Valentina Visani¹, Mattia Veronese¹, Francesca B Pizzini¹, Annalisa Colombi¹, Valerio Natale¹, Corina Marjin¹, Agnese Tamanti¹, Julia J Schubert¹, Noha Althubaity¹, Inés Bedmar-Gómez¹, Neil A Harrison¹, Edward T Bullmore¹, Federico E Turkheimer¹, Nima Consortium¹, Massimiliano Calabrese¹, and Marco Castellaro¹

¹Affiliation not available

February 15, 2024
ASCHOPLEX: a generalizable approach for the automatic segmentation of choroid plexus


Abstract—The Choroid Plexus (ChP) plays a vital role in brain homeostasis, serving as part of the Blood-Cerebrospinal Fluid Barrier, contributing to brain clearance pathways and being the main source of cerebrospinal fluid. Since the involvement of ChP in neurological and psychiatric disorders is not entirely established and currently under investigation, accurate and reproducible segmentation of this brain structure on large cohorts remains challenging. This paper presents ASCHOPLEX, a deep-learning tool for the automated segmentation of human ChP from structural MRI data that integrates existing software architectures like 3D U-Net, UNETR, and DynUNet to deliver accurate ChP volume estimates. Here we trained ASCHOPLEX on 128 T1-w MRI images comprising both controls and patients with Multiple Sclerosis. ASCHOPLEX’s performances were evaluated using traditional segmentation metrics; manual segmentation by experts served as ground truth. To overcome the generalizability problem that affects data-driven approaches, an additional fine-tuning procedure (ASCHOPLEX_tune) was implemented on 77 T1-w PET/MRI images of both controls and depressed patients. ASCHOPLEX showed superior performance compared to commonly used methods like FreeSurfer and Gaussian Mixture Model both in terms of Dice Coefficient (ASCHOPLEX 0.86, ASCHOPLEX_tune 0.78) and estimated ChP volume error (ASCHOPLEX_tune 9.22%, ASCHOPLEX_tune 9.23%). These results highlight the high accuracy, reliability, and reproducibility of ASCHOPLEX ChP segmentations.

Index Terms—Choroid Plexus, Deep Neural Networks, Ensemble, Magnetic Resonance Imaging, Semantic Segmentation.

Date of current version: January 31st, 2024. Asterisk indicates corresponding author.

V. Visani, M. Veronese, and *M. Castellaro are with the Department of Information Engineering, University of Padova, Padova, Italy (e-mail: valentina.visani@phd.unipd.it; mattia.veronese@unipd.it; marco.castellaro@unipd.it). M. Veronese is also with the Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK.

F. B. Pizzini is with the Department of Engineering for Innovation Medicine, Verona University, Verona, Italy (e-mail: francescabenedetta.pizzini@univr.it).

A. Colombi is with Unit of Neurology, Fondazione Poliambulanza, Brescia, Italy (e-mail: annalisa.colombi@poliambulanza.it).

V. Natale is with the Department of Diagnostic and Public Health, University of Verona, Verona, Italy (e-mail: valerio.natale01@gmail.com).

C. Majin, A. Tamanti, and M. Calabrese are with the Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy (e-mail: corinamajin@gmail.com; agnese.tamanti@univr.it; massimiliano.calabrese@univr.it).

J. J. Schubert, N. Althubait, I. Bedmar-Gómez, and F. E. Turkheimer are with the Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK (e-mail: julia.schubert@kcl.ac.uk; noha.althubait@kcl.ac.uk; ines.bg.med@gmail.com; federico.turkheimer@kcl.ac.uk). N. Althubait is also with the Department of Radiological Sciences, College of Applied Medical Science, King Saud bin Abdulaziz University for Health Sciences, King Abdullah International Medical Research Center (KAIMRC), Riyadh, SA.

N. A. Harrison is with the Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University, Cardiff, UK (e-mail: harissonN4@cardiff.ac.uk).

E. T. Bullmore is with the Department of Psychiatry, School of Clinical Medicine, University of Cambridge, UK, and with the Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, UK, and also with the Immuno-Psychiatry, Immuno-Inflammation Therapeutic Area Unit, GlaxoSmithKline R&D, Stevenage, UK (e-mail: etb23@medschl.cam.ac.uk).

NIMA Consortium information are presented in the Appendix in the Supplementary Materials.

I. INTRODUCTION

THE Choroid Plexus (ChP) is a brain anatomical structure situated within the cerebral ventricular system across all four ventricles, which constitutes a significant component of the Blood - Cerebrospinal Fluid Barrier (BCSFB). ChP primary function involves the production of the Cerebrospinal Fluid (CSF) [1], but it also serves as a mediator for brain clearance pathways, thereby contributing to the maintenance of brain homeostasis [2], [3] and lymphatic function [4]. The functional role of the ChP has been primarily characterized in animals [5] but recent literature in clinical cohorts has evidenced its role in the mediation of peripheral and central inflammation [6]; importantly, increased ChP volume has been validated as a robust marker of neuroinflammation across diagnostic neurological and psychiatric cohorts [7]. Due to its clinical relevance, several quantitative neuroimaging modalities, including Diffusion Weighted Imaging [8], perfusion imaging [9], and Positron Emission Tomography (PET) [10], [11], have been used to investigate both its function and morphology. ChP investigations based on neuroimaging techniques require precise segmentation from structural data and are typically performed on T1-weighted (T1-w) MRI sequence since this modality is routinely acquired and demonstrates acceptable contrast. Note that the gold standard MRI sequence to image the ChP is the T1-w MRI sequence enhanced with contrast injection (cT1-w) [12], [13]. However, due to its invasive nature [14] and limited routine acquisition, T1-w without contrast agent injection are instead used [15] given that estimates of ChP volume (ChPV) using T1-w images compared to cT1-w...
are fairly correlated [12], [16]. Direct ChP segmentation and volume estimation have been used extensively in neurological and neurodegenerative disorders. In Alzheimer’s disease an increase in ChPV is related to worse cognitive impairment, as well as reduced clearance of CSF proteins like amyloid-beta that leads to protein accumulation in tissue [2], [12], [17], [18]. Concerning Multiple Sclerosis (MS) an alteration of ChPV is related to neuroinflammation processes [5], [7], [19]–[21]. Particularly, a recent study [22] has also demonstrated a potential correlation between the ChP enlargement and both increased CSF albumin concentration and the clinical disability grade of MS patients suggesting that ChPV could serve as a valuable biomarker for comprehending disease progression [23]. In psychiatric cohorts, an enlarged ChP is associated with lower cognition in psychosis and reduction of BCSFB permeability in depression [10], [24], [25]. Therefore, the robustness, generalizability and reliability of ChP segmentation are important to guarantee the validity and reproducibility of ChP biomarkers in these cohorts [26]. Segmentation of ChP on T1-weighted MRI is routinely done manually but the process is time-consuming and prone to inter- and intra-operator variability [12], [27]. These variabilities are further exacerbated by the high heterogeneity of ChP in terms of size and tortuosity, as well as by the limited contrast between the ChP and the ventricles [14], [27]. Whereas manual segmentation is still the gold standard and considered the ground truth (GT MSeg), these weaknesses make the study of ChP on large amounts of data problematic. Few automatic tools for ChP segmentation have been proposed in the literature, although with sub-optimal results. To enable studies on large patient cohorts, few automatic tools for ChP segmentation have been proposed in the literature, although with sub-optimal results. FreeSurfer (FS) [28], an open-source software routinely used for the automatic segmentation of brain images, was the first to be tested, but it poorly correlates with the GT MSeg [12], [29]. The Gaussian Mixture Model (GMM), an add-on method applied by Tadayon et al. [12] that starts from the FS ChP segmentation, has limited applicability to diseased subjects with morphological abnormalities [29]. Recently, the single 3D U-Net method [30] was proven to return better spatial overlapping than FS, but its validation was limited as the study was carried out only on ten subjects. A recent study exploited the use of 2-step 3D U-Net emphasizing the potential of this simple architecture [14]. The GMM has been recently improved to incorporate other conventional MRI sequences, but the improvement in the segmentation performance seems to lack robustness [29]. As of today, a reliable and generalizable automatic segmentation of the ChP for population studies involving large cohorts of both patients and controls is still unavailable, though highly desirable. The aim of this study is to propose ASCHOPLEX (Automatic Segmentation of CHOroid PLEXus), an automatic Deep Learning tool to obtain a reproducible and reliable ChP segmentation on T1-w MRI data minimizing the error in ChPV estimation. Differently from other applications of deep learning for ChP segmentation, ASCHOPLEX increases the robustness of the ChP segmentation through the ensemble of the predictions made by the most performant trained models and overcomes the generalizability problem that affects data-driven approaches by including a fine-tuning step to learn dataset-specific features. ASCHOPLEX implementation relies on the MONAI Auto3dSeg pipeline [31] to include the architecture template of 3D U-Net [32], UNETR (UNETr Transformers) [33] and DynUNet, a variant of the nnU-Net (no new net) [34], with different path-size and loss functions. 3D U-Net is widely recognized as the workhorse in medical imaging segmentation tasks; however, we also included UNETR and nnU-Net which outperformed the simpler U-Net architecture in recent Medical Imaging Segmentation challenges [33], [34]. ASCHOPLEX was tested on three independent datasets composed of controls and patients with different brain disorders (MS and Depression) acquired from three different scanners. All data were manually segmented by experts. Two datasets were used to train the core networks, while the third dataset was used to test the generalizability of the tool implementing a fine-tuning procedure. The proposed approach has been compared to the commonly used literature approaches, FS and GMM.

II. MATERIALS AND METHODS

A. Datasets

Data were obtained from three studies for a total of 205 subjects. Datasets 1 and 2 were provided by the Multiple Sclerosis Centre of the University Hospital of Verona and were acquired prospectively between March 2019 and October 2021. All subjects gave their written informed consent prior to participating in the study. All procedures were performed in accordance with the Declaration of Helsinki (2008) and the study protocols were approved by the local Ethical Committee. Below is a detailed description of the two datasets.

1) Dataset 1: It comprises 67 subjects (M/F 21/46): 24 healthy controls (HC) (age: 37.2 ± 9.5 years) and 43 Relapsing-Remitting Multiple Sclerosis (RRMS) (age: 40.9 ± 9.9 years) patients. T1-w MRI images were acquired on a Philips Achieva TX with 8-channels head coil (Software version R3.2.3.2). Parameters of 3D T1-w MPRAGE sequences were: resolution 1x1x1 mm; SENSE acceleration factor: 2.5; TE/TR: 3.7/8.4 ms; FA: 9°; total acquisition time: 4min 50s.

2) Dataset 2: It comprises 61 RRMS patients (M/F 13/48, age: 36.7 ± 10.1 years). T1-w MRI images were acquired on a Philips Elition S with 32-channels head coil (Software version R5.7.2.1). Parameters of 3D T1-w MPRAGE sequences were: resolution: 1x1x1 mm; compressed SENSE acceleration factor: 4; TE/TR: 3.7/8.4 ms; FA: 8°; total acquisition time: 3min 20s.

Dataset 3 was collected in the United Kingdom as part of the Biomarkers in Depression Study (BIODEP, NIMA consortium, https://www.neuroimmunology.org.uk/). All procedures were approved by the local ethics committee and performed in accordance with the Declaration of Helsinki. Participants gave their written informed consent prior to participating. The BIODEP study was approved by the NRES Committee East of England Cambridge Central (REC reference:15/EE/0092). In brief, Dataset 3 comprises 77 subjects (M/F 26/51): 51 depressed participants (age: 36.2 ± 7.3 years) and 26 matched HC (age: 37.3 ± 7.8 years) (Full details are reported in
the study protocol [10], [35]). Each participant underwent a structural MRI simultaneously to PET imaging with a GE SIGNA PET/MR scanner. Parameters of 3D T1-w Fast SPGR sequences were: resolution: 1x1x1 mm; TE/TR: 2.99/6.96 ms; FA: 12°; total acquisition time: 6 min.

B. Pre-processing

Images of Datasets 1 and 2 did not undergo any preprocessing before being passed as input to ASCHOPLEX. Images of Dataset 3 underwent a rigid-body transformation to the MNI Talairach ICBM 152 2006c template provided via MIAKAT v4.2.6 software (http://www.miakat.org/MIAKAT2/index.htm). To obtain the FS segmentation, the T1-w sequence of Datasets 1 and 2 have undergone a lesion filling procedure based on Lesion Prediction Algorithm available in the Lesion Segmentation Toolbox for SPM12 (www.statistical-modelling.de/lst.html) [36].

C. Choroid Plexus Manual Segmentation

The GT MSeg of each subject was obtained in T1-w native space. Chp segmentation was performed only in the two lateral ventricles because the ChP is often challenging to reliably distinguish in the third and fourth ventricles using conventional 3T MRI imaging [15]. For Datasets 1 and 2 the ChP was segmented by consensus of a junior and a senior neuroradiologist for each dataset using ITK-snap [37]. The same was done for Dataset 3 using Analyze software v.12 (https://analyzedirect.com). The GT was used as reference to compare the proposed method with both FS and GMM. Moreover, it was also employed during the training of the deep neural network (DNN) models and as reference to test their performances in the validation/testing phase.

D. ASCHOPLEX implementation

ASCHOPLEX (Automatic Segmentation of CHOroid PLEXus) has been implemented (Fig.1) in MONAI version 1.0.1 [31]. MONAI (Medical Open Network for Artificial Intelligence) is a collection of open-source, freely available collaborative frameworks. Particularly, MONAI has a streaming loading modality to improve the training efficiency, many DNNs already implemented, as domain specific metrics and loss functions. Specifically, ASCHOPLEX is a modified version of Auto3DSeg (https://github.com/Project-MONAI/tutorials/tree/main/auto3dseg), a MONAI application for 3D medical data that exploits cutting-edge DNN architectures with an ensemble strategy to propose a flexible tool to perform image segmentation in the medical imaging domain. Auto3DSeg is agnostic to the image domain of application and exploits structured format that generates the algorithm folders for each selected DNN based on the MONAI bundle. A bundle is a self-descriptive network that combines its architecture definition with the metadata, like the global information on the data, and the scripts to train and infer the model network. The main advantages of Auto3DSeg are easier distribution of the networks, user-friendly approach (the user only needs to provide the data folder and its dataset), and reproducibility of the results. A further feature of Auto3DSeg is the ensemble module that ranks trained algorithms based on checkpoints validation accuracy, selects the top N models, and generates predictions through ensemble methods, implementing both average and major voting. The output segmentations derived from various DNN models might exhibit significant variability. Therefore, an effective solution to this challenge is the implementation of an ensemble procedure, which consolidates and integrates all individual predictions.

1) ASCHOPLEX core networks: training and ensemble: Datasets 1 and 2 were used in the initial stage to train the core networks. We will refer to the merge of Datasets 1 and 2 as a single dataset renamed “core networks training” dataset. In the training/validation/testing splitting process we balanced the phenotype (RRMS/HC) and the scanner (Table I). The training procedure was implemented following a

![ASCHOPLEX core network: training and ensemble](image-url)
five-fold cross-validation. The testing set was left untouched until the end of the study to compare the performance of ASCHOPLEX with the commonly used literature approaches. We started comparing DNNs of varying architectures and hyperparameters. We selected three DNN architectures already implemented in MONAI: 3D U-Net [32]; DynUnet, MONAI implementation of nnU-Net [34], and UNETR [33]. All the configurations were implemented with 3D patch inputs. The 3D U-Net was configured with a single input channel. It consisted of 5 layers, with a number of filters equal to 16 for the first layer and doubled for the following ones. Data were downsampled/upsampled in the encoder/decoder part using 2-strided convolutions residual units with a down/up-sampling kernel size of $3 \times 3 \times 3$ voxels, followed by instance normalization and LeakyReLU activation blocks. The DynUnet is part of a new generation of self-configuring DNNs. It accepted single 3D images as input with a single input channel. It was configured with four blocks: input, downsample, bottleneck, upsample. Data were down/up-sampled using $[1, 2, 2, 1]$ strides convolutions with a kernel size of $3 \times 3 \times 3$ voxels, followed by instance normalization and LeakyReLU activation. UNETR is an adaptation of vision transformers. It is composed by a vision transformer in the encoder part connected to the decoder part via skip connections to merge the major benefit of both transformers and U-Net, which inspired this architecture: respectively, great capability in learning global information and properly capturing localized features. The configuration accepted single 3D inputs. The dimension of the network feature size was set to 16, the hidden layer size to 768, and the feedforward layer size to 3072. The transformer in the encoder branch used 12 attention heads. To preserve spatial information of the patches, a perceptron positional embedding layer was used. At every level of resolution, the reshaped tensors from the embedding space were projected into the input space through a series of successive $3 \times 3 \times 3$ convolutional layers, which were subsequently followed by normalization layers. As hyperparameters we considered: three patch size $(64 \times 64 \times 64, 96 \times 96 \times 96, 128 \times 128 \times 128)$ and three loss functions (Generalized Dice (Dice), CrossEntropy (CE), combination of Dice and CE (DiceCE)), for a total of twenty-seven DNNs. MONAI Auto3DSeg bundle was extended to include each DNN among the twenty-seven selected configurations. We trained each configuration separately in a five-fold cross-validation fashion overfitting. Each DNN was trained using the Adam-Weighted optimizer [38] with parameters: learning rate $(1 \times 10^{-4})$, weight decay $(1 \times 10^{-5})$, maximum number of iterations $(2 \times 10^{4})$, batch size (1). The training of the configurations was executed using a 16 GB NVIDIA Tesla V100 GPU. The overall time required for this initial training step was 20 days (483 hours). Following the cross-validation scheme, we use the Dice Coefficient to select the best configuration among the twenty-seven configurations trained for each fold, obtaining the five best models to perform the ensemble by majority voting to increase the robustness of the predicted segmentation on the testing set.

2) Study of the generalizability of ASCHOPLEX: The principal limitation of data-driven methodologies lies in the limited capacity for model generalization when applied to datasets distinct from those employed during the training phase with potentially different patterns of acquisition due to either the scanner (e.g. MR scanner hardware, MR sequence parameters or MR software version) or to the brain morphology of the subjects involved. Changing the MRI dataset, even if the MR sequence contrast is the same, means showing the DNNs images with untrained features. This is a generalizability problem that especially affects deep learning when it deals with MR images [39]. This issue can be in some cases overcome with a fine-tuning procedure (Fig.1). Dataset 3, denoted as the “fine-tuning” dataset, was employed to assess the tool’s generalizability. The dataset was partitioned into distinct subsets for training, validation, and testing, as outlined in Table I. The validation set, composed by five subjects, served the purpose of model validation throughout the training process, ensuring its robustness. The testing set remained unaltered until the completion of the study, allowing for the evaluation of ASCHOPLEX’s performance on a novel dataset. This evaluation also facilitated a comparative analysis with conventional approaches found in the existing literature. Firstly, to test the generalizability of the tool, ASCHOPLEX was inferred on the 62 testing subjects of the fine-tuning dataset with a no fine-tuned network ensemble. This approach was named ASCHOPLEX$_0$. Secondly, a fine-tuning step is carried out. This approach, called ASCHOPLEX$_{fine}$, aimed to gauge the tool’s performance across a spectrum of training scenarios, providing insights into its robustness and versatility. To better understand the minimum number of subjects that might be needed when adapting ASCHOPLEX to a new unseen dataset, we started from the five selected core networks as pretrained models and we varied the number of subjects in the fine-tuning training procedure from one to ten. The number of necessary manually labeled subjects must be reasonable: if the number is too low, the model struggles to learn the characteristics of the dataset; using more data, the manual segmentation becomes time-consuming. As before, the MONAI bundle was modified to consider the new algorithm segmentation templates with the addition of the fine-tuning step. The training parameters were the same except for the maximum number of iterations $(1 \times 10^{4})$. During this phase, the five-fold cross-validation methodology was not employed; a single validation step was introduced, eliminating the need of dividing the training set into folds. Similarly to previous iterations, the ultimate prediction for each testing subject was generated by conducting an ensemble process, employing a majority voting mechanism among the segmentations derived from the five fine-tuned models. The total time taken for this step, fine-tuning and segmentation saving, was 20 hours.

E. Performance evaluation

We conducted a comparative analysis between the segmentations generated by ASCHOPLEX and those produced by FS and GMM. This comparison was performed on the testing sets of both the core networks and the fine-tuning dataset. We generated segmentations using the FS
recon-all pipeline (version 7.1.1 for core networks training dataset, version 6.0.0 for fine-tuning dataset) [28], and applied the GMM segmentation improvement method to the FS output. The GMM pipeline is freely available on GitHub (https://github.com/EhsanTadayon/choroid-plexus-segmentation) [12]. The performance of the compared approaches were evaluated using the following metrics: Dice Coefficient $\text{Dice} = \frac{2\sum_{i} \min(x_i, y_i)}{\sum_{i} x_i + \sum_{i} y_i}$; 95% Hausdorff Distance (95% HD);Pearson’s Volume Correlation ($r$); Percentage Volume Difference $\Delta V \text{ol}\% = \frac{\sum_{i} |x_i - y_i| \cdot r}{\sum_{i} y_i}$; Absolute Percentage Volume Difference ($\Delta V \text{ol}\%$); Recall $\text{Recall} = \frac{\sum_{i} x_i \cdot y_i}{\sum_{i} y_i}$; Precision $\text{Precision} = \frac{\sum_{i} x_i \cdot y_i}{\sum_{i} x_i}$, where $x_i, y_i \in X, Y$ are the predicted segmentation and the GT MSeg, respectively.

Dice Coefficient gives an estimate of the similarity of the compared segmentations, just like Recall and Precision. On the contrary, 95% HD focused on how far the compared segmentations are: to reach good performance, the segmentation approach must maximize the Dice and minimize the 95% HD. In addition, $|\Delta V \text{ol}|\%$, $\Delta V \text{ol}\%$, and $r$ focus on the ChPV estimate.

Concerning ASCHOPLEX training, one-way ANOVA and post-hoc t-test ($\alpha = 0.05$) were performed for Absolute Volume between all available segmentations for the core network training dataset (ASCHOPLEX, FS, GMM and GT MSeg). Concerning ASCHOPLEX_{tune} and ranging the number of subjects of ASCHOPLEX_{tune} from one to ten, one-way ANOVA and post-hoc t-test ($\alpha = 0.05$) were performed on the fine-tuning dataset for Dice Coefficient, 95% HD, and Absolute Volume to identify the minimum number of training subjects for the fine-tuning step. Once the optimal number of subjects for ASCHOPLEX_{tune} was identified, one-way ANOVA and post-hoc t-test ($\alpha = 0.05$) were performed for Absolute Volume calculated on the fine-tuning dataset for the tested methods: ASCHOPLEX_{tune}, FS, GMM and GT MSeg. A Bonferroni correction for multiple comparisons was used.

F. Software availability

The tool proposed in this paper is freely available, published as a user-friendly package on GitHub (https://gitlab.dei.unipd.it/fair/aschoplex) [1]. The end user can directly infer the segmentation with pre-trained models based on the core network training dataset or perform the manual segmentation for a subsample of a new dataset running the fine-tuning step, adapting the model to the new dataset features to improve the reliability of the predicted segmentations.

III. RESULTS

A. ASCHOPLEX core networks: training and ensemble

The five selected configurations were characterized by UNETR and DynUNet architectures. All models were 128 patch-based with Dice or DiceCE loss function. Table II reports the results for the segmentation metrics on the 36 testing subjects of the core network training dataset. The compared approaches were ASCHOPLEX, FS, GMM and GT MSeg. Dice Coefficient is higher for the ensemble approach (0.80 ± 0.07) than other methods proposed by the literature. Concerning 95% HD, ASCHOPLEX prediction location (2.50 ± 2.89) is nearest to GT MSeg compared to FS and GMM. Focusing on the volume analysis, Pearson’s correlation analysis between each segmentation reports a positive correlation between the volume of the ASCHOPLEX and GT MSeg (0.95) that is higher than that between FS and GT MSeg (0.36), while GMM does not show any statistically significant correlation. One-way ANOVA reveals a statistically significant main effect on the absolute volume among the tested methods ($p_{\text{value}} < 0.0001$). FS and GMM have statistically lower Absolute Volume when compared to both GT MSeg and ASCHOPLEX (GMM $p_{\text{value}} = 0.003/0.009$, FS $p_{\text{value}} < 0.0001$), while there are no statistically significant differences between ASCHOPLEX and GT MSeg volumes ($p_{\text{value}} = 1.0$) (Fig. 2). ASCHOPLEX provides lower $|\Delta V \text{ol}|\%$ (9.22 ± 6.59 %) compared to both GMM and FS. Automatic methods tend to underestimate ChPV as shown in Fig. 2(a), where the distributions of ASCHOPLEX and GT MSeg volumes look similar. Fig. 2(b) highlights the linear regression model fit between ChPV predicted by ASCHOPLEX and that calculated on the GT MSeg follows the points with strict confidence bounds, while FS and GMM provide a strong underestimation of the ChPV. Regarding the other metrics, Recall (0.80 ± 0.08) and Precision (0.80 ± 0.09) of ASCHOPLEX are higher than those of the other two methods. Fig. 3 shows the automatic segmentation results overlapped to the T1-w MRI image for two representative patients with GT Mseg reported as reference.

B. Study of the generalizability of ASCHOPLEX

Table III reports the results for the segmentation metrics on the testing set of the fine-tuning dataset. The compared approaches were: ASCHOPLEX_{0}, ASCHOPLEX_{tune}, FS, GMM and GT MSeg.

Firstly, we have compared ASCHOPLEX_{0} and ASCHOPLEX_{tune}. Fig. 4 reports the boxplot of the

<table>
<thead>
<tr>
<th>Metrics</th>
<th>ASCHOPLEX</th>
<th>FS</th>
<th>GMM</th>
<th>GT MSeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>2432 ± 668</td>
<td>1319 ± 469</td>
<td>1928 ± 580</td>
<td>2480 ± 837</td>
</tr>
<tr>
<td>Dice</td>
<td>0.80 ± 0.07</td>
<td>0.32 ± 0.09</td>
<td>0.54 ± 0.09</td>
<td>-</td>
</tr>
<tr>
<td>95% HD</td>
<td>2.50 ± 2.89</td>
<td>9.11 ± 2.79</td>
<td>10.52 ± 3.99</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta V \text{ol}%$</td>
<td>0.70 ± 11.31</td>
<td>43.05 ± 22.12</td>
<td>-15.08 ± 14.93</td>
<td>-</td>
</tr>
<tr>
<td>$\Delta V \text{ol}%$</td>
<td>9.22 ± 6.59</td>
<td>44.82 ± 18.26</td>
<td>29.80 ± 23.66</td>
<td>-</td>
</tr>
<tr>
<td>Recall</td>
<td>0.80 ± 0.08</td>
<td>0.25 ± 0.11</td>
<td>0.50 ± 0.12</td>
<td>-</td>
</tr>
<tr>
<td>Precision</td>
<td>0.80 ± 0.09</td>
<td>0.46 ± 0.13</td>
<td>0.63 ± 0.14</td>
<td>-</td>
</tr>
</tbody>
</table>

All values (except for Absolute Volume and Pearson’s Volume Correlation Coefficient) are presented in the form mean ± standard deviation. Absolute Volume is expressed in [mm$^3$], while $\Delta V \text{ol}\%$ in [%]. Legend: n.s.- non significant.
TABLE III

RESULTS OF THE SEGMENTATION METRICS ON THE 62 TESTING SUBJECTS OF THE FINE-TUNING DATASET. COMPARISON OF ASCHOPLEX₀ (INFERENCEx WITHOUT FINE-TUNING), ASCHOPLEX₉ (FINE-TUNING VARYING THE NUMBER OF SUBJECTS FROM ONE TO TEN), FS (FREESURFER) AND GMM (GAUSSIAN MIXTURE MODEL) WITH THE REFERENCE MANUAL SEGMENTATION (GT MSEG).

| GT MSEG | Volume [mm³] | Dice | 95% HD | ∆Vol% [%] | |∆Vol% [%] | Recall | Precision | r |
|---------|-------------|------|--------|-----------|--------|--------|-----------|---------|
| ASCHOPLEX₀ | 1267 ± 479 | 0.59 ± 0.16 | 4.75 ± 6.42 | -33.80 ± 26.59 | 15.60 ± 24.12 | 0.51 ± 0.19 | 0.78 ± 0.07 | n.s. |
| ASCHOPLEX₉ 1 SJ | 2238 ± 599 | 0.69 ± 0.06 | 4.11 ± 1.69 | 16.96 ± 24.40 | 25.13 ± 15.81 | 0.75 ± 0.09 | 0.65 ± 0.09 | 0.63 |
| ASCHOPLEX₉ 2 SJ | 2509 ± 670 | 0.71 ± 0.07 | 3.88 ± 1.99 | 25.17 ± 24.65 | 28.85 ± 20.22 | 0.79 ± 0.06 | 0.65 ± 0.11 | 0.70 |
| ASCHOPLEX₉ 3 SJ | 2464 ± 601 | 0.75 ± 0.05 | 2.94 ± 1.50 | 22.68 ± 17.45 | 23.09 ± 15.09 | 0.83 ± 0.06 | 0.68 ± 0.08 | 0.83 |
| ASCHOPLEX₉ 4 SJ | 2375 ± 600 | 0.75 ± 0.05 | 2.54 ± 1.29 | 17.98 ± 16.44 | 20.44 ± 13.25 | 0.82 ± 0.06 | 0.71 ± 0.08 | 0.84 |
| ASCHOPLEX₉ 5 SJ | 2191 ± 506 | 0.78 ± 0.03 | 1.87 ± 1.03 | 9.23 ± 1.68 | 12.46 ± 9.98 | 0.82 ± 0.05 | 0.75 ± 0.06 | 0.87 |
| ASCHOPLEX₉ 6 SJ | 2287 ± 499 | 0.78 ± 0.03 | 2.02 ± 1.23 | 14.37 ± 13.93 | 16.25 ± 11.68 | 0.83 ± 0.05 | 0.73 ± 0.06 | 0.87 |
| ASCHOPLEX₉ 7 SJ | 2237 ± 506 | 0.78 ± 0.03 | 2.15 ± 1.81 | 12.63 ± 12.69 | 14.72 ± 10.20 | 0.83 ± 0.05 | 0.73 ± 0.06 | 0.88 |
| ASCHOPLEX₉ 8 SJ | 2327 ± 506 | 0.78 ± 0.03 | 1.86 ± 1.87 | 16.26 ± 12.99 | 17.47 ± 11.19 | 0.84 ± 0.05 | 0.73 ± 0.05 | 0.89 |
| ASCHOPLEX₉ 9 SJ | 2344 ± 514 | 0.78 ± 0.03 | 1.95 ± 1.48 | 17.07 ± 13.09 | 18.30 ± 11.29 | 0.80 ± 0.05 | 0.73 ± 0.05 | 0.89 |
| ASCHOPLEX₉ 10 SJ | 2176 ± 513 | 0.79 ± 0.03 | 1.93 ± 1.85 | 8.19 ± 11.13 | 10.83 ± 8.59 | 0.82 ± 0.05 | 0.76 ± 0.05 | 0.92 |
| FS | 1037 ± 290 | 0.81 ± 0.08 | 11.94 ± 5.29 | -47.95 ± 11.58 | 47.95 ± 11.58 | 0.24 ± 0.07 | 0.40 ± 0.10 | 0.64 |
| GMM | 1581 ± 430 | 0.61 ± 0.05 | 8.08 ± 5.04 | 20.92 ± 13.32 | 22.53 ± 12.80 | 0.55 ± 0.07 | 0.70 ± 0.09 | 0.75 |

All values (except for Absolute Volume and Pearson’s Volume Correlation Coefficient) are presented in the form mean ± standard deviation. Legend: X SJ - number of subjects used during training of ASCHOPLEX₉; n.s.- non significant.

In this paper we proposed a new approach called ASCHOPLEX to obtain an accurate, reliable, and fast semantic segmentation of the human ChP. ASCHOPLEX is a fully data-driven method that has proven to outperform the two state-of-the-art methods for brain segmentation (namely FS and GMM) in multiple datasets. Deep Learning based methods have become the state of the art for medical image processing tasks [40], [41], with performance equivalent to human trained operators. Recently, the list of DNNs to choose from for medical imaging segmentation tasks has grown. Particularly, self-attention mechanisms of transformers brought interesting results. The UNETR is a novel architecture where a vision transformer is used as encoder to capture long range dependencies, while the U-Net convolutional architecture is employed as a decoder to learn local features [33]. In addition, the nnUNet is a self-configuring net that raised high performance in the Medical Decathlon Segmentation Challenge [34]. Having too many possibilities makes it difficult to prefer one architecture over the others. From this evidence, ensemble strategies have been proposed. The ensemble consists of agreement procedures (e.g. major voting) between predictions derived from different trained DNNs, making the final segmentation more robust and reliable. Therefore, we followed this evidence

4In Supplementary Materials - Supplementary Tables (Table A, B, C) are reported the multiple comparison results for, Dice Coefficient, 95% HD, and Absolute Volume, respectively.

5In Supplementary Materials - Supplementary Figures (Figure B) are shown the performance metrics for ASCHOPLEX₉ using five subjects in the training set, FS and GMM.
to design a novel and generalizable ChP segmentation tool based on DNNs ensemble. Twenty-seven DNNs configurations have been selected, varying both hyperparameters, like the network size and the cost function, and architectures: 3D U-Net, a fundamental tool for medical imaging segmentation, UNETR, that combine transformers with the U-Net, and the self-configuring net nnU-Net. The models selected to be part of the ASCHOPLEX ensemble have common characteristics. These were 1) the DNNs architecture type, 2) the dimension of the net and 3) the chosen loss function. The selected DNNs...
architectures are Dynunet and UNETR, new generations of DNNs architectures for Medical Image Segmentation, suggesting that the use of self-configuring nets or attention mechanisms improves the results for this segmentation task than simpler 3D U-Net. Regarding the network dimension, only the bigger patch size (128x128x128) was selected. Large networks lead on one hand to higher performance, and on the other hand, to increased costs, both in terms of increased training time and minimum hardware requirements [34]. The most selected loss function is DiceCE, concordantly as reported in previous studies [33], [34]. Including Dice calculation in the loss function helps when the number of background and foreground voxels is unbalanced and, consequently, improves similarity between GT MSeg and prediction [42]. ASCHOPLEX improves ChP predictions on the testing set both in terms of similarity indices and ChPV estimate. The mean error committed by ASCHOPLEX in ChPV estimate is 9%, that is 20% and 35% lower than, respectively, GMM and FS. In addition, the volumes of the GT MSeg and the proposed method have similar distributions and the statistical tests revealed there are no statistically significant differences between the two resulting populations. These findings highlight the ability of ASCHOPLEX to emulate, better than FS and GMM, the neuroradiologist in the manual segmentation task, and improve the manual approach in speed and reproducibility. The mediocre ability of FS to emulate the neuroradiologist has already been verified [12], [14]. Despite suboptimal performance, the main advantage of deterministic methods like FS and GMM compared to data-driven ones (e.g. DNNs) is the generalizability. However GMM’s design was optimized on datasets of healthy controls (Human Connectome Project [43]), mild cognitive impairment, and Alzheimer Disease (Alzheimer’s Disease Neuroimaging Initiative [44]) and therefore it might be that different protocols and different pathologies (such as MS or Depression) require a different parameter optimization. ASCHOPLEX has overcome the generalizability issue by
implementing a fine-tuning step, ASCHOPLEX_tune, making the already trained models adaptable to an unseen dataset. The challenge was to investigate if a short additional training combined with a small dataset is sufficient to adapt the tool to a new dataset obtaining better performance than FS and GMM. The comparison between the direct application of ASCHOPLEX, and ASCHOPLEX_tune has shown that the direct inference of the trained model over an unseen dataset with a different pattern of characteristics brings a drop of performance. A fine-tuning step applied before application on a new dataset improves the results. In particular, results suggest that the best compromise is to have a sufficiently large training set to be able to increase the model’s knowledge of the characteristics of the new dataset, and to minimize the number of manual segmentations required to perform the fine-tuning step; here we set the number of subjects in the ASCHOPLEX_tune training set to five. The main disadvantage of implementing a fine-tuning step is the cost, both in terms of hardware requirements (GPU), and the extra time required to run the short training. Based on our investigations, the extra fine-tuning procedure requires manual segmentations of at least ten subjects (five for the training set and five for the validation). However, the added value of the fine-tuning is the invaluable capability of ASCHOPLEX_tune to achieve the same performance of ASCHOPLEX without training the models from scratch [45]. Although the number of subjects and the pathological status are different between the two testing sets (36 MS subjects for ASCHOPLEX, 62 depressed subjects for ASCHOPLEX_tune), segmentation metrics are comparable. Considering the above results, the fine-tuning procedure is recommended when ASCHOPLEX has to be inferred on a new dataset; particularly, ASCHOPLEX_tune with five subjects in the training set guarantees an optimal trade-off between the number of extra subjects to be included and the overall segmentation performance, that are comparable to ASCHOPLEX on the original training dataset. This study has limitations. Firstly, the entire data sample has been acquired with MRI scanners with magnetic field strength of 3T at a resolution of 1x1x1 mm. Future directions to be addressed might be the enrichment of the dataset by varying field strength or resolution. Nevertheless, the three tested datasets comprehend different neurological disorders (MS, depression) and have been acquired with different scanners (MRI and PET/MRI) derived from different manufacturers (Philips and GE). Both these variables contribute to boost the confidence in the generalizability of the tool. Moreover, the inter- and intra-rater variability of the neuroradiologists when performing the ChP manual segmentation task was not investigated as it is out of the scope of the paper and has been previously addressed [16]. Lastly, the impact of neuroimaging preprocessing steps (e.g. brain extraction, N4 bias field correction) on ASCHOPLEX performances has not been investigated, as preprocessing steps are not required for ASCHOPLEX to run. Recent studies suggest avoiding preprocessing steps for medical imaging segmentation tasks [46], [47], but we did not explicitly test it in our samples. We can conclude that ASCHOPLEX is a valid tool for obtaining an accurate, and reliable segmentation of the ChP with performance that overcomes the literature reference standard in estimating ChPV. The addition of a fine-tuning step improves generalizability and enables accurate ChPV estimates for unseen datasets. ASCHOPLEX is a transdiagnostic, generalizable, reliable, tool for ChP segmentation, which may further facilitate the use of ChPV as a biomarker.

ACKNOWLEDGMENT

The BIODEP study was sponsored by the Cambridgeshire and Peterborough NHS Foundation Trust and the University of Cambridge, and funded by a strategic award from the Wellcome Trust (104025) in partnership with Janssen, Glaxo-SmithKline, Lundbeck and Pfizer. Recruitment of participants was supported by the National Institute of Health Research (NIHR) Clinical Research Network: Kent, Surrey and Sussex & Eastern. Additional funding was provided by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London, and by the NIHR Cambridge Biomedical Research Centre (Mental Health). ETB is supported by a Senior Investigator award from the NIHR. Study data were collected and managed using REDCap electronic data capture tools hosted at the University of Cambridge. We would like to gratefully thank all study participants, research teams and laboratory staff, without whom this research would not have been possible.

J. J. Schubert and F. E. Turkheimer are supported by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London.

REFERENCES


5All members of the NIMA Consortium at the time of data collection are thanked and acknowledged in the Appendix of the Supplementary Materials.


