Received: 3 August 2023 | Revised: 21 March 2024 | Accepted: 5 April 2024

DOI: 10.1002/nbm.5169

RESEARCH ARTICLE

NMR IN BIOMEDICINE WILEY

# Comparative assessment of established and deep learning-based segmentation methods for hippocampal volume estimation in brain magnetic resonance imaging analysis

Hsi-Chun Wang[1] | Chia-Sho Chen[1] | Chung-Chin Kuo[1] | Teng-Yi Huang[1] | Kuei-Hong Kuo[2,3] | Tzu-Chao Chuang[4] | Yi-Ru Lin[5] | Hsiao-Wen Chung[6] | for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[2]Division of Medical Image, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[3]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[4]Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan

[5]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[6]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

**Correspondence**
Teng-Yi Huang, Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan.
Email: tyhuang@mail.ntust.edu.tw

**Funding information**
This study was supported by the National Science and Technology Council, Taiwan (112-2314-B-011-002-MY2, 110-2221-E-002-026-MY3, 110-2221-E-110-006-MY2, 111-2221-E-011-034-MY3).

## Abstract

In this study, our objective was to assess the performance of two deep learning-based hippocampal segmentation methods, SynthSeg and TigerBx, which are readily available to the public. We contrasted their performance with that of two established techniques, FreeSurfer-Aseg and FSL-FIRST, using three-dimensional T1-weighted MRI scans ($n = 1447$) procured from public databases. Our evaluation focused on the accuracy and reproducibility of these tools in estimating hippocampal volume. The findings suggest that both SynthSeg and TigerBx are on a par with Aseg and FIRST in terms of segmentation accuracy and reproducibility, but offer a significant advantage in processing speed, generating results in less than 1 min compared with several minutes to hours for the latter tools. In terms of Alzheimer's disease classification based on the hippocampal atrophy rate, SynthSeg and TigerBx exhibited superior performance. In conclusion, we evaluated the capabilities of two deep learning-based segmentation techniques. The results underscore their potential value in clinical and research environments, particularly when investigating neurological conditions associated with hippocampal structures.

**KEYWORDS**
deep learning, hippocampal segmentation

## 1 | INTRODUCTION

Recent advancements in neuroimaging, such as high-resolution magnetic resonance imaging (MRI) scans, have enabled accurate and reproducible measurements of hippocampal volume in vivo. Both hippocampal volume and the atrophy rate have been associated with numerous neurological conditions, including Alzheimer's disease (AD),[1,2] autism spectrum disorder,[3,4] major depressive disorder,[5] and temporal lobe epilepsy.[6–8] For example, patients with AD exhibit significantly faster hippocampal volume loss compared with healthy aging individuals.[9] Furthermore, patients with major depressive disorder have been found to possess smaller hippocampal volumes.[5] Accurate hippocampal segmentation allows

**Abbreviations:** AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; ASSD, average symmetric surface distance; CANDI, Child and Adolescent NeuroDevelopment Initiative; CNN, convolutional neural network; DSC, Dice similarity coefficient; HVLR, hippocampal volume loss rate; MCI, mild cognitive impairment; mHV, mean hippocampal volume; MMSE, Mini-Mental State Examination; NL, normal aging; RAVD, relative absolute volume difference; RVD, relative volume difference; SIMON, Single Individual Volunteer for Multiple Observations across Networks; T1w, T1-weighted.

  

researchers to assess brain anatomy, and to monitor disease progression. A swift and precise segmentation procedure can facilitate early identification and diagnosis of these diseases.

As advancements in computer algorithms continue, researchers have increasingly utilized automatic software to delineate bilateral hippocampal regions and estimate their volumes from high-resolution three-dimensional (3D) T1-weighted (T1w) MRI scans.[10] For the investigation of AD, these derived hippocampal volumes can then be employed to differentiate AD from mild cognitive impairment (MCI) and to longitudinally investigate volume reductions in the hippocampus associated with cognitive decline. FreeSurfer[11] and FSL[12] are two widely adopted software packages for brain MR image analysis. Numerous studies have validated the reliability and accuracy of these research tools,[13–16] demonstrating their effectiveness in estimating hippocampal volumes with 3D T1w MR images. Both tools rely on processing in template spaces, making them dependent on accurate registration algorithms. Consequently, segmentation procedures can be time consuming because of the iterative processes involved in the accurate registration algorithm.

Recently, deep learning-based techniques have demonstrated their efficacy for image segmentation tasks. In brain MRI applications, deep learning has proven its effectiveness in automatically segmenting various structures, including tumors,[17,18] subcortical regions,[19,20] stroke lesions,[21,22] as well as gray and white matter.[23,24] For example, Billot et al. introduced SynthSeg,[19,25] a contrast-agnostic segmentation model that has been integrated into the FreeSurfer software suite. Weng and Huang[20] and Wang[26] developed an open-source tool, TigerBx, which employs deep learning-based techniques and a large-scale imaging database to accurately segment subcortical brain structures. Both SynthSeg and TigerBx offer efficient execution times, typically less than 1 min, and are publicly accessible, making them practical tools for facilitating investigations involving hippocampal segmentation. Despite their potential, these newly introduced tools warrant thorough analysis and validation to ensure their efficacy and dependability in estimating hippocampal volumes.

In this study, our objective is to evaluate and compare the performance of deep learning-based tools, specifically SynthSeg and TigerBx, with the well-established tools, FreeSurfer and FSL, focusing on hippocampal volume estimation and AD classification. Through this comparison, we aim to contribute to the ongoing efforts in enhancing the investigation of brain hippocampus structure by utilizing advanced segmentation techniques in brain MRI analysis.

## 2 | MATERIALS AND METHODS

Figure 1 presents an overview of the methodology implemented in this study. We compiled 3D T1w MRI scans from four public databases, amassing a total of 1447 datasets. These 3D T1w images were processed using four methods to execute subcortical segmentations. Following

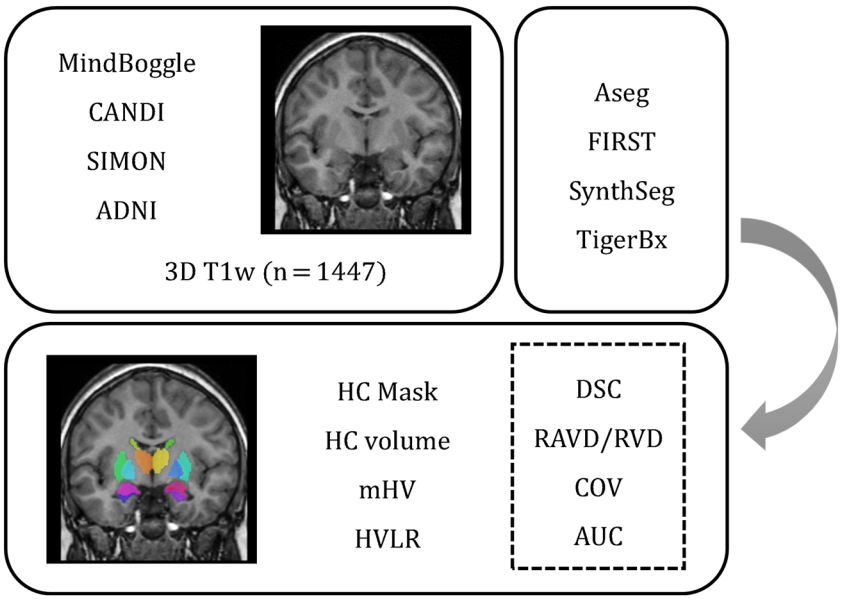


**FIGURE 1** Schematic representation of the study methodology. The process began with the collection of 3D T1w MRI scans from four public databases, totaling 1447 datasets. These 3D volumes were then segmented using four distinct tools to perform hippocampal segmentation and relevant hippocampal (HC) volume indices were calculated. 3D, three-dimensional; ADNI, Alzheimer's Disease Neuroimaging Initiative; AUC, area under the curve; CANDI, Child and Adolescent NeuroDevelopment Initiative; COV, coefficient of variation; DSC, Dice similarity coefficient; HVLR, hippocampal volume loss rate; mHV, mean hippocampal volume; MRI, magnetic resonance imaging; RAVD, relative absolute volume difference; RVD, relative volume difference; SIMON, Single Individual Volunteer for Multiple Observations across Networks; T1w, T1-weighted.

this, we extracted bilateral hippocampal masks from the segmentations and computed associated hippocampal volume indices. The final stages of our study involved evaluating the performance of the segmentation tools, assessing the accuracy of the volumetric measurements, exploring their reproducibility, and determining the accuracy of pathology group classification based on the hippocampal volume indices. The complete procedures are elaborated in the subsequent sections.

## 2.1 | Datasets

We gathered datasets from various public databases, as detailed in Table 1, and employed these datasets in a series of experiments. The databases comprised Mindboggle-101,[27] Child and Adolescent NeuroDevelopment Initiative (CANDI),[28] Single Individual Volunteer for Multiple Observations across Networks (SIMON),[29,30] and Alzheimer's Disease Neuroimaging Initiative (ADNI).[2] All these datasets featured 3D T1w MR images. Each of these databases acquires informed consent from their respective participants during data collection, adhering to ethical guidelines.

We employed the MindBoggle-101 and CANDI databases to evaluate the segmentation accuracy of the four tools. The MindBoggle-101 database consists of 20 T1w MRI scans featuring manual segmentations according to the BrainCOLOR protocol,[27] whereas the CANDI database includes T1w MRI data for 103 participants, with manual tracings performed in adherence to the Center for Morphometric Analysis protocol.[28] We utilized the manual hippocampus segmentation masks from both databases as a reference to assess the segmentation accuracy of the four tools.

SIMON offers a longitudinal MRI dataset of a healthy male from the age of 29 to 46 years, scanned in 73 sessions across various sites using 35 different scanner models. In some sessions, multiple T1w imaging scans were available. One scan of the 48th session experienced an unknown error during the execution of the FreeSurfer pipeline. As a result, we incorporated 94 scans into our experiments in this study.

Data used for evaluating AD classification were obtained from the ADNI database (https://adni.loni.usc.edu/). ADNI, initiated in 2003, is a public–private partnership led by Principal Investigator Dr. Michael W. Weiner. The ADNI study collected various data types for quantitative analysis, including MRI, positron emission tomography, genetics, cognitive tests, cerebrospinal fluid, and blood biomarkers. The primary goal is to investigate whether imaging and biological markers can be combined effectively to monitor the progression of MCI and early AD. For up-to-date information, see https://adni.loni.usc.edu/. It consists of databases such as ADNI 1, 2, 3, and ADNI-GO. In this study, we gathered 615 subjects from the ADNI 1 database, which included AD ($n = 125$), MCI ($n = 301$), and normal aging (NL; $n = 189$) participants. In addition

**TABLE 1** Summary of databases employed in this study.

| Database | Subject | Datasets | Evaluation | Vendors | TR/TE (ms) | Sequence |
|---|---|---|---|---|---|---|
| MindBoggle 101 | 20 | 20 | Segmentation accuracy | Siemens | 9.7/4 | 3D T1w |
| CANDI | 103 | 103 | Segmentation accuracy | GE | 10/3 | 3D T1w |
| SIMON | 1 | 94 | Longitudinal reproducibility | Philips<br>GE<br>Siemens | 7.3/3.3<br>6.7/2.9<br>2300/2.98 | 3D T1w |
| ADNI | 615 | 1230 | Classification of AD | Philips<br>GE<br>Siemens | 8.6/4<br>8.5–10.4/3.8–4.1<br>2400–3000/3.5–3.87 | 3D T1w |

Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; CANDI, Child and Adolescent NeuroDevelopment Initiative; SIMON, Single Individual Volunteer for Multiple Observations across Networks; T1w, T1-weighted.

**TABLE 2** Demographic data summary from the ADNI database.

| Summary of baseline data | | | |
|---|---|---|---|
| | NL | MCI | AD |
| *n* | 189 | 301 | 125 |
| Women (%) | 49.2 | 35.9 | 48.8 |
| Age (years) | 76.1 ± 5.0 | 74.9 ± 7.0 | 74.8 ± 7.6 |
| MMSE | 29.1 ± 1.0 | 27.0 ± 1.8 | 23.5 ± 1.9 |

Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; NL, normal aging.

to the initial baseline data, we also chose the datasets acquired during the 12-month follow-up visit. This enabled us to assess the short-term loss in hippocampal volume among the AD, MCI, and NL groups, using the four segmentation tools. Table 2 provides a summary of the demographic data from the ADNI at baseline, which include cognitive impairment scores as assessed by the Mini-Mental State Examination (MMSE).

## 2.2 | Hippocampal segmentation tools

We utilized four tools for hippocampal segmentation: Aseg, FIRST, SynthSeg, and TigerBx. The specifics of each tool are outlined in Table 3. We employed the default parameters and used raw T1w images without preprocessing for all four segmentation methods in our study, thereby ensuring consistency and reproducibility in our analysis. Aseg refers to a specific type of brain segmentation result generated by FreeSurfer's pipeline "recon-all" (version 7.3.2; https://surfer.nmr.mgh.harvard.edu/). The pipeline involves several processing steps for brain segmentation, including intensity normalization, skull-stripping, gray and white matter segmentation, normalization to template space, and refinement of the pial surface. It employs a Bayesian framework to calculate the likelihood of each voxel being part of a specific brain structure. The pipeline produces a wide range of brain segmentation outputs and surface information. For our analysis, we extracted the bilateral hippocampal masks from the aseg.mgz output file, with the resulting segmentation referred to as Aseg.

FSL (version 6.0; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL) is a versatile neuroimaging analysis tool,[12] and FIRST (FMRIB's Integrated Registration and Segmentation Tool) is one of its pipelines dedicated to subcortical segmentation.[31] The pipeline (*run_first_all*) consists of a series of image preprocessing and registration steps, with the subcortical segmentation based on a Bayesian framework built using manually annotated masks. FIRST generates 15 masks of brain regions, including subcortical regions, brain stem, and the fourth ventricle, from which we extracted bilateral hippocampal masks. The obtained segmentation is referred to as FIRST.

SynthSeg[19,25] and TigerBx[20,26] both employ the 3D U-Net model,[32] a convolutional neural network (CNN) extensively used for segmentation tasks in the medical imaging field.[17,22,33] These tools perform segmentation in the native space of brain volume, thereby circumventing the time-consuming computation involved in image registration. SynthSeg is trained with synthetic brain data, generated using a Gaussian mixture generative model, and employs a domain randomization strategy. This allows for contrast-agnostic segmentation of 3D brain volumes. SynthSeg's effectiveness in analyzing heterogeneous clinical data was demonstrated in a previous study.[25] It is now incorporated in the recent release of FreeSurfer (version 7.3.2; https://surfer.nmr.mgh.harvard.edu/fswiki/SynthSeg). We used the built-in command line tool (*mri_synthseg*) to perform the segmentation.

On the other hand, TigerBx (tissue mask generation for brain extraction, version 0.1.12a; https://github.com/htylab/tigerbx) is trained on large-scale databases, which exclusively comprise T1 images, collected from publicly available cohorts.[20] It can work as a standalone application or serve as a module in the Python software environment. We used the deep gray matter segmentation option (*tigerbx -d*) to generate masks containing hippocampal regions. The segmentation masks produced by all four methods were resampled to align with the original voxel size of the 3D T1w volumes for subsequent evaluation processes.

## 2.3 | Evaluation: segmentation performance and reproducibility

This study used two metrics to evaluate each model's performance: (i) the Dice similarity coefficient (DSC); and (ii) average symmetric surface distance (ASSD).[34] The DSC and ASSD values are employed to measure the similarity between the ground-truth label (manually traced label) and the segmentation results. The DSC value ranges from 0 to 1, with a higher value indicating more accurate segmentation. The ASSD value, ranging from 0 to infinity, quantifies the average surface distance between segmented masks and their corresponding ground-truth masks. A lower ASSD value indicates better segmentation accuracy. To evaluate the longitudinal variations in hippocampal volumes within the SIMON datasets, we employed the coefficient of variation (COV) as a measure of variability for each segmentation tool. This involved calculating the COV by dividing the standard deviation of the hippocampal volumes by their respective mean values for each tool.

**TABLE 3** Summary of hippocampal segmentation tools.

| Tool | Method | Skull-stripping | Registration to template |
|---|---|---|---|
| Aseg | Bayesian | Yes | Yes |
| FIRST | Bayesian | Yes | Yes |
| SynthSeg | 3D U-Net | No | No |
| TigerBx | 3D U-Net | No | No |

### 2.4 | Evaluation: hippocampal volume and loss rate

We performed hippocampal segmentation on datasets using the four segmentation tools and calculated hippocampal volumes by summing the segmented hippocampal voxels and adjusting for the corresponding voxel size. Both left and right hippocampal volumes were estimated, and the average of these two volumes is referred to as the mean hippocampal volume (mHV) in this study. To estimate the accuracy of hippocampal volume estimation, we calculated two key metrics: the relative absolute volume difference (RAVD) and the relative volume difference (RVD). These metrics were computed using the following equations: (i) RAVD = |(Vp - Vm)|/ Vm; and (ii) RVD = (Vp − Vm)/Vm. In these equations, Vp denotes the predicted hippocampal volume as determined by the segmentation tools, and Vm refers to the volume estimated from the manual tracing, which serves as the ground truth. From the ADNI1 database, we selected two scans per subject (at baseline and 12 months) and calculated the short-term mHV loss rate (i.e., the hippocampal volume loss rate [HVLR]) for the NL, MCI, and AD groups. The HVLR is calculated as the annual percentage reduction in the mHV. The equation used is: $HVLR = (V_{0m} - V_{12m})/V_{0m}$, where $V_{0m}$ represents the baseline mHV, and $V_{12m}$ is the mHV estimated from a follow-up scan 12 months later.

## 3 | RESULTS

### 3.1 | Segmentation performance

Figure 2 displays the hippocampal segmentation results on the test datasets superimposed on T1w images. For each participant, we calculated the mean DSC values across the four methods. We then sorted the samples according to the mean DSC values and selected the *n*th quartile for illustration in this figure. The number beneath each image represents the corresponding DSC value. Although all tools produced reliable segmentation outcomes, there were 14 instances from which FIRST derived empty masks. The worst result (i.e., Q0) of FIRST did not generate labels on both hippocampi, while the remaining three methods obtained DSC values from 0.73 to 0.79. Figure 3A presents the DSC values of bilateral hippocampi (123 datasets × 2 hippocampi = 246 samples) obtained using the four segmentation tools (4 methods × 246 samples = 984 points). Scattered FIRST results (orange) with DSC values of zero indicate failure cases. Figure 3B displays an enlarged view with a zoomed y-axis. It becomes evident that the highest DSC values for individual samples are predominantly generated by FIRST and TigerBx. This suggests that FIRST has the potential to be highly accurate if users can manually address the exceptional failures caused by registration issues.

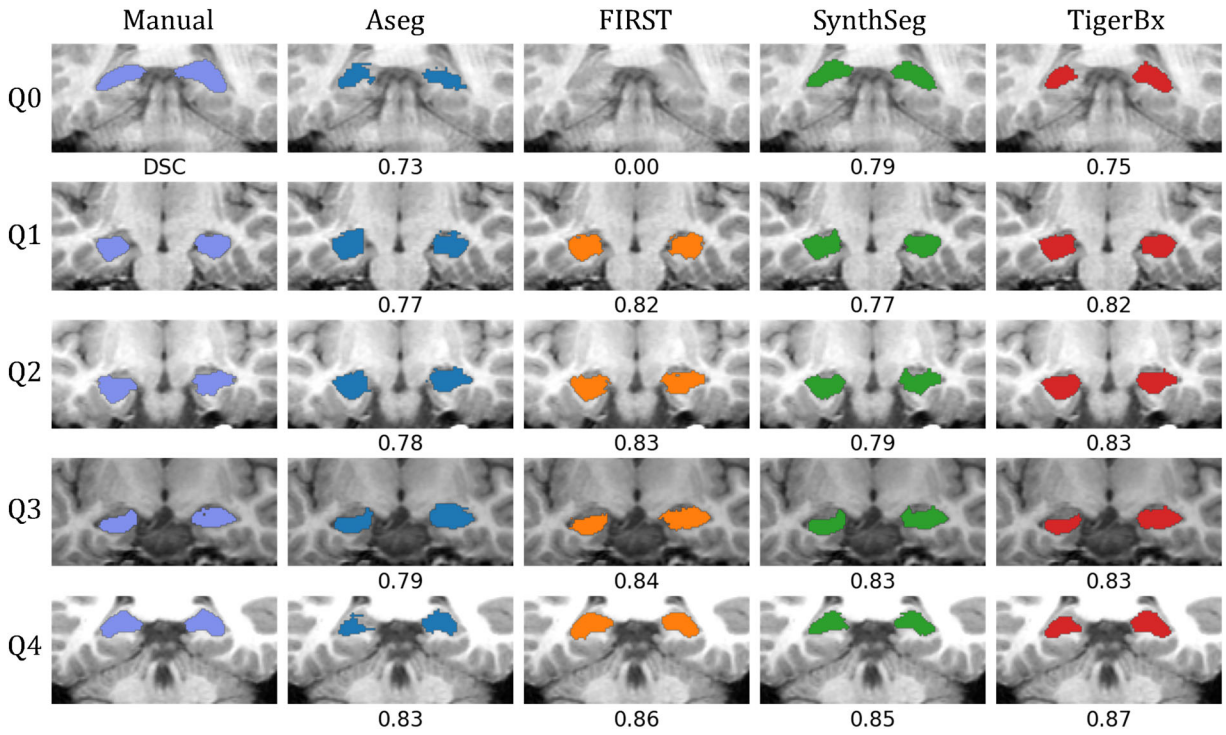


**FIGURE 2** Hippocampal segmentation results on test datasets superimposed on T1-weighted images. The *n*th quartile of mean DSC values was selected for illustration. The number beneath each image represents the corresponding DSC value. DSC, Dice similarity coefficient.

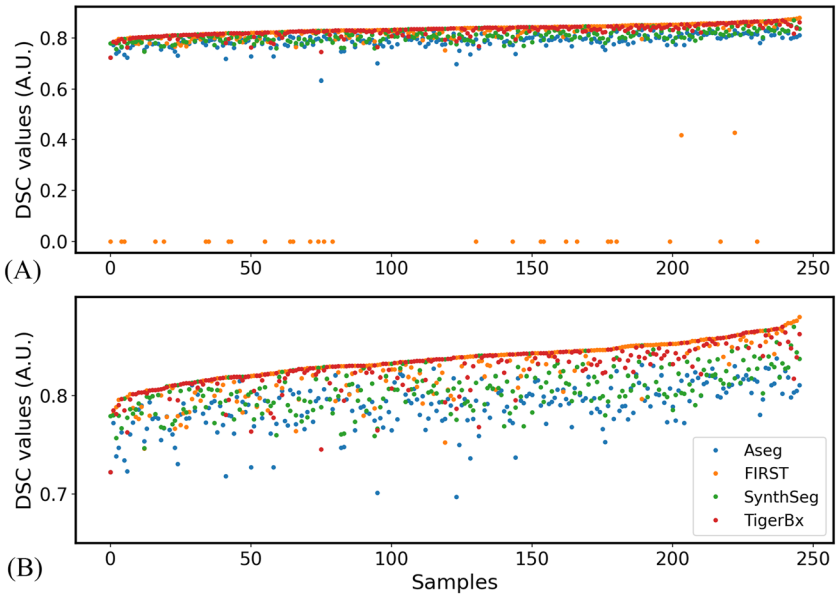


**FIGURE 3** (A) DSC values of bilateral hippocampi obtained using the four segmentation tools on 123 datasets (4 methods × 123 datasets × 2 hippocampi = 984 points). Scattered FIRST results (orange) with DSC values of zero indicate failure cases. (B) The enlarged view with a zoomed y-axis. The highest DSC values for individual samples are predominantly generated by FIRST and TigerBx. DSC, Dice similarity coefficient.

**TABLE 4** Average DSC values for segmentation performance.

| | DSC All ($n = 123$) | | DSC Valid cases ($n = 109$) | |
|---|---|---|---|---|
| | **LHC** | **RHC** | **LHC** | **RHC** |
| Aseg | 0.79 ± 0.03 | 0.79 ± 0.02 | 0.79 ± 0.03*§† | 0.79 ± 0.02*§† |
| FIRST | 0.73 ± 0.23 | 0.73 ± 0.27 | 0.83 ± 0.05 | 0.82 ± 0.05 |
| SynthSeg | 0.81 ± 0.02 | 0.81 ± 0.02 | 0.81 ± 0.02*† | 0.81 ± 0.02*† |
| TigerBx | 0.83 ± 0.03 | 0.83 ± 0.02 | 0.83 ± 0.02 | 0.83 ± 0.02 |

*Note*: The assessment of differences was exclusively conducted on valid cases ($n = 109$) using the Wilcoxon signed-rank test.
Abbreviations: DSC, Dice similarity coefficient; LHC, left hippocampus; RHC, right hippocampus.
†Significantly lower than FIRST ($p < 0.01$).
§Significantly lower than SynthSeg ($p < 0.01$).
*Significantly lower than TigerBx ($p < 0.01$).

Tables 4 and 5 display the average DSC and ASSD values for hippocampal segmentation performance across the four tools, accounting for both the complete dataset ($n = 123$) and the valid cases only ($n = 109$). When considering all datasets, both SynthSeg and TigerBx outperformed Aseg and FIRST in terms of DSC and ASSD values. TigerBx notably demonstrated higher DSC and lower ASSD values compared with the other three methods. After excluding the failed samples, FIRST presented both metrics significantly superior to Aseg and SynthSeg ($p < 0.01$, Wilcoxon signed-rank test), while showing results on a par with TigerBx.

Table 6 displays the average RAVD and RVD values of the test datasets including only the valid samples ($n = 109$). Consistent with the evaluation of DSC values, FIRST and TigerBx produced significantly lower RAVD values ($p < 0.01$, $t$-test) than Aseg and SynthSeg. The RVD values are all positive, suggesting that the tools tend to overestimate hippocampal volumes compared with the manual drawing method. When excluding the failed segmentation cases of FIRST, it actually produces the highest hippocampal volumetric accuracy among the four tools.

## 3.2 | Long-term reproducibility of heterogeneous datasets

We utilized the SIMON datasets ($n = 94$), obtained from a healthy male from the age of 29 to 46 years in 35 MRI scanners, to assess the reproducibility of the automatic segmentation methods across multiple scanner models and institutions. An experienced radiologist (KKH) delineated

**TABLE 5** ASSD values for segmentation performance.

| | ASSD All ($n = 123$) | | ASSD Valid cases ($n = 109$) | |
|---|---|---|---|---|
| | **LHC** | **RHC** | **LHC** | **RHC** |
| Aseg | 0.73 ± 0.13 | 0.71 ± 0.09 | 0.73 ± 0.13*§† | 0.71 ± 0.09*§† |
| FIRST | 3.74 ± 9.79 | 3.66 ± 9.41 | 0.63 ± 0.26 | 0.65 ± 0.30 |
| SynthSeg | 0.68 ± 0.10 | 0.68 ± 0.09 | 0.68 ± 0.10*† | 0.68 ± 0.09*† |
| TigerBx | 0.65 ± 0.11 | 0.64 ± 0.08 | 0.65 ± 0.11† | 0.64 ± 0.08 |

*Note*: The assessment of differences was exclusively conducted on valid cases ($n = 109$) using the Wilcoxon signed-rank test.
Abbreviations: ASSD, average symmetric surface distance; LHC, left hippocampus; RHC, right hippocampus.
†Significantly higher than FIRST ($p < 0.01$).
§Significantly higher than SynthSeg ($p < 0.01$).
*Significantly higher than TigerBx ($p < 0.01$).

**TABLE 6** Average RAVD and RVD values for hippocampal volume estimation.

| | **RAVD (%) ($n = 109$)** | | **RVD (%) ($n = 109$)** | |
|---|---|---|---|---|
| | **LHC** | **RHC** | **LHC** | **RHC** |
| Aseg | 18 ± 10*§† | 17 ± 8*§† | 18 ± 10*§† | 17 ± 9*§† |
| FIRST | 9 ± 7 | 10 ± 8 | 5 ± 10 | 7 ± 10 |
| SynthSeg | 14 ± 9*† | 12 ± 9*† | 14 ± 9*† | 12 ± 9*† |
| TigerBx | 9 ± 7 | 10 ± 6 | 8 ± 8† | 10 ± 7† |

*Note*: Statistical analysis was carried out exclusively on valid cases ($n = 109$). Abbreviations: LHC, left hippocampus; RAVD, relative absolute volume difference; RHC, right hippocampus; RVD, relative volume difference.
†Significantly higher than FIRST ($p < 0.01$, Wilcoxon signed-rank test).
§Significantly higher than SynthSeg ($p < 0.01$, Wilcoxon signed-rank test).
*Significantly higher than TigerBx ($p < 0.01$, Wilcoxon signed-rank test).

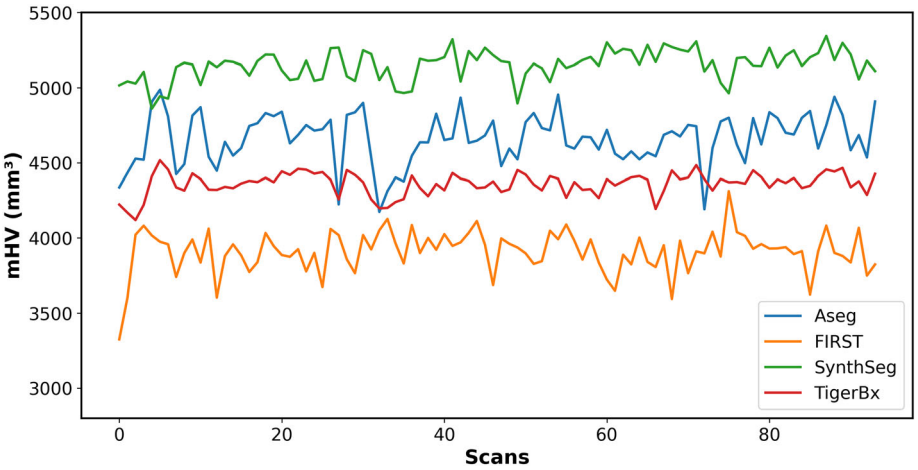


**FIGURE 4** Long-term reproducibility of hippocampal volume estimation using the SIMON dataset. The time-series plot shows the mHV values obtained for 94 scans from a single healthy male participant over 17 years, segmented using Aseg, FIRST, SynthSeg, and TigerBx. Despite variations in average mHV values across the methods, all four tools demonstrated low coefficients of variation (less than 4%), indicating their consistency in hippocampal volume estimation over time. mHV, mean hippocampal volume; SIMON, Single Individual Volunteer for Multiple Observations across Networks.

hippocampal masks in select SIMON datasets, focusing on sessions 1, 15, 30, 45, 60, and 73. The COV across the mHV values of these sessions was at 2.6%, suggesting a minimal change in hippocampal volume for this healthy individual over the course of 17 years. We then performed automatic hippocampal segmentation on the SIMON datasets using the four tools and calculated the corresponding mHV values. Figure 4 displays the time series of mHV values obtained from the 94 scans. The average mHV values across the time series are 4660 ± 168, 3909 ± 141,

5150 ± 101, and 4361 ± 75 $mm^3$ for Aseg, FIRST, SynthSeg, and TigerBx, respectively. The resultant COVs were 3.60%, 3.60%, 1.95%, and 1.71% for each tool, respectively. Although the average mHV values vary among the four methods (ranging from 3909 to 5150 $mm^3$), they all produced mHV variations of less than 4%. The reproducibility of the deep learning-based methods outperformed Aseg and FIRST.

### 3.3 | AD classification: mHV

We employed the four methods to segment the collected ADNI datasets ($n = 615$), which comprised two scans per subject (baseline and 12 months). We subsequently calculated the baseline mHV for the three groups (NL, MCI, and AD). Figure 5A displays the boxplot of baseline mHV values obtained from the four methods, highlighting a distinct stepped trend in the order of the NL, MCI, and AD groups. The mHV values showed significant differences ($p < 0.01$, $t$-test) in all comparisons between groups. Figure 6A presents the receiver operating curve analysis of classification performance between groups using mHV. The area under the curve (AUC) values for the three comparisons, namely, NL-MCI (NL vs. MCI), NL-AD (NL vs. AD), and MCI-AD (MCI vs. AD), ranged from 0.71 to 0.74, from 0.83 to 0.87, and from 0.66 to 0.68, respectively. The results demonstrated that the mHV values obtained using the four tools can all effectively distinguish the groups, with the classification between NL and AD exhibiting the highest accuracy. SynthSeg and TigerBx generate results comparable with the two established methods.

### 3.4 | AD classification: HVLR

Using the longitudinal two scans (baseline and 12-month) of each participant, we calculated the mHV for both time points and determined the HVLR for each participant based on the reduction in mHV. Figure 5B presents the results obtained from the four tools. The boxplots exhibit a similar stepped trend as observed in the mHV results, indicating an increasing mHV reduction rate in the order of NL, MCI, and AD. The HVLR values

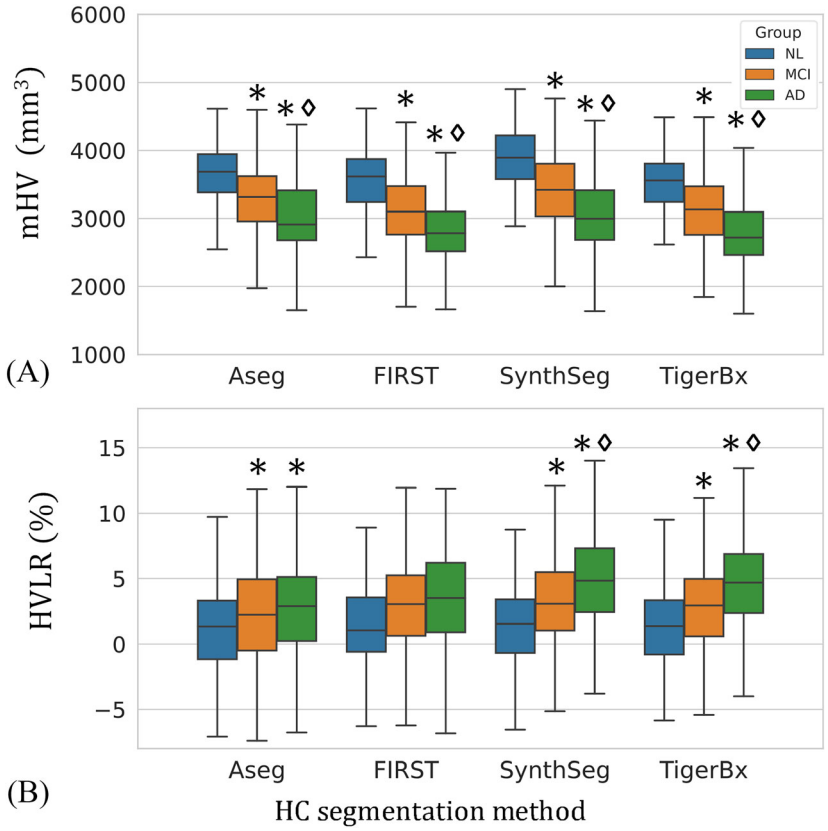


**FIGURE 5** Boxplots of (A) Baseline mHV, and (B) HVLR values for the NL, MCI, and AD groups obtained from the four segmentation methods (Aseg, FIRST, SynthSeg, and TigerBx). Both plots reveal a noticeable stepped trend in the sequence of the NL, MCI, and AD groups. The mHV values demonstrated significant differences ($p < 0.01$, $t$-test) in all comparisons between groups. Significant differences from the NL group are marked with asterisks ($p < 0.01$, $t$-test), and significant differences from the MCI group are marked with diamonds ($p < 0.01$, $t$-test). AD, Alzheimer's disease; HVLR, hippocampal volume loss rate; MCI, mild cognitive impairment; mHV, mean hippocampal volume; NL, normal aging.

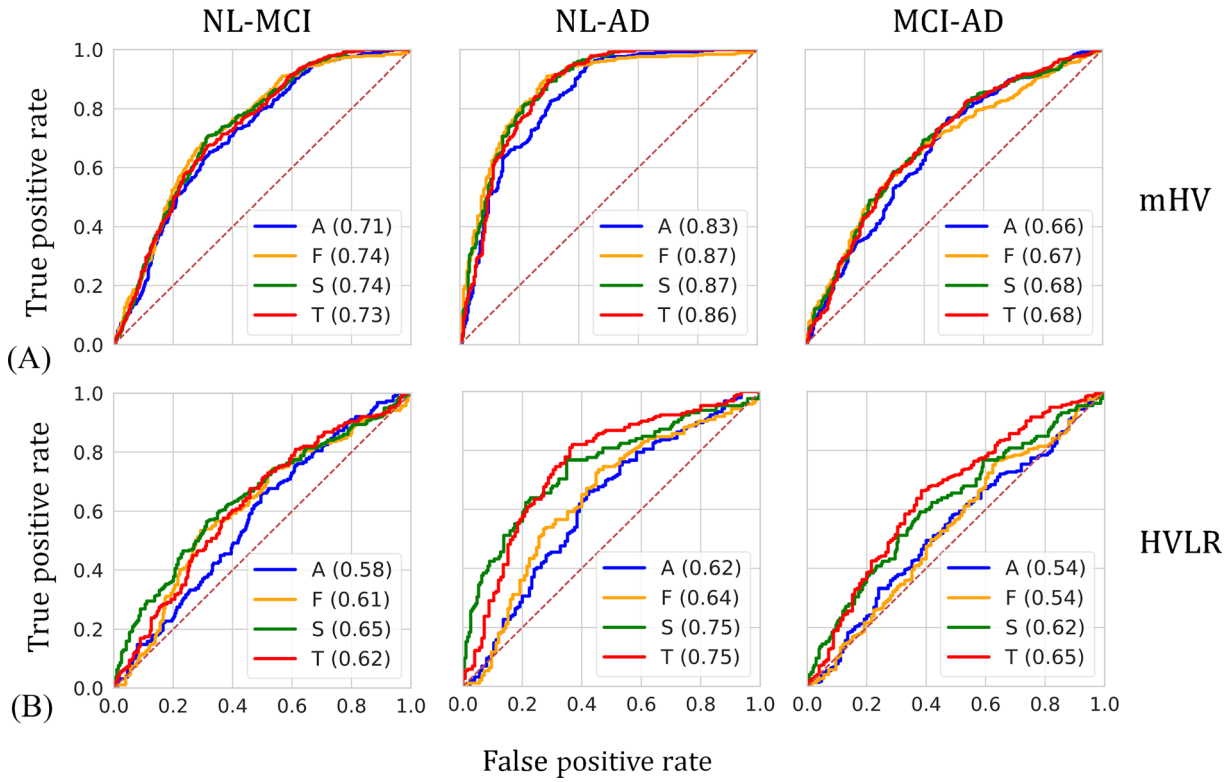


**FIGURE 6** Receiver operating curve analysis of classification performance between the groups using (A) mHV, and (B) HVLR values obtained from the four segmentation methods. A, F, S, and T denote Aseg, FIRST, SynthSeg, and TigerBx, respectively. The AUC values for the three comparisons (NL-MCI, NL-AD, and MCI-AD) are shown. FIRST, SynthSeg, and TigerBx attained comparable AUC values across all classification tasks relying on mHV, with SynthSeg and TigerBx notably surpassing Aseg and FIRST when using HVLR for the classification tasks. AD, Alzheimer's disease; AUC, area under the curve; HVLR, hippocampal volume loss rate; MCI, mild cognitive impairment; mHV, mean hippocampal volume; NL, normal aging.

obtained by SynthSeg and TigerBx revealed significant differences ($p < 0.01$, $t$-test) in all comparisons between the groups. Figure 6B illustrates the receiver operating curve analysis of the classification performance between groups using HVLR. Among the four tools, SynthSeg and TigerBx display the superior AUC values in all the group comparisons.

### 3.5 | Computation time

In this study, we compared the computation time of the four tools using 10 randomly selected datasets from the ADNI1 database. The evaluations were conducted on a personal computer equipped with an Intel i7-9700K CPU, 64 GB RAM, running Ubuntu 20.04.2 LTS operating system. All tools were operated in CPU-only mode, processing each dataset individually. The average computation time per dataset was 188 ± 7 min, 122 ± 9 s, 58 ± 1 s, and 28 ± 0 s for Aseg, FIRST, SynthSeg, and TigerBx, respectively. Notably, SynthSeg and TigerBx achieved hippocampal segmentation considerably faster than Aseg and FIRST.

## 4 | DISCUSSION

In the domain of deep-learning applications within medical imaging, because of the limited size of training datasets, the reliability of the models in handling cross-institutional datasets remains a question. A significant challenge is the difficulty in fairly comparing performances across different studies, as each is often validated on distinct datasets. This study focuses on conducting an equitable comparison across the four methods, encompassing both established and deep learning-based tools, using a uniform set of validation datasets from multiple institutions processed automatically without any preprocessing of the original data. In this study, we evaluated and compared the performance of the segmentation tools using the MindBoggle-101, CANDI, and SIMON datasets. Moreover, we assessed their ability to accurately estimate mHV and HVLR for classifying participant groups in the ADNI datasets.

Our results indicate that SynthSeg and TigerBx exhibited superior performance in terms of DSC values, outperforming both Aseg and FIRST. It is important to note that the lower performance of FIRST was largely due to a number of failed segmentations. Based on our data, we observed that issues with the FIRST algorithm often stemmed from image registration steps. When these failed cases were excluded, FIRST's accuracy was comparable with the other three tools. In this condition, the RVD and RAVD analyses suggested that FIRST and TigerBx provided significantly lower RAVD values (9%–10%) compared with Aseg and SynthSeg (12%–18%), indicating enhanced accuracy in volume estimation. However, all tools showed a tendency to overestimate hippocampal volume, as demonstrated by RVD values ranging from 5% to 18%.

We also evaluated the long-term reproducibility of the methods using the SIMON dataset. Despite some variations in the average mHV values produced by the four tools, all methods exhibited COVs less than 4%, indicating their ability to generate consistent results over time, a crucial factor for longitudinal studies. The deep learning-based tools, SynthSeg and TigerBx, demonstrated COVs of less than 2%, indicating that the consistency of these new tools is noteworthy and has potential for further application in hippocampal segmentation. The variation in average mHVs, which ranges from 3909 to 5150 $mm^3$ across the four segmentation methods, as well as the overestimation observed in the RVD analysis mentioned above, could be attributed to differences in optimization procedures and the reference ground truths used in creating these techniques. Although identifying the exact reasons for these discrepancies is challenging, our study's head-to-head comparison provides a benchmark for understanding measurement differences across tools, emphasizing the importance of recognizing these variations in hippocampal volume assessments in healthy subjects.

A prominent feature of AD is the reduction in hippocampal volume, which is closely related to the decline in cognitive function in the disease.[35] In our analysis of ADNI datasets, the mHV values obtained using the four tools were capable of distinguishing between the groups, with the comparison between the groups showing a significant difference ($p < 0.01$, $t$-test). The AUC values for classification between groups are in the order of NL-AD, NL-MCI, and MCI-AD. Among the four tools, SynthSeg achieved the highest AUC values for all classification tasks (NL-MCI: 0.74, NL-AD: 0.87, MCI-AD: 0.68), indicating its superior performance in classifying different cognitive impairment levels based on hippocampal volume.

The assessment of short-term hippocampal volume loss in the ADNI dataset demonstrated an increasing HVLR trend from the NL to the MCI and AD groups across all tools, aligning with prior findings that the hippocampal atrophy rate correlates with cognitive decline and the progression of AD.[35] In the HVLR analysis, SynthSeg and TigerBx demonstrated superior AUC values when comparing the three groups: NL-MCI (AUC = 0.65 and 0.62), NL-AD (AUC = 0.75 and 0.75), and MCI-AD (AUC = 0.62 and 0.65), respectively. This suggests that SynthSeg and TigerBx could potentially offer superior performance in detecting longitudinal changes in hippocampal volume among the participant groups. This finding aligns with the reproducibility assessment of the four tools, in which the COV values indicate that SynthSeg and TigerBx generate consistent mHV values. Accurate estimation of hippocampal volume loss over time is crucial for monitoring disease progression, thus suggesting that both of the deep learning-based methods are suitable segmentation tools for future studies investigating the progression of AD. Comparing HVLR and mHV in the current study, we found that despite HVLR exhibiting higher values in AD datasets, it was mHV that showed a superior performance in AD classification, with the best NL-AD AUC scores being 0.87 for mHV and 0.75 for HVLR. The underperformance of HVLR might be due to its calculation method, which, based on the difference between two mHV measurements, can introduce additional noise influenced by the stability and precision of each mHV measurement.

It is worth mentioning that our study has some limitations. The analysis was performed on datasets collected from four databases, which may not be representative of other populations or imaging protocols. Additionally, the performance of these tools may be influenced by factors such as MRI hardware types, imaging parameters, and image processing procedures. Therefore, it is essential to validate the results on other cohorts, such as OASIS[36] or HCP.[37] An additional limitation is that we did not include other established tools (e.g., ANTs[38] and STAPLE[39]) or deep learning-based tools.[40–43] Additionally, there are several alternative approaches for hippocampal segmentation using FreeSurfer, such as SamSeg.[44] The comparative analysis of these tools, along with their evaluations on additional cohorts, could provide a crucial reference for future researchers considering a transition from well-established tools to deep-learning methods in hippocampal segmentation studies. These comparisons and evaluations merit further investigation.

In conclusion, our study demonstrates the solid performance of the deep learning-based tools, SynthSeg and TigerBx, in the estimation of hippocampal volume and the analysis of short-term hippocampal volume loss. These methods, not only exhibiting comparable segmentation performance with established tools, but also offering faster computational speeds, indicated their potential for hippocampal segmentation. This is particularly relevant in the clinical applications of AD, planning epilepsy treatments, assessing psychiatric disorders, and other neurological conditions involving hippocampal changes. While further validation in larger and more diverse cohorts remains crucial, this study points to the promise of deep learning-based methods for efficient and accurate hippocampal segmentation in neuroimaging studies, achievable in less than 1 min. These findings suggest that deep learning-based tools are becoming increasingly reliable and can serve as suitable segmentation tools for future research focusing on monitoring disease progression using hippocampal volume as a biomarker.

## ACKNOWLEDGMENTS

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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.

We are grateful to the National Center for High-performance Computing for computer time and facilities. We acknowledge the use of Grammarly and OpenAI's ChatGPT-4 for its assistance in refining the grammar in this manuscript. We respectfully acknowledge the participants and the investigators of the open-access datasets that were adopted in this work. CANDI was supported by NIH-NIMH R01 MH083320. SIMON is supported by CIMA-Q (www.cima-q.ca), the CCNA. This study was supported by the National Science and Technology Council, Taiwan (112-2314-B-011-002-MY2, 110-2221-E-002-026-MY3, 110-2221-E-110-006-MY2, 111-2221-E-011-034-MY3).

ADNI datasets collection for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's 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 Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern (California.www.ccna-ccnv.ca), the ONDRI (ondri.ca), and courtesy scans at MR manufacturers.

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no financial interests or potential conflicts of interest that relate to the research described in this paper.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in websites (1) https://mindboggle.info/data, (2) https://www.nitrc.org/projects/candi_share, (3) http://fcon_1000.projects.nitrc.org/indi/retro/SIMON.html and (4) https://adni.loni.usc.edu.

## ORCID

*Tzu-Chao Chuang* https://orcid.org/0000-0001-5115-1958

## REFERENCES

1. Convit A, De Leon M, Tarshish C, et al. Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. *Neurobiol Aging*. 1997;18(2):131-138. doi:10.1016/S0197-4580(97)00001-8
2. Petersen RC, Aisen P, Beckett LA, et al. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*. 2010;74(3):201-209. doi:10.1212/WNL.0b013e3181cb3e25
3. Goldman S, O'Brien LM, Filipek PA, Rapin I, Herbert MR. Motor stereotypies and volumetric brain alterations in children with autistic disorder. *Res Autism Spectr Disord*. 2013;7(1):82-92. doi:10.1016/j.rasd.2012.07.005
4. di Martino A, Yan C-G, Li Q, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*. 2014;19(6):659-667. doi:10.1038/mp.2013.78
5. Schmaal L, Veltman DJ, van Erp TG, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol Psychiatry*. 2016;21(6):806-812. doi:10.1038/mp.2015.69
6. Wu WC, Huang CC, Chung HW, et al. Hippocampal alterations in children with temporal lobe epilepsy with or without a history of febrile convulsions: evaluations with MR volumetry and proton MR spectroscopy. *Am J Neuroradiol*. 2005;26(5):1270-1275.
7. Keihaninejad S, Heckemann RA, Gousias IS, et al. Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic MRI segmentation. *PLoS ONE*. 2012;7(4):e33096. doi:10.1371/journal.pone.0033096
8. Hosseini M-P, Nazem-Zadeh M-R, Pompili D, Jafari-Khouzani K, Elisevich K, Soltanian-Zadeh H. Comparative performance evaluation of automated segmentation methods of hippocampus from magnetic resonance images of temporal lobe epilepsy patients. *Med Phys*. 2016;43(1):538-553. doi:10.1118/1.4938411
9. Schuff N, Woerner N, Boreta L, et al. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*. 2009;132(4):1067-1077. doi:10.1093/brain/awp007
10. Sánchez-Benavides G, Gómez-Ansón B, Sainz A, Vives Y, Delfino M, Peña-Casanova J. Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. *Psychiatry Res Neuroimaging*. 2010;181(3):219-225. doi:10.1016/j.pscychresns.2009.10.011
11. Fischl B. FreeSurfer. *Neuroimage*. 2012;62(2):774-781. doi:10.1016/j.neuroimage.2012.01.021
12. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage*. 2012;62(2):782-790. doi:10.1016/j.neuroimage.2011.09.015

13. Mulder ER, de Jong RA, Knol DL, et al. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. *Neuroimage*. 2014;92:169-181. doi:10.1016/j.neuroimage.2014.01.058
14. Brown EM, Pierce ME, Clark DC, et al. Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *Neuroimage*. 2020;210:116563. doi:10.1016/j.neuroimage.2020.116563
15. Yao Z, Fu Y, Wu J, et al. Morphological changes in subregions of hippocampus and amygdala in major depressive disorder patients. *Brain Imaging Behav*. 2020;14(3):653-667. doi:10.1007/s11682-018-0003-1
16. Sämann PG, Iglesias JE, Gutman B, et al. FreeSurfer-based segmentation of hippocampal subfields: a review of methods and applications, with a novel quality control procedure for ENIGMA studies and other collaborative efforts. *Hum Brain Mapp*. 2022;43(1):207-233. doi:10.1002/hbm.25326
17. Chang YJ, Huang TY, Liu YJ, Chung HW, Juan CJ. Classification of parotid gland tumors by using multimodal MRI and deep learning. *NMR Biomed*. 2021;34(1):e4408. doi:10.1002/nbm.4408
18. Chan H-W, Weng Y-T, Huang T-Y. Automatic classification of brain tumor types with the MRI scans and histopathology images. *Springer International Publishing*. 2020;353-359. doi:10.1007/978-3-030-46643-5_35
19. Billot B, Greve DN, Puonti O, et al. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med Image Anal*. 2023;86:102789. doi:10.1016/j.media.2023.102789
20. Weng J-S, Huang T-Y. Deriving a robust deep-learning model for subcortical brain segmentation by using a large-scale database: Preprocessing, reproducibility, and accuracy of volume estimation. *NMR Biomed*. 2023;36(5):e4880. doi:10.1002/nbm.4880
21. Juan C-J, Lin S-C, Li Y-H, et al. Improving interobserver agreement and performance of deep learning models for segmenting acute ischemic stroke by combining DWI with optimized ADC thresholds. *Eur Radiol*. 2022;32(8):5371-5381. doi:10.1007/s00330-022-08633-6
22. Li Y-H, Lin S-C, Chung H-W, et al. The role of input imaging combination and ADC threshold on segmentation of acute ischemic stroke lesion using U-Net. *Eur Radiol*. 2023;33(9):6157-6167. doi:10.1007/s00330-023-09622-z
23. Moeskops P, de Bresser J, Kuijf HJ, et al. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *NeuroImage Clin*. 2018;17:251-262. doi:10.1016/j.nicl.2017.10.007
24. la Rosa F, Abdulkadir A, Fartaria MJ, et al. Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage Clin*. 2020;27:102335. doi:10.1016/j.nicl.2020.102335
25. Billot B, Magdamo C, Cheng Y, Arnold SE, Das S, Iglesias JE. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proc Natl Acad Sci U S A*. 2023;120(9):e2216399120. doi:10.1073/pnas.2216399120
26. Wang H-C. Automatic subcortical brain segmentation based on deep learning: The effect of image resolution on accuracy and reproducibility. Master. National Taiwan University of Science and Technology; 2022. https://hdl.handle.net/11296/74esq3
27. Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. *Front Neurosci*. 2012;6:171. doi:10.3389/fnins.2012.00171
28. Kennedy DN, Haselgrove C, Hodge SM, Rane PS, Makris N, Frazier JA. CANDIShare: a resource for pediatric neuroimaging data. *Springer Neuroinformatics*. 2012;10(3):319-322. doi:10.1007/s12021-011-9133-y
29. Duchesne S, Chouinard I, Potvin O, et al. The Canadian Dementia Imaging Protocol: harmonizing national cohorts. *J Magn Reson Imaging*. 2019;49(2):456-465. doi:10.1002/jmri.26197
30. Duchesne S, Dieumegarde L, Chouinard I, et al. Structural and functional multi-platform MRI series of a single human volunteer over more than fifteen years. *Scientific Data*. 2019;6(1):245. doi:10.1038/s41597-019-0262-8
31. Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*. 2011;56(3):907-922. doi:10.1016/j.neuroimage.2011.02.046
32. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Springer*. 2016;424-432. doi:10.1007/978-3-319-46723-8_49
33. Sandfort V, Jacobs M, Arai AE, Hsu L-Y. Reliable segmentation of 2D cardiac magnetic resonance perfusion image sequences using time as the 3rd dimension. *Eur Radiol*. 2021;31(6):3941-3950. doi:10.1007/s00330-020-07474-5
34. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29. doi:10.1186/s12880-015-0068-x
35. Convit A, de Asis J, de Leon MJ, Tarshish CY, De Santi S, Rusinek H. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiol Aging*. 2000;21(1):19-26. doi:10.1016/s0197-4580(99)00107-4
36. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*. 2007;19(9):1498-1507. doi:10.1162/jocn.2007.19.9.1498
37. van Essen D, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: an overview. *Neuroimage*. 2013;80:62-79. doi:10.1016/j.neuroimage.2013.05.041
38. Tustison NJ, Cook PA, Klein A, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*. 2014;99:166-179. doi:10.1016/j.neuroimage.2014.05.044
39. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*. 2004;23(7):903-921. doi:10.1109/TMI.2004.828354
40. Akudjedu TN, Nabulsi L, Makelyte M, et al. A comparative study of segmentation techniques for the quantification of brain subcortical volume. *Brain Imaging Behav*. 2018;12(6):1678-1695. doi:10.1007/s11682-018-9835-y
41. Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *Neuroimage*. 2018;170:434-445. doi:10.1016/j.neuroimage.2017.02.035
42. Huo Y, Xu Z, Xiong Y, et al. 3D whole brain segmentation using spatially localized atlas network tiles. *Neuroimage*. 2019;194:105-119. doi:10.1016/j.neuroimage.2019.03.041
43. Coupé P, Mansencal B, Clément M, et al. AssemblyNet: a large ensemble of CNNs for 3D whole brain MRI segmentation. *Neuroimage*. 2020;219:117026. doi:10.1016/j.neuroimage.2020.117026

44. Cerri S, Greve DN, Hoopes A, Lundell H, Siebner HR, Mühlau M, Leemput KV, An open-source tool for longitudinal whole-brain and white matter lesion segmentation. *NeuroImage Clin*. 2023;38:103354. doi:10.1016/j.nicl.2023.103354

**How to cite this article:** Wang H-C, Chen C-S, Kuo C-C, et al. Comparative assessment of established and deep learning-based segmentation methods for hippocampal volume estimation in brain magnetic resonance imaging analysis. *NMR in Biomedicine*. 2024; e5169. doi:10.1002/nbm.5169