://doi.org/10.1016/S0896-6273(02)00569-X).
- Frisoni, G. B., R. Ganzola, E. Canu, et al. 2008. "Mapping Local Hippocampal Changes in Alzheimer's Disease and Normal Ageing With MRI at 3 Tesla." *Brain* 131, no. 12: 3266–3276.
- Frisoni, G. B., and C. R. Jack. 2011. "Harmonization of Magnetic Resonance-Based Manual Hippocampal Segmentation: A Mandatory Step for Wide Clinical Use." *Alzheimer's & Dementia* 7, no. 2: 171–174. <https://doi.org/10.1016/j.jalz.2010.06.007>.
- Frisoni, G. B., and C. R. Jack. 2015. "HarP: The EADC-ADNI Harmonized Protocol for Manual Hippocampal Segmentation. A Standard of Reference From a Global Working Group." *Alzheimer's & Dementia* 11, no. 2: 107–110. <https://doi.org/10.1016/j.jalz.2014.05.1761>.
- Goubran, M., E. E. Ntiri, H. Akhavein, et al. 2020. "Hippocampal Segmentation for Brains With Extensive Atrophy Using Three-Dimensional Convolutional Neural Networks." *Human Brain Mapping* 41, no. 2: 291–308. <https://doi.org/10.1002/hbm.24811>.
- Griffanti, L., G. Gillis, M. C. O'Donoghue, et al. 2022. "Adapting UK Biobank Imaging for Use in a Routine Memory Clinic Setting: The Oxford Brain Health Clinic." *NeuroImage: Clinical* 36: 103273. <https://doi.org/10.1016/j.nicl.2022.103273>.
- Henschel, L., S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter. 2020. "FastSurfer—A Fast and Accurate Deep Learning Based Neuroimaging Pipeline." *NeuroImage* 219: 117012. <https://doi.org/10.1016/j.neuroimage.2020.117012>.
- Iglesias, J. E., J. C. Augustinack, K. Nguyen, et al. 2015. "A Computational Atlas of the Hippocampal Formation Using Ex Vivo, Ultra-High Resolution MRI: Application to Adaptive Segmentation of In Vivo MRI." *NeuroImage* 115: 117–137. <https://doi.org/10.1016/j.neuroimage.2015.04.042>.
- Jack, C. R., R. C. Petersen, Y. C. Xu, et al. 1999. "Prediction of AD With MRI-Based Hippocampal Volume in Mild Cognitive Impairment." *Neurology* 52, no. 7: 1397. <https://doi.org/10.1212/WNL.52.7.1397>.
- Jack, C. R., R. C. Petersen, Y. C. Xu, et al. 1997. "Medial Temporal Atrophy on MRI in Normal Aging and Very Mild Alzheimer's Disease." *Neurology* 49, no. 3: 786–794. <https://doi.org/10.1212/WNL.49.3.786>.
- Jack, C. R., Jr., M. A. Bernstein, N. C. Fox, et al. 2008. "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods." *Journal of Magnetic Resonance Imaging* 27, no. 4: 685–691. <https://doi.org/10.1002/jmri.21049>.
- Jia, J., M. Staring, and B. C. Stoel. 2024. "Seg-Metrics: A Python Package to Compute Segmentation Metrics." arXiv:2403.07884 arXiv. <https://doi.org/10.48550/arXiv.2403.07884>.
- Konrad, C., T. Ukas, C. Nebel, V. Arolt, A. W. Toga, and K. L. Narr. 2009. "Defining the Human Hippocampus in Cerebral Magnetic Resonance Images—An Overview of Current Segmentation Protocols." *NeuroImage* 47, no. 4: 1185–1195. <https://doi.org/10.1016/j.neuroimage.2009.05.019>.
- Kulaga-Yoskovitz, J., B. C. Bernhardt, S.-J. Hong, et al. 2015. "Multi-Contrast Submillimetric 3Tesla Hippocampal Subfield Segmentation Protocol and Dataset." *Scientific Data* 2: 150059. <https://doi.org/10.1038/sdata.2015.59>.
- Mackay, C. E., J. A. Webb, P. R. Eldridge, D. W. Chadwick, G. H. Whitehouse, and N. Roberts. 2000. "Quantitative Magnetic Resonance Imaging in Consecutive Patients Evaluated for Surgical Treatment of Temporal Lobe Epilepsy." *Magnetic Resonance Imaging* 18, no. 10: 1187–1199. [https://doi.org/10.1016/S0730-725X\(00\)00220-4](https://doi.org/10.1016/S0730-725X(00)00220-4).
- Miller, K. L., F. Alfaro-Almagro, N. K. Bangerter, et al. 2016. "Multimodal Population Brain Imaging in the UK Biobank Prospective Epidemiological Study." *Nature Neuroscience* 19, no. 11: 1523–1536. <https://doi.org/10.1038/nn.4393>.
- Nadel, L., and J. O'Keefe. 1978. *The Hippocampus as a Cognitive Map*. Oxford University Press.
- O'Donoghue, M. C., J. Blane, G. Gillis, et al. 2023. "Oxford Brain Health Clinic: Protocol and Research Database." *BMJ Open* 13, no. 8: e067808. <https://doi.org/10.1136/bmjopen-2022-067808>.
- Patenaude, B., S. M. Smith, D. N. Kennedy, and M. Jenkinson. 2011. "A Bayesian Model of Shape and Appearance for Subcortical Brain Segmentation." *NeuroImage* 56, no. 3: 907–922. <https://doi.org/10.1016/j.neuroimage.2011.02.046>.
- Poiet, C., A. Bouyeure, S. Patil, et al. 2023. "A Fast and Robust Hippocampal Subfields Segmentation: HSF Revealing Lifespan Volumetric Dynamics." *Frontiers in Neuroinformatics* 17: 1130845. <https://doi.org/10.3389/fninf.2023.1130845>.
- Thom, M. 2014. "Review: Hippocampal Sclerosis in Epilepsy: A Neuropathology Review." *Neuropathology and Applied Neurobiology* 40, no. 5: 520–543. <https://doi.org/10.1111/nan.12150>.
- Thyreau, B., K. Sato, H. Fukuda, and Y. Taki. 2018. "Segmentation of the Hippocampus by Transferring Algorithmic Knowledge for Large Cohort Processing." *Medical Image Analysis* 43: 214–228. <https://doi.org/10.1016/j.media.2017.11.004>.
- Voets, N. L., R. A. L. Menke, S. Jbabdi, et al. 2015. "Thalamo-Cortical Disruption Contributes to Short-Term Memory Deficits in Patients With Medial Temporal Lobe Damage." *Cerebral Cortex* 25, no. 11: 4584–4595. <https://doi.org/10.1093/cercor/bhv109>.
- Wickham, H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer International Publishing.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.